Proceedings of LFG12

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Contents

1 Dedication

We dedicate this volume of the LFG proceedings to the memory of Prof. Yehuda N. Falk (1958–2012).
He will be missed by all of us.

גליון זה מוקדש לזכרו של פרופ׳ יהודה נ׳ פאלק (תשי"ט - תשע"ב).
חסרונו יורגש.

2 Editors' Note

The program committee for LFG'12 were Mary Dalrymple and Dag Haug. We would like to thank them for making an extremely efficient team and doing a wonderful job on putting together the program that gave rise to this collection of papers. Thanks also go to the executive committee and the reviewers, without whom the conference would not have been possible. This is especially true for I Wayan Arka, David Gil, I Nyoman Arya Wibawa, Meladel Mistica and the local organizing committee, who put together a superb conference! Finally, as always, we would like to thank Dikran Karagueuzian for his and CSLI's unfailing support.

HIERARCHIES AND COMPETING GENERALIZATIONS IN SERBO-CROATIAN HYBRID AGREEMENT

Alex Alsina and Boban Arsenijević

Pompeu Fabra University CSIC/CCHS

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

Within an approach in which agreement relations can target either the syntactic features or the semantic features of the agreement trigger, the goal of this paper is to explain the distribution of these two types of agreement relations in Serbo-Croatian, focusing on the cases in which the agreement trigger is a hybrid noun. Of particular interest are the situations in which a given word class (e.g. relative pronouns or personal pronouns) shows a split between the two types of agreement relations such that some forms of the paradigm present semantic agreement and others present syntactic agreement. We propose that the distribution of the two types of agreement relations is regulated by two generalizations. These generalizations create a conflict in some words, which we propose is resolved through the application of a principle stating that the more oblique a case form is, the likelier the form is to show semantic agreement.*

1 Introduction

Mixed, or hybrid, agreement is the phenomenon that arises when a noun triggers different agreement forms on its agreement targets depending on a number of syntactic and semantic factors. Usually, a competition arises in the specification of agreement features between the semantic properties and the syntactic features of the noun that governs agreement. We will refer to nouns that display the relevant syntax-semantics mismatches as hybrid nouns, of which there are several in Serbo-Croatian, the language that provides the data for the present analysis. This paper assumes the view of hybrid agreement proposed in Alsina and Arsenijević 2012 (A&A), according to which there is only one set of syntactic features relevant for agreement, in addition to the semantic properties. This is in contrast with the view of hybrid agreement in Wechsler and Zlatić 2000, 2003 (W&Z), which posits two sets of syntactic features (*concord* and *index*), besides the semantic features, in order to explain the facts of agreement.

A&A argue that, in hybrid agreement, the deciding factor in the choice between semantic and syntactic agreement is the grammatical category of the agreement target: adjectives, determiners, and adnominal modifiers show syntactic agreement with a hybrid noun; finite verbs show semantic agreement with a hybrid noun; however, pronouns (relative and personal pronouns) show a mixed pattern in which the choice between the two types of agreement depends on the case form of the pronoun. The goal of this paper is to describe this mixed pattern and to explain why it occurs the way it does, assuming only one set of syntactic agreement features, as in A&A.

The main claim of the paper is that there are two generalizations that determine whether a word class will show semantic or syntactic agreement and a principle that resolves those cases in which a conflict arises between

* The research presented in the paper is supported by the research projects FFI2011-23046, FFI2010-15006 by MICINN, and OI178014 by MPNRS.

the two generalizations. One generalization states that syntactic agreement is dependent on the presence of case morphology; the other generalization makes semantic agreement dependent on agreement with expressions with potentially marked person values. A conflict arises in forms that have both case morphology and agree with expressions with potentially marked person values: the conflict is resolved through the application of a principle stating that the more oblique a case form is, the likelier the form is to show semantic agreement.

No conflicts arise in finite verbs, which do not agree in case, or in attributive adjectives, which cannot agree with an expression bearing a marked person value. Most prominent among the forms in which this conflict occurs are pronouns, which are the word classes in Serbo-Croatian that show both case morphology and agreement with different person values. Different pronouns show a different cut-off point between semantic and syntactic agreement. Whereas personal pronouns can show semantic agreement in all case forms and syntactic agreement is only possible in nominative forms, as an alternative to the semantically agreeing form, non-restrictively used relative pronouns show syntactic agreement in all forms, while semantic agreement is fully grammatical only in the more oblique case forms, such as genitive, dative, and instrumental. What these different patterns have in common is that syntactic agreement is more likely to occur in the less oblique forms and semantic agreement in the more oblique forms.

In section 2, we compare the two views of hybrid agreement proposed in A&A and in W&Z and recapitulate the arguments in A&A in favor of having only one set of syntactic features for agreement. Within this approach, in section 3, we state the two generalizations governing the choice of semantic vs. syntactic agreement and show that a conflict arises between these two generalizations. In section 4, we propose a way to resolve this conflict and to explain the mixed pattern of semantic and syntactic agreement found in pronouns and, interestingly, in predicative adjectives. In section 5, we try to explain why there is a correlation between obliqueness of case forms and semantic vs. syntactic agreement. In section 6, we present a formalization of our proposal in LFG. And, finally, we present our conclusions.

2 Two views of agreement

The standard analysis of hybrid agreement has been to assume that linguistic expressions in an agreement relation are sensitive sometimes to the syntactic features such as gender and number of the agreement target (i.e. the noun or the noun phrase) and sometimes to the corresponding semantic properties. The assumption is that the syntactic agreement features of a nominal expression have a default semantic correlate; for example, the syntactic feature "singular" corresponds to an entity not composed of more than one unit of counting, where the unit of counting is given by the meaning of the noun, whereas the syntactic feature "plural" corresponds to an entity composed of more than one such unit. In exceptional cases, this default correspondence is broken and we find nouns that, for example, are syntactically singular, but denote entities composed of more than one unit of counting. In Serbo-Croatian (S-C), there are such nouns and their syntax-semantics mismatches are visible in the forms that are selected in agreeing expressions in the sentence and discourse, as we see in (1):[1]

(1)	Starija	braća	su	puno	vikala.
	old.FSg	brothers	Aux.Pl	much	shouted.NPl/FSg
	'Older brothers shouted a lot.'				

The noun *braća* 'brothers' requires adnominal adjectives to be in the feminine singular form, but selects the plural form of the auxiliary and other finite verb forms.[2] The traditional view (e.g. Corbett 1979, 1983, 1991) assumes that the adnominal adjective *starija* agrees with the hybrid noun in the syntactic features of gender and number (as well as case), whereas the auxiliary agrees with it in the semantic feature of number (as well as person).

Contrasting with this traditional view, W&Z have proposed a different framework, in which there are two sets of syntactic features relevant to agreement, instead of one, in addition to the semantic features. These two sets are called *concord* and *index*. Concord is made up of three features: gender, number, and case. And index consists of the features gender, number, and person. There are, therefore, two different features of gender and number, given that concord gender is different from index gender and concord number is different from index number. While, as a default, concord gender and number have the same values as index gender and number, respectively, in marked situations, they have different values. This is what W&Z assume to be the case with hybrid nouns such as *braća*. This word would have the following specifications for the various manifestations of gender and number:

(2) *braća* 'brothers'

CONCORD:	[[NUMBER: Sg],	[GENDER: F]]
INDEX:	[[NUMBER: Pl],	[GENDER: N]]
SEMANTICS:	[[NUMBER: Pl],	[GENDER: M]]

Assuming that adnominal adjectives agree with the head noun in concord features, this explains why the adjective *starija* in (1) is feminine singular: *braća* is feminine singular in concord. On the assumption that finite verb forms, such as auxiliaries, agree with their subject in index features, the

[1] The following abbreviations are used: F (feminine), M (masculine), N (neuter), Sg (singular), and Pl (plural). And combinations of them: FSg (feminine singular), MPl (masculine plural), etc.

[2] The participial form *vikala* in (1) is potentially ambiguous between a neuter plural and a feminine singular form, as it is in the nominative case and in S-C there is a homonymy between neuter plural and feminine singular in all nominative case forms. The form by itself cannot tell us whether it is agreeing with a feminine singular expression or with a neuter plural one: it is only through an analysis that we can decide which of the two is right in (1). So, we will leave aside this form for the moment.

choice of the plural form *su* in (1) is explained: the noun phrase headed by *braća* is plural, as well as neuter, in index. The participial form *vikala* in (1), which could be analyzed as showing either concord agreement (hence, feminine singular) or index agreement (hence, neuter plural), is assumed in W&Z for theory-internal considerations to show index agreement and therefore to be neuter plural. As for the semantic features, they are assumed to be needed in W&Z for explaining the facts of pronominal coreference: a personal pronoun in the nominative form referring back to *braća* can be either *ona* or *oni* (see ex. (7a)), since a pronoun agrees with its antecedent either in index features or in semantic features. *Ona* is taken to agree in index features and thus to be neuter plural, although it is homophonous with the feminine singular form. The semantic features of *braća* are reflected in the agreeing pronoun *oni*, which is unambiguously masculine plural.

Although the W&Z framework is quite successful in providing an explanation for the complex facts of hybrid agreement in S-C, A&A argue against the idea that there are two sets of syntactic features for agreement, in addition to their semantic counterparts—the Dual Syntactic Agreement Hypothesis, or 2SAH. The main arguments against this hypothesis are the following:

1. A framework incorporating the 2SAH is much more complex than one that assumes only one set of syntactic agreement features. The latter, for a language like S-C, with two values for number (singular and plural) and three values for gender (feminine, masculine, and neuter), predicts the existence of six classes of words with different combinations of gender and number features. All of these classes have members in them. The 2SAH framework predicts the existence of thirty-six classes of words with different combinations of gender and number features (six combinations of concord gender and number features multiplied by six combinations of index gender and number features). Of these, according to W&Z, only eight have any members in them. Furthermore, once we remove those words whose index features can be derived from their semantic information (i.e. there is no mismatch between index and semantics), we are left with two words in S-C that have some feature mismatch for concord, index, and semantics: *braća*, as shown in (2), and *deca* 'children'. Thus, not only is the 2SAH considerably more complex than the alternative, but the added complexity is motivated by only two words.
2. The 2SAH does not simplify the explanation of the facts. The two words that motivate the 2SAH are claimed by W&Z to be feminine singular (FSg) in concord and neuter plural (NPl) in index. This claim rests on the language-particular syncretism of FSg and NPl in nominative forms. Within a framework with only one set of syntactic features of number and gender, the facts are explained by assuming that certain agreement phenomena are sensitive to the syntactic features (e.g. the head-modifier relation or the agreement of predicative adjectives) and certain others are sen-

sitive to the semantic features (e.g. finite verb agreement). Having the distinction between concord and index does not make the explanation of the facts any simpler.

3. An analysis within the 2SAH framework makes some incorrect predictions. A clear case of this is provided by (3) (from A&A):

(3) Pričamo o deci.
talk.1Pl about children
Ona se danas smatraju gladnom / *gladnim.
they.NPl/FSg Refl today consider.Pl hungry.InstFSg hungry.InstPl
'We're talking about children. They are considered hungry today.'

The pronoun *ona*, although lexically ambiguous between NPl and FSg, has to be NPl in (3) according to W&Z, because pronouns agree with their antecedents in index features and *deci*, like *braća*, is NPl in index. But then the prediction would be that the predicate adjective should agree with *ona* in the plural and yet that is not possible and the grammatical form is in the singular. For the analysis in A&A with a single set of syntactic features, this is expected: *ona* is FSg in (3), agreeing with *deci* in syntactic features, and therefore the predicate adjective is also expected to be FSg.

Within a framework that assumes only one set of syntactic features for agreement, A&A propose that agreement phenomena in S-C split between syntactic and semantic agreement as follows:

- Adnominal modifiers, predicative adjectives,[3] and participles agree in syntactic features; hence, are FSg when agreeing with a hybrid noun.
- Finite verb forms show semantic agreement with a subject headed by a hybrid noun; hence, are plural.
- Pronouns, both personal and relative, agree either semantically or syntactically (hence, are sometimes FSg and sometimes plural), although the choice depends on the case form.

3 Two generalizations governing the choice of agreement

In order to explain the fact that agreement targets in S-C sometimes show semantic agreement and sometimes syntactic agreement with their (hybrid agreement) triggers, we posit the following two generalizations:

(4) a. **Generalization 1**: words that inflect for case show syntactic agreement with their agreement triggers.

b. **Generalization 2**: words that agree with expressions of different (i.e. marked) person values show semantic agreement with their agreement triggers.

The first consequence that follows from these generalizations is that adnominal words, whether adjectives, demonstratives, or other word classes,

[3] In fact, predicative adjectives display a split behavior, as we shall see in section 4.

agree with the hybrid head noun in syntactic features, as shown in (1), repeated here as (5a), and in (5b):

(5) a. Starija braća su puno vikala.
old.FSg brothers Aux.Pl much shouted.NPl/FSg
'Older brothers shouted a lot.'

b. Sreo sam stariju braću.
met.MSg Aux.1Sg old.FSg.Acc brothers.Acc
'I met the older brothers.'

There are two properties of adnominal words that explain this: first, they inflect for case, as we see with words such as *starija* and *stariju* in (5), which makes them subject to generalization 1; and, second, they are restricted to modifying third person expressions. An attempt to restrictively modify a first or second person pronoun results in an ill-formed expression, which slightly improves if the agreement trigger is reanalyzed as third person:

(6) Stariji ti ?je/??si zanimljiviji.
old.CmprtMSg you Aux.3Sg/2Sg interesting.CmprtMSg
~'The older you is more interesting.'

Since adnominal words cannot agree with expressions of marked person values, they are not subject to generalization 2 and thus are not expected to show semantic agreement. Adnominal elements are thus correctly predicted to show syntactic agreement only.

The second consequence that follows from the generalizations in (4) concerns finite verbs.[4] Finite verbs agree with expressions of different person values, but do not inflect for case, which means they are subject to generalization 2, but not to generalization 1. Consequently, they show semantic agreement with the target, which is always the subject: as seen in (5a), the finite auxiliary chosen to agree with the NP headed by *braća* is the plural form *su*. While *braća* is syntactically feminine singular, it is semantically plural, as it denotes a group of individuals.

The third consequence concerns pronouns. Pronouns inflect for case. Therefore, by generalization 1, they should show syntactic agreement. Non-restrictively used relative pronouns and personal pronouns also agree with expressions of different person values (the latter even carry different person values). Therefore, by generalization 2, they should show semantic agreement. Since they are subject to both generalizations, a conflict arises when they agree with hybrid nouns. In such contexts, pronouns show a mixture of semantic and syntactic agreement, which is what this paper wants to explain.

[4] When other triggers of mixed agreement enter the picture, such as pluralia tantum nouns and the honorific form *vi* 'you.Pl', generalization 2 may need to be modified along the following lines: words that agree with expressions with marked person values agree preferentially in the marked features of those expressions. And it does not matter whether it is a syntactic feature or a semantic feature: if it is marked (as in plural number or second person), the agreeing word picks this feature.

As for personal pronouns, in the nominative form they allow either syntactic or semantic agreement with their antecedent, as seen in (7a), whereas non-nominative forms of personal pronouns strongly favor semantic agreement with their antecedent, as in (7b).

(7)		Starija	braća$_i$	su	stroga.	
		elder.FSg	brothers	Aux.Pl	strict.FSg/NPl	
		‘Older brothers are strict.’				
	a.	Oni$_i$	/ Ona$_i$	puno	viču.	
		pro.NomMPl	pro.NomFSg/NPl	much	shout.Pl	
		‘They shout a lot.’				
	b.	Njima$_i$	/??Njoj$_i$	se	puno	viče.
		pro.DatPl	pro.DatFSg	SE	much	shout.Sg
		‘People shout a lot to them. ’				

As for relative pronouns of the *koji* series in their non-restrictive use, there is a contrast between oblique forms, as in (8a), and nominative and especially accusative forms, as in (8b): oblique forms such as instrumental allow either syntactic or semantic agreement with the head noun (hence the choice between FSg *kojom* and Pl *kojima* in (8a)); nominative and accusative forms only allow syntactic agreement, as we see with the accusative *koju* in (8b).

(8)	a.	moja	braća,	sa	kojom	/kojima	se	igram
		my.Fsg	brothers	with	wh.InstFSg	wh.InstPl	SE	play.1Sg
		‘my brothers, with whom I play’						
	b.	moja	braća,	koju	/*koje	viđam	češće	
		my.Fsg	brothers	wh.AccFSg	wh.AccPl	see.1Sg	often.Comp	
		‘my brothers, whom I see more often’						

An interesting fact about relative pronouns that follows from the present analysis is that the split we see between oblique and non-oblique forms of relative pronouns only occurs in non-restrictive clauses: in restrictive clauses, relative pronouns agree syntactically in all cases. Compare (8) with (9):

(9)	a.	deca	sa	kojom	/*kojima	se	igram
		children	with	wh.InstFSg	wh.InstPl	SE	play
		‘the children (that) I play with’					
	b.	deca	koju	/*koje	viđam	češće	
		children	wh.AccFSg	wh.AccPl	see.1Sg	often.Comp	
		‘the children that I see more often’					

This is expected: relative pronouns in non-restrictive clauses, like personal pronouns, can take antecedents of any person value. Relative pronouns in restrictive clauses, like adnominal modifiers, only take 3rd person antecedents. Therefore, relative pronouns are subject to generalization 2 only in non-restrictive clauses and only in this case are they expected to agree semantically. Relative pronouns in restrictive clauses are never expected to agree semantically.

To summarize, we find a split behavior in pronouns with respect to syntactic or semantic agreement. Personal pronouns agreeing with hybrid nouns of the *deca* type can be either plural or singular (i.e. both semantic and syntactic agreement are available) in the nominative form; in all other cases, they only take plural forms (i.e. enter semantic agreement), as shown in (7). Non-restrictively used relative pronouns agreeing with hybrid nouns appear only in feminine singular (i.e. only syntactic agreement is possible) in the nominative and accusative case; in all other cases, both options are available: singular and plural (i.e. allowing either semantic and syntactic agreement). The fact that pronouns are the word classes that show an alternation and a split between these two types of agreement is to be expected under the present analysis, as they are subject to both generalizations in (4). We still need to explain why pronouns split the way they do with respect to semantic and syntactic agreement.

4 Explaining the split

The facts of personal and relative pronouns with respect to whether semantic or syntactic agreement is available are explained by the following principle:

(10) **Semantic agreement and case obliqueness (SACO)**:
The more oblique a case form is in the obliqueness hierarchy (11), the likelier the form is to show semantic agreement.

(11) **Obliqueness hierarchy of case forms**:
nominative < accusative < oblique cases

The SACO correlates case obliqueness with semantic vs. syntactic agreement and its effects vary for each grammatical category that both has case and shows agreement (with expressions of different person values).

In personal pronouns, the split is between forms that allow both syntactic and semantic agreement and forms that only allow semantic agreement. The cut-off point is between nominative and accusative: any case form more oblique than nominative strongly favors semantic agreement. Since nominative forms are left out of this restriction, they are allowed to show either semantic or syntactic agreement. In example (7), for the pronoun that refers back to *braća*, there is a choice between the semantically agreeing *oni* and the syntactically agreeing *ona* in the nominative, but in the dative the FSg *njoj* is not allowed and only the Pl *njima* is possible.

In relative pronouns, the split is between forms that only allow syntactic agreement and forms that allow either syntactic or semantic agreement. The cut-off point is between accusative and oblique cases such as genitive or dative: the less oblique cases (i.e., nominative and accusative) are required to show syntactic agreement. This allows the oblique cases to show either syntactic or semantic agreement. This is consistent with the SACO, as semantic agreement is only found in the more oblique case forms of the relative pronoun. As shown in (8), modifying *braća*, either the FSg *kojom* or the Pl form *kojima* is possible in the instrumental but only the FSg form *koju* is allowed in the accusative case.

An unexpected consequence of the SACO is that it explains a contrast found in predicative adjectives. Predicative adjectives occur in two constructions in S-C: as complements to copular and auxiliary verbs, where the adjective is in the nominative case, as in (12a), and as complements to semantically heavy verbs like *smatrati* 'consider', where the adjective is in the instrumental case, as in (12b). Whereas syntactic agreement is the preferred option in both constructions, semantic agreement is completely out in the nominative case, but only somewhat degraded in the instrumental case, as illustrated in (11).[5]

(12) a. Gospoda su došla / *došli.
gentlemen-Nom are come.Fsg come.MPl
'The gentlemen came.'

b. Gospodu smatram brzom / ?brzim.
gentlemen-Acc consider.1Sg fast.Fsg fast.MPl
'I consider gentlemen fast.'

The SACO predicts that, if one of the two predicative adjectives should allow semantic agreement, that would be the oblique case form. This is what we see here: although the semantically agreeing form is not perfect in either case, it is much better in the instrumental case than in the nominative case.

5 Implications of the proposal

A well-known alternative proposal to explain the distribution of syntactic and semantic agreement in different constructions across languages is found in Corbett's (1979, 1991, 2006) work. According to Corbett's hypothesis, the further right in the hierarchy in (13) an agreement target is, the likelier semantic agreement is to occur.

(13) **Corbett's hierarchy of agreement targets**:
attributive < predicate < relative pronoun < personal pronoun

The claim is that semantic agreement should affect a continuous segment of the categories specified in the hierarchy in (13), so that, if, for example, relative pronouns in a particular language show semantic agreement, personal pronouns should also show semantic agreement. Or, if predicates show semantic agreement, so too should relative pronouns.

In the case of S-C, we find that both relative pronouns and personal pronouns show a mix of semantic and syntactic agreement that depends on the specific case form. In fact, as there are more case forms of personal

[5] The hybrid noun used in (12) is *gospoda* 'gentlemen', which, like the other hybrid noun seen in this paper *braća*, is syntactically feminine singular and semantically masculine and plural.

pronouns than case forms of relative pronouns that show semantic agreement and fewer case forms of personal pronouns than of relative pronouns that show syntactic agreement, these facts are quite consistent with the predictions of Corbett's hypothesis. However, when we take predicates into consideration, it is not so clear that the facts are consistent with Corbett's predictions. It is reasonable to assume that finite verbs are predicates. However, since finite verbs consistently show semantic agreement with a subject headed by a hybrid noun, the expectation derived from Corbett's hypothesis is that relative pronouns and personal pronouns should also show consistent semantic agreement with a hybrid noun. That is not so. What we have seen is that relative and personal pronouns do not show consistent semantic agreement with a hybrid noun, but sometimes show syntactic agreement given the appropriate case form.

Our explanation based on two competing generalizations, (4), and a principle to solve the situations where conflict arises, (10), incorporating the case obliqueness hierarchy, does not have this problem:

- predicative adjectives, relative pronouns, and personal pronouns all show a split, with semantic agreement in some cases and syntactic agreement in others, as they are subject to both generalizations;
- finite verbs do not have case morphology and hence are not required to show syntactic agreement.

Moreover, we derive the facts of hybrid agreement from two more primitive properties: case marking and restriction to agreement triggers in third person. Case marking is fully dependent on the syntactic position of the agreement trigger (i.e. its grammatical function), and hence the agreeing item is pulled towards a general syntactic agreement. Case marking appears only if the agreeing item also has a nominal nature (in the sense in which nouns and adjectives are [+n]), which means that all the nominal features subject to agreement are available for syntactic agreement. (The nominal features are gender, number, and case.)

Syntactic expressions that agree with marked person values either agree with full fledged referential expressions, as is the case of finite verbs and predicative phrases, or are themselves referential expressions, as with personal pronouns. Items agreeing with full fledged referential expressions have access to the actual referents, and hence can establish semantic agreement (they may still have access to the syntactic features as well, thus having both options available).

As for the logic in principle (10), SACO, we have to bear in mind a distinction often made between two groups of case features: nominative and accusative are often referred to as the structural cases, while the other case features are labeled inherent. "Structural" refers here to the idea that the case feature is dependent on the grammatical function, and not on the semantic relation, while "inherent" reflects the idea that the case feature depends on the semantic relation of the syntactic expression. Given this, it makes sense that structural cases show a stronger tendency towards syntactic agreement, while inherent cases, which are less dependent on syntactic relations, should leave their bearers more open for semantic agreement. Within the structural cases, it is reasonable to view nominative as less dependent on semantic role than accusative, as it is the subject case and the subject can correspond to any semantic role.

6 An LFG encoding of the proposal

At this point, one should ask how these ideas can be translated into the LFG framework, if at all. The generalizations in (4) and the principle (10) should not be seen as principles of a formal grammar, regardless of the framework adopted, but as principles that constrain a formal grammar. Let us consider several situations that are relevant for our analysis: (a) adjectival modification, (b) finite verb agreement, (c) personal pronouns, and (d) relative pronouns. The analysis that follows is sketchy, but—we hope—clear enough that it can be adapted to different versions of the framework.

One of the properties of adjectival modification in S-C is that the syntactic agreement features of the adjective have to be identical to those of the head noun that the adjective modifies. We can capture this idea by assuming that the features of case, gender and number of nominal expressions such as nouns and adjectives are grouped as the f-structure value of the feature AGR (for agreement) and that adjectival modification requires identification, or sharing, of the AGR of the NP with the AGR of the AP modifier. Thus, an AP daughter of NP is required to have the grammatical function MOD (for modifier, or ADJ) and is specified as sharing its AGR feature with that of the f-structure it is a feature of.[6] Let us assume that an adjective like *starija* and a noun like *braća* have the c- and f-structure information as part of their lexical entries in (14).[7] Notice that, at this point we are not concerned about the semantic information, which is where we would encode the idea that *braća* denotes a group of male individuals.

(14) a. starija:

$$A_1 \quad \begin{bmatrix} \text{PRED} & \text{‘old’} \\ \text{AGR} & \begin{bmatrix} \text{CASE} & \text{Nom} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix}_1$$

[6] There are different ways of representing Adj-Noun agreement in addition to the one presented here, as can be seen in Dalrymple, Dyvik, and King 2004, and any will work fine for our purposes as long as the features involved are f-structure features and not semantic features.
[7] Coindexation signals correspondence between pieces of structure: in (14a) it indicates that the categorial information A corresponds to the f-structure with the same index.

b. braća: N_2 $\left[\begin{array}{ll} \text{PRED} & \text{‘brothers’} \\ \text{AGR} & \left[\begin{array}{ll} \text{CASE} & \text{Nom} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{array}\right] \end{array}\right]_2$

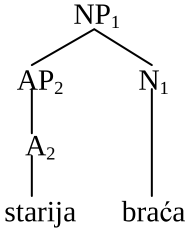
The rule for modifier-head agreement would be stated as in (15):

(15) Modifier-head agreement:

$$\left[\begin{array}{ll} \text{AGR} & [\ ]_1 \\ \text{MOD} & [\ \text{AGR} \quad [\ ]_1\ ] \end{array}\right]$$

On the assumption that every f-structure is uniquely identified by its index, having two f-structures with the same index, as in (15), means that they are the same f-structure. So, (15) says that the agreement features of a modifier are the same as those of the structure in which it belongs. In this way, the NP *starija braća*, from (1), has the c-structure and f-structure shown in (16):

(16)



$$\left[\begin{array}{ll} \text{PRED} & \text{‘brothers’} \\ \text{AGR} & \left[\begin{array}{ll} \text{CASE} & \text{Nom} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{array}\right]_3 \\ \text{MOD} & \left[\begin{array}{ll} \text{PRED} & \text{‘old’} \\ \text{AGR} & [\ ]_3 \end{array}\right]_2 \end{array}\right]_1$$

Given that the agreement features that matter in this construction are the syntactic ones, it is irrelevant if the semantic features of the two words involved are not the same. If we should choose the form of the adjective with the features that correspond to masculine plural—*stariji*—the result would be ungrammatical (**stariji braća*), even though the adjective and the noun would be in semantic agreement.

Finite verb agreement. As argued in A&A and as shown in examples like (1) and (5a), the finite verb form agrees in person and in semantic number with its subject. We can assume, as is standard practice in LFG to explain subject-verb agreement, that the lexical entry of a finite verb in S-C specifies certain features of its subject. What is special about S-C is that one of these features—the number feature—is not an f-structure feature, but an s-structure (or semantic structure) feature. For example, the auxiliary form *su*, as in the examples just cited, has the lexical specifications in (17). Here, to distinguish f-structure features from s-structure features, we prefix a feature structure belonging to f-structure with an “f” and a feature structure belonging to s-structure with an “s”.

(17) su: I_1 ${}_{f}\left[\begin{array}{ll} \text{SUBJ} & [\text{PERS} \quad 3]_2 \\ \text{TENSE} & \text{pres} \end{array}\right]_1$

$${}_{s}[\text{NUM} \quad \text{Pl}]_2$$

The f-structure information specified for the auxiliary *su* is compatible with the subject NP *starija braća*, whose f-structure is shown in (16). The number feature specified in the lexical entry (17) is an s-structure feature of the subject. Since the NP *starija braća* is semantically plural (i.e. it would have the s-structure feature [NUM Pl]), this NP can function as the subject of the auxiliary *su*, as seen in (1) and (5a).

Personal pronouns. The salient facts to explain are that accusative and oblique case forms of the personal pronoun agree with their antecedents in semantic features, whereas the nominative forms are free to agree either semantically or syntactically. One way to explain this observation is to assume that in accusative and oblique case forms, the semantic correlates of their syntactic agreement features are lexically specified, whereas they are only optionally specified for nominative forms. The dative plural *njima* and the dative FSg *njoj*, from (7b), would have the lexical information in (18), and the nominative forms *oni*, MPl, and *ona*, FSg, from (7a), would have the lexical entries in (19), where the parentheses around the semantic structures indicate that they are optional:

(18) a. njima: N_1

$$
{}_{f}\left[\begin{array}{ll} \text{AGR} & \left[\begin{array}{ll} \text{CASE} & \text{Dat} \\ \text{NUM} & \text{Pl} \end{array}\right] \\ \text{PRED} & \text{‘pro’} \end{array}\right]_{1}
\qquad {}_{s}[\text{NUM} \quad \text{Pl}]_{1}
$$

b. njoj: N_2

$$
{}_{f}\left[\begin{array}{ll} \text{AGR} & \left[\begin{array}{ll} \text{CASE} & \text{Dat} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{array}\right] \\ \text{PRED} & \text{‘pro’} \end{array}\right]_{2}
\qquad {}_{s}\left[\begin{array}{ll} \text{NUM} & \text{Sg} \\ \text{GEND} & \text{F} \end{array}\right]_{2}
$$

(19) a. oni: N_3

$$
{}_{f}\left[\begin{array}{ll} \text{AGR} & \left[\begin{array}{ll} \text{CASE} & \text{Nom} \\ \text{GEND} & \text{M} \\ \text{NUM} & \text{Pl} \end{array}\right] \\ \text{PRED} & \text{‘pro’} \end{array}\right]_{3}
\qquad \left({}_{s}\left[\begin{array}{ll} \text{NUM} & \text{Pl} \\ \text{GEND} & \text{M} \end{array}\right]_{3} \right)
$$

b. ona: N_4

$$
{}_{f}\left[\begin{array}{ll} \text{AGR} & \left[\begin{array}{ll} \text{CASE} & \text{Nom} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{array}\right] \\ \text{PRED} & \text{‘pro’} \end{array}\right]_{4}
\qquad \left({}_{s}\left[\begin{array}{ll} \text{NUM} & \text{Sg} \\ \text{GEND} & \text{F} \end{array}\right]_{4} \right)
$$

What we need to assume in order to explain the relevant facts is that the semantic agreement features of the pronoun must be shared with those of the antecedent; in the absence of relevant semantic features, syntactic features are shared. This explains why a dative pronoun referring to an NP headed by a hybrid noun like *braća* has to be the plural *njima* and cannot be the feminine singular *njoj*, as seen in (7b): the specified semantic features have to match those of the antecedent, which is plural. With a nominative pronoun, the option of having the semantic number and gender features specified explains the choice of pronoun in a case like (7a): if the semantic agreement features are specified, they must be shared with the antecedent and, therefore, with *braća* as the antecedent, the masculine plural pronoun *oni* is required; if those features are not specified, the syntactic gender and number features must be shared and then it is the feminine singular *ona* that is chosen.

Relative pronouns. The facts are that relative pronouns in the nominative and accusative cases must agree in syntactic features with the head noun, whereas oblique case relative pronouns (such as dative, genitive, or instrumental) in non-restrictive clauses can show either syntactic or semantic agreement. We can explain this by assuming that the non-oblique case forms only specify syntactic features, while the oblique case forms specify either the semantic features or the syntactic features of gender and number. So, whereas the accusative form *koju* has the lexical entry shown in (20a), the instrumental *kojima* has either of the entries in (20b,c):

(20) a. koju: N_1

$${}_{f}\left[\begin{array}{ll} \text{AGR} & \left[\begin{array}{ll}\text{CASE} & \text{Acc}\\ \text{NUM} & \text{Sg}\\ \text{GEND} & \text{F}\end{array}\right] \\ \text{PRED} & \text{‘relpro’}\end{array}\right]_{1}$$

b. kojima: N_2

$${}_{f}\left[\begin{array}{ll} \text{AGR} & \left[\begin{array}{ll}\text{CASE} & \text{Inst}\\ \text{NUM} & \text{Pl}\end{array}\right] \\ \text{PRED} & \text{‘relpro’}\end{array}\right]_{2}$$

c. kojima: N_3

$${}_{f}\left[\begin{array}{ll} \text{AGR} & [\text{CASE} \quad \text{Inst}] \\ \text{PRED} & \text{‘relpro’}\end{array}\right]_{3}$$

$${}_{s}[\text{NUM} \quad \text{Pl}]_{3}$$

Assuming that a relative pronoun has to agree with its antecedent in all gender and number features, whether syntactic or semantic, we explain the facts illustrated in (8). The accusative, as well as the nominative, relative pronoun has its syntactic gender and number features specified, but not its semantic counterparts. Consequently, the accusative form *koju* is chosen when agreeing with the hybrid noun *braća*. The instrumental form *kojima* can be chosen when agreeing with the head noun *braća*, because both have semantic plural number. Alternatively, the instrumental form *kojom* can also be chosen because it too has two lexical entries and in one of them it has the features of feminine gender and singular number as part of the f-structure, just like the agreeing head noun *braća*.

Although the grammar of S-C does not include the generalization in (4) or the principle in (10), it does comply with these principles. These principles are thus *metagrammatical principles*, which constrain how a grammar of a particular language can be.

7 Conclusions

In this paper we have argued that hybrid agreement in S-C is best explained by a theory in which linguistic expressions have two types of agreement features (only): one syntactic and one semantic. We have also proposed that whether an agreeing item will reflect the syntactic or the semantic features (yielding syntactic or semantic agreement) depends on two properties of the agreeing item: whether it is marked for case and whether it combines with agreement triggers of marked person values. These correlations are captured by the two generalizations in (4): items with case morphology tend to show syntactic agreement and items that agree with expressions of different person values tend to show semantic agreement. These generalizations impose conflicting requirements on forms that both have case morphology and agree with expressions of different person values. When such a conflict arises, a split emerges within the paradigm of the word classes affected: some case forms show syntactic agreement, some show semantic agreement, and some show either. Whether a case form behaves one way or another is not a completely random fact, but is constrained by principle (10), which correlates the likelihood of a form showing semantic agreement with its position in the case obliqueness hierarchy.

These principles can be compared to Corbett's (1979, 1991, 2006) explanation of how semantic and syntactic agreement is distributed across constructions in different languages. We show that some of the facts of hybrid agreement in S-C may constitute counterexamples to the predictions made by Corbett's explanation, whereas they are consistent with the proposal in this paper.

The status of these principles is similar to that of Corbett's explanation: they are not principles of the grammar of a natural language, but constraints on possible grammars. So, when the grammar of a language is formalized, there is no principle or constraint that can be identified with the principles in (4) or (10). Nevertheless, the grammar conforms to these principles.

References

Alsina, Alex and Boban Arsenijević. 2012. The two faces of agreement. *Language* 88: 369-380.

Corbett, Greville. 1979. The agreement hierarchy. *Journal of Linguistics* 15: 203–224.

Corbett, Greville. 1983. *Hierarchies, targets, and controllers: Agreement patterns in Slavic*. London: Croom Helm.

Corbett, Greville. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville. 2006. *Agreement*. Cambridge: Cambridge University Press.

Dalrymple, Mary, Helge Dyvik, and Tracy Holloway King. 2004. Copular complements: Closed or open?. In M. Butt and T. H. King (eds.) *Proceedings of the LFG04 Conference*, 188-196. CSLI Publications, Stanford University

Wechsler, Stephen and Larisa Zlatić. 2000. A theory of agreement and its application to Serbo-Croatian. *Language* 76: 799–832.

Wechsler, Stephen and Larisa Zlatić. 2003. *The many faces of agreement***.** Stanford, CA: CSLI Publications.

VERBAL NUMBER, ARGUMENT NUMBER, AND PLURAL EVENTS IN MARORI

I Wayan Arka

Australian National University
Udayana University

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper discusses number and plurality in the nominal and verbal domains of Marori (isolate, Trans New Guinea). Marori shows evidence that verbal number and argument/nominal number should be distinguished, even though they are integrated in a complex way, with some parallelism in the constructed mode of expressing limited plural (paucal). The complexity of the syntax and semantics of verbal number in relation to argument number, aspect, and other constructions such as reciprocals in this language calls for a sophisticated precise unified analysis. I propose that verbal and nominal number have the same composite number features (+/–SG, +/–PL, and +/–AUG) and demonstrate that their intricacy can be straightforwardly captured within a unification-based LFG framework.

1 Introduction [*]

Verbal number is a category of number related to events, reflecting the plurality of events (i.e., the number of times an action/state happens) (Durie 1986; Corbett 2000; Veselinova 2008). Plurality of events can be conceptualised as iterated events involving the same participants or as distributive events involving different participants. Verbal number is very common in the languages of North America, but it also found in South American and Papuan languages (Veselinova 2008). This paper demonstrates that Marori (isolate, Trans New Guinea) has two kinds of verbal number distinct from argument (nominal) number and that verbal number and argument number are integrated into the overall number system in Marori in an intricate way.

Verbs showing verbal number are often suppletive in nature. Mithun (1988: 213) reports the alternation of roots showing the opposition of singular (SG) vs. non-singular (NSG) verbal number in North American languages with intransitive verbs such as 'sit' and 'stand' (reflecting the number of subject participants) and transitive verbs such as 'kill' (reflecting the number of object participants) (equivalent to the distinction between *kill* and *massacre* in English). Verbal number in Marori shows this property, but it will be shown later that verbal number in Marori is not simply an

[*] Research reported in this paper was supported by ARC Discovery Grant 110100307 (2011–2015). I thank the following people for their comments and questions: Ash Asudeh, Mary Dalrymple, Nick Evans, Louisa Sadler, and Rachel Nordlinger. Special thanks must go to Verman Singerin, Agustinus Mahuze, Bapak Paskalis Kaize, Esebyus Basik-basik, Bapak Amandanus, Mama Veronika, and Bapak Willem Gebze for their hospitality while I was in the field and their help with the data and the transcriptions.

alternation that is lexically determined, but also subject to grammatical constraints such as finiteness (section 4.2).

Verbal number should be distinguished from argument number. The latter is related to the number of entities/event participants. Argument number is often realised on the nominal unit of a clause, hence, also called nominal number. Nominal number within an NP is typically associated with the noun head—e.g., *book* vs. *books*, *girl* vs. *girls* in English. There may be number agreement within the NP between the noun and its determiner—e.g., *this girl* vs. *these girls* in English—or agreement within the clause between the subject NP and the verb—e.g., *the girl is …* vs. *the girls are …* in English. In short, argument number shows a systematic opposition of participant number possibly realised on the nominal phrase (determiner, pronoun, possessive, and adjective) and on the verb.

The Marori data presented in this paper provides support to what has been noted in the literature: namely, that verbal number, while related to aspect and argument number, should be treated as a distinct independent category (cf. Corbett 2000). The complexity of the syntax and semantics of verbal number and also the parallelism between verbal and argument number in this language calls for a precise unified analysis. I demonstrate that this can be straightforwardly captured within a unification-based LFG framework. I will show that plurality applies to both nominal and verbal domains and that the same mechanism is used, e.g., the same constructed strategy to express paucal in both nominal and verbal number.

The paper is structured as follows: An overview of Marori morphosyntax and nominal number is outlined in section 2, while evidence for two types of verbal number—namely, Actor/Subject verbal number (A-vn) Object and verbal number (O-vn)—is given in section 3. The two relate to different conceptions of event plurality, with A-vn used to express distributive plural. The interplay between verbal number and other grammatical phenomena such as finiteness and reciprocity is discussed in section 4. The important point discussed in this section is the parallelism between argument and verbal number in encoding constructive number. An LFG analysis is outlined in 5, and the conclusion is given in 6.

2 Marori morphosyntax in brief

Marori is a non-configurational verb-final language. Subject and object NPs typically come before the verb, without a fixed order, but they can be scrambled, including appearing after the verb. The predicate unit typically consists of a lexical verb and a light or auxiliary verb. The lexical verb immediately precedes the light/auxiliary verb.

Grammatical relations are encoded by verbal agreement as well as by marking on the argument NPs. In general, A(ctor) receives suffix verbal agreement, whereas U(ndergoer) receives prefix verbal agreement. Free NPs do not come with a case marking, but definite U NPs may be marked by the =*i* clitic. In a transitive structure, only one =*i* is possible. In a ditransitive structure, =*i* marks the recipient object NP. In an intransitive structure, the sole U NP receives =*i* as in (1)a below. A transitive/intransitive actor never gets marked by =*i*. In short, grammatical relations in Marori are semantically marked: undergoer marking. Below are several examples.[1]

(1) Intransitives

a.	*na=i*	*patar*	*yu-nggo-f*	b.	*efi*	*ramon(*=i)*	*kundo-f*
	1SG=U	cold	1SG-AUX-NrPST		that	woman	run.3SG-NrPST
	'I suffered from being cold.'				'She/the woman ran off.'		

(2) Ditransitives

Nawa	*tamba*	*Albert=i*	*nji=me-ben*	*bosik*	*sokodu.*
1SG	already	Albert=U	3.give=AUX-1NPL.NrPST	pig	one
'I already gave Albert a pig.'					

Nouns are not marked for number. Pronouns and their corresponding pronominal affixes on the verb do show number distinctions, e.g., *na* '1SG' vs. *nie* 'NSG' for free pronouns.

Pronominal suffixes are portmanteau forms showing person, number, tense, aspect, and mood information. They can be grouped into two classes as shown in Tables 1 and 2, depending on the aspect they encode in their past tenses: the completive and durative classes.[2]

	(1a)			(1b)			(1c)		
	IRR/FUT			NrPST (Completive)			RmPST (Completive)		
	1	2	3	1	2	3	1	2	3
SG	*-ru*	*-Ø*	*-Ø*	*-ben*	*-f*	*-f*	*-fori*	*-fi*	*-fi*
DU	*-ren*	*n--Ø*	*-Ø*	*-ben*	*n- -f*	*-f*	*-fori*	*n- -fi*	*-fi*
PL	*-men*	*n-(ri)m*	*-(ri)m*	*-freben* *-frendu*	*n- -(fre)f*	*(fre)f*	*-mbrofori*	*-mbrofi*	*mbrofi*

Table 1: Class 1 Argument suffixes in Marori

[1] Abbreviations: 1,2,3 (First, Second, Third Person); A (Actor); COMP (Complementiser); DEIC (Deictic); DU (Dual); DUR (Durative); F (Feminine); FUT (future); INT (Intensifier); LOC (Locative); M (Masculine); NF (Non Finite); NrPST (Near Past); NSG (Non Singular); NPL (Non Plural); PRES (Present); RECIP (Reciprocal); REDUP (Reduplication); SG (Singular); U (Undergoer).

[2] The formatives *–re/-ro/-ri* are, strictly speaking, not part of pronominal argument suffixes but are of Actor verbal number (A-vn, see Figure 1). They are included here to show that they serve to encode the general opposition of underspecified NSG vs. PL.

	(2a) REAL/MacroPRES (Completive/extended)			(2b) NrPST (Durative)			(2c) RmPST (Durative)		
	1	2	3	1	2	3	1	2	3
SG	*-du*	*-Ø*	*-Ø*	*-men*	*-m*	*-m*	*-maf*	*-maf*	*-maf*
DU	*-den*	*n-Ø*	*-Ø*	*-men*	*n- -m*	*-m*	*-maf*	*n- -maf*	*-maf*
PL	*-men*	*n--Ø*	*-Ø*	*-ben*	*n- -b/-m*	*-b/-im*	*-baf*	*n- -baf*	*-baf*

Table 2: Class 2 Argument suffixes in Marori

3 Two types of verbal number

3.1 Marori verbal template

The verbal template in Marori showing two kinds of verbal number, called O- and A- verbal number for simplicity, is shown in Figure 1. The O-verbal number (O-vn) shows alternate forms expressing plurality of events due to the plurality of transitive objects[3] or the plurality of the intransitive subjects (typically, but not restricted to, patientive or unaccusative verbs). O-vn is encoded by suppletive root alternations: e.g., *nde* 'bring.SG.O' vs. *kei* 'bring.PlO', *mara* 'fly.SG' vs. *merfe-* 'fly.PL'. The A-verbal number (A-vn) expresses plural distributive events associated with the plurality of transitive/intransitive subject A/S.[4] It is marked by *-ro* (and its variants *-ri, -re, -ra*), occupying the position immediately after the verbal root.

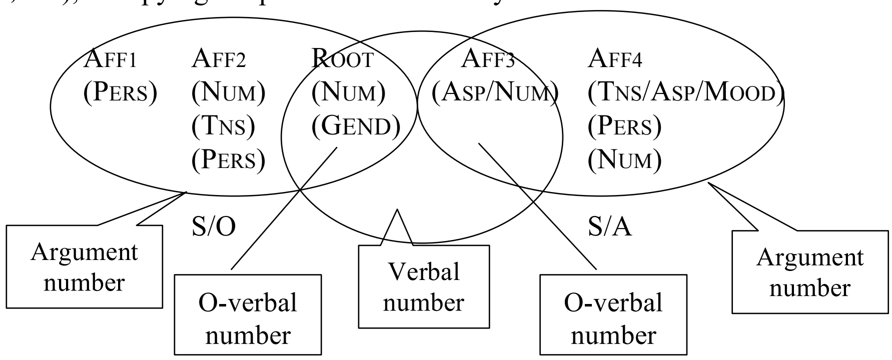


Figure 1

[3] Note that the case of plural objects in a single event is possible, e.g., with the verb *kei* 'bring.PlO', i.e., a 'carrying' situation where a single actor carries plural objects in one go.
[4] The morpheme *-ro* cannot be simply labeled as a distributive marker, because it is also used to mark the durative/progressive aspect. Marking both the durative aspect and distributive plural is a common function associated with verbal number.

The template also shows slots associated with argument number. The prefix encodes S/O agreement, whereas the suffix encodes S/A agreement.[5] The circles indicate that number information is distributed across different exponents with overlapping space.

The intricacies of how argument number and verbal number interact will be described in the subsequent sections.

3.2 Argument number vs. verbal number

Verbal number and argument number—while intertwined, as seen from Figure 1—are distinct categories in Marori. The evidence comes from the fact that the two are encoded differently and that they serve different functions in the grammar.

3.3 Different coding

In terms of formal coding, argument number is realised by distinct agreement affixes, depending on the grammatical functions of the arguments. The suffixes mark S/A arguments and consist of two classes, as shown in Table 1 and Table 2. The S/A agreement suffixes carry complex agreement information (person and TAM).

The prefixes mark and agree with S/O arguments. They are *y*- '1', *k*- '2', and ∅- '3'. They may come with additional formatives expressing other information such as tense and number, e.g., *y-ar*-'1-1/2.NSG.PST'.

O-verbal number is realised by suppletive alternates exemplified in (3). Certain adjectival stative predicates form their NPL vs. PL number opposition by -*on* and *nde*, exemplified in (4). These two may appear together in a clause with the regular plural A-vn morpheme -*ro* (or -*re*/-*ri*/-*ra*), exemplified by *bring* in (5).

(3) Suppletive roots expressing O-verbal number

nde	'bring.SG.O'	vs.	*kei*	'bring.PL.O'
tr	'hit.NPL.O'	vs.	*ksw*	'hit.PL.O',
kunonjo	'go.NPL'	vs.	*kurfenj*	'go.PL',
anep	'big.SG'	vs.	*kofe*	'big.NSG'.

(4) Verbal number: stative predicates

	SG/NPL	PL
'red'	*paraw-on*	*paraw-nde*
'short'	*sor-on*	*sor-nde*
'tall'	*nggworow-on*	*nggworow-nde*

[5] The abbreviations S, A, and O follow the tradition in typological linguistics: S (intransitive subject), A (transitive subject), and O (transitive object).

(5) Verbal number *bring*

			O-vn:	
			SG.O	NSG.O
	NPL:	SG	*nde*	*kei*
A-vn		DU	*nde*	*kei*
	PL		*nde-re*	*kei-re*

The following are worth noting in terms of coding and status of number categories. First, the two kinds of verbal number (O-vn and A-vn) are clearly distinct because they cross-cut the verbal number space, as exemplified by the formation of the verb *bring* in (5). The two give rise to cases showing plural A-vn with singular objects and plural A-vn with plural objects. That is, distributivity/plurality for the subject is independent from plurality for the object.

Second, the verbal template shows that the A-vn formative -*ro* occupies a slot different from the slot of the S/A argument number morpheme. -*ro* is adjacent to the verbal root, whereas the S/A morpheme is in the outermost position. A deictic morpheme -*n* can intervene between the A-vn and the S/A argument number suffix, as seen in the following example.[6]

(6) *kurfenj-re-n-du*
return.PL-PL-DEIC-1PRES
'We (three or more) return here.'

Third, as seen in (4), adjectives also show alternates to encode event plurality. The adjectives can be predicative, e.g., *soron/sorde,* as seen in (7). The predicative part is structurally distinct from the verbal part (with its own verbal number, e.g., -*re*). Thus, the term *predicative number* is perhaps better than the term *verbal number*, as such number opposition does not solely apply to the verbal part of the predicate. In addition, the term *predicate number* is appropriate if we want to highlight the two kinds of numbers, contrasting them with the other kind of number, namely, argument number.[7]

(7) a.

Na	*tanamba*	*sor-on*	*to-mbo-du*
1SG	now	short-NPL	be-NPL-1SG.PRES

'I am short now.'

b.

Nie	*yanadu*	*tanamba*	*sor-on*	*to-mbo-den*
1NSG	two	now	short-NPL	be-NPL-1DU.PRES

'We (2) are short now.'

[6] The distributive plural event marked by –*ro* can be simultaneous or not.
[7] It appears that a state involving a single participant with a stative predicate such as *short* is counted as one event in this language. Hence, plural participants/subjects are necessarily associated with plural states/events.

c. *Nie* *usindu* *tanamba* *sor-de* *te-re-men*
1NSG all now short be-PL-1PL.PRES
'We (2<) are all short now.'

Fourth, while encoded by affixation, the adjectival number shown in (4) is also lexically determined. That is, only certain adjectives allow the alternation.[8] This lexical constraint makes the adjectival number alternation similar to that of the O-vn. Hence, the adjectival number in (4) can be classified as O-vn. That is, it is associated with the lexical predicate, as is the case with other (suppletive) O-vn in (3), distinct from the A-vn (-*ro*).

In addition, the predicate is stative, with the sole argument being O-like. At first, it may not be immediately clear whether *–de* is a verbal number suffix. However, given the overall system of the grammar in Marori where argument number agreement only occurs in the verbal auxiliary part of the verb complex, then the PL suffix -*de* must be analysed as predicative/verbal number marking, rather than argument number. In addition, as mentioned earlier, its encoding is lexically determined. This is a typical property of predicative/verbal number. Argument number is, in contrast, typically part of a highly regular inflectional system, although there may be a number of irregular plural verbs.

Finally, in larger syntax, the predicative number must respect (verbal/argument) number agreement with the auxiliary. Thus, the plural *sorde* must appear with plural verbal number and argument number, as seen in (7)c.

3.3.1 Different but intertwined functions

Verbal and argument number serve different functions in the grammar.

Grammatically, argument number on the verb is part of transitivity and the agreement mechanism, tracking participant roles, e.g., Actor-Undergoer identification. Verbal number, in contrast, is not directly part of the argument-tracking mechanism. It is part of an event-tracking mechanism, where event conceptions such as repetitive, durative, and distributive are relevant. Thus, it is grammatically related to the TAM system.

However, complexity arises due to the fact that the relevant information associated with verbal and argument number in Marori is distributed across different typically portmanteau morphemes. The verbal suffixes -*m* vs. -*f,* for example, are argument agreement suffixes, but they also carry aspect and tense information relevant for the eventualities.

Both verbal number and argument number encode plurality. Argument number expresses an aggregate of three or more entities essentially within the

[8] It remains to be investigated whether the affixation depends on certain semantic properties such as lexical aspect.

nominal domain.[9] It is also encoded on the verb due to grammatical verbal agreement. In contrast, plurality in verbal number expresses a complex concept of aggregate eventualities, where event multiplicity/distribution and aspectual properties such as punctuality vs. non-punctuality are important.

Evidence that verbal and argument number have different functions comes from the fact that the plural verbal marker does not impose plural argument agreement. Example (8)a shows the verbal number *-ra* with a singular argument. *-ra* expresses extended aspect. The absence of it in (8)b signifies a non-extended event. Both are past events. The first highlights the duration (of the whole day), whereas the second highlights the point at which that event had already been completed from the moment of speaking.

(8) a. *Fis* *na=i* *kara* *yu-ngg-ra-m*
yesterday 1SG=U sick 1SG-AUX-PL.NrPST.DUR
'I was sick yesterday.'

b. *na-i* *patar* *yu-nggo-f*
1SG=U cold 1SG-AUX-NrPST.NonDUR
'I suffered from a cold.'

Verbal number and argument number are intertwined. As seen in (8), the presence and absence of *-ra* may trigger different argument suffixes, *-m* 'DUR' and *-f* 'NonDUR'. (These are always correlated in this way when *–ra* expresses durative aspect.) In addition, for certain types of verbs, there is also a requirement that both verbal and argument number must have the same value. Thus, the intransitive verb *return* (here), which has the alternates *kurfenj-* (PL) and *kunonjo-* (SG) in Marori must have the plural A-vn marker *–re* when the sole subject argument is plural, as in (9)a. The A-vn *–re* is absent for singular or dual subjects, as seen in (9)b.

(9) a. *kurfenj-re-n-du*
return.PL-PL.ACT-DEIC-1PRES
'We (three or more) return here.'

b. *kunonjo-n-du*
return.NPL-DEIC-1PRES
'I or we (2) return here.'

For other eventualities, however, there is no such requirement. The transitive verb *hit*, for instance, has alternates showing object number distinction: *trm* 'hit.NPL.O' vs. *kswm* 'hit.PL.O'. With this verb, plural verbal number can be used to encode progressive aspect, in which case no plural object is required. This has to mean multiple hitting events. Thus, the plural O-vn form *kswm* can take a singular object, as seen in example (10)b. As the translation shows, the verb is aspectually extended (iterative,

[9] This meaning of plural in Marori is independent of the coding of number, which shows a three-way marking for the first person but a two-way marking for the second and third person.

progressive). Note that the argument suffix must also be synchronized for this, *-m* in (10)b but *-f* in (10)a.

(10) a. *Tomas* *Jon-i* *ter=me-f* (*sokodu/yanadu* *ngge*)
Thomas John=U hit.SG-AUX.2/3 one/two times
'Thomas hit John (once or twice).' (now or yesterday)

b. *Tomas* *Jon-i* *keswe=mi-m* (*nggujen ngge*)
Thomas John=U hit.PL-AUX.M-2/3NrPST.DUR(several times)
'Thomas hit John several/many times', or 'Thomas was hitting John.'

3.4 Event plurality and extended aspect

It has been discussed in the literature that the number of objects measures out the aspectuality of transitive verbs (Tenny 1992, 1994), e.g., *built one house* is aspectually bounded, whereas *built houses* is not. In languages like Marori where verbal number is encoded by alternative roots showing the number of objects, it is not surprising that a type of its verbal number, namely, the O-vn, serves as a resource for encoding aspect. The singular O-vn is for completive aspect and the plural O-vn for durative aspect, as seen in (10)a-b. Note that the plural O-vn expressing durative/progressive aspect as in (10)b does not require a plural object.

Of course, the plural O-vn can also appear with a plural object in durative aspect as in (11)a and in non-durative aspect as in (11)b. The two require different argument suffixes: *-m* (durative) vs. *-f* (non-durative). There seems to be no clear difference in meaning between the two other than that the first appears to be more extended than the latter. The grammar of Marori, however, treats them differently in terms of marking.

(11) a. *Tomas* *emnde* *usindu=i* *kaswa-ma-m*
Thomas 3NSG all=U hit.PL-AUX.2/3-2/3NrPST.DUR
'Thomas hit them all.'

b. *Tomas* *nie* *yanadu=i* *kasaw-ri-ma-f*
Thomas 1NSG two=U hit.PL-1U-AUX.2/3-NrPST.NonDUR
'Thomas hit both of us.'

For intransitive predicates, since there is no Object, the plurality of subject arguments is related to the plurality of events. Hence, it is not surprising that the same marker, in this case the A-vn morpheme *-ro* (or its variants such as *-ri*), is used to encode extended aspect for intransitive predicates.[10] Thus, with the dynamic root *nggV*, all of the forms (i.e., SG, DU, and PL) for the durative aspect have a variant of *-ro*, as seen in Table 3.

[10] Note that *ri-* in (11)b is the first person plural object prefix *-i* with thematic *r-*.

However, for the non-durative aspect, the A-vn marker *-ro/-ri* only marks the plural events/arguments.

For certain inherently durative intransitive predicates with verbal number such as *sit* whose forms are *kuye*- 'sit.NPL' vs. *minggri*- 'sit.PL'[11] (Present/Near Past), both the singular and plural verbal forms are used in durative aspect, as seen in (12). Because of this, both verbs must appear with the durative suffix -*m*.

(12) a.

John	*ndu*	*fis*	*kuye-m*	*keke*
John	INT	yesterday	sit.NPL-NrPST.DUR	here

'Only John sat here yesterday.'

b.

Usin	*purfam=ndu*	*fis*	*keke*	*minggri-m*
all	person=INT	yesterday	here	sit-NrPST.DUR

'All persons sat here yesterday.'

To conclude, verbal number and aspect are related. Plural verbal number is naturally used for durative aspect. However, the verbal number and aspect are distinct categories in the grammar. There is no one-to-one correlation: e.g., singular verbal number can be also durative, as in (12)a, or plural verbal number can also be non-durative, e.g., *kei-fre-f* 'bring.PLO-PL-PST.NonDUR'.

3.5 Distributive plural

The notion of distributive plural (also called 'pluractional') expressed by the A-vn is important to note. Plural A-vn signifies multiple occurrences of events simultaneously or in overlapping temporal/spatial points by different Actor participants grammatically A/S in Marori. For example, the plural A-vn verb of *hitting* in Marori means hittings by three or more agents (typically simultaneously), irrespective of whether the object is singular or plural. If the multiple hittings are done by a non-plural agent (one or two), then the plurality of hitting is not conceptualised as distributive. In such a situation, the plural event is expressed by plural O-vn only, without plural A-vn. Before examining this point further, let us look first at the distribution of the A-vn.

The A-vn formative *-ro* shows a rather complex distribution, depending on the transitivity of the verb, tense-aspectuality (durative or non-durative, past or non-past), and the number of the object if the verb is transitive. In Marori, the aspectual type of a predicate determines the selection of the light verb or auxiliary that it can co-occur with. In what

[11] The plural verb itself, namely, *minggri*, in fact consists of three formatives *mi*-, *ngg*-, and *ri*- , with *-ri* being plural A-vn. The plural verb is formed by the stacking of formatives. This appears to be common in Marori

follows, we discuss distributive plural in intransitive predicates first, followed by distributive plural in transitive predicates.

Intransitive inchoative/action predicates take the auxiliary root *nggV* (IRR or REAL (PRES/PST)), whereas states take different auxiliary roots, depending on whether they are static or dynamic. If static, *te* 'NonPST' is used. If dynamic, related to positional posture, *mi* 'IRR/FUT' or *kuye* 'REAL' is used.

For simplicity, let us consider the A-vn *-ro* and its variants (*-ra/ri*) when they appear with *ngg* for the dynamic/inchoative predicates. This is shown in Table 3, but for the third person only. As noted, *-ro/-ra/-ri* are present in forms showing durative aspect (past or present), i.e., including non-plural arguments. In the non-durative aspect (shown in the last two columns), the plural A-vn *-ro/-ra/-ri* only occurs with plural argument number. In other words, distributive plurals are only possible when plural (S) participants are involved.

	REALIS: Extended Aspect			REALIS: NonDUR Aspect	IRR
	R.PST	Nr.PAST	PRES	PST: –f RPST:-fi	FUT
3sM	*ngg(u)-**ri**-maf*	*nggu-**ri**-m*	*nggu-ri*	*nggu-f*	*nggu*
F	*nggo-ra-mof*	*nggo-ra-m*	*nggo-ra*	*nggwo-f*	*nggwo*
DU	*nggo-ra-mof*	*nggwa-ra-m*	*nggo-ra*	*nggwo-f*	*nggwo*
PL	*nggo-ra-bof*	*nggo-ra-b*	*nggo-**ro***	*nggorfo-**ro**-f*	*nggu-**ri-m***

Table 3

The same holds for transitive predicates: distributive plural events encoded by the plural A-vn *-re* are possible with plural subjects. The verb *bring* in (5), for example, can have plural events with singular or plural O-vn roots: *nde-re* and *kei-re*. The plural verbal forms must have a plural argument suffix, however, e.g., *-men*, as exemplified in (13)a. When *-re* is absent, as in (13)b, the event might be construable as plural but not distributive, e.g., a situation where the bringing of plural coconuts is a shared action.

(13) a. *nie* *usindu* *sajer-sajer* *sokodu* *poyo=i*
1NSG all day-REDUP one coconut=U
nde-re-men *pambe*
bring.SG.O-PL-1PL.PRES there
'We all (three or more), each of us, every day bring one coconut there.'

b. *nie* *usindu* *pa* ***keyi-men*** *pambe* *poyo=i.*
1NSG all FUT bring.NSG.O-1PL.PRES there coconut=U
'We all (three or more) will bring the coconuts (>1) there.'

The notion of distributive plural events must include three or more events. Two events are not counted as distributive plural in Marori, in spite of the plurality of objects involved, as seen in (14)b where A-vn *-re* is not possible with a dual subject. A plural subject would require *-re* (i.e., *keif-re-f)*.

(14) *Emnde* *yanadu* *poyo=i* *kei-f* *nggambe*
3NSG two coconut=U bring.PL.O -2/3NrPST there
'They (2) brought coconuts there.'

Likewise, completive hittings (with a singular object) by two people assumed to be two events do not constitute a distributive plural, as seen in (15)a. In contrast, hitting by a few people does constitute a plural distributive event, in which case the plural A-vn marker *-re* is used, as in (15)b. (Note that *yanadu* glossed 'two' in (15)b is used in constructed paucal number.)

(15) a. *Yanadu* *purfam=ndu* *tembok=i* *ter-me-f*
two person=FOC wall=U hit.SG-AUX-NonDUR.NrPST
'Only the two people hit the wall.'

b. *Yanadu* *purfam=ndu* *tembok=i* *ter-mbe-re-f*
two person=FOC wall=U hit.SG-AUX-PL-NonDUR.NrPST
'Only a few people hit the wall.'

However, two actors can trigger a distributive plural meaning when the events involve at least two objects (i.e., giving rise to a total of more than three events). For example, the verb root *kick* showing no O-vn can take the A-vn *-ra* in a situation involving a dual subject with a dual (or plural) object:

(16) *Nie* *yanadu* *turpungg-ra-bon* *emnde* *yanadu=i*
1NSG two kick-PL-1NrPST.NDUR 3NSG two=U
'We two kicked them two.'

Of course, a singular actor can trigger a distributive plural meaning in individuated (i.e., telic/punctual) events involving plural objects:

(17) *Na* *emnde* *usindu=i* *turpungg-ra-bon* *fis*
1SG 3NSG all=U kick-PL-1NrPST.NPL.NDUR yesterday
'I kicked them all yesterday.'

To conclude, distributive plural must include multiple events. Distributive events marked by the A-vn morpheme *-ro* can be durative or non-durative, requiring a distinct S/A argument suffix (e.g., *-m* vs. *-f*). This serves as evidence that the A-vn morpheme (or the O-vn) is not an aspect marker and that verbal number and aspect are two distinct categories.

4 Verbal number and its interaction in the grammar

4.1 Plurality and parallelism between verbal and argument number

There is interesting parallelism between argument and verbal number in terms of the plural meaning and coding, especially for the constructed paucal, as well as number reference in negation.

As discussed in the preceding section, plural in Marori means 'three or more'. In the nominal domain, the argument number agreement dictates that an argument NP referring to 'three or more' must be encoded by plural agreement. The referent of 'two' must have either dual agreement or non-plural agreement with the verb, never plural agreement.

Likewise, in the verbal domain, events taking place 'three times or more' are encoded by the plural verbal number. This has been exemplified with the verbs *hit* and *kick* in their distributive plural meanings in the preceding section. The following show that repetitive events occurring 'twice' with a singular actor cannot take the plural A-vn *–ri*, whereas events occurring several times can.

(18) a.

Albert	*yanadu*	*ngge=du*	*turpengge-f*	*John=i*	*fis*
Albert	two	times=INT	kick.3M-3NDUR	John=U	yesterday

'Albert kicked John only two times yesterday.'

b.

Albert	*turpengg-ri-m*	*John=i*	*nggunjendumba*	*fis*
Albertkick.3M-PL-3DUR		John=U	several	yesterday

'Albert kicked John several times yesterday.'

Constructed number in the expression of paucal is observed in both verbal and nominal domains by means of the same strategy. In both domains, the constructed paucal 'several, few' is achieved by constructively augmenting *yanadu* 'two' by plural verbal number on the verb. The constructed number in the nominal domain is exemplified by (15)b. Note that without the plural verbal morphology, the argument number is dual, as seen in (15)a.

The same augmentation strategy holds in the verbal domain to mean 'several, few' events. Thus, *yanadu ngge* is not augmented in (19)a when it comes with the verb without the plural *–ri*. It means 'two times'. The same adverbial *yanadu ngge* is augmented to mean 'few/several times' (i.e., necessarily three or more times) when it comes with the plural *–ri* (19)b.

(19) a. *Thomas fek yanadu ngge nggu-f*
Thomas nod two time AUX-2/3NrPST.NonDUR
'Thomas nodded two times.'

b. *Thomas fek yanadu ngge nggu-ri-m*
Thomas nod two times AUX-PL-2/3NrPST.DUR
'Thomas nodded a few/several times.'

4.2 Verbal number and finiteness

There is no verbal number distinction in Marori non-finite clauses. The non-finite verb is typically a morphologically invariant form. Verbs that show O-vn may have distinct finite forms. For example, the finite O-vn forms for *bring* are *nde* 'SG.O' and *kei* 'NSG.O', whereas the non-finite form is *ndow*. The finite forms for *come* are *umo* and *ya* (IRR)/*seri* (REAL), whereas the non-finite form is *embiw*. The A-vn suffix -*ro* is also absent in the non-finite form.

Examples showing non-finite invariant forms of *bring* are given in (20).

(20) a. *Na Maria=i tirfo=nggo-bon*
1SG Maria=U ask.SG-AUX-1SG.NrPST

[*sokodu buku ndow mbe*]
one book bring.NF MBE
'I asked Maria to bring one book.'

b. *Na maria=i tirfo=nggo-bon*
1SG Maria=U ask.SG-AUX-1SG.NrPST

[*usin buku ndow mbe*]
many book bring.NF MBE
'I asked Maria to bring many books.'

The invariant non-finite forms without the A-vn -*ro* in Marori suggests that verbal number alternation in Marori is grammatically constrained. It is not purely lexical of the English type *kill* vs. *massacre*.

4.3 Verbal number and reciprocal

Reciprocals in Marori are expressed by the affix -*n*-. It is affixed/infixed (phonologically conditioned) to a verb with non-singular O-vn, as expected, since the reciprocal is necessarily conceived as more than one object.

For example, for the verb *hit*, the reciprocal verb takes the non-singular O-vn root *ksw*-. Consider (21)a with the plural A-vn –*ro* and (21)b without it.

The first one signifies distributive reciprocals (i.e., more than one pair involved), whereas the second means only one pair is involved.

(21) a. *Ka=yofo! Emnde usindu koswo-n-mb-ro-∅*
2=see 3NSG all hit.PL-RECIP-AUX.PL.O-PL-3
'Look. They all (>2) are hitting each other.'

b. *Ka=yofo! Emnde yanadu kaswa-n-ma-∅*
2=see 3NSG two hit.PL-RECIP-AUX.NPL.O-3
'Look. They (2) are hitting each other.'

Non-finite reciprocal clauses have their verbs marked by the reciprocal marker -*n*-, but both the A-vn morpheme -*ro* and the argument agreement suffix are absent. Consider (22) where the invariant form with the reciprocal *koswonmow* is used irrespective of whether the argument is dual as in (22)a, or plural as in (22), a third person as in (22)a-b, or a second person as in (22)c.

(22) a. *Na tir-ngga-bon emde yanadu=i koswo<n>mow mbe*
1SG ask.NSG-AUX-SG.NrPST 3NPL two=U hit-RECIP-NF COMP
'I asked them two to hit each other.'

b. *Anton kie usindu=i tir-ngga-f*[12] *koswo<n>mow mbe*
Anton 2NSG two=U ask.NSG-AUX-PST-3 hit<RECIP>NF COMP
'Anton asked them all to hit each other.'

c. *Anton kie yanadu=i tir-ngga-f koswo<n>mow mbe*
Anton 2NSG two=U ask.NSG-AUX-PST-3 hit<RECIP>NF COMP
'Anton asked you two to hit each other.'

Note that a finite reciprocal verb requires an argument suffix. It is -∅ for the third person in (21). For the first person, it is –*bon*, as in (23)a (dual, non-distributive without -*ro*) and (23)b (distributive, with –*ro*). In contrast to (23)b, the non-finite distributive reciprocal equivalent in (23)c shows no argument suffix -*bon* and no A-vn -*ro*.

(23) a. *Na Thomas=fi tafa<n>ja-bon*
1SG Thomas=and meet<RECIP>-1NrPST
'Thomas and I met (each other).'
(reciprocal, dual: no A-verbal number morpheme -ro)

[12] alternatives: *tiranggraf*, *tiringraf*

b. *Nie* *usindu* *tamba* *tofo<n>j-ro-bon* *pasar=ku* *fis*
1NSG all already meet.PL<RECIP>-1NrPST market=LOC yesterday
'We all met each other in the market yesterday.'

c. *John* *nie* *usindu=i* *tirir-ngga-f* *tofo<n>jow* *mbe*
John 1NSG all=U ask.NSG-AUX-NrPST meet.PL<RECIP>NF COMP
'John asked us all to meet each other.'

To conclude, the status of reciprocal and verbal number marking in Marori is not the same. The reciprocal marker is purely morphosemantic in nature, not grammatically constrained by finiteness. In contrast, verbal number (O- or A-vn) marking is morphosyntactic in nature, grammatically constrained by finiteness. In the absence of plural coding, plural meanings in embedded non-finite clauses in both verbal and nominal domains can only be arrived at by means of larger context in relation to the main clause.

5 LFG Analysis

While verbal number and its interaction with argument number within the TAM system in Marori is quite complex, its constraints can be straightforwardly analysed within a unification-based framework such as LFG (Bresnan 2001; Dalrymple 2001). The essence of the analysis is to capture the two kinds of number (argument/nominal and verbal) and their aspectual properties as part of an integrated system in the grammar of Marori.

One of the challenges is how to handle the parallelism between nominal and verbal domains, particularly in capturing constructed number that applies to both domains in the same manner. An outline of the analysis proceeds as follows.

I propose that the relevant NUM features are the same features for both nominal and verbal number. Building on earlier work on argument number in Marori (Arka 2011) and studies on underspecification (Dalrymple and Kaplan 2000; Dalrymple, King, and Sadler 2009; Sadler 2010), I adopt a composite NUM feature analysis, with [+/–SG], [+/–PL], and [+/–AUG], as shown in Table 4. [+/–AUG] (see also Harbour 2007) is to capture the augmentation strategy employed in constructing paucal in Marori and other languages (Arka 2011).

Questions remain as to the precise meaning of these features, the extent of their universality, as well as their locus in LFG's model of parallel structures. Discussing all of these questions in detail is beyond the scope of this paper. What is clear is that [+PL] in Marori means an aggregate of 'three or more entities or events', whereas [–PL] means 'either one or two entities/events'. [+SG] means 'a single individuated entity/event', whereas [–SG] means 'an aggregate of two or more'. [+AUG] means 'augmentation of the semantic space of the [SG, PL] number features'. Thus, [–SG, –PL,

+AUG] means ‘augmentation of the semantic space of [–SG, –PL] (i.e., ‘two’)’, conceptually referring to as ‘few, several, relatively small in number’. In contrast, [–SG, –PL, –AUG] means that there is no augmentation of [–SG, –PL], i.e., precisely ‘two’ (dual).

	FEATURES		
CATEGORY	[+/– SG]	[+/–PL]	[+/–AUG]
SINGULAR	+	–	–
DUAL	–	–	–
LIMITED PLURAL	–	–	+
GENERIC PLURAL	–	+	–
LARGE PLURAL	–	+	+

Table 4: Number features

For simplicity, I assume that these are *f*-structure features. In Marori, two of these features, namely, the [PL] and [SG] features, are also morphological features available in this language. (I assume lexical integrity where morphological features are part of word-internal information, not visible to syntax.) Thus, the pronominal forms that show a singular vs. non-singular distinction carry [+/–SG] (interpreted as morphological as well as *f*-str features). The plural A-vn morpheme *-ro* carries the morphological and syntactic [+PL] feature as well as a syntactic feature [+AUG]. This [+AUG] feature is visible in syntax; that is, it can interact with other number features coming from a node outside the verbal unit.

Other relevant features, such as ASPECT and FINITE, should be entered into the system, too. The feature ASP(ECT) captures the aspectual properties of verbal number. (It is a syntactic (f-str) and semantic feature.) The ASP feature has binary values, [+/–DUR]. [+DUR] means durative aspect, whereas [–DUR] means completive or telic aspect. The FINITE feature is to capture the finiteness constraint of verbal number, as discussed in section 4.2.

Morphemes participating in constructions involving argument and verbal number, carry rich relevant information in their entries. The agreement suffixes *-m* ‘2/3.NPL.DUR.NrPST’ and *-f* ‘2/3.NonDUR.NrPST’, for example, can be formulated to have the following entries.

(24) a. *-m*
(↑SUBJ NUM PL) = −
(↑SUBJ PERS) = {2|3}
(↑ASP DUR) = +
(↑TNS) = NrPST

b. *-f*
(↑ASP DUR) = −
(↑SUBJ PERS) = {2|3}
(↑TNS) = NrPST

To account for constructed verbal number, as exemplified in (25), we also need the entry of the A-vn formative *-ri,* which is partially shown in (26).

(25) *Thomas fek yanadu ngge nggu-ri-m*
Thomas nod two times AUX-PL-2/3NrPST.DUR
'Thomas nodded a few/several times.'

(26) *-ri*

~(↑OBJ)
{(↑SUBJ NUM PL)=+
|
(↑SUBJ NUM PL)= − ⇒ (↑ASP DUR)= +
(↑SUBJ NUM SG)= + ⇒ (↑SUBJ GEND)= M
(↑SUBJ PERS)=3
}
({↑ADJ ∈ SPEC NUM AUG)=+ | (↑SUBJ NUM AUG)=+})

The entry in (26) says that if *-ri* is in an intransitive structure (i.e., the grammatical function constraint of ~(↑OBJ)), it can take either a plural or non-plural subject. The specifications contain a conditional 'if then' rule indicated by the arrow (⇒), e.g., (↑SUBJ NUM PL)= − ⇒ (↑ASP DUR)=+ means that *-ri* used with a non-plural subject triggers a durative interpretation. It optionally carries an augmented feature: [AUG +] associated with either an ADJ(unct) or SUBJ path. This allows it to interact with other features in the unification process, e.g., with *yanadu* 'two' to create an augmented constructed 'paucal' number. The augmentation can be captured as follows.[13]

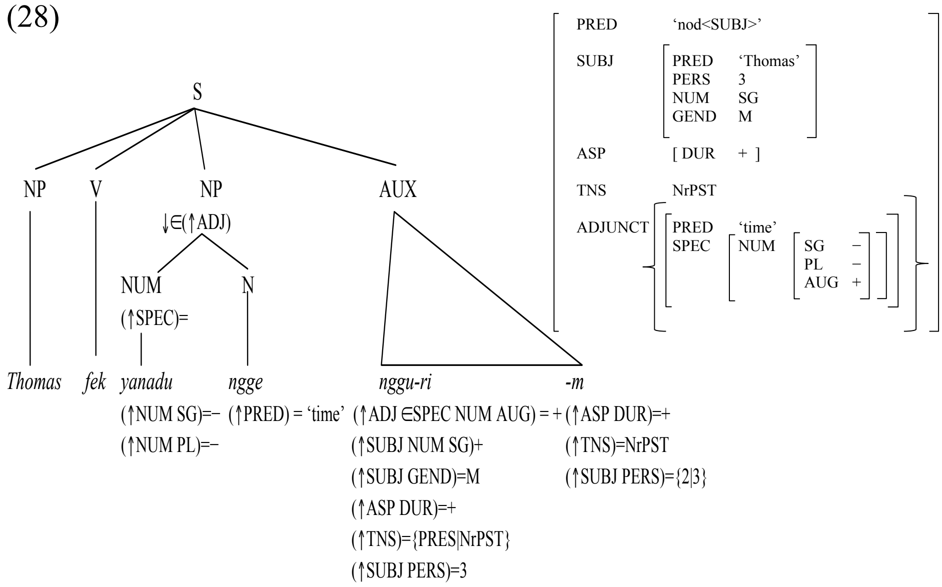
(27) Augmentation:

yanadu		*-ri*		'few, several'
[−SG, −PL]	U	[+AUG]	=	[−SG, −PL, +AUG]

The c-structure and f-structure of sentence (25) showing constructed paucal verbal number can be shown in (28).

[13] Note that I analyse the augmentation as belonging to *f*-str, i.e., the relevant features come from separate nodes, an NP argument/adjunct, and predicate head, in syntax. However, constructed number (dual, but not paucal) is also possible word-internally in Marori. It remains debatable whether it is desirable to have a different analysis of the constructed number, e.g., with underspecified semantic analysis without syntactic ambiguity as proposed in this paper. I leave this for future research.

(28)



Since the subject is singular, then the plural A-vn *-ri* is associated with durative aspect (i.e., due to the conditional rule of *-ri*). The durative aspect then requires the durative argument suffix *-m*, carrying (↑ASP DUR)=+. Given the c-str annotation of the adjunct NP *yanadu ngge*, the relevant number features of *yanadu* (↑NUM SG)= −, (↑NUM PL)=− end up as the values of ADJUNCT and therefore unify correctly with the feature (↑ADJ SPEC NUM AUG) = + of *–ri* in the same ADJUNCT path. This results in the intended reading, namely, paucal in relation to the verbal number: 'few occurrences of the event of nodding'.

6 Conclusion

This paper has discussed how verbal number and argument number are distinguished in Marori as well as how they interact with each other and with other grammatical phenomena such as aspect, finiteness, and reciprocals. Two kinds of verbal number in Marori, the O-vn and A-vn, can be distinguished. The O-vn shows suppletive alternations with the verbal roots and morphological alternation with *-on/-nde* with adjective roots. The A-vn is morphologically encoded by *-ro* (and its variants). The A-vn is used to encode distributive plural showing multiple events, involving multiple actors, and/or multiple objects. It is also used to express aspectually extended/durative events. Aspect (and also tense) and number information is also carried by the argument suffix. Verbal number alternations, therefore, impose a co-occurrence constraint on argument suffixes: e.g., plural O-vn/A-vn encoding durative aspect must have a durative subject argument suffix.

This compatible requirement can be easily captured in LFG's unification-based architecture. It has also been demonstrated in this paper that more complex issues of number in this language, including the parallelism between verbal and nominal domains in paucal, can be straightforwardly captured in LFG. There remain theoretical and empirical issues, however. Theoretically, the precise nature and analysis of number features remain to be worked out: to what extent the features are morphological, syntactic, and semantic. Empirically, more research is needed to map out the variation in number systems across languages, in particular in the meaning of plural.

References

Arka, I Wayan. 2011. Constructive number systems in Marori and beyond. In t*he Proceedings of th*e *International Lexical Functional Grammar (LFG2011) conference*, edited by M. Butt and T. H. King. The University of Hong Kong, Hong Kong, 19 July 2011: CSLI, http://cslipublications.stanford.edu/LFG/16/lfg11-toc.html.

Bresnan, Joan. 2001. *Lexical functional syntax*. London: Blackwell.

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press.

Dalrymple, Mary. 2001. *Lexical functional grammar*: *syntax and semantics*. San Diego: Academic Press.

Dalrymple, Mary, and Ronald M. Kaplan. 2000. Feature indeterminacy and feature resolution. *Language* 76 (4):759-798.

Dalrymple, Mary, Tracy Holloway King, and Louisa Sadler. 2009. Indeterminacy by underspecification. *Journal of Linguistics* 45:31-68.

Durie, Mark. 1986. The grammaticization of number as a verbal category. In *Proceedings of the 12th annual meeting of the Berkeley Linguistics Society*, edited by V. Nikiforidou, M. VanClay, M. Niepokuj and D. Feder.

Harbour, Daniel. 2007. *Morphosemantic number: from Kiowa noun classes to UG number features*. Dordrecht, the Netherlands: Springer.

Mithun, Marianne, ed. 1988. *Lexical categories and the evolution of number marking*. Edited by M. Hammond and M. Noonan, *Theoretical morphology: approaches in modern linguistics*. San Diego: Academic Press.

Sadler, Louisa. 2010. Indeterminacy, complex features and underspecification. *Morphology* 21 (2):379-417.

Tenny, C. 1992. The Aspectual interface hypothesis. In *Lexical Matters*, edited by I. Sag and A. Szabolcsi. Stanford: CSLI.

Tenny, C. 1994. *Aspectual roles and the syntax-semantics interface*. Dordrecht: Kluwer.

Veselinova, Ljuba. 2008. Verbal number and suppletion. In *The World Atlas of Language Structures Online*, edited by M. Haspelmath, M. Dryer, D. Gil and B. Comrie. Munich: Max Planck Digital Library.

AFFECTED EXPERIENCERS AND MIXED SEMANTICS IN LFG/GLUE

Doug Arnold and Louisa Sadler
University of Essex

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

Bosse, Bruening and Yamada (2012) (BBY) provides a study of several constructions involving 'non-selected' arguments, and outlines an approach to the syntax and semantics of one such construction: the Affected Experiencer (AE) construction. The syntactic analysis relies on abstract functional projections and particular assumptions about configurational syntax. We show how an account may be given without these syntactic assumptions. Semantically, BBY argue that AEs may contribute both at-issue content and conventional implicatures, which raises interesting issues for the approach of e.g. Potts (2005). We explore some consequences of their semantic analysis and show that it faces a number of difficulties.

1 Introduction

In a number of recent papers Bosse and others have presented analyses of a variety of constructions involving 'non-selected' arguments (i.e. complements that do not intuitively fill lexical argument slots), including affected experiencer (AE) constructions, external possessor constructions, and benefactives, arguing for the existence of a number of subtypes (see e.g. Bosse et al., 2012; Bosse and Bruening, 2011; Bosse, 2011), and providing relatively detailed syntactic and semantic analyses. In particular, Bosse et al. (2012) presents an appealing analysis of an affected experiencer (AE) dative construction in German, exemplified in (1), below. Semantically, the approach is based on that of Potts (2005), though it purports to raise some fundamental problems for Potts. Syntactically, the approach relies on abstract/functional projections, and particular assumptions about configurational syntax.

In the first part of this paper, we explore whether the insights of Bosse et al.'s analysis can be expressed, without these syntactic assumptions, in an LFG/glue-based implementation of Potts' ideas – specifically the approach presented in Arnold and Sadler (2010), Arnold and Sadler (2011). We will see that the answer here is positive. However, it turns out that when the analysis is explored in more detail, the initial appeal of the approach evaporates. The second part of the paper demonstrates this, and shows that some of the theoretical points that Bosse et al. seek to make about Potts's approach do not bear close scrutiny.

In more detail, the paper is structured as follows. Section 2 presents Bosse et al.'s account of the AE construction in German, with some observations about other languages, including Hebrew and Japanese. We will pay particular attention to the syntax that Bosse et al. propose, which is highly configurational, and involves a rich array of functional categories, and the semantics, which Bosse et al. believe motivates some interesting modifications of Potts' ideas.

†We are grateful to several people for insightful comments and stimulating discussion, notably, Boban Arsenijević, Ash Asudeh, Miriam Butt, Mary Dalrymple, Gianluca Giorgolo, Dag Haug, Tracy Holloway King, Helge Lødrup, Chris Potts, and Adam Przepiórkowski, as well as several anonymous referees, and other participants at LFG 2012 in Denpasar, Bali. But none of these people can be blamed for deficiencies in what follows.

Section 3 presents an implementation of Bosse et al.'s descriptive insights in the framework of LFG, specifically a variant of the 'Pottsian LFG' approach presented in Arnold and Sadler (2010), Arnold and Sadler (2011). This framework is briefly summarized in Section 3.1; the actual analysis is presented in 3.2. We will see that while Bosse et al.'s insights can be adequately captured in this framework, close examination shows there is reason to think that the modifications to Potts' framework that they propose are problematic.

2 AE Constructions: Bosse *et al*'s Analysis

2.1 Basic Properties

A basic example of the AE construction in German can be seen in (1). Like its English equivalent, *zerbrechen* ('break') can occur with a subject (denoting the agent), and a direct object (denoting the patient), as in (2). In (1) it occurs with an additional complement, *Chris*, which denotes an entity which is some way affected by the breaking event. This is normally glossed with the preposition *on*, presumably because it has some similarity with the usage of *on* in examples like *My laptop has just died on me*, *They have closed the local shop on us*. Though it is not obvious from (1), because a proper noun like *Chris* does not show case marking, this extra complement is in fact dative, as can be seen when a pronoun is used, as in (3).[1]

(1) Alex zerbrach Chris Bens Vase.
    Alex broke Chris.DAT Ben's vase
    Alex broke Ben's vase 'on Chris'. (i.e. and this affected Chris)

(2) Alex zerbrach Bens Vase.
    Alex broke Ben's vase
    Alex broke Ben's vase.

(3) Alex zerbrach mir Bens Vase.
    Alex broke me.DAT Ben's vase
    Alex broke Ben's vase 'on me'.

The AE dative complement must be both a potential experiencer (thus, e.g. sentient) and actually affected. For example, Bosse et al. point out that (4) is unacceptable if Paul was

[1]Superficially, AE constructions are often similar to, and can be confused with, instances of the external possessor construction. An example like the following is ambiguous – *mir* can be interpreted as an AE, giving the meaning 'She cleaned the suit on me' (i.e. she cleaned it, and the cleaning affected me), but it can also be interpreted as an 'external possessor' (EP), in which case the interpretation will be just 'She cleaned my suit'.

Sie säuberte mir den Anzug.
she clearned me.DAT the suit
She cleaned the suit 'on me'. AE
She cleaned my suit. EP

In this paper, examples are always intended to be instances of the AE construction.

already dead when his mother died,

(4) *Dann starb ihm auch seine Mütter. (Context: Paul died first)
Then died *him*.DAT also his mother
Then his mother died 'on him', too.

Bosse et al. suggest that in some languages (e.g. French, Hebrew), what we will call the 'AE content' (i.e. with respect to (1), the assertion that the vase breaking affected Chris) is not part of the 'at issue' content at all, contributing only to what Potts calls 'conventionally implicated' (*ci*) content (Potts, 2005). For example, it cannot be questioned or negated, and is generally rather strictly separated from the normal *at-issue* content. However, they suggest that in other languages (including Japanese, Albanian and German) AE datives contribute **both** *at-issue* content and *ci*-content. For example in (5), the assertion that the vase-breaking matters to Chris appears to be *ci* content, since it can escape the negation – (5) conveys the idea that though the breaking did not occur, it would have mattered to Chris.

(5) Alex zerbrach Chris Bens Vase nicht.
Alex broke Chris.DAT Ben's vase not
Alex didn't break Ben's vase 'on Chris'. (But it would matter to Chris.)

Similarly, consideration of (6) suggests that the AE content is not part of the question: notice in particular, that it would be wrong to answer 'Nein' ('No') to (6) if Alex did break Ben's vase, but Chris does not care. This information cannot be conveyed in response to (6) with any simple answer – it requires a fuller explanation.

(6) Zerbrach Alex Chris Bens Vase?
broke Alex Chris.DAT Ben's vase
Did Alex break Ben's vase 'on Chris'?

But in other ways the AE seems to contribute *at-issue* content, as witness the way the AE itself can be questioned as in (7) (which is not generally possible with *ci* content), can contribute to the truth conditions of a conditional as in (8), and can bind an argument in the at-issue domain, as in (9).[2]

(7) Wem hat Alex Bens Vase zerbrochen?
who.DAT has Alex Ben's vase broken
On whom did Alex break Ben's vase?

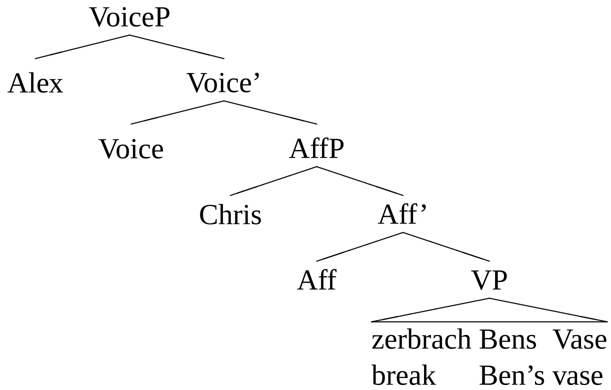
(8) Wenn Lisa ihrem Mann den Anzug lobt, dann bekommt Jan €100 von ihm.
if Lisa her.DAT husband the suit praises, then get Jan €100 from him
If Lisa praises the suit 'on her husband', then Jan will get €100 from him.

[2]Bosse et al. claim that the truth conditions of (8) are such that Jan will only get the €100 if Lisa praises the suit *and* her husband is affected by the praising. The praising alone is insufficient.

(9) Ich habe jedem Jungen$_i$ seine$_i$ Vase zerbrochen.
I have every.DAT boy his vase broken
I broke his vase 'on every boy.'

2.2 Bosse *et al*'s Analysis

Bosse et al.'s account of these data involves a number of functional projections, as in (10), notably VoiceP and AffP ('Aff' for *affected*).

(10)


The semantics of Aff' and Voice' are derived by applying the semantics of Aff and Voice to the semantics of their sisters, the semantics of AffP and VoiceP are derived by applying the semantics of Aff' and Voice' to their NP sisters. That is, schematically:

(11) a. $[\![VoiceP]\!]^{M,g} = [\ [\![Voice]\!]^{M,g}(\ [\![AffP]\!]^{M,g})\ ](\ [\![Alex]\!]^{M,g})$
b. $[\![AffP]\!]^{M,g} = [\ [\![Aff]\!]^{M,g}(\ [\![VP]\!]^{M,g})\ ](\ [\![Chris]\!]^{M,g})$

This is most easily appreciated by way of an example, making the (false) assumption that AE content is contributed to the *at-issue* dimension of meaning (we will correct this directly below). Suppose the interpretation of the lowest VP *zerbrach Bens Vase* is as in (12) (intuitively, it denotes the set of breaking events that involve Ben's vase as Theme – the set of events where Ben's vase gets broken).

(12) $[\![VP]\!]^{M,g} = \lambda e.break(e)\&Thm(Ben's\ vase)(e)$

The interpretation of AffP is derived from this as in (13) (ignoring for the moment the distinction between *ci* and *at-issue* dimensions of meaning).

(13) $[\![AffP]\!]^{M,g} = \ [\![Aff]\!]^{M,g}(\ [\![VP]\!]^{M,g})](\ [\![Chris]\!]^{M,g})$
$= [\lambda P_{vt}.\lambda x.\lambda e.P(e)\&\exists e'(exp(e')\&Exp(x)(e'))$
$\quad \forall e''(P(e'') \rightarrow Source(e'')(e')](\lambda e.break(e)\&Thm(Ben's\ vase)(e))](Chris)$
$= \lambda e.break(e)\&Thm(Ben's\ vase)(e)\&\exists e'(exp(e')\&Exp(Chris)(e'))$
$\quad \forall e''((break(e'')\&Thm(Ben's\ vase)(e'')) \rightarrow Source(e'')(e'))$

Intuitively, this adds (i) the assertion that there exists an experiencing event e' where Chris is the experiencer and (ii) the assertion that if any event at all is a breaking event involving

Ben's vase, then that event will cause (be the source of) the aforesaid experiencing event. Very roughly, it adds the information that Chris would care about Ben's vase getting broken, and that an event of Chris experiencing this emotion actually occurred.

The interpretation of VoiceP is as in (14).

(14) $[\![VoiceP]\!]^{M,g} = [\![[Voice]\!]^{M,g}([\![AffP]\!]^{M,g})]([\![Alex]\!]^{M,g})$
$= \lambda e.break(e)\&Thm(Ben's\ vase)(e)\&Agt(Alex)(e)\&\exists e'((exp(e')\&Exp(Chris)(e'))$
$\forall e''((break(e'')\&Thm(Ben's\ vase)(e'')) \rightarrow Source(e'')(e'))$

Intuitively, this just adds the information that Alex is the agent of the breaking.

For simplicity, this explanation has assumed that all the content is contributed to the *at-issue* dimension. This is incorrect, but it is easily corrected. Bosse et al. follow Potts in separating *at-issue* and *ci* content with an uninterpreted operator; in the case of Bosse et al., this is a colon. The proper meaning derivation is then as follows (15), with the colon highlighted at line endings.

(15) $[\![VP]\!]^{M,g} = \lambda e.break(e)\&Thm(Ben's\ vase)(e)$

(16) $[\![AffP]\!]^{M,g} = [\![Aff]\!]^{M,g}([\![VP]\!]^{M,g})]([\![Chris]\!]^{M,g})$
$= [\lambda P_{vt}.\lambda x.\lambda e.P(e)\&\exists e'(exp(e')\&Exp(x)(e'))\ :$
$\forall e''(P(e'') \rightarrow Source(e'')(e')](\lambda e.break(e)\&Thm(Ben's\ vase)(e))](Chris)$
$= \lambda e.break(e)\&Thm(Ben's\ vase)(e)\&\exists e'(exp(e')\&Exp(Chris)(e'))\ :$
$\forall e''((break(e'')\&Thm(Ben's\ vase)(e'')) \rightarrow Source(e'')(e'))$

(17) $[\![VoiceP]\!]^{M,g} = [\![[Voice]\!]^{M,g}([\![AffP]\!]^{M,g})]([\![Alex]\!]^{M,g})$
$= \lambda e.break(e)\&Thm(Ben's\ vase)(e)\&Agt(Alex)(e)\&\exists e'((exp(e')\&Exp(Chris)(e'))\ :$
$\forall e''((break(e'')\&Thm(Ben's\ vase)(e'')) \rightarrow Source(e'')(e'))$

The effect of tense interpretation will be to existentially bind the 'main' event variable *e* (as well as adding information about time reference, which we ignore), giving a two dimensional interpretation as in (18).

(18) $\exists e(break(e)\&Thm(Ben's\ vase)(e)\&Agt(Alex)(e)\&$
$\exists e'(exp(e')\&Exp(Chris)(e'))\ :$
$\forall e''((break(e'')\&Thm(Ben's\ vase)(e'')) \rightarrow Source(e'')(e')))$

The *at-issue* content here asserts the existence of (i) a breaking event *e* where Alex is the Agent, and Ben's vase is the Theme, and (ii) an experiencing event e', where Chris is the experiencer. The *ci* content asserts that any such breaking event (i.e. any breaking event involving Ben's vase) would be the source of e'.

This semantics is plausible, so far as it goes, and seems to reflect the basic intuition about the meaning of this example (viz that Alex broke Ben's vase, and that Chris is affected by this).

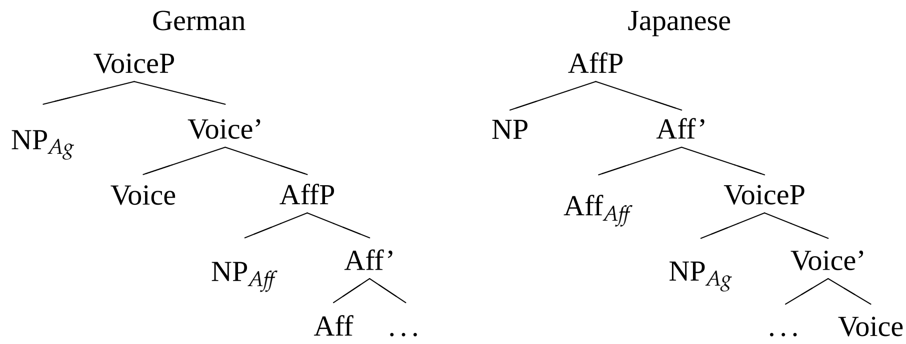
At this point, at least two points are worth developing further. The first relates directly to cross-linguistic variation. Notice that with respect to these German examples, 'any such breaking event' means any breaking of Ben's vase (by Chris, or anyone else). Bosse et al.

suggest that this reflects one parameter of cross linguistic variation. They suggest that for example in Japanese, an example like (19), which roughly corresponds to (1), conveys the more precise meaning that any breaking of Ben's vase *by Alex* would affect Chris.[3]

(19) Chris-ga Alex-ni Ben-no kabin-o kowas-are-ta. (Japanese).
Chris-Nom Alex-Dat Ben-Gen vase-Acc break-Aff-Past
Alex broke Ben's vase 'on Chris'.

That is, in Japanese, the *ci* content includes the agent. To deal with this, Bosse et al. assume that there is parametric variation in the height at which the Aff head attaches. In German VoiceP is higher than AffP, in Japanese it is the other way round. Because VoiceP is responsible for introducing the Agent into the semantics, this captures the variation with respect to whether the Agent part of the *ci* content or not.

(20)



The second point is also a matter of cross-linguistic variation, but the real interest is the formal issue raised by this construction in a language like German. Bosse et al. claim that in some languages AE content is contributed entirely in the *ci* dimension. This is formally unproblematic for the two-dimensional, Pottsian, approach. But German, where some content seems to be in the *at-issue* dimension, and some in the *ci* dimension, and where, in particular, there is variable binding across the dimensions (cf. in (18) the variable e' is associated with a quantifier in the *at-issue* dimension, and also appears after the colon in the *ci* dimension), is a serious challenge for Potts.[4]

The questions are: Can we provide an LFG implementation that deals with this data, without the projections? Can we get a neat account of the parametric variation across languages? What are we to make of the challenge this construction in German seems to pose for the Pottsian enterprise?

[3]Example (19) is similar to one given by Bosse et al., and there are some complications, which we will ignore. In particular, there is a potential complication due to voice (where the German example is in the active voice, the Japanese is morphologically passive, with the Agent (*Alex*) marked with *ni*, and the Experiencer (*Chris*) marked as nominative: Bosse et al. assume the 'passive' morphology is spell out of the Voice head.

[4]Bosse et al. note that their analysis also involves a single item introducing elements of meaning in both *ci* and *at-issue* tiers of meaning, which Potts had claimed was not possible. Since we think Potts' claim has been convincingly challenged elsewhere, (e.g. McCready, 2010; Sawada, 2011), we will not pursue this issue here.

3 An LFG Implementation

3.1 Basic Framework

In this section we introduce the formal and conceptual framework in which we will investigate these questions. On the morpho-syntactic side, our assumption are entirely conventional LFG (e.g. Dalrymple, 2001). On the semantic representation side, we will assume a Discourse Representation Theory (DRT) style semantics (using a version of DRT augmented with a λ operator).[5] The syntax-semantics interface uses the standard LFG/Glue logic approach (e.g. Dalrymple, 2001; Asudeh, 2004, 2012), as modified by Arnold and Sadler (2010, 2011) to provide a Potts style account of appositive constructions.[6]

For the sake of familiarity, we exemplify with reference to non-restrictive ('appositive') relative clauses (ARCs), as in (21a).

(21) a. Kim believes that linguists, who dislike Maths, are stupid. [ARC]
b. Kim believes that linguists who dislike Maths are stupid. [RRC]

Compared to a restrictive relative clause (RRC), as in (21b), ARCs display a number of distinctive syntactic and semantic properties. Most obviously, in (21b), the relative clause *who dislike Maths* is used to restrict the denotation of *linguists* so that the NP *linguists who dislike Maths* denotes an intersection. By contrast, (21a) is about all linguists, not some subset thereof. This provides a useful test, since in the former, but not the latter, one can infer the existence of a 'contrast' set (linguists who do not dislike Maths) and pick this out anaphorically with an expression like 'other kinds'. Compare:

(22) a. Kim believes that linguists, who dislike Maths, are stupid. #Other kinds she regards as cool. [ARC]
b. Kim believes that linguists who dislike Maths are stupid. Other kinds she regards as cool. [RRC]

A less obvious, but none-the-less well known, property of ARCs is that they generally appear to be semantically scopeless, or interpreted with wide scope.[7] This can be seen with respect to (21b)/(21a). Notice that in the case of the RRC the interpretation involves Kim having a belief that '(some) linguists don't understand first order predicate calculus (FOPC)', or something equivalent, and must therefore involve Kim having, in the widest sense, some notion of what FOPC is. This is not required in the case of the ARC, where the (false) assertion that 'linguists do not understand FOPC' is associated with the speaker, and need not form any part of Kim's beliefs. A natural account of this is that in the case of the ARC, the content of the ARC is interpreted outside the scope of the belief operator.

[5]For DRT, see e.g. Kamp and Reyle (1993). For versions of DRT that have a λ operator, see e.g. Muskens (1996).

[6]See Giorgolo and Asudeh (2011) for an alternative approach to these issues.

[7]It is now clear that though this is generally true, it is not invariably true, and there are many situations where ARCs and other appositives display narrow scope. See Arnold and Sadler (2011) and references there.

(23) a. Kim believes that linguists, who don't understand FOPC, are stupid.
b. Kim believes that linguists who don't understand FOPC are stupid.

This phenomenon is not restricted to propositional verbs, but can be observed with respect to a wide range of scope related phenomena. To take just two other examples: in (24a) the issue of linguists' understanding of FOPC is not part of the question (which is, essentially, 'Are linguists stupid'), but the content of the RRC is part of the question in (24b). Similarly, (25a), where there is a negative polarity item (*any*) inside the ARC, is ungrammatical. Plausibly this is because the ARC, and hence the negative polarity item, is outside the scope of negation. Compare the fully acceptable (25b), where *any* is in an RRC, and in the scope of negation.

(24) a. Are linguists, who understand FOPC, stupid? [ARC]
b. Are linguists who understand FOPC stupid? [RRC]

(25) a. *We did not write to the customers, who had any complaints.
b. We did not write to the customers who had any complaints.

Potts' account of these phenomena involves having two dimensions of meaning: a dimension of 'normal' '*at-issue*' meaning, and a second dimension of 'conventionally implicated' (*ci*) content. The content of RRCs belongs to the *at-issue* dimension, the content of ARCs belongs to the *ci* dimension. Potts' account involves a very strict separation of these dimensions of meaning. In particular, the way material in the two dimensions is assigned semantic types guarantees that nothing in the *at-issue* dimension can access anything in the *ci* dimension. Hence, *ci* content is always outside the scope of *at-issue* operators (e.g. negation, question operators, propositional verbs).

The LFG/Glue implementation of these ideas presented in Arnold and Sadler (2011) (which is a refinement of that in Arnold and Sadler (2010)) differs from Potts' in two main ways. First, it uses the projection architecture of LFG so that the separation of semantic content into two types (*at-issue*, and *ci*) is not necessary. The second difference is that Potts assumes that *at-issue* and *ci* content are always entirely separate: the only commonality is that they are interpreted in the same model. Arnold and Sadler (2011) point out that on standard LFG/Glue assumptions about anaphora (Dalrymple, 2001; Asudeh, 2004, 2012, e.g.), this should make ARCs and other appositives anaphoric islands, which they clearly are not, as witness the following, where one can see anaphora into and out of ARCs:

(26) a. Pissarro, *who Matisse$_i$ met in 1898*, encouraged him$_i$ greatly.
b. Matisse$_i$ was greatly encouraged by Pissarro, *who he$_i$ met in 1898*.

To deal with this, Arnold and Sadler (2011) propose that *at-issue* and *ci* content should be integrated 'at the top' (i.e. the final representation of a sentence should be a representation where *at-issue* and *ci* content is conjoined).[8]

The basic ideas of Arnold and Sadler (2011, 2010) can be seen in Figure 1, which repre-

[8]Conjunction is empirically the correct interpretation: *Kim, who Sam dislikes, left* means roughly the same as the conjunction *Kim left,* and *Sam dislikes Kim*.

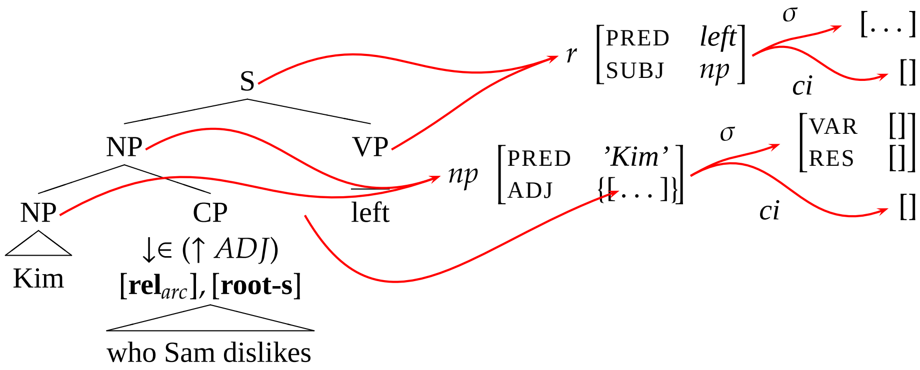


Figure 1

sents (27). The c- and f-structures are entirely conventional, and in fact identical to what one would have for a restrictive relative (though one would not normally have a restrictive relative with a proper noun, of course). In particular, the ARC is a normal adjunct, fully integrated into the c- and f-structures.

(27) Kim, who Sam dislikes, left.

The semantics is more interesting. Notice that as well as the standard σ projection, which introduces normal *at-issue* semantic resources, there is an additional *ci* projection, which introduces 'conventionally implicated' content. Thus one has resources np_σ and np_{ci}, corresponding to the two semantic projections of the f-structure np.

The resources associated with the projections of *Kim*, and *left* are standard:

(28) [**Kim**] $Kim : np_\sigma$

(29) [**leave**] $\lambda X.left(X) : np_\sigma \multimap r_\sigma$

We associate two semantic resources with the ARC: [**rel**$_{arc}$], and [**root-ci**]. The latter is given in (30). Its role is to combine the *ci* resource that the ARC introduces with the semantics of the root S, which we will designate as r_σ, conjoining the associated meanings, as $p \wedge q$.[9]

(30) [**root-ci**] $\lambda q.\lambda p.(p \wedge q) : np_{ci} \multimap [r_\sigma \multimap r_\sigma]$

We will not present [**rel**$_{arc}$] here, because it is not relevant, what is relevant is the resource that it produces when it combines with the resource associated with the host NP, [**Kim**]. This is given in (31).

(31) $Kim \times dislikes(Sam, Kim) : np_\sigma \otimes np_{ci}$

[9] The resource r_σ appears in both (29), and (30), but in the former r is the f-structure of *leave* (e.g. the value of ↑ with respect to the subject NP, which is the host of the ARC) but in (30) it is the topmost (root) f-structure (which can be picked out by an inside-out functional uncertainty expression): conjoining *ci* content to the root is what gives it wide scope. In a monoclausal structure like (27), these are the same, but this would not be the case in examples like (21a), where the ARC is in a subordinate clause.

This is a 'tensor' resource, consisting of two components, np_σ and np_{ci}, corresponding to, respectively, the *at-issue* content of *Kim, who Sam dislikes*, namely, just *Kim*, and its *ci* resource, whose meaning is the proposition $dislikes(Sam, Kim)$.

The general strategy for dealing with tensor resources is to create a context in which the components can be simultaneously consumed. This in general involves the use of hypothetical reasoning. Here the idea is that one does not have to have all the resources one needs available before one starts a proof, or makes a particular move (which would require the resources to become available in a rather strict order). Instead, one can at any point hypothesise the resource(s) one needs: the proof will be successful so long as one can at some later time discharge those hypotheses.

At a certain point in the semantic derivation of the content of *Kim, who Sam dislikes, left*, we will have the resources in (32), which were introduced above.

(32) a. $Kim \times dislikes(Sam, Kim) : np_\sigma \otimes np_{ci}$
b. $\lambda X.left(X) : np_\sigma \multimap r_\sigma$
c. $\lambda q.\lambda p.(p \wedge q) : np_{ci} \multimap [r_\sigma \multimap r_\sigma]$

There is nothing to be done with these, as they stand. However, if we hypothesise a resource H_1 corresponding to the *at-issue* content of the subject NP, we can produce a hypothetical proof as in (33).

(33) $$\frac{[H_1 : np_\sigma]^2 \quad \lambda X.left(X) : np_\sigma \multimap r_\sigma}{left(H_1) : r_\sigma}$$

If we similarly hypothesise a resource corresponding to the *ci* content of the subject NP, we can produce a partial proof as in (34):

(34) $$\frac{[H_2 : np_{ci}]^1 \quad \lambda q.\lambda p.(p \wedge q) : np_{ci} \multimap [r_\sigma \multimap r_\sigma]}{\lambda p.(p \wedge H_2) : r_\sigma \multimap r_\sigma}$$

Abbreviating (33), which shows that hypothesizing $H_1 : np_\sigma$ allows us to derive $left(H_1) : r_\sigma$, and (34), which shows that hypothesizing $H_2 : np_{ci}$ allows us to derive $\lambda p.(p \wedge H_2) : r_\sigma \multimap r_\sigma$, we can produce the derivation in (35).

(35) $$\cfrac{\cfrac{Kim \times dislikes(Sam, Kim) : np_\sigma \otimes np_{ci} \qquad \cfrac{\begin{matrix}[H_1] \\ \vdots \\ left(H_1) : r_\sigma\end{matrix} \quad \begin{matrix}[H_2] \\ \vdots \\ \lambda p.(p \wedge H_2) : r_\sigma \multimap r_\sigma\end{matrix}}{left(H_1) \wedge H_2 : r_\sigma}[a]}{\textbf{let}\ Kim \times dislikes(Sam, Kim)\ \textbf{be}\ H_1 \times H_2\ \text{in}\ left(H_1) \wedge H_2 : r_\sigma}[b]}{left(Kim) \wedge dislikes(Sam, Kim) : r_\sigma}[c]$$

Up to step [*a*], the hypothetical proofs from above are used. At [*a*] itself there is simple function application. At step [*b*] the result of this function application is combined with the tensor resource associated with *Kim, who Sam dislikes* into a 'let' expression. This is simplified by pair-wise substitution at step [*c*]. Notice this gives the intuitively correct

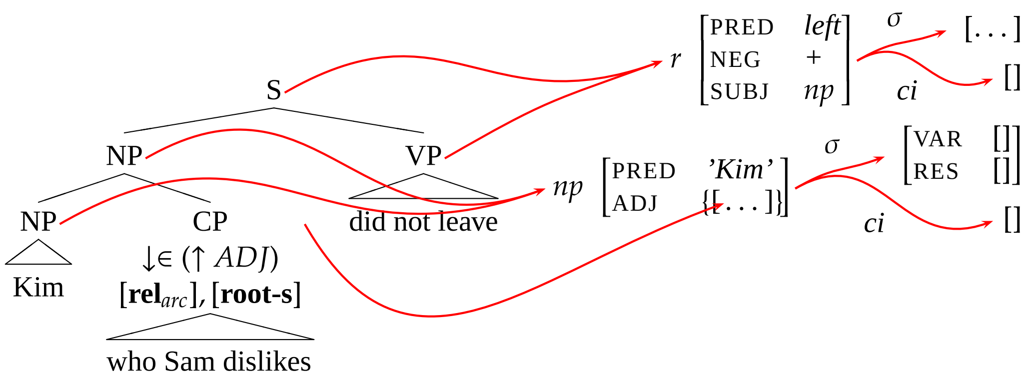


Figure 2

interpretation: issues of focus, backgrounding, etc. aside, example (27) means the same as *Sam dislikes Kim* and *Kim left*.

In this case, there is only one level of embedding, so conjoining the *ci* content with the root content (r_σ) is the same as conjoining it with the *at-issue* content of *dislikes* (s_σ). But this will not always be the case: the *at-issue* content of *dislikes* may be consumed by another operator, e.g. a propositional verb, question operator, or negation. In this case, the *ci* content will escape the scope of that operator (i.e. will get wide scope). We will demonstrate this using negation.

Consider example (36). For our purposes, the c- and f-structures involved are not very different. The main difference will be the existence of an additional resource, corresponding to sentential negation. We assume this to be of the form (37), which, intuitively, consumes a resource associated with the sentence and produces another resource associated with the sentence, but with the difference that the output resource has a negative meaning.

(36) Kim, who Sam dislikes, did not leave.

(37) **[Neg]** $\lambda p.(\neg p) : r_\sigma \multimap r_\sigma$

The proofs are also almost the same as above. (38) differs from (33) only in using the negative resource just mentioned, and in producing an appropriately different result (cf. the meaning is $(\neg left(Kim))$ instead of $left(Kim)$).

(38) $$\dfrac{\lambda p.(\neg p) : r_\sigma \qquad \dfrac{[H_1 : np_\sigma]^2 \qquad \lambda X.left(X) : np_\sigma \multimap r_\sigma}{left(H_1) : r_\sigma}}{(\neg left(H_1)) : r_\sigma}$$

For the rest, the premises and the structure of the proofs are identical. But notice that the result of the proof is that the scope of negation is restricted to the *at-issue* content.[10]

[10] The careful reader will notice that while we show how the *ci*-content *can* get wide scope with respect to negation, we have not show that it *must*: as things stand, there is an equally valid glue derivation that applies [**root-ci**] before [**Neg**], putting the *ci*-content in the scope of negation – which we do not want. There is no

$$\text{(39)}\quad \dfrac{Kim \times dislikes(Sam, Kim) : np_\sigma \otimes np_{ci} \qquad \dfrac{\begin{matrix}[H_1] \\ \vdots \\ (\neg left(H_1)) : r_\sigma\end{matrix} \quad \begin{matrix}[H_2] \\ \vdots \\ \lambda p.(p \wedge H_2) : r_\sigma \multimap r_\sigma\end{matrix}}{(\neg left(H_1)) \wedge H_2 : r_\sigma}}{\dfrac{\textbf{let}\ Kim \times dislikes(Sam, Kim)\ \textbf{be}\ H_1 \times H_2\ \text{ in }\ (\neg left(H_1)) \wedge H_2 : r_\sigma}{(\neg left(Kim)) \wedge dislikes(Sam, Kim) : r_\sigma}}$$

3.2 AE Constructions: LFG Analysis

In this section, we will show how the facts and basic insights of Bosse et al.'s analysis can be expressed using the grammatical apparatus of LFG, as discussed in the previous section.

Our assumptions about c- and f-structure are entirely conventional. Figure 3 shows the sort of c- and f-structure we assume for example (1), repeated here as (40). As previously noted, the meaning representation language will be a version of DRT augmented with an abstraction operator.

(40) Alex zerbrach Chris Bens Vase.
Alex broke Chris.DAT Ben's vase
Alex broke Ben's vase 'on Chris'. (i.e. and this affected Chris)

The basic approach we assume is lexical. We posit a derived verb $zerbrechen_{AE}$ 'break', whose entry is just like that of the normal *zerbrechen* except that (i) it allows an extra OBJ_{dat} complement; and (ii) it introduces the semantic resources in (41) and (42).

---

shortage of technical fixes for this problem, but making a motivated choice among them is not easy, and would be a distraction here. Perhaps the simplest is to assume a rule like the following, taking S_{root} to be the start symbol of the grammar:

$$S_{root} \rightarrow \begin{matrix} S \\ (\uparrow ROOT) = \downarrow \\ \lambda p.p : \downarrow_\sigma \multimap \uparrow_\sigma \end{matrix}$$

The effect of this is to distinguish the root f-structure (and hence the corresponding glue resources) that is involved in combining *at-issue* and *ci* content (i.e. instances of [**root-ci**]) from the one that is negated. Let us call these respectively r_σ and r'_σ. The glue type associated with the rule above is thus $r'_\sigma \multimap r_\sigma$. The glue type of [**Neg**] will be $r'_\sigma \multimap r'_\sigma$, so it cannot apply after this resource has been used. Similarly instances of [**root-ci**] will only be able to operate on the output of this resource, conjoining *ci*-content outside the scope of negation. Adding this would slightly complicate the proofs, but would be otherwise unproblematic.

Alternative solutions to this problem might involve the use of semantic features (like the book-keeping features VAR, RESTR used in treatments of quantification), or the logical type system – e.g. suppose the final goal of a semantic derivation is an object of type T (for 'text'), rather than t. Semantic negation would be of type $\langle t, t\rangle$, and [**root-ci**] would have the logical type $\langle ci, T, T\rangle$.

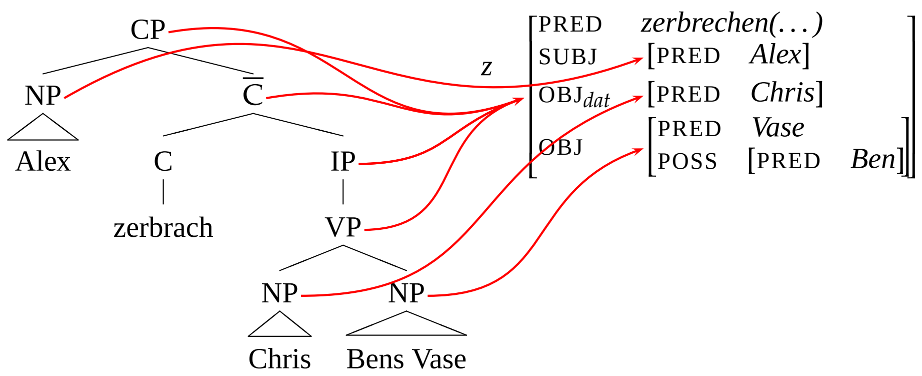


Figure 3

(41) [**zerbrechen**$_{AE}$]

$$\lambda C \lambda B \lambda A. \left(\boxed{\begin{array}{l} E, A, B, C, E' \\ \hline breaking(E) \\ Theme(E,B) \\ Agent(E,A) \\ experience(E') \\ Experiencer(E',C) \end{array}} \times \boxed{\begin{array}{c} \\ \hline \boxed{\begin{array}{l} E'' \\ \hline breaking(E'') \\ Theme(E'',B) \end{array}} \Rightarrow \boxed{\begin{array}{l} \\ \hline source(E',E'') \end{array}} \end{array}} \right) :$$

$(\uparrow OBJ_{dat})_\sigma \multimap [(\uparrow OBJ)_\sigma \multimap [(\uparrow SUBJ)_\sigma \multimap [\uparrow_\sigma \otimes \uparrow_{ci}]]]$

(42) [**ci-root**] $\lambda q.\lambda p.(p \sqcup q) : \uparrow_{ci} \multimap [r_\sigma \multimap r_\sigma]$

The [**ci-root**] resource in (42) is just as in the previous discussion. As before, it is is responsible for merging the *at-issue* and *ci* content associated with the root f-structure (r). The only difference is that instead of $\wedge$ (conjunction), the meaning representation involves $\sqcup$ (discourse merge, which has the same semantic effect as conjunction).

The meaning constructor [**zerbrechen**$_{AE}$] in (41) will consume, in order:

- the ‘affected experiencer object’ (OBJ$_{dat}$) resource (corresponding to *Chris*);
- the direct object resource (corresponding to *Bens Vase*)
- and the subject resource (corresponding to *Alex*) ;

It will produce a pair resource with glue type $\uparrow_\sigma \otimes \uparrow_{ci}$, as in (43), consisting of:

- the ordinary (*at-issue*) content of the verb and its arguments ; and
- a resource associated with the *ci* projection of the verb’s f-structure.

If we denote the outermost f-structure in Figure 3 as z, these will be respectively z_{ci} and z_σ, and we will have (43).

(43) $$\begin{array}{|l|}\hline \textit{E, Alex, BV, Chris, E'} \\ \hline \textit{breaking(E)} \\ \textit{Theme(E,BV)} \\ \textit{Agent(E,Alex)} \\ \textit{experience(E')} \\ \textit{Experiencer(E',Chris)} \\ \hline \end{array} \times \begin{array}{|c|}\hline \\ \hline \begin{array}{|l|}\hline \textit{E''} \\ \hline \textit{breaking(E'')} \\ \textit{Theme(E'',BV)} \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \textit{source(E',E'')} \\ \hline \end{array} \\ \hline \end{array} \; : \; z_\sigma \otimes z_{ci}$$

In words, the *at-issue* content asserts the existence of a breaking event E, two individuals ($Alex$, and BV – Ben's vase), who are respectively the Agent and Patient of E, as well as an additional individual (*Chris*), and event E', of which *Chris* is the Experiencer. That is, roughly, Alex broke Ben's vase, and Chris experienced something. The *ci* content is that every breaking event E'' would be a cause of E'. (For any event E'', if E'' is a breaking event involving Ben's Vase, then it is a source (cause) of E' – the experiencing event. This is essentially identical to the representation Bosse et al. gave in (18) above, expressed in different notation).

To present the glue proof, we will abbreviate (43) as (44) (i.e. $\mathcal{A}$ is the meaning language representation of the *at-issue* content of (43)):

(44) $\mathcal{A} \times \mathcal{B} : z_\sigma \otimes \ z_{ci}$

We can now produce a hypothetical derivation as in (45).

(45) $$\dfrac{\dfrac{\mathcal{A} \times \mathcal{B} : z_\sigma \otimes \ z_{ci} \qquad \dfrac{[H_2 : r_\sigma]^2 \qquad \dfrac{[H_1 : z_{ci}]^1 \quad \lambda q.\lambda p.(p \sqcup q) : z_{ci} \multimap [r_\sigma \multimap r_\sigma]}{\lambda p.(p \sqcup H_1) : r_\sigma \multimap r_\sigma}[a]}{(H_2 \sqcup H_1) : r_\sigma}[b]}{\textbf{let}\ \mathcal{A} \times \mathcal{B}\ \textbf{be}\ H_2 \times H_1 \ \text{ in } \ (H_2 \sqcup H_1) : r_\sigma}[c]}{\mathcal{A} \sqcup \mathcal{B} : r_\sigma}[d]$$

At [a], we hypothesize a resource corresponding to the *ci*-content of *zerbrechen*, which can be consumed by [**root-ci**]. We then hypothesize a resource corresponding to the *at-issue* content of the root f-structure (r_σ)[11], which can be consumed to produce a resource ($H_2 \sqcup H_1$) associated with the root f-structure r (r_σ). This provides an environment into which the pair resource associated with our example, (43), abbreviated in (44), can be substituted (at [c]). This produces a **let** expression which can be simplified, as at [d].

For our purposes, the discourse merger operation notated as $\sqcup$ can be taken to be simple merger of universes (discourse variables) and conditions of DRSs, which in this case produces (46), which has the truth conditions we want.

[11]Though notice that in this case, where *zerbrechen* is the main verb and there are complications involving negation etc., z_σ and r_σ are identical.

(46)
$$\begin{array}{|l|}\hline \textit{E, Alex, BV, Chris, E'} \\ \hline \textit{breaking(E)} \\ \textit{Theme(E,BV)} \\ \textit{Agent(E,Alex)} \\ \textit{experience(E')} \\ \textit{Experiencer(E',Chris)} \\ \begin{array}{|l|}\hline \textit{E''} \\ \hline \textit{breaking(E'')} \\ \textit{Theme(E'',BV)} \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \textit{source(E',E'')} \\ \hline \end{array} \\ \\ \hline \end{array} : r_\sigma$$

It is easy to see that this approach will allow *ci*-content to escape negation. Suppose we abbreviate (47), which we assume is the resource associated with sentential negation, as $\lambda p.(\neg p)$: $z_\sigma \multimap z_\sigma$.

(47)
$$\lambda p.\begin{array}{|c|}\hline \\ \hline \neg\ p \\ \hline \end{array} : z_\sigma \multimap z_\sigma$$

The proof that derives the interpretation of (48) can procede as in (49). The only important difference between this and (45) is at $[a']$. Here sentence negation has been applied to the resource we hypothesized for the *at-issue* content of the sentence. For the rest, the derivation is identical, except that this resource, and ultimately the non-hypothetical resource that discharges it, are thereafter in the scope of negation. Notice, however, that the *ci*-content is not the scope of negation, which is the result that we want.

(48)

Alex	zerbrach	Chris	Bens	Vase	nicht.
Alex	broke	Chris.DAT	Ben's	vase	not

Alex didn't break Ben's vase 'on Chris'. (But it would matter to Chris.)

(49)
$$\dfrac{\dfrac{\mathcal{A}\times\mathcal{B} : z_\sigma \otimes z_{ci} \qquad \dfrac{\dfrac{\lambda p.(\neg p) : z_\sigma \multimap z_\sigma \quad [H_2 : z_\sigma]^2}{(\neg H_2) : z_\sigma}[a'] \qquad \dfrac{[H_1 : z_{ci}]^1 \quad \lambda q.\lambda p.(p \sqcup q) : z_{ci} \multimap [r_\sigma \multimap r_\sigma]}{\lambda p.(p \sqcup H_1) : r_\sigma \multimap r_\sigma}[a]}{((\neg H_2) \sqcup H_1) : r_\sigma}[b]}{\textbf{let } \mathcal{A}\times\mathcal{B} \textbf{ be } H_2 \times H_1 \text{ in } ((\neg H_2) \sqcup H_1) : r_\sigma}[c]}{(\neg\mathcal{A}) \sqcup \mathcal{B} : r_\sigma}[d]$$

To make this more concrete, notice that the structure $(\neg\mathcal{A}) \sqcup \mathcal{B}$ is an abbreviation for the structure in (50), which, when discourse merge has applied, gives rise to (51), where the *ci*-content is clearly outside the scope of negation.

(50)

$$\begin{array}{|l|}\hline \\ \hline \neg \begin{array}{|l|}\hline \textit{E, Alex, BV, Chris, E'} \\ \hline \textit{breaking(E)} \\ \textit{Theme(E,BV)} \\ \textit{Agent(E,Alex)} \\ \textit{experience(E')} \\ \textit{Experiencer(E',Chris)} \\ \hline \end{array} \\ \hline \end{array} \sqcup \begin{array}{|l|}\hline \\ \hline \begin{array}{|l|}\hline \textit{E''} \\ \hline \textit{breaking(E'')} \\ \textit{Theme(E'',BV)} \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \textit{source(E',E'')} \\ \hline \end{array} \\ \hline \end{array} : r_\sigma$$

(51)

$$\begin{array}{|l|}\hline \\ \hline \neg \begin{array}{|l|}\hline \textit{E, Alex, BV, Chris, E'} \\ \hline \textit{breaking(E)} \\ \textit{Theme(E,BV)} \\ \textit{Agent(E,Alex)} \\ \textit{experience(E')} \\ \textit{Experiencer(E',Chris)} \\ \\ \hline \end{array} \\ \begin{array}{|l|}\hline \textit{E''} \\ \hline \textit{breaking(E'')} \\ \textit{Theme(E'',BV)} \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \textit{source(E',E'')} \\ \hline \end{array} \\ \hline \end{array} : r_\sigma$$

In fact, there is rather more to say about (51), because it is less satisfactory than it at first seems. However, before we pursue this, we should address the issue of cross-linguistic variation with respect to the interpretation of the agent in relation to *ci* and *at-issue* content.

Recall that according to Bosse et al., there is a difference between Japanese and German, in that in the former (but not the latter), the agent is part of the *ci* content. Bosse et al. propose to capture this by variation of the relative heights of VoiceP and AffP. The question naturally arises as to whether our approach can accommodate this variation.

The answer to this question is positive, and the method almost trivial. All that is required is a very small change to the output of the lexical rule that we posit for Japanese verbs, as compared to their German counterparts. The lexical entry for the Affected Experiencer version of *kowas-are-ru* (‘break’) should be as in (52). Ignoring syntactic details, the sole difference between this and German *zerbrechen* is highlighted. Demonstrating that this has the desired effect is left as a (trivial) exercise for the reader. The difference between Japanese and German can be simply captured by a small variation in the respective lexical rules.

(52) [**kowas-are-ru**$_{AE}$]

$$\lambda C \lambda B \lambda A.\left(\begin{array}{|l|}\hline \mathit{E, A, B, C, E'} \\ \hline \mathit{breaking(E)} \\ \mathit{Theme(E,B)} \\ \mathit{Agent(E,A)} \\ \mathit{experience(E')} \\ \mathit{Experiencer(E',C)} \\ \hline \end{array} \times \begin{array}{|c|}\hline \\ \hline \begin{array}{|l|}\hline \mathit{E''} \\ \hline \mathit{breaking(E'')} \\ \mathit{Agent(E'',A)} \\ \mathit{Theme(E'',B)} \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \mathit{source(E',E'')} \\ \hline \end{array} \\ \hline \end{array}\right) :$$

$$(\uparrow SUBJ)_\sigma \multimap [(\uparrow OBJ)_\sigma \multimap [(\uparrow OBJ_{ni})_\sigma \multimap [\uparrow_\sigma \otimes \uparrow_{ci}]]]$$

The question we started out with was whether an LFG/glue implementation could be provided that deals with the data that Bosse et al. present: specifically, whether an account can be found that eschews abstract functional projections like VoiceP and AffP. We see that such an account is indeed possible: the flexibility provided by LFG-glue semantics is sufficient, and allows us to operate with a simpler, and far less abstract, syntax.

4 Discussion

We appear to have replicated Bosse et al.'s analysis in the current framework, which would seem to be an entirely positive result. Unfortunately, things are not so simple. In fact, because the replication is close, it shows up some troubling problems with Bosse et al.'s approach.

Consider again the DRS of the example involving negation in (51), which correctly shows the AE content outside the scope of negation. The problem is that this is not a well-formed DRS: it is *improper*. This is because the AE content contains a variable E' (in the consequent) which is, intuitively, unbound. Notice in particular that it is not in the scope of the instance of E' which is introduced in the *at-issue* content, because this is in a more deeply embedded sub-DRS (because it is in the scope of negation).

It is important to stress that this is not some artefact of the DRT representation we have adopted, or some arbitrary piece of formalization that can be evaded by some minor reformulations.

As regards the first point, exactly the same problem would arise with the predicate logic based account that Bosse et al. present. Suppose we modify Bosse et al.'s representation (18) from above so that the *at-issue* content is in the scope of negation, and the *ci* content is outside the scope of negation. We will have the following (the brackets delimiting the scope of negation are highlighted):

(53) $\neg\ (\ \exists e(break(e)\&Thm(Ben's\ vase)(e)\&Agt(Alex)(e)\&\exists e'(exp(e')\&Exp(Chris)(e'))\ )\ :$
$\forall e''((break(e'')\&Thm(Ben's\ vase)(e''))\ \rightarrow\ Source(e'')(e')))$

There is nothing syntactically wrong with this, as a piece of predicate logic. The problem is its interpretation. Notice that here the variable e' in the *ci*-content is free – in particular, it is not bound by the existential quantifier that binds the instance of e' in the *at-issue* content. (We have followed Bosse et al.'s use of variable names, but where they write $Source(e'')(e')$ we write $source(E', E'')$). Since it is unbound, one can freely replace it with any other variable (x, say) without changing the meaning. Formalizations of the semantics of predicate logic differ in how they deal with the interpretation of unbound variables, but it is clear that (53) will not mean what Bosse et al. want it to mean.

As regards DRT, there is a very good reason why we would want representations like (51) to be ill-formed or in some way illicit, because this is at the heart of the DRT account of what is wrong with examples like (54), which is represented by a DRS like (55).

(54) #John doesn't have a coat. If he goes out, he wears it.

(55)

$$\begin{array}{|l|}\hline \textit{John} \\ \hline \neg \begin{array}{|l|}\hline X \\ \hline coat(X) \\ own(John,X) \\ \hline \end{array} \\ \\ \begin{array}{|l|}\hline \\ \hline go\text{-}out(John) \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline wear(John,X) \\ \hline \end{array} \\ \hline \end{array}$$

Notice that in (55), as in (18) there is a variable (X in (55)) that appears in a DRS condition in the consequent of a conditional without an instance in a 'higher' DRS to bind it. The fact that such structures are ruled out reflects an important piece of theory for DRT.

There is, in short, a serious problem for the Bosse et al. analysis here. Notice also that the problem arises precisely because of the way Bosse et al. assume that variable binding can occur across the *at-issue*/*ci* boundary, which was the challenge that the construction seemed to pose for Potts' approach. The question of how to deal with the facts that lead Bosse et al. to propose that this should occur remains open.

References

Doug Arnold and Louisa Sadler. Pottsian LFG. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG10 Conference*, pages 43–63, Stanford, Ca., 2010. CSLI Publications.

Doug Arnold and Louisa Sadler. Resource splitting and reintegration with supplementals. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of LFG11*, pages 26–46, Stanford, Ca., 2011. CSLI Pubs.

Ash Asudeh. *Resumption as Resource Management*. PhD thesis, Stanford University, 2004.

Ash Asudeh. *The Logic of Pronominal Resumption*. Oxford University Press, Oxford, 2012.

Solveig Bosse. *The syntax and semantics of applicative arguments in German and English*. PhD thesis, University of Delaware, 2011.

Solveig Bosse and Benjamin Bruening. Benefacive versus Experiencer Datives. In Mary Byram Washburn, Katherin McKinney-Bock, Erika Varis, Ann Sawyer, and Barbara Tomaszewicz, editors, *Proceedings of the 28th West Coast Conference on Formal Linguistics*, pages 69–77, Somerville, MA, 2011. Cascadilla Proceedings Project.

Solveig Bosse, Benjamin Bruening, and Masahiro Yamada. Affected experiencers. *Natural Language & Linguistic Theory*, 30(4):1185–1230, 2012.

Mary Dalrymple. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press, New York, 2001.

Gianluca Giorgolo and Ash Asudeh. Multidimensional semantics with unidimensional glue logic. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of LFG11*, pages 236–256, Stanford, Ca., 2011. CSLI Pubs.

Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht, 1993.

Eric McCready. Varieties of conventional implicature. *Semantics and Pragmatics*, 3:1–57, 2010.

Reinhard Muskens. Combining Montague semantics and discourse representation. *Linguistics and Philosophy*, 19:143–186, 1996.

Christopher Potts. *The Logic of Conventional Implicatures*. Oxford University Press, Oxford, 2005.

Osamu Sawada. *Pragmatic aspects of scalar modifiers*. PhD thesis, University of Chicago, 2011.

FLEXIBLE COMPOSITION FOR OPTIONAL AND DERIVED ARGUMENTS

Ash Asudeh and Gianluca Giorgolo
Carleton University & University of Oxford
King's College, London & University of Oxford

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

A simple but insightful analysis of optional and derived arguments at the syntax–semantics interface is provided, based on established features of LFG with Glue Semantics (optionality and templates in lexical entries and flexible, resource-sensitive semantic composition).

1 Introduction[1]

There is broad agreement in linguistic theory that arguments and adjuncts must be distinguished, but there is substantial disagreement as to how the distinction is to be represented and how borderline cases should be captured. There are a number of representational options, of which we list some illustrative examples. In Principles and Parameters Theory (Chomsky 1981, 1995), an argument is either the complement or specifier of a head, whereas an adjunct is adjoined at the XP level. In some versions of Head-Driven Phrase Structure Grammar, an adjunct is distinguished by being a member of the DEPS list but not a member of the VALENCE lists or of the ARG-ST list (Bouma et al. 2001). In LFG, we see a hybrid approach. Adjuncts are distinguished at f-structure by being a member of a predicate's ADJUNCT set, whereas arguments fill specific grammatical functions, such as SUBJ, OBJ, etc. However, given the structure-function mapping principles proposed by Bresnan (2001) and developed further by Toivonen (2001, 2003) (see also Bresnan et al. 2013), adjuncts normally appear in distinguished c-structural positions.

In this paper, we present the initial developments in a theory of adjuncts and arguments, building on recent work by Needham and Toivonen (2011), that uses LFG and Glue Semantics (Dalrymple 1999, 2001, Asudeh 2012) to treat the argument/adjunct distinction not narrowly as an issue of syntactic representation, but rather as a distinction that primarily concerns semantic composition.[2]

The main questions that we seek to answer are the following:

1. What are the implications of optional and derived arguments for the mapping from syntax to semantics?

2. How can lexical generalizations about optional and derived arguments best be captured?

[1]This research was supported by an Early Researcher Award from the Ontario Ministry of Research and Innovation (Asudeh), NSERC Discovery Grant #371969 (Asudeh), SSHRC Standard Research Grant #410-2010-1841 (PI: I. Toivonen; Collaborator: A. Asudeh) and by a grant from the John Fell Oxford University Press Research Fund (Asudeh). We thank the following for helpful comments and questions: Doug Arnold, Boban Arsenijević, Miriam Butt, Mary Dalrymple, Dag Haug, Ron Kaplan, Tracy Holloway King, Stefano Quaglia, Louisa Sadler, Sebastian Sulger, Ida Toivonen, and the participants at LFG12. Any remaining errors are our own.

[2]In the companion piece to this paper (Giorgolo and Asudeh 2012), which also appears in these proceedings, we take a distinct formal approach that uses monads, building on Giorgolo and Asudeh (2011), but we maintain the key insight that the argument/adjunct distinction is an issue of semantic composition.

We attempt initial answers to these questions by looking at three cases:

1. Optional objects of semantically relational verbs (e.g., *drink*, *eat*)
2. Passive *by*-phrases
3. Instrumental *with*-phrases

The paper is organized as follows. Section 2 sets out the phenomena we are interested in and the problems and challenges they constitute. Section 3 presents the key ideas of our analysis informally. Section 4 presents our formal analysis. Section 5 discusses the contribution that templates can make to the analysis. Section 6 concludes.

2 Optional Arguments and Borderline Cases

The problematic cases of interest can be divided into two classes. First, there is the case of predicates that semantically denote a relation (i.e., take two arguments), but which do not require the second argument to be syntactically expressed:

(1) Any child of Kim's is unfortunately likely to drink __.

(2) Kim ate __ at noon.

Clearly one has to drink or eat something, so these verbs are semantically relational, yet the object argument can be unexpressed.

It is typical to contrast verbs like these with similar verbs that do not allow the object to be unexpressed:

(3) a. Isak quaffed his milk at lunch.
 b. *Isak quaffed __ at lunch.

(4) a. Thora devoured her cake after dinner.
 b. *Thora devoured __ after dinner.

The distinctions between *drink*/*quaff* and *eat*/*devour* need to be captured lexically somehow — in other words, it is part of what we know as language speakers that *drink* can drop its object argument but that *quaff* does not.[3] We refer to these sorts of cases as 'optional arguments'.

Needham and Toivonen (2011) review a number of other cases in which a syntactic phrase seems to be an adjunct in some ways (e.g., it is optional; it is a PP instead of a direct argument), but which seems to be an argument in other ways (e.g., it expresses some entailed participant in the event that the verb denotes). Here are some examples with Needham and Toivonen's labels:

[3] It has been noted (e.g., Jackendoff 2002) that this may be predictable based on semantic factors, since devouring/quaffing is a particular manner of eating/drinking, etc., but this would just seem to mean that the lexical generalization may be stated in a more general fashion, perhaps in a hierarchically organized lexicon, not that it is not part of lexical knowledge.

(5) The hole was plugged by Kim. Passive *by*-phrase

(6) Kim plugged the hole with a cork. Instrumental

(7) Kim's obstruction of the hole Possessive phrase in event nominal

(8) Kim plugged the hole for them. Benefactive

(9) The hole crawled with bugs. Displaced theme

(10) It seemed to Kim like the hole could be plugged. Experiencer

(11) The bugs crawled from the hole. Directional

We follow Needham and Toivonen (2011) in referring to these sorts of cases as 'derived arguments', although it should be fairly obvious that no bright line can be drawn between optional and derived arguments.

2.1 The Problem

The basic intuition behind the argument/adjunct distinction is that arguments are "semantically necessary" in some way that adjuncts are not. As pointed out by Needham and Toivonen (2011), despite the intuitive appeal of this claim, it fails spectacularly because many clear adjuncts, such as those involving time and place, are also clearly semantically necessary: every event that we refer to linguistically happens at some time, in some place. This points to a different understanding of the intuition, which Needham and Toivonen call 'verb specificity': arguments are 'semantically distinctive' in that they are associated with particular verb classes, such that these are distinguished from other classes. Thus, time and place are generally poor arguments, *because* they are ubiquitous and fail to distinguish between verb classes.

The semantic function that arguments play is typically tied to their obligatory realization in syntax, with optionality often taken to be a hallmark of adjuncts. However, there are cases of clear arguments, according to any plausible semantic criterion, which are nevertheless syntactically optional, such as the objects of *drink* and *eat* in English. Similarly, there are argument-like functions ('derived arguments'), such as instrumentals, that distinguish verb classes according to verb specificity, but which seem to always be optional.

Most solutions to this problem can be characterized as some version of the solution of Bresnan (1978), which proposes two distinct versions of, e.g., the verb *eat*.

(12) *eat:* V, [___ NP], NP_1 'eat' NP_2
[___], ($\exists$ y) NP_1 'eat' y

However, this kind of approach is clearly unappealing, because it basically posits an ambiguity for each relevant verb and misses the generalization that, e.g., the 'eating' is the same sort of thing in both cases.

Lastly, it has been noted (e.g., Fillmore 1986) that there may be restrictions on implicit arguments that are absent for their explicit counterparts:

(13) a. Fido ate this morning.
$\Rightarrow$ Whatever Fido ate counts as food for Fido

b. Fido ate my homework.
$\not\Rightarrow$ My homework counts as food for Fido

(14) a. Kim drank last night.
$\Rightarrow$ Whatever Kim drank last night is alcoholic/intoxicating

b. Kim drank milk last night.
$\not\Rightarrow$ Milk is alcoholic/intoxicating

In sum, the challenge is to capture the core argument structure of verb classes that display optional or derived arguments in a way that:

1. Doesn't simply treat distinct valencies as accidentally related (homonymous).

2. Supports a systematic semantic treatment of optional and derived arguments.

3. Enables semantic restrictions on optional arguments to be stated.

4. Captures commonalities between derived arguments and adjuncts

In the next section, we informally sketch our way of meeting this challenge.

3 An Informal Sketch of Our Approach

Our main claim is that a simple but insightful analysis of optional and derived arguments at the syntax–semantics interface can be provided based on established features of Lexical-Functional Grammar with Glue Semantics:

1. **Optionality**, offered by the regular language of LFG's functional descriptions in lexical entries (Kaplan and Bresnan 1982, Dalrymple 2001).

2. **Flexible semantic composition**, offered by the commutative glue logic of Glue Semantics (Dalrymple 1999, 2001, Asudeh 2012).

3. **Resource-sensitive semantic composition**, again offered by the glue logic.

4. **Generalizations over descriptions**, offered by templates (Dalrymple et al. 2004, Asudeh et al. 2008, Asudeh 2012).

The basic strategy will be to break apart lexical information in such a way that, for example, a transitive verb with an optional object can supply semantic information about the implicit object just in case the object is unexpressed. However, a single lexical entry for the verb handles both the intransitive and transitive instantiation of the verb.

We can exemplify the general approach with the following schematized lexical entry for *eat* — as it occurs in the analysis of a sentence like (15) — with most formal details suppressed for now:

(15) Kim ate at noon.

(16) *ate* V ($\uparrow$ PRED) = 'eat'

F-structure constraints

Obligatory Glue meaning constructor;
encodes general semantic information that is
common to transitive and intransitive uses

(**Optional Glue meaning constructor;**
encodes semantic information that is
specific to the intransitive use)

The PRED feature of this lexical entry does not encode whether it is transitive or instransitive. We assume that subcategorization of grammatical functions other than expletives is not represented at f-structure, but is rather captured by resource-sensitive semantic composition (Kuhn 2001, Asudeh 2012). If this were not the case, the formal f-structure description language would force a disjunctive lexical entry — with the attendant issues discussed in section 2.1 — but for theoretically uninteresting reasons (see Giorgolo and Asudeh 2012 for further discussion).

The lexical entry in (16) is different from the disjunctive lexical entries suggested by Bresnan (1978), shown in (12) above, in an important respect. The two Glue meaning constructors in (16) do not stand in a purely disjunctive relationship, whereas the two options in (12) do.[4] In other words, the entry in (16) does not treat the two subcategorizations of *eat* as coincidentally homophonous, but rather posits a single lexical entry with an obligatory meaning constructor that captures the fact that the verb is semantically relational (i.e., it takes two arguments) and posits an optional meaning constructor that existentially closes the second semantic argument if and only if the object is unexpressed. The core meaning of *eat* is thus maintained across the two cases and associated with a single form, in a principled fashion, whereas in (12) it is treated as purely coincidental that the two subcategorizations share the core of their meaning.

4 Analysis

Butt et al. (1997) treat argument structure (a-structure) as a level interpolated between constituent structure and functional structure, such that the correspondence

[4]The logic of the relevant part of the entry in (16) can be represented as $A \veebar (A \wedge B)$, where A is the obligatory meaning constructor, B is the optional meaning constructor, and $\veebar$ is exclusive disjunction. In contrast, the logic of the lexical entry in (12) is purely exclusive disjunction: $A \veebar B$, where A is the transitive option and B is the intransitive option. It is easy to verify that $(A \veebar (A \wedge B)) \not\equiv (A \veebar B)$.

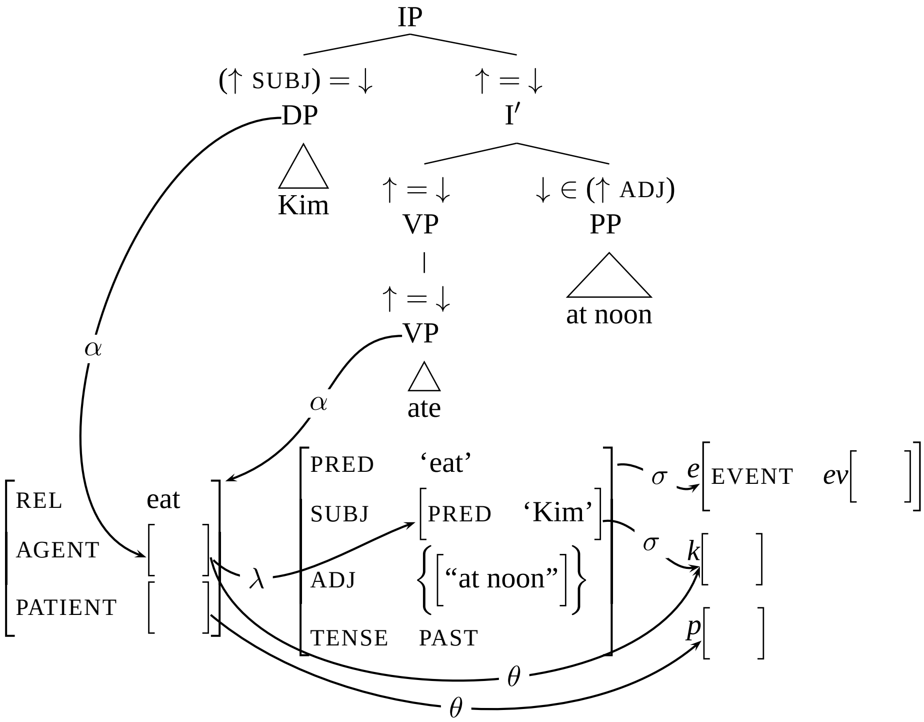


Figure 1: Relevant structures and correspondences for *Kim ate at noon*

function ϕ can be understood as the composition of the correspondence functions α (from c-structure to a-structure) and λ (from a-structure to f-structure). If we adopt this approach, it is necessary to postulate a direct correspondence function θ from argument structure to semantic structure, where the function θ is not the composition of λ and σ.

The example in Figure 1, which adopts the Butt et al. (1997) architecture, illustrates why this is so. The c-structure maps to a-structure, via the α correspondence function. The a-structure maps to f-structure, via the λ correspondence function. Lastly, the f-structure maps to semantic structure, via the σ correspondence function. The PATIENT argument in a-structure must map to an element of semantic structure, since this element provides the resource for the corresponding argument in the semantics in resource-sensitive semantic composition (Dalrymple 2001, Asudeh 2012). However, the PATIENT does not map to an OBJECT grammatical function at f-structure, because this occurrence of *drank* is syntactically intransitive. Therefore, it is not possible to get to the semantic structure correspondent of the PATIENT by going through f-structure, because there is no f-structure correspondent of the PATIENT. Moreover, because the semantic structure is normally unconnected in Glue Semantics (Dalrymple 2001, Asudeh 2012), it is also not possible to get to the semantic structure correspondent of the PATIENT by passing from the outermost a-structure to the outermost f-structure to the semantic structure, since there is no relation expressed in this semantic structure between the semantic structure correspondent of the a-structure containing the PATIENT and the semantic structure correspondent of the PATIENT.

It would be possible to circumvent this problem by positing a null pronominal OBJECT at f-structure, but this is empirically problematic.[5] Standard syntactic tests such as pronominalization, ellipsis, and secondary predication do not support an OBJECT at f-structure when the second argument is unexpressed:

(17) a. Kim drank a beer, but it turned out to be Sandy's.
 b. *Kim drank, but it turned out to be Sandy's.

(18) a. Kim is eating a cake, and so is Sandy. (strict or sloppy)
 b. Kim is eating, and so is Sandy. (sloppy only)

(19) a. Kim drank the whiskey neat.
 b. *Kim drank neat.

Therefore, in order to express the correspondence between the PATIENT and a semantic structure resource, we would have to add a new correspondence function, which we have called θ, to the Butt et al. (1997) architecture.

We instead assume an alternative architecture that does away with the λ-projection, the λ correspondence function, and the θ correspondence function. Argument structure is captured in semantic structure instead. Some of the benefits of this approach are as follows:

1. We achieve a simplified architecture, which eliminates a separate a-structure projection, without losing information.
2. We do not lose linking relations and they are still post-constituent structure.[6]
3. We remove the non-determinacy that results from the presence of both the λ and θ correspondence functions.
4. Many of the meaning constructors for semantic composition are more elegant and simplified.
5. We regain the simple, traditional ϕ mapping from c-structure to f-structure.
6. We gain a connected semantic structure.

Figure 2 shows relevant structures and correspondences for our alternative representation of the example in Figure 1. Since we will be assuming an event semantics for our meaning language, such that thematic roles are functions from events to individuals (Parsons 1990), we avoid redundancy in the argument structure by using

[5]These implicit arguments are therefore not analyzable as cases of "pro-drop", unlike the typologically common case of subject arguments which are not realized in c-structure but which are realized in f-structure.

[6]Feeding argument structure from c-structure is motivated by Butt's (1995) work on Urdu complex predicates, in which she argues that the complex predicates can be syntactically complex in c-structure but nevertheless express a single argument structure like that of a non-complex predicate.

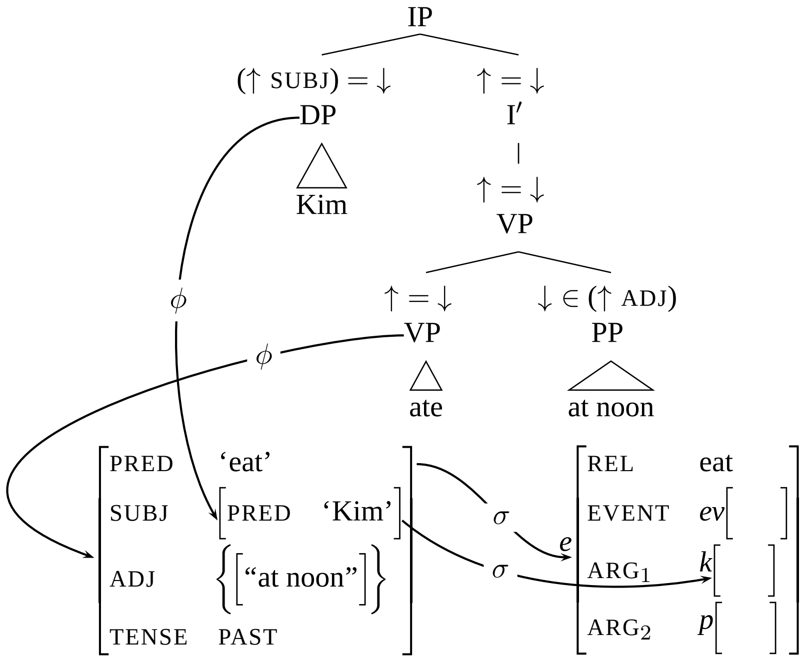


Figure 2: Alternative analysis of *Kim ate at noon.*

attributes like ARG_1 instead of AGENT, etc. If an alternative meaning language that does not encode thematic roles is used, it may be preferable to represent the nature of the arguments directly in a-structure, using the more specific attributes.

4.1 Optional Transitives

Let us begin with the case of ‘optional arguments’, which are semantic arguments that can be syntactically unexpressed, as exemplified by the optional transitivity of *eat* and *drink* versus *devour* and *quaff*. In semantic composition, our analysis simultaneously existentially closes the argument that is alternatively expressed by the object — capturing the fact that even though the argument is unexpressed, it is still an understood argument — and appropriately restricts the existentially closed argument (Fillmore 1986). For example, the existentially closed argument of *drink* is an alcoholic beverage and that of *eat* is food.[7] Moreover, the predicate that expresses this in the semantics must be a relation that also takes the subject as an argument. That is, it is not enough, e.g., for the unexpressed argument to be edible, it must be edible *for* the subject. Contrast the following:

(20) My cousin Kim ate with gusto last night.

(21) My cow Kim ate with gusto last night.

[7]This information is perhaps better treated as a presupposition or conventional implicature than a straight entailment, but we leave this aside here, since it would be straightforward to augment the analysis in standard ways to capture this aspect.

My cousin Kim and my cow Kim eat different sorts of things, and our understanding of these sentences reflects that.[8]

The lexical entry for *ate* is shown in (23) and the Glue proof for example (22) is shown in Figure 6,[9] assuming other standard premises as appropriate and with premises instantiated as per Figure 2 above.

(22) Kim ate at noon.

(23) *ate* V

$(\uparrow \text{PRED}) = \text{'eat'}$
$(\uparrow \text{TENSE}) = \text{PAST}$

$(\uparrow \text{SUBJ})_\sigma = (\uparrow_\sigma \text{ARG}_1)$
$(\uparrow_\sigma \text{ARG}_2)$

$\lambda y \lambda x \lambda e.eat(e) \wedge agent(e) = x \wedge patient(e) = y :$
$(\uparrow_\sigma \text{ARG}_2) \multimap (\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma$

$$\left(\begin{array}{l} \lambda P \lambda y \exists x.[P(x)(y) \wedge food.for(x, y)] : \\ [(\uparrow_\sigma \text{ARG}_2) \multimap (\uparrow_\sigma \text{ARG}_1) \multimap \uparrow_\sigma] \multimap [(\uparrow_\sigma \text{ARG}_1) \multimap \uparrow_\sigma] \end{array} \right)$$

The predicate $food.for(x, y)$ is interpreted such that x is food for y.

The order of arguments of a function can be easily swapped in a Glue proof:

(24)
$$\cfrac{\cfrac{\cfrac{\cfrac{\cfrac{\lambda y \lambda x.f(x, y) : a \multimap b \multimap c \quad [v : a]^1}{\lambda x.f(x, v) : b \multimap c} \quad [u : b]^2}{f(u, v) : c}}{\lambda v.f(u, v) : a \multimap c}\multimap_{\mathcal{I},1}}{\lambda u \lambda v.f(u, v) : b \multimap a \multimap c}\multimap_{\mathcal{I},2}}{\lambda x \lambda y.f(x, y) : b \multimap a \multimap c}\Rightarrow_\alpha$$

We therefore adopt the convention of choosing a version of the lexically specified function in question that is convenient for the larger proof, abbreviating the function as eat', etc., until the final line of proofs, when the abbreviation is unpacked.

The same lexical entry in (23) is used for the analysis of an example like this:

(25) Kim ate the cake at noon.

However, in this case the resource sensitivity of Glue Semantics (Asudeh 2004, 2012) ensures that the optional premise cannot be selected. The obligatory premise is the only consumer of the object resource in the relevant resource pool. If the

[8]This is not obvious for *drink*, but it seems to be equally the case. For example if Dr. McCoy from *Star Trek* utters "Every subject drank", referring to a group of alien beings in his lab, we expect that each subject drank something compatible with its biology (see also Giorgolo and Asudeh 2012). The editors have mentioned to us that another example is the trolls in Terry Pratchett's *Discworld* novels, who drink stuff which is drinkable only to them.

[9]In order to save space, we gather all Glue proofs at the end of the paper, after the references.

optional premise is also in the resource pool, then the optional premise acts as a modifier of the obligatory premise, as shown in Figure 6 above, such that there is no longer a consumer for the object premise. Therefore, selection of the optional premise leads to a successful Glue proof if and only if there is no object resource. If the object is expressed and therefore contributes a resource, the optional premise is not selected and the obligatory premise consumes its object as per usual. The proof for (25) is shown in Figure 7.

Lastly, let us consider obligatory transitives, such as *devour* and *quaff*, which do not allow their objects to be unexpressed (see (3) and (4) above). The lexical entries for these verbs lack the optional, modificational premise:

(26) *devoured* V

$(\uparrow \text{PRED}) = \text{'devour'}$
$(\uparrow \text{TENSE}) = \text{PAST}$

$(\uparrow \text{SUBJ})_\sigma = (\uparrow_\sigma \text{ARG}_1)$
$(\uparrow \text{OBJ})_\sigma = (\uparrow_\sigma \text{ARG}_2)$

$\lambda y \lambda x \lambda e.devour(e) \;\wedge\; agent(e) = x \;\wedge\; patient(e) = y :$
$(\uparrow_\sigma \text{ARG}_2) \multimap (\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma$

Resource-sensitive composition ensures that predicates like this must have an expressed object that contributes the ARG_2 resource; otherwise the dependency on this resource is not properly discharged and there is no valid Glue proof.

4.1.1 Scope

Fodor and Fodor (1980) note that a quantifier in subject position must take wide scope over the existentially closed implicit argument of a syntactically intransitive but semantically relational verb:[10]

(27) Every student ate.
$\Rightarrow$ For every student x, there is some thing y such that x ate y.
$\not\Rightarrow$ There is some thing y such that, for every student x, x ate y.

Our analysis captures this scope generalization. The quantifier and the optional premise contributed by the verb *ate* both constitute dependencies on a dependency on the subject. That is, both the quantifier and the optional premise are consumers of a premise that can be schematized as $subj \multimap predicate$. There is only one such premise (the verb's premise, having consumed the implicit argument's resource). The optional premise, however, is a modifier-type premise that outputs the same dependency again. Therefore, the quantifier can consume the output of the optional premise. In contrast, the quantifier does not output a premise of this type, but rather one of a propositional type. Therefore, the optional premise cannot consume the

[10]This claim has been refined by Lasersohn (1993), based on distributed readings, but he does not seem to have found the correct generalization. This is discussed further in Giorgolo and Asudeh (2012).

output of the quantifier. This means that the quantifier must come later in the proof, which entails that it scopes wide. The successful proof for the wide scope reading is shown in Figure 8.[11]

4.2 Passives

We assume that, in the absence of a *by*-phrase, the suppressed argument of a passive is not represented at f-structure, but is represented at semantic structure. A short passive is thus semantically relational, but syntactically intransitive, much like our previous cases. We again propose a lexical entry that has an obligatory semantic component and an optional semantic component. The suppressed argument is again optionally existentially closed. Given Glue's resource-sensitive semantic composition, this option only leads to a well-formed proof in the absence of a *by*-phrase. If both the meaning constructor contributed by a *by*-phrase and the optional existential meaning constructor were present, there would be two dependencies on the resource corresponding to the suppressed highest role of the predicate (e.g., the $\textsc{arg}_1$ of *eaten*, which is mapped to the *by*-phrase, if one is present), but only one such dependency could be satisfied. Resource sensitivity similarly guarantees that the optional premise contributed by the passive verb *must* be realized in the absence of a *by*-phrase, because otherwise the dependency on the highest role (i.e., $\textsc{arg}_1$) is not discharged. We thus correctly predict that there is existential closure of the suppressed argument if and only if there is no *by*-phrase.

Consider the examples of a short passive and a *by*-passive in (28) and (29).

(28) Kim was eaten last night.

(29) Kim was eaten by Godzilla last night.

Lexical entries for the passive predicate, *eaten*, and the passive *by* are shown in (30) and (31).[12]

(30) *eaten* V

$(\uparrow \text{PRED}) = \text{‘eat’}$
$(\uparrow \text{VOICE}) = \text{PASSIVE}$

$(\uparrow \text{SUBJ})_\sigma = (\uparrow_\sigma \text{ARG}_2)$
$(\uparrow_\sigma \text{ARG}_1)$

$\lambda x \lambda y \lambda e.eat(e) \wedge agent(e) = x \wedge patient(e) = y :$
$(\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{ARG}_2) \multimap (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma$

$\Big(\lambda P \exists x.[P(x)] : [(\uparrow_\sigma \text{ARG}_1) \multimap \uparrow_\sigma] \multimap \uparrow_\sigma \Big)$

[11] Our approach allows the subject quantifier and existential event closure to scope freely with respect to each other, since examples like (27) are ambiguous between a single event of every student eating and separate events of each student eating. The proof in Figure 8 captures only the first of these readings.

[12] There has been some inconsistency in the LFG literature regarding the realization of the *by*-phrase at f-structure: Is it an ADJ or an OBL? (See Needham and Toivonen 2011 for discussion and references.) This choice does not substantively affect our analysis, but we assume the *by*-phrase is an OBL here. Otherwise, change the occurrences of ($\uparrow$ OBL) in (31) to (ADJ $\in \uparrow$).

(31) *by* P $(\uparrow \textsc{pred}) = \text{`by'}$
$((\textsc{obl}\uparrow)\ \textsc{voice}) =_c \textsc{passive}$
$(\uparrow \textsc{obj})_\sigma = ((\textsc{obl}\uparrow)_\sigma\ \textsc{arg}_1)$
$\lambda x \lambda P.[P(x)] : (\uparrow \textsc{obj})_\sigma \multimap [\uparrow_\sigma \multimap (\textsc{obl}\uparrow)_\sigma] \multimap (\textsc{obl}\uparrow)_\sigma$

Rather than existentially closing the suppressed argument of its passive verb, the *by*-phrase saturates the corresponding argument of the passive with the OBJ in the *by*-phrase (e.g., *Godzilla* in *by Godzilla*). Needham and Toivonen (2011) also note that the nominal in the *by*-phrase must fill the role of whatever was the highest/suppressed argument of the verb that it modifies; this is accomplished through the equation $(\uparrow \textsc{obj})_\sigma = ((\textsc{obl}\uparrow)_\sigma\ \textsc{arg}_1)$ in the lexical entry in (31).

Figures 3 and 4 respectively show analyses of examples (28) and (29). From this point on, for reasons of space, we do not show c-structures, since they can be inferred from f-structures. Figures 9 and 10 show Glue proofs for the examples.

4.3 Instrumentals

The last case we consider is instrumental *with*-phrases:

(32) a. Robin killed Sandy.
b. Robin killed Sandy with dynamite.

(33) a. An explosion killed Sandy.
b. #An explosion killed Sandy with dynamite.

Instrumental *with*-phrases, like passive *by*-phrases, are instances of Needham and Toivonen's 'derived arguments'.

Following Reinhart (2002), Needham and Toivonen (2011: 415) note that instrumental *with*-phrases are only well-formed with "agent verbs", cf. (32b) vs. (33b). Stated as such, a generalization seems to be missed, because we seem to require two verbs *kill*: agentive *kill*$_1$ and non-agentive *kill*$_2$. Our analysis avoids this undesirable outcome while properly capturing the empirical generalization.

We capture the contrast through the same kind of standard restrictive semantics used for *eat* above, by imposing a requirement of animacy on the subject argument while simultaneously adding the information that the object of the *with*-phrase is

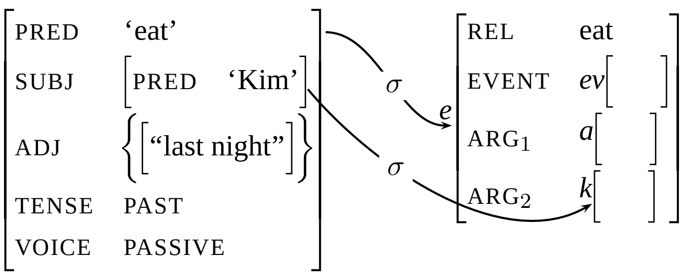


Figure 3: Relevant structures and correspondences for *Kim was eaten last night.*

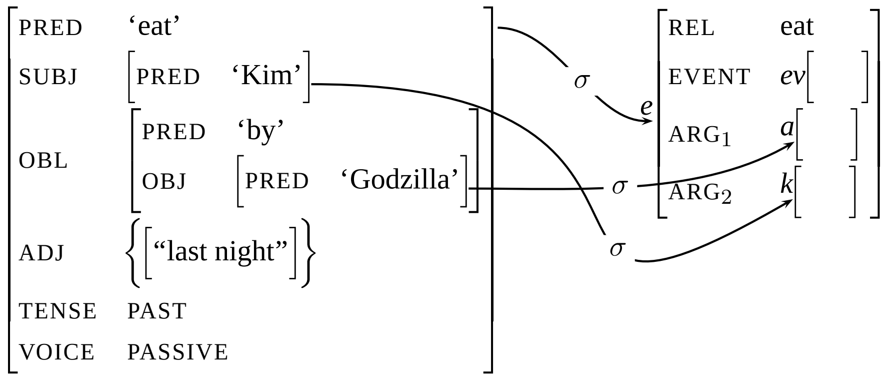


Figure 4: Relevant structures and correspondences for *Kim was eaten by Godzilla last night.*

an instrument in the event. We assume that, unlike passive *by*-phrases, instrumental *with*-phrases add an argument that is not otherwise linguistically represented for the predicate, thus constituting central cases of 'derived arguments'. However, this is accomplished through lexical information associated with instrumental *with*, rather than by directly modifying the lexical entry of the verb.[13] Consider example (34) in light of the lexical entry in (35).

(34) Kim tapped Sandy with Excalibur.

(35) *with* P

$(\uparrow \text{PRED}) = \text{`with'}$

$(\uparrow \text{OBJ})_\sigma = ((\text{OBL} \uparrow)_\sigma \text{ INSTRUMENT})$

$\lambda y \lambda P \lambda x \lambda e.[P(x)(e) \ \wedge \ \mathit{animate}(x) \ \wedge \ \mathit{instrument}(e) = y] :$
$(\uparrow \text{OBJ})_\sigma \multimap$
$[((\text{OBL} \uparrow) \text{ SUBJ})_\sigma \multimap ((\text{OBL} \uparrow)_\sigma \text{ EVENT}) \multimap (\text{OBL} \uparrow)_\sigma] \multimap$
$((\text{OBL} \uparrow) \text{ SUBJ})_\sigma \multimap ((\text{OBL} \uparrow)_\sigma \text{ EVENT}) \multimap (\text{OBL} \uparrow)_\sigma$

No mention is made of the thematic role of the subject, allowing it to be the same role whether the instrumental is present or not. The f-structure and semantic structure for (34) are shown in Figure 5 and the Glue proof is shown in Figure 11.

5 Further Capturing Lexical Generalizations

An LFG template or macro is an abbreviation for a set of equations or constraints (Dalrymple et al. 2004, Asudeh et al. 2008, Crouch et al. 2011, Asudeh 2012). A template is referenced in a lexical entry, as denoted by the '@' prefix. The semantics of template invocation is simple substitution: any template in a lexical

[13] Once again, for consistency with Needham and Toivonen (2011), we treat the *with*-phrase as an OBL, but once again this does not substantively affect our analysis; see footnote 12.

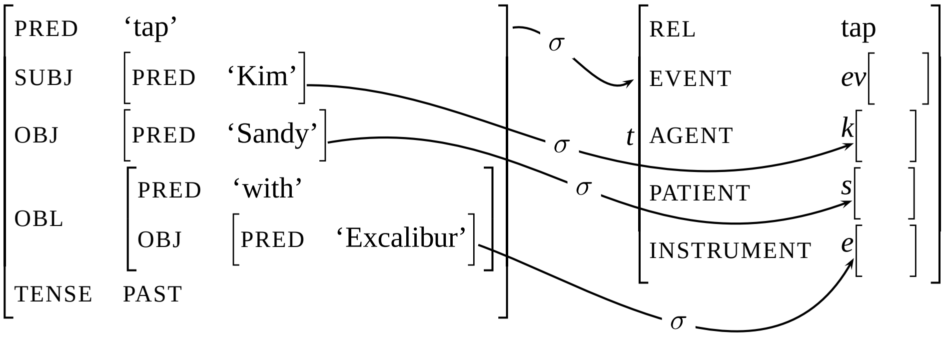


Figure 5: Relevant structures and correspondences for *Kim tapped Sandy with Excalibur.*

entry can be equivalently replaced by the contents of template. Even though they are purely abbreviatory devices, templates can capture linguistic generalizations, since they cross-classify the lexical entries that contain the same templates. Thus, even though a grammar with templates is extensionally equivalent to a grammar with all template calls substituted with the contents of the templates, the former grammar might express generalizations that the latter does not.

The cases that we have examined demonstrate this. It is clear that there is something common to semantically relational verbs — e.g., *eat*, *drink*, *devour*, and *quaff* — and it is also clear that these verbs further subcategorize into the optionally transitive — e.g., *eat* and *drink* — versus the obligatorily transitive — e.g., *devour*, and *quaff*. The following templates and lexical entries demonstrate how templates can capture such generalizations:

(36) PAST = $(\uparrow \text{TENSE}) = \text{PAST}$

(37) AGENT-PATIENT-VERB =

$(\uparrow \text{SUBJ})_\sigma = (\uparrow_\sigma \text{ARG}_1)$

$\lambda P \lambda y \lambda x \lambda e.P(e) \ \wedge \ agent(e) = x \ \wedge \ patient(e) = y :$
$[(\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma] \multimap (\uparrow_\sigma \text{ARG}_2) \multimap (\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma$

(38) OPTIONAL-TRANSITIVE = $(\uparrow_\sigma \text{ARG}_2)$

$\lambda P \exists x.[P(x)] : [(\uparrow_\sigma \text{ARG}_2) \multimap \uparrow_\sigma] \multimap \uparrow_\sigma$

(39) *ate* V $(\uparrow \text{PRED}) = \text{'eat'}$

@PAST

@AGENT-PATIENT-VERB

$$\left(\begin{array}{l} \text{@OPTIONAL-TRANSITIVE} \\ \lambda P \lambda y \lambda x \lambda e.P(y)(x)(e) \wedge food.for(y,x) : \\ [(\uparrow_\sigma \text{ARG}_2) \multimap (\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma] \multimap \\ \quad (\uparrow_\sigma \text{ARG}_2) \multimap (\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma \end{array} \right)$$

$\lambda e.eat(e) : (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma$

(40) *devoured* V $(\uparrow \text{PRED}) = \text{`devour'}$
$@\text{PAST}$
$@\text{AGENT-PATIENT-VERB}$
$(\uparrow \text{OBJ})_\sigma = (\uparrow_\sigma \text{ARG}_2)$
$\lambda e.devour(e) : (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma$

(41) $\text{PASSIVE} =$ $(\uparrow \text{VOICE}) = \text{PASSIVE}$
$(\uparrow \text{SUBJ})_\sigma = (\uparrow_\sigma \text{ARG}_2)$
$(\uparrow_\sigma \text{ARG}_1)$
$(\ \lambda P \exists x.[P(x)] : [(\uparrow_\sigma \text{ARG}_1) \multimap \uparrow_\sigma] \multimap \uparrow_\sigma\ )$

(42) *eaten* V $(\uparrow \text{PRED}) = \text{`eat'}$
$@\text{PASSIVE}$
$\lambda x \lambda y \lambda e.eat(e)\ \wedge\ agent(e) = x\ \wedge\ patient(e) = y :$
$(\uparrow_\sigma \text{ARG}_1) \multimap\ (\uparrow \text{SUBJ})_\sigma\ \multimap\ (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma$

Reasons of space preclude us from discussing these entries carefully, but it should be evident that much information has been moved out of particular lexical entries into templates that generalize across lexical entries. We do note that the Glue meaning constructors have been modified as a result of the new distribution of information, which further highlights the flexibility of resource-sensitive composition in Glue Semantics. One welcome result of this modification is that the core, obligatory information of verbs is now just a predicate on events, such that each verb adds very little parochial information. This can also be readily extended for verbs like *devour* such that the core meaning involves an eating event and an appropriate intensifying adverbial that is a manner modifier of the eating event.

6 Conclusion

We sought to answer two main questions about optional and derived arguments, and to meet a number of challenges constituted by these phenomena. The first question concerned implications for the syntax–semantics interface. We have presented an analysis which merges argument structure and semantic structure and which depends on flexible composition in Glue Semantics. This flexibility ultimately derives from the fact that Glue is a type-logical approach that separates syntax and semantics, very much in the spirit of LFG, such that the logic of composition is commutative. The second question concerned how to properly capture lexical generalizations about the relevant cases. Flexible composition again featured here: lexical entries can contribute obligatory, core meaning constructors as well as optional, modificational meaning constructors, where the optionality is captured by LFG's normal language of lexical specification. LFG templates can capture yet further lexical generalization.

Our analysis meets various general challenges to analyses of these phenomena. First, we do not treat the distinct valencies of the predicates in question as ambiguities (accidental homonymy), but rather as involving core information and modificational information, which interacts properly with optionality. This modificational information is intuitively and formally adjunct-like, which perhaps sheds some light on why these cases have adjunct-like behaviour. It also enables semantic restrictions on optional arguments to be captured. Lastly, resource-sensitive semantic composition in Glue Semantics ensures that the obligatory and optional information interact properly.

In conclusion, this sort of approach can constitute the first step toward a more general theory of arguments and adjuncts, although we are not necessarily wed to all of the formal details. In the companion paper in this volume Giorgolo and Asudeh (2012), we present a different way to deal with the issue of unexpressed arguments in a resource-sensitive semantics.

References

Asudeh, Ash. 2004. Resumption as Resource Management. Ph.D. thesis, Stanford University.

—. 2012. *The Logic of Pronominal Resumption*. Oxford: Oxford University Press.

Asudeh, Ash, Mary Dalrymple, and Ida Toivonen. 2008. Constructions with Lexical Integrity: Templates as the Lexicon-Syntax Interface. In Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG08 Conference*, 68–88. Stanford, CA: CSLI Publications.

Bouma, Gosse, Robert Malouf, and Ivan A. Sag. 2001. Satisfying Constraints on Extraction and Adjunction. *Natural Language and Linguistic Theory* 19: 1–65.

Bresnan, Joan. 1978. A Realistic Transformational Grammar. In Morris Halle, Joan Bresnan, and George A. Miller, eds., *Linguistic Theory and Psychological Reality*, 1–59. Cambridge, MA: MIT Press.

—. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2013. *Lexical-Functional Syntax*. Oxford: Wiley-Blackwell, 2nd edn. Forthcoming.

Butt, Miriam. 1995. *The Structure of Complex Predicates in Urdu*. Stanford, CA: CSLI Publications.

Butt, Miriam, Mary Dalrymple, and Anette Frank. 1997. An Architecture for Linking Theory in LFG. In Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG97 Conference*. Stanford, CA: CSLI Publications.

Butt, Miriam, and Tracy Holloway King, eds. 2011. *Proceedings of the LFG11 Conference*. Stanford, CA: CSLI Publications.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

—. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Crouch, Richard, Mary Dalrymple, Ronald M. Kaplan, Tracy King, John T. Maxwell, III, and Paula Newman. 2011. *XLE Documentation*. Palo Alto Research Center (PARC), Palo Alto, CA.

Dalrymple, Mary, ed. 1999. *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*. Cambridge, MA: MIT Press.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. San Diego, CA: Academic Press.

Dalrymple, Mary, Ronald M. Kaplan, and Tracy Holloway King. 2004. Linguistic Generalizations over Descriptions. In Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG04 Conference*, 199–208. Stanford, CA: CSLI Publications.

Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, eds. 1995. *Formal Issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.

Fillmore, Charles J. 1986. Pragmatically Controlled Zero Anaphora. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*, 95–107. Berkeley: Berkeley Linguistics Society.

Fodor, Jerry A., and Janet Dean Fodor. 1980. Functional Structure, Quantifiers, and Meaning Postulates. *Linguistic Inquiry* 11(4): 759–770.

Giorgolo, Gianluca, and Ash Asudeh. 2011. Multidimensional Semantics with Unidimensional Glue Logic. In Butt and King 2011, 236–256.

—. 2012. Missing Resources in a Resource-Sensitive Semantics. In Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG12 Conference*, This volume. Stanford, CA: CSLI Publications.

Jackendoff, Ray. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Kaplan, Ronald M., and Joan Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*, 173–281. Cambridge, MA: MIT Press. Reprinted in Dalrymple et al. (1995: 29–135).

Kuhn, Jonas. 2001. Resource Sensitivity in the Syntax-Semantics Interface and the German Split NP Construction. In W. Detmar Meurers and Tibor Kiss, eds., *Constraint-Based Approaches to Germanic Syntax*. Stanford, CA: CSLI Publications.

Lasersohn, Peter N. 1993. Lexical Distributivity and Implicit Arguments. In *Proceedings from Semantics and Linguistic Theory III*, 145–161.

Needham, Stephanie, and Ida Toivonen. 2011. Derived Arguments. In Butt and King 2011, 401–421.

Parsons, Terence. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. Cambridge, MA: MIT Press.

Reinhart, Tanya. 2002. The Theta System — An Overview. *Theoretical Linguistics* 28: 229–290.

Toivonen, Ida. 2001. The Phrase Structure of Non-Projecting Words. Ph.D. thesis, Stanford University.

—. 2003. *Non-Projecting Words: A Case Study of Swedish Verbal Particles*. Dordrecht: Kluwer.

$$
\dfrac{\dfrac{
\begin{array}{l}\textbf{PAST}\\ \lambda P\exists e.[P(e)\wedge past(e)]:\\ (ev\multimap e)\multimap e\end{array}
\quad
\dfrac{
\begin{array}{l}\textbf{at noon}\\ \lambda P\lambda e''.[P(e'')\wedge at.noon(e'')]:\\ (ev\multimap e)\multimap(ev\multimap e)\end{array}
\quad
\dfrac{\dfrac{\dfrac{
\begin{array}{l}\textbf{ate (opt.)}\\ \lambda P\lambda y\exists x.[P(x)(y)\ \wedge\ food.for(x,y)]:\\ [p\multimap k\multimap e]\multimap k\multimap e\end{array}
\quad
\dfrac{\begin{array}{l}\textbf{ate}\\ eat':\\ ev\multimap p\multimap k\multimap e\end{array}\quad [e':ev]^1}{eat'(e'):p\multimap k\multimap e}
}{\lambda y\exists x.[eat'(e')(x)(y)\wedge food.for(x,y)]:k\multimap e}
\quad
\begin{array}{l}\textbf{Kim}\\ kim:\\ k\end{array}
}{\exists x.[eat'(e')(x)(kim)\wedge food.for(x,kim)]:e}
}{\lambda e'\exists x.[eat'(e')(x)(kim)\wedge food.for(x,kim)]:ev\multimap e}\multimap_{\mathcal{I},1}
}{\lambda e''\exists x.[eat'(e'')(x)(kim)\wedge food.for(x,kim)\wedge at.noon(e'')]:ev\multimap e}
}{\exists e\exists x.[eat'(e)(x)(kim)\wedge food.for(x,kim)\wedge at.noon(e)\wedge past(e)]:e}
}{\exists e\exists x.[eat(e)\wedge agent(e)=kim\wedge patient(e)=x\wedge food.for(x,kim)\wedge at.noon(e)\wedge past(e)]:e}\Rightarrow_{\beta}
$$

Figure 6: Proof for *Kim ate at noon.*

$$
\dfrac{\dfrac{
\begin{array}{l}\textbf{PAST}\\ \lambda P\exists e.[P(e)\wedge past(e)]:\\ (ev\multimap e)\multimap e\end{array}
\quad
\dfrac{
\begin{array}{l}\textbf{at noon}\\ \lambda P\lambda e'.[P(e')\wedge at.noon(e')]:\\ (ev\multimap e)\multimap(ev\multimap e)\end{array}
\quad
\dfrac{\dfrac{
\begin{array}{l}\textbf{ate}\\ eat':\\ c\multimap k\multimap ev\multimap e\end{array}
\quad
\begin{array}{l}\textbf{the cake}\\ \iota x.[cake(x)]:\\ c\end{array}
}{eat'(\iota x.[cake(x)]):k\multimap ev\multimap e}
\quad
\begin{array}{l}\textbf{Kim}\\ kim:\\ k\end{array}
}{eat'(\iota x.[cake(x)])(kim):ev\multimap e}
}{\lambda e'.[eat'(\iota x.[cake(x)])(kim)(e')\wedge at.noon(e')]:ev\multimap e}
}{\exists e.[eat'(\iota x.[cake(x)])(kim)(e)\wedge at.noon(e)\wedge past(e)]:e}
}{\exists e.[eat(e)\wedge agent(e)=kim\wedge patient(e)=\iota x.[cake(x)]\wedge at.noon(e)\wedge past(e)]:e}\Rightarrow_{\beta}
$$

Figure 7: Proof for *Kim ate the cake at noon.*

$$
\dfrac{
\dfrac{
\begin{array}{l}\textbf{PAST}\\ \lambda P \exists e.[P(e) \wedge past(e)] :\\ (ev \multimap e) \multimap e\end{array}
\quad
\dfrac{
\dfrac{
\begin{array}{l}\textbf{every student}\\ \lambda P \forall z.[student(z) \rightarrow P(z)] :\\ \forall X.[(s \multimap X) \multimap X]\end{array}
\quad
\dfrac{
\begin{array}{l}\textbf{ate (opt.)}\\ \lambda P \lambda y \exists x.[P(x)(y) \wedge food.for(x,y)] :\\ [p \multimap s \multimap e] \multimap s \multimap e\end{array}
\quad
\dfrac{\begin{array}{l}\textbf{ate}\\ eat' :\\ ev \multimap p \multimap s \multimap e\end{array} \quad [e' : ev]^1}{eat'(e') : p \multimap s \multimap e}
}{\lambda y \exists x.[eat'(e')(x)(y) \wedge food.for(x,y)] : s \multimap e}
}{\forall z.[student(z) \rightarrow \exists x.[eat'(e')(x)(z) \wedge food.for(x,z)]] : e} \forall_{\mathcal{E}}[e/X]
}{\lambda e' \forall z.[student(z) \rightarrow \exists x.[eat'(e')(x)(z) \wedge food.for(x,z)]] : ev \multimap e} \multimap_{\mathcal{I},1}
}{\exists e.[\forall z.[student(z) \rightarrow \exists x.[eat'(e'')(x)(z) \wedge food.for(x,z)]] \wedge at.noon(e'') \wedge past(e)] : e}
}{\exists e.[\forall z.[student(z) \rightarrow \exists x.[eat(e) \wedge agent(e) = z \wedge patient(e) = x \wedge food.for(x,z)]] \wedge at.noon(e) \wedge past(e)] : e} \Rightarrow_{\beta}
$$

Figure 8: Proof for subject wide scope reading of *Every student ate*.

$$
\dfrac{
\dfrac{
\begin{array}{l}\textbf{was}\\ \lambda P \exists e.[P(e) \wedge past(e)] :\\ (ev \multimap e) \multimap e\end{array}
\quad
\dfrac{
\begin{array}{l}\textbf{last night}\\ \lambda P \lambda e''.[P(e'') \wedge last.night(e'')] :\\ (ev \multimap e) \multimap (ev \multimap e)\end{array}
\quad
\dfrac{
\dfrac{
\begin{array}{l}\textbf{eaten (opt.)}\\ \lambda P \exists x.[P(x)] :\\ (a \multimap e) \multimap e\end{array}
\quad
\dfrac{
\dfrac{\begin{array}{l}\textbf{eaten}\\ eat' :\\ ev \multimap k \multimap a \multimap e\end{array} \quad [e' : ev]^1}{eat'(e) : k \multimap a \multimap e}
\quad
\begin{array}{l}\textbf{Kim}\\ kim :\\ k\end{array}
}{eat'(e')(kim) : a \multimap e}
}{\exists x.[eat'(e')(kim)(x)] : e}
}{\lambda e' \exists x.[eat'(e')(kim)(x)] : ev \multimap e} \multimap_{\mathcal{I},1}
}{\lambda e'' \exists x.[eat'(e')(kim)(x) \wedge last.night(e'')] : ev \multimap e}
}{\exists e \exists x.[eat'(e)(kim)(x) \wedge last.night(e) \wedge past(e)] : e}
}{\exists e \exists x.[eat(e) \wedge agent(e) = x \wedge patient(e) = kim \wedge last.night(e) \wedge past(e)] : e} \Rightarrow_{\beta}
$$

Figure 9: Proof for *Kim was eaten last night.*

was $\lambda P \exists e.[P(e) \wedge past(e)] : (ev \multimap e) \multimap e$

last night $\lambda P \lambda e''.[P(e'') \wedge last.night(e'')] : (ev \multimap e) \multimap (ev \multimap e)$

by $\lambda x \lambda P.[P(x)] : g \multimap (a \multimap e) \multimap e$

Godzilla $godzilla : g$

eaten $eat' : ev \multimap k \multimap a \multimap e$ $\quad [e' : ev]^1$

Kim $kim : k$

$$eat'(e) : k \multimap a \multimap e$$

$$eat'(e')(kim) : a \multimap e$$

$$\lambda P.[P(godzilla)] : (a \multimap e) \multimap e$$

$$eat'(e')(kim)(godzilla) : e$$

$$\lambda e'.[eat'(e')(kim)(godzilla)] : ev \multimap e \quad \multimap_{\mathcal{I},1}$$

$$\lambda e''.[eat'(e')(kim)(godzilla) \wedge last.night(e'')] : ev \multimap e$$

$$\exists e.[eat'(e)(kim)(godzilla) \wedge last.night(e) \wedge past(e)] : e$$

$$\exists e.[eat(e) \wedge agent(e) = godzilla \wedge patient(e) = kim \wedge last.night(e) \wedge past(e)] : e \quad \Rightarrow_\beta$$

Figure 10: Proof for *Kim was eaten by Godzilla last night.*

PAST $\lambda P \exists e.[P(e) \wedge past(e)] : (ev \multimap t) \multimap t$

with $\lambda y \lambda P \lambda x \lambda e.[P(x)(e) \wedge animate(x) \wedge instrument(e) = y] : e \multimap (k \multimap ev \multimap t) \multimap k \multimap ev \multimap t$

Excalibur $excalibur : e$

tapped $tap' : s \multimap k \multimap ev \multimap t$

Sandy $sandy : s$

Kim $kim : k$

$$\lambda P \lambda x \lambda e.[P(x)(e) \wedge animate(x) \wedge instrument(e) = excalibur] : (k \multimap ev \multimap t) \multimap k \multimap ev \multimap t$$

$$tap'(sandy) : k \multimap ev \multimap t$$

$$\lambda x \lambda e.[tap'(sandy)(x)(e) \wedge animate(x) \wedge instrument(e) = excalibur] : k \multimap ev \multimap t$$

$$\lambda e.[tap'(sandy)(kim)(e) \wedge animate(kim) \wedge instrument(e) = excalibur] : ev \multimap t$$

$$\exists e.[tap'(sandy)(kim)(e) \wedge animate(kim) \wedge instrument(e) = excalibur] : t$$

$$\exists e.[tap(e) \wedge agent(e) = kim \wedge patient(e) = sandy \wedge animate(kim) \wedge instrument(e) = excalibur \wedge past(e)] : t \quad \Rightarrow_\beta$$

Figure 11: Proof for *Kim tapped Sandy with Excalibur.*

EXPLORING THE TREATMENT OF SELECTED TYPOLOGICAL CHARACTERISTICS OF TSWANA IN LFG

Ansu Berg
North-West University

Rigardt Pretorius
North-West University

Laurette Pretorius
University of South Africa

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King
(Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu

Abstract

Tswana is a Bantu language in the south eastern zone of Bantu languages and one of the eleven official languages of South Africa. From a linguistic as well as a computational point of view the language is a lesser-studied and under-resourced language. Recently a project was undertaken to describe the syntactic structure of Tswana in the LFG formalism and to develop parser support for this using the XLE parser toolkit. In this paper a short overview of selected distinctive typological characteristics of Tswana, including the word order, agglutinative character, disjunctive orthography and agreement, is presented. The aim is to demonstrate how selected instances of agreement in Tswana can be modeled in LFG and XLE.

1 Introduction

Tswana is a Bantu language in the south eastern zone (zone S in Doke's classification) of Bantu languages and is one of the three languages in the Sotho language group (Cole, 1959; Guthrie, 1971). The other two Sotho languages are Northern Sotho (Sepedi) and Southern Sotho. Tswana is predominantly spoken in South Africa and Botswana. It is one of the eleven official languages of South Africa with approximately 3,272,720 (8.2% of the population) first language speakers (Statistics South Africa Census 2001, 2004: 9). In Botswana, Tswana is the only national language and approximately 1,070,000 (79,06% of the population) in Botswana speak Tswana as first language (Central Statistics Office, 2009:14).

Existing grammatical descriptions of Tswana, mainly focussing on the morphology, are to a large extent based on the structural functional approach (Cole, 1955; Krüger, 2006). Recently a description of the syntactic structure of Tswana in the constraint-based LFG formalism (Dalrymple, 2001) was commenced.

In recent years the following core technologies in Natural Language Processing (NLP) were developed for Tswana:

- a proposed word-class tagset (Van Rooy and Pretorius, 2003)
- a lemmatiser (Brits *et al.*, 2005; Brits, 2006)
- a morphological analyser and a tokeniser (Pretorius *et al.*, 2009)

This work is part of a bigger project to develop a syntactic parser for Tswana, which will in due course form part of the NLP pipeline for Tswana. An LFG description of the syntactic structures of Tswana is undertaken, which will serve as basis for the parser development, using the XLE parser toolkit (Crouch *et al*., 2011).

Agreement is a typical typological characteristic of Tswana. The aim of this paper is to describe the employment of selected instances of subject-verb and noun phrase internal agreement in LFG and XLE. For this purpose an overview of the agglutinative character, word order, and disjunctive orthography of Tswana is provided.

2 Typological features of Tswana

2.1 Agglutinative language

Tswana is an agglutinative language. It is characterised by a complex morphology where affixation is prominent. Affixes modify or extend the meaning of words (Krüger, 2006:40-41).

The noun class prefixes provide essential information regarding class and number features of nouns. Noun suffixes extend the meaning of nouns and provide information regarding certain characteristics (Krüger, 2006: 73-96). In example (1) the meaning of the noun **setlhare** is extended by adding the diminutive and locative suffixes:

(1) **(mo) setlharenyaneng** '(here) in the little tree'
setlharenyaneng
se- + -tlhare + -ana + -ing
NPre7+tree+Dim+Loc

The verbal prefixes and suffixes provide essential information regarding type, tense, aspect and mood (Krüger, 2006: 198-243). The verb in example (2) comprises a verbal root **-thus-** to which prefixes (subject and object agreement morphemes) and suffixes (causative, perfect, verbal ending) are added:

(2) **ba re thusitse (go fetsa tiro ya rona)**
'they helped us (to finish our work)'
ba re thusitse
ba + re + thus- + -is- + -il- + -e
AgrSubj-Cl2+AgrObj-p1-Pl+help+Caus+Perf+VerbEnd

2.2 Word classes

Tswana words are divided into word classes on the basis of similarities between certain words. Words in Tswana have autonomous word status. This has been proven through the application of one or more of the word tests, namely isolatability, separability, transposability and replaceability (Van Wyk, 1967:230-261; Krüger, 2006:12-16; Louwrens and Poulos, 2006:392). The major word categories in Tswana are nouns, verbs, pronouns, particles, adverbs, idiophones and interjections. Within these categories nouns and verbs are open word classes on the basis of their productive morphology. Pronouns, particles, adverbs, interjections and idiophones are closed word classes because they can be exhaustively listed and are morphologically unproductive.

Sub-categories of the word classes are also distinguished on the grounds of similarities between words within a specific word category. The sub-categories of the nouns include amongst others basic nouns, adjectival nouns, deverbative nouns and locative nouns. An adjective is a special kind of noun because it is characterised by a class prefix which changes in accordance with the class of the qualified noun (Lombard *et al*., 1993:57). Proper verbs, auxiliary verbs and copulative verbs are sub-categories of verbs (Krüger, 2006:24). Absolute pronouns, demonstrative pronouns, quantitative pronouns and possessive pronouns are sub-categories of pronouns (Krüger, 2006:24). The associative particle, instrumental particle, locative particle, possessive particles, qualificative particles as well as the conjunctions are sub-categories of the word class particle (Krüger, 2006:25).

Basic Tswana nouns consist morphologically of a class prefix and a root. Twenty noun classes are distinguished in Tswana (Krüger, 2006:57-70). In classes 1 to 14 the even classes contain singular nouns and the odd classes contain plural nouns. Classes 15 to 20 do not refer to singular or plural (Krüger, 2006:57-124).

Example (3) shows a singular and plural noun with the same root. The singular noun class prefix ***mo-*** indicates a class 1 noun, while the plural class prefix ***ba-*** indicates a noun in class 2:

(3)	**monna** 'man'	>	**banna** 'men'
	mo- *+ -nna*		***ba-*** *+ -nna*
	NPre1+man		**NPre2**+men

The agreement system in Tswana is based on the noun class prefixes (Louwrens, 1994:9-10). Words in a syntactic relation to a specific noun exhibit formal similarities with the class prefix of that noun. For example, the class prefix of the subject noun determines the form of the subject agreement morpheme on the verb:

(4) **Banna ba bua Setswana.** 'The men speak Tswana.'

banna	**ba** bua	Setswana
ba- *+-nna*	***ba*** *+ bu-+ -a*	*se- + -tswana*
NPre2+men	**AgrSubj-Cl2**+speak+VerbEnd	NPre7+tswana

Nouns are also used in relation to other words which modify them. The noun class of the head noun determines the agreement affixes on modifiers within the NP as illustrated in example (5):

(5) **banna ba** ‘these men’ (‘men these’)

banna	**ba**
ba- + -*nna*	***ba***
NPre2+men	**DemPro-Cl2**

Words in a syntactic relation to nouns therefore present formal similarities to that noun (Louwrens, 1994:9-10). Various morphemes and words are derived from the noun class prefixes. This applies to, for example, the forming of subject agreement morphemes, object agreement morphemes, pronouns such as absolute pronouns, possessive pronouns, demonstrative pronouns and particles such as possessive particles and qualificative particles, etc. In example (6) agreement regarding noun class 8 is illustrated:

(6) An example of agreement in noun class 8:

• ***noun class prefix***

di: ditlhako ‘shoes’

ditlhako
di- + -*tlhako*
NPre8+shoes

• ***subject agreement morpheme***

di: Ditlhako di latlhegile. ‘The shoes got lost.’

ditlhako	di latlhegile
di- + -*tlhako*	***di*** + *latlheg*- + -*il*- + -*e*
NPre8+shoes	**AgrSubj-Cl8**+lose+Perf+VerbEnd

• ***consecutive subject agreement morpheme***

tsa: Ditlhako tsa latlhega. ‘The shoes then got lost.’

ditlhako	tsa latlhega
di- + -*tlhako*	***tsa*** + *latlheg*- + -*a*
NPre8+shoes	**AgrSubjCons-Cl8**+lose+VerbEnd

• ***object agreement morpheme***

di: Basadi ba a di reka. ‘The women buy it.’

basadi	ba a di reka
ba- + -*sadi*	*ba* + *a* + ***di*** + *rek*- + -*a*
NPre2+women	AgrSubj-Cl2+AspPr+**AgrObj-Cl8**+buy+VerbEnd

• ***absolute pronoun***

tsone: Ditlhako tsone re di rekile. ‘As for the shoes, we bought them.’

ditlhako	tsone	re di rekile
di- + -*tlhako*	***tsone***	*re* + *di* + *rek*- + -*il*- + -*e*
NPre8+shoes	**AbsPro-Cl8**	AgrSubj-p1-Pl+AgrObj-Cl8+buy +Perf+VerbEnd

• ***demonstrative pronoun (distance 1)***

tse: ditlhako tse 'these shoes'

ditlhako	tse	
di- + -tlhako	***tse***	
NPre8+shoes	**DemPro-Cl8-d1**	

• ***demonstrative pronoun (distance 2)***

tseo: ditlhako tseo 'those shoes'

ditlhako	tseo	
di- + -tlhako	***tseo***	
NPre8+shoes	**DemPro-Cl8-d2**	

• ***demonstrative pronoun (distance 3)***

tsele: ditlhako tsele 'those shoes "over there" '

ditlhako	tsele	
di- + -tlhako	***tsele***	
NPre8+shoes	**DemPro-Cl8-d3**	

• ***possessive particle***

tsa: ditlhako tsa basadi 'women's shoes'

ditlhako	tsa	basadi
di- + -tlhako	***tsa***	*ba- + -sadi*
NPre8+shoes	**PosPart-Cl8**	NPre2+women

• ***possessive pronoun***

tsone: mebala ya tsone 'their colours' ('colours of them')

mebala	ya	tsone
me- + -bala	*ya*	***tsone***
NPre4+colours	PosPart-Cl4	**PosPro-Cl8**

• ***qualificative particle***

tse: ditlhako tse dintsi 'many shoes' ('shoes that are many')

ditlhako	tse	dintsi
di- + -tlhako	***tse***	*di- + -ntsi*
NPre8+shoes	**QualPart-Cl8**	NPre8+many

2.3 Disjunctive orthography

A disjunctive orthography is used for Tswana verbs (Kosch, 1993:43). The prefixes are usually written disjunctively but the suffixes are written conjunctively. The consequence of this writing style is that Tswana words cannot be tokenised only on white space (Pretorius *et al.*, 2009). The correct identification of Tswana word boundaries is essential in the identification of the constituents of Tswana sentences. This is illustrated and explained in examples (7) and (8):

(7) **o a e reka** 'she buys it'
o + a + e + rek- + -a
AgrSubj-Cl1+Asp+AgrObj-Cl9+buy+VerbEnd

The verb **o a e reka** in example (7) represents one token and this token consists of four orthographic items.

(8) **Mosadi o badile dibuka.** 'The woman read the book.'

mosadi	o badile	dibuka
mo- + -sadi	*o + bal- + -il- + -e*	*di- + -buka*
NPre1+woman	AgrSubj-Cl1+read+Perf+VerbEnd	NPre10+book

The Tswana sentence in example (8) consists of three tokens, namely *mosadi /o badile / dibuka*. While the verb **o badile** consists of two orthographic items it represents only one token.

In Tswana a linguistic verb can be a sequence of orthographic items that together function as members of that verb. These orthographic items are also referred to as orthographic words (Louwrens and Poulos, 2006:393).

2.4 Word order

2.4.1 Word order in a simple Tswana sentence

The basic word order in simple Tswana sentences is SVO where the subject precedes a verb and an object appears post verbally (Krüger, 2006:11-12). This word order is illustrated in the following example:

(9) **Banna ba bua Setswana.** 'The men speak Tswana.'

Subject	Verb	Object
banna	ba bua	Setswana
ba- + -nna	*ba + bu- + -a*	*se- + -tswana*
NPre2+men	AgrSubj-Cl2+speak+VerbEnd	NPre7+tswana

The phrase structure rule indicating the word order followed in a simple Tswana sentence is as follows:

$$\begin{array}{lll} S \rightarrow & NP & VP \\ & (\uparrow SUBJ)=\downarrow & \uparrow=\downarrow \end{array}$$

Tswana phrases are head initial. The Tswana verb phrase (VP) can be made up of a verb and a noun phrase (NP) and the following phrase structure rule is then followed (Department of African Languages and Literature, 2000:10):

$$\begin{array}{lll} VP \rightarrow & V & NP \\ & \uparrow=\downarrow & (\uparrow OBJ)=\downarrow \end{array}$$

One or more obliques or adjuncts referring to place, time, manner, etc. can be incorporated in a simple Tswana sentence. An SVOX word order is then followed, where 'X' represents the obliques and adjuncts (Creissels, 2000:250-252).

2.4.2 Word order in Tswana noun phrases

In a Tswana NP the head appears in initial position and it is followed by a variable number of modifiers (determiners) (Creissels, 2000:232). Several nominal words or word phrases can act as the head of the NP. If the head is a noun, the following phrase structure rule is used (King and Dalrymple, 2004:71):

NP →	N	Det
	↑=↓	↑=↓

Within the NP, all of the modifiers follow the head noun and modify the head regarding some quality or characteristic (Krüger, 2006:301). Examples of noun modifiers are:

- **a pronoun (personal, absolute, demonstrative, quantitative)**

(10) **banna ba** 'these men'

banna	ba
ba- + -nna	*ba*
NPre2+men	DemPro-Cl2-d1

- ***a possessive phrase (consists of a possessive particle and a complement such as a noun)***

(11) **sekolo sa basimane** 'the boys' school'

sekolo	sa	basimane
se- + -kolo	*sa*	*ba- + -simane*
NPre7+school	PosPart-Cl7	NPre2+boys

- **a qualificative phrase (consists of qualificative particle and a complement such as an adjective)**

(12) **banna ba bagolo** 'big men'

banna	ba	bagolo
ba- + -nna	*ba*	*ba- + -golo*
NPre2+men	QualPart-Cl2	NPre2+big

2.5 Agreement

Agreement in Tswana is observed in the relationship between verbs and nouns and all instances where nouns occur in relationship with other words that modify (qualify) them (Watters, 2000:202).

2.5.1 Subject-verb agreement in a Tswana sentence

Subject verb agreement is established through the subject agreement morpheme on the verb. Subject agreement morphemes agreeing with nouns indicate class. The class feature implies number (singular or plural). Subject agreement morphemes agreeing with personal pronouns indicate person and number (Krüger, 2006: 171-175).

The following sentence (example (13)) has a noun as subject. The CLASS feature is included in both the subject (noun) and the verb. The subject agreement morpheme agrees with the class of the noun and therefore obtains a valid f-string:

(13) **Basadi ba reka dijo.** ‘The women buy food.’

basadi	**ba** reka	dijo
ba- *+ -sadi*	***ba*** *+ rek- + -a*	*di- + -jo*
NPre2+women	**AgrSubj-Cl2**+buy+VerbEnd	NPre8+food

basadi:	(↑PRED) = ‘BASADI’
	(↑CLASS) = 2
ba reka:	(↑PRED) = ‘BA REKA <SUBJ OBJ>’
	(↑SUBJ CLASS) = 2
dijo:	(↑PRED) = ‘DIJO’
	(↑CLASS) = 8

2.5.2 NP-internal agreement in Tswana

Modifiers (determiners) in a NP agree with the head noun (Louwrens, 1994:52). Apart from locative noun phrases where the demonstrative pronoun precedes the noun, all other nominal modifiers are post modifiers (Louwrens, 1994:10). They are the different pronouns, the possessive phrase, the adjectival phrase and the verbal relative phrase.

The CLASS feature is used to validate NP internal agreement at the level of functional structure. If the specification from the head noun can unify with the specification of the determiner then a valid f-structure is obtained.

Examples (14), (15) and (16) show NP internal agreement:

- ***Noun (Head) + demonstrative pronoun***

In a NP in which the head noun is modified by a demonstrative pronoun, both the noun and demonstrative pronoun must show the same class information:

(14) **sekolo seo** ‘that school’ (‘school that’)

sekolo	seo
se- *+ -kolo*	***seo***
NPre7+school	DemPro-Cl7

sekolo:	(↑PRED) = ‘SEKOLO’
	(↑CLASS) = 7
seo:	(↑PRED) = ‘SEO’
	(↑CLASS) = 7

- ***Noun (Head) + possessive phrase***

Tswana only has a long possessive construction (Krüger, 2006: 139-145). The possessive particle agrees with the head noun (example (15)). Should the CLASS specification from the head unify with that of the possessive particle then a valid f-structure is obtained.

(15)	**sekolo sa basimane**	'the boys' school'	('school of boys')
	sekolo	sa	basimane
	se- + *-kolo*	*sa*	*ba-* + *-simane*
	NPre7+school	**PosPart-Cl7**	NPre2+boys

sekolo: (↑PRED) = 'SEKOLO'
(↑CLASS) = 7
sa: (↑PRED) = 'SA'
(↑CLASS) = 7
bone: (↑PRED) = 'BASIMANE'
(↑CLASS) = 2

- ***Noun (Head) + adjectival phrase***

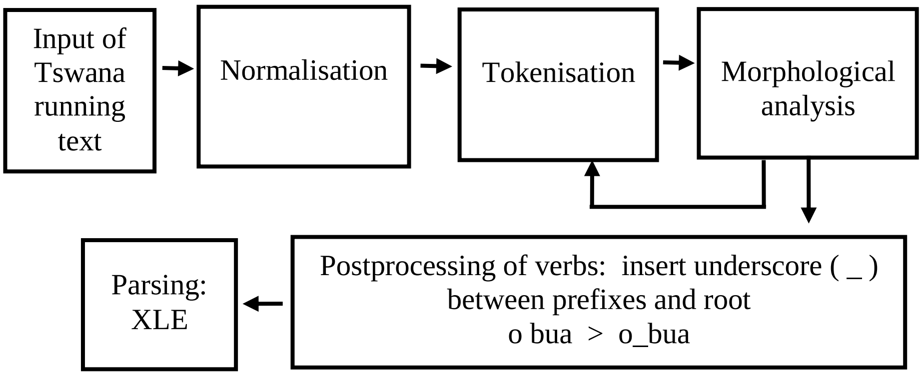
NP's in which an adjective modifies a head noun in Tswana differ in internal structure from the equivalent NP's in English (example (16)) (Krüger, 2006:150). In these phrases both the qualificative particle and the adjective agree with the head noun. The head noun, qualificative particle and the adjective must show the same class information.

(16)	**banna ba bagolo**	'big men' ('men which are big')	
	banna	ba	bagolo
	ba- + *-nna*	***ba***	***ba-*** + *-golo*
	NPre2+men	QualPart-Cl2	NPre2+big

banna: (↑PRED) = 'BANNA'
(↑CLASS) = 2
ba: (↑PRED) = 'BA'
(↑CLASS) = 2
bagolo: (↑PRED) = 'BAGOLO'
(↑CLASS) = 2

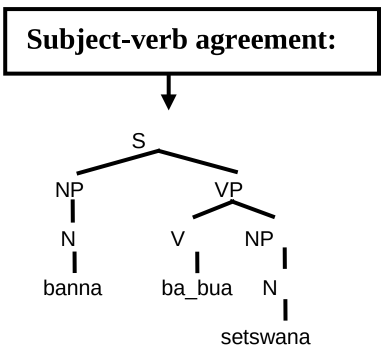
3 The treatment of Tswana agreement in XLE

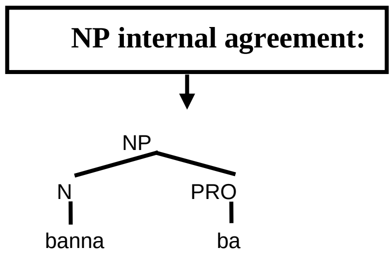
The Tswana NLP pipeline can be presented schematically as follows:



(Crouch *et al.*, 2011; Pretorius *et al.*, 2009)

The following section shows the LFG description of examples (9) and (10), and their modelling in XLE. Agreement is validated by the CLASS specification of the subject and the verb (example 9) and that of the head and modifier in (example 10):





1 valid f-structure

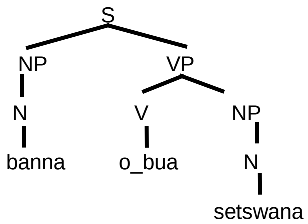
'banna ba_bua Setswana"

PRED 'BU<[NNA], [TSWANA]>'
SUBJ [PRED 'NNA' CLASS c2]
OBJ [PRED 'TSWANA' CLASS c7]
CLASS c2, TENSE pres

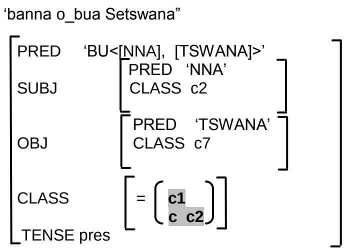
1 valid f-structure

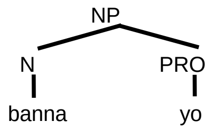
'banna ba"

PRED 'NNA'
CLASS c2, PRON-TYPE 'Demon'

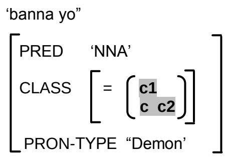


0 valid f-structure (inconsistent)





0 valid f-structure (inconsistent)



4 Conclusion

Tswana has distinctive typological characteristics regarding agglutinative character, word classes, disjunctive orthography, word order and agreement. Agreement is based on the noun classes and personal pronouns. Only a limited number of instances of agreement are presented in this paper. The agreement features of LFG are found to be appropriate for the modelling of some instances of subject-verb and NP-internal agreement in Tswana. Their implementation in XLE could also be done satisfactorily.

As a next step a description of the structure of the simple Tswana sentence will be attempted. All instances of subject-verb and NP-internal agreement will then be described in LFG and XLE.

References

Brits, Karien. 2006. Outomatiese Setswana Lemma-identifisering. Potchefstroom: NWU. (Dissertation – MA). 205p.

Brits, Karien, Pretorius, Rigardt and Van Huyssteen, Gerhard. 2005. Automatic Lemmatization in Setswana: Towards a Prototype. *South African Journal of African Languages*, 25(11):37-47.

Central Statistics Office. 2009. Botswana Demographic Survey 2006. Gaborone: Government Printer. 144p.

Cole, Desmond. 1955. An Introduction to Tswana Grammar. Cape Town: Longman. 473p.

Cole, Desmond. 1959. Doke's classification of Bantu Languages. *African Studies*, 18(4):197-213.

Creissels, Dennis. 2000. Typology. (*In* Heine, Bernd and Nurse, Derek, *eds*. African Languages. An Introduction. Cambridge: Cambridge University Press. p.231-258.)

Crouch, Dick, Dalrymple, Mary, Kaplan, Ron, King, Tracy Holloway, Maxwell, John and Newman, Paula. XLE Documentation. California: PARC. http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html.

Dalrymple, Mary. 2001. Lexical Functional Grammar. Syntax and Semantics Series, Volume 34. New York: Academic Press. 473p.

Department of African Languages and Literature (University of Botswana). 2000. The structure of Setswana sentences. An introduction. Gaborone: Lightbooks. 73p.

Guthrie, Malcolm. 1971. Comparative Bantu: An introduction to the Comparative Linguistics and Prehistory of the Bantu languages. Volume 2. Farnborough: Greg International Publishers. 121p.

King, Tracy Holloway and Dalrymple, Mary. 2004. Determiner agreement and noun conjunction. *Journal of Linguistics*, 40:69-104.

Kosch, Inge. 1993. A historical perspective on Northern Sotho linguistics. Pretoria: UP. (Thesis – PhD). 216p.

Krüger, Caspar. 2006. Introduction to the Morphology of Tswana. München: Lincom. 314p.

Lombard, Daan, Van Wyk, Egidius and Mokgokong, Pothinus. 1993. Introduction to the grammar of Northern Sotho. Pretoria: J.L. van Schaik. 203p.

Louwrens. Louis. 1994. Dictionary of Northern Sotho Grammatical Terms. Pretoria: Via Afrika. 257p.

Louwrens, Louis and Poulos, George. 2006. The status of the word in selected conventional writing systems - the case of disjunctive writing. *Southern African Linguistics and Applied Language Studies*, 24(3):389-401.

Pretorius, Rigardt, Berg, Ansu, Pretorius, Laurette and Viljoen, Biffie. 2009. Setswana Tokenisation and Computational Verb Morphology: Facing the Challenge of a Disjunctive Orthography. (*In* AfLaT2009. Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages. Hosted by the European Association for Computer Linguistics. Athens, Greece. p.66-73.)

Statistics South Africa. 2004. Census 2001: Primary tables South Africa. http://www.statssa.gov.za/census01.html.

Van Rooy, Bertus and Pretorius, Rigardt. 2003. A word-class tagset for Setswana. *Southern African Linguistics and Applied Language Studies*, 21(4):203-222.

Van Wyk, Egidius. 1967. The word classes of Northern Sotho. *Lingua*, 17: 230-261.

Watters, John. 2000. Syntax. (*In* Heine, Bernd and Nurse, Derek, *eds*. African Languages: An introduction. Cambridge: Cambridge University Press. p.194-230.)

Appendix: Tags in the text

Tag	**Meaning**
AbsPro	Absolute pronoun
AgrObj	Object agreement morpheme
AgrSubj	Subject agreement morpheme
AgrSubjCons	Consecutive subject agreement morpheme
AspPr	Aspectual prefix
Caus	Causative suffix
Cl	Noun class
DemPro	Demonstrative pronoun
Dim	Diminutive suffix
d1	Distance 1
d2	Distance 2
d3	Distance 3
Loc	Locative suffix
NPre	Noun prefix
Perf	Perfect
Pl	Plural
PosPart	Possessive particle
QualPart	Qualificative particle
VerbEnd	Verbal ending

THE P-DIAGRAM – A SYLLABLE-BASED APPROACH TO P-STRUCTURE

Tina Bögel
University of Konstanz

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper introduces a formal account of p-structure via a new conception, the p-diagram. It is based on the phonological unit of the syllable, which constitutes the phonological string. On top of this basic entity, prosodic layers are added 'as needed' and tied to the respective syllables via vectors. The result is a theory-independent and compact description of the speech signal itself, which allows for easy extraction of relevant information for other correspondence structures and can be adjusted to language- and theory-specific needs. An example of how this approach works is given by describing a solution to the syntactically ambiguous, but prosodically unambiguous phenomenon of constituent grouping in coordination.

1 Introduction

Current prosodic research is very much driven by the prosodic hierarchy theory, originally proposed by Selkirk (1984), which consists of hierarchically ordered prosodic units, e.g., *intonational phrase, phonological phrase, prosodic word* etc. However, these prosodic units are only descriptions of events in the speech signal, like stress, intensity, duration of elements, or breaks. The literature on the question of how these units should be defined and which information of the speech signal should be used for their calculation is huge and quite controversial. A majority of the community (e.g Selkirk, 1984; Nespor and Vogel, 1986) takes, e.g., the prosodic word to *minimally* be of the same size than the the morphosyntactic stem or even the syntactic word. While some researchers of this group only allow for prosodically deficient items (clitics/function words (e.g Selkirk, 1984)) to be included into the prosodic word, others believe that, under certain circumstances the prosodic word can contain the whole sentence (Kleinhenz, 1998). On the other hand, there are also researchers who propose that a prosodic word starts with the stressed syllable of a trochaic foot encountered in the speech signal (Lahiri and Plank 2010, see also Dalrymple and Mycock 2011). A consequence of this approach is that the prosodic word is not necessarily identical to a syntactic word; instead, the morphosyntactic stem can in principle be split by a prosodic word boundary, separating the syllables of one morphosyntactic stem into two different prosodic words.

Apart from the controversial discussion on the exact size and nature of the different prosodic units, a further problem is posed by the fact that the status of the prosodic speech units might change in relation to external factors like speech tempo. Thus, the same (written) string might be realized by several variants of prosodic grouping.

In order to avoid these problems, my approach is based on the largest 'reliable' unit, the syllable. Syllables are rhythmic units, which consist of at least a vowel or

[†]I would like to thank the audience of LFG 2012, especially Ron Kaplan and Mary Dalrymple, and Louise Mycock for comments on a preliminary version of this approach. Furthermore, I would like to thank Melanie Seiss for her math coaching.

a syllabic consonant (the nucleus), and mostly include consonants grouped around this nucleus according to certain rules. Although syllables are considered to be part of the prosodic hierarchy, the approach introduced in this paper is not built upon the general notion of prosodic units and their respective grouping according to hierarchical restrictions, but encodes the events encountered in the speech signal in a linear order. Since prosodic units are determined by different indicators within the speech signal, this syllable-based approach is in principle able to account for them as well, because the required information (on, e.g., breaks or boundary tones) is encoded in the p-structure representation. The syllable as the basic unit also allows for independence from external factors like speech tempo, since the basic unit of the syllable remains the same (in contrast to a time-based or a prosodic grouping approach, for example).[1]

Apart from the notion of the syllable as the basic unit, this paper also introduces a new way of representing prosodic information: The p-diagram. This representation allows, in principle, for all relevant elements of the speech signal to be encoded. It is thus broadly structured in the sense of being able to encode the speech signal's structure independently of the researcher's theoretical assumptions, but allows for a much finer-grained analysis of the speech signal in comparison to other approaches. The information stored in the p-diagram is not conveyed to the overall grammar in general; instead, its content is accessible via the correspondence structure of LFG.

In order to to demonstrate this idea concretely, the phenomenon of constituent grouping in coordination is analyzed within this system. Coordination grouping is interesting for the syntax-prosody interface in that syntax alone provides no basis for a decision between several possible constituent groupings, but that a consideration of prosody yields the information necessary for disambiguation.

This paper is structured as follows: First, an account of the p-diagram and the different elements involved in the speech signal representation, i.e., the lexicon and the strings and the exact shape of these representations, are given. In the second part, the implementation of grouping in coordination shows the implementation of a phenomenon, where the prosody-syntax interface is at its best.

2 The Lexicon

Following (Levelt et al., 1999, p. 4), I assume that a lexical entry consists of several parts.

1. The *concept*: The concept describes the semantic concept of a lexical item; i.e., the *idea* of an entity in the world, which we have in our minds before we add a lemma (a morphosyntactic item) and a form (a phonological representation) to this idea.

[1] There might be some syllable reduction or deletion depending on the respective speech tempo, but either a syllabic consonant remains or the syllable is completely deleted. This does not affect the overall representation.

2. The *lemma*: The lemma is the morphosyntactic representation of the concept.

3. The *form*: The form describes the phonological representation of the concept.

Following Dalrymple and Mycock (2011), I will refer to the lemma as *s(yntactic)-form* and to the form as *p(honological)-form*. The s-form represents the morphosyntactic representation of a concept; its realisation is a terminal node of the c-structure. It carries morpho-syntactic information, e.g., word-class, tense or subcategorization frames, which is subsequently processed by the grammar. The p-form, on the other hand, encodes information about syllable structure, word stress (if applicable) and an IPA transcription of its phonetic representation.[2] In (1), (ˈ) indicates the nucleus of the primarily stressed syllable; syllable structure is indicated by periods.

(1) The p-form of the lexical entry for *'übersetzen'* (German, 'to translate')

concept	**p-form**
translate	/yː.bɐ.zˈɛ.tsn̩/

S-form and p-form are two different aspects of the same concept; they represent two sides of one coin, and while they may look very different and encode very different information, they still represent the same core – their concept. This close relationship has been similarly described before, recently in LFG by Dalrymple and Mycock (2011). I would like to build on that work and show how these two-dimensional lexical entries are essential (but not omnipotent) for the disambiguation of ambiguous p-forms and s-forms and how this relationship works in both directions.

2.1 Disambiguating s-forms via word stress

As can be seen in (1), the phonological form includes information on word stress. Word stress is especially helpful for the disambiguation of elements, which have an identical morphosyntactic representation, but belong to different concepts. In English, for example, the word accent can differentiate between lexical categories (e.g., *pˈermit* (noun) vs. *permˈit* (verb)); thus, depending on the position of the word stress, a specific lexical entry is chosen and further processed by syntax as eiher a noun or a verb with its respective c- and f-structure representations. The grammar does not rely on a specific representation of phonological structure to differentiate

[2]Note that one lexical entry can include many p-forms, as they may differ according to dialect, speech register and other external reasons. I follow Lahiri and Reetz (2002, 2010), who assume the p-form of the lexical entry to be an abstract and underspecified representation of phonological features, which allows them a flexible treatment of different phonetic representations of the same concept. However, in order to simplify the p-form representation in the lexicon, I will depict the IPA 'standard' pronunciation.

between the two concepts, because this distinction is already represented by the syntactic structure, which encodes information on the respective lexical category extracted from the lexicon.

However, consider German, where the difference in word stress does not necessarily represent a difference in lexical categories. Here, a (phonological) representation of word stress in a grammar projection is essential in order to recognize the exact meaning of the sentence. Consider the German verb *übersetzen*, which can mean ‘to translate’ or ‘to cross over’. The first verb is a standard verb in German; the second one, on the other hand, is a particle verb, which can be split in certain syntactic environments. However, there are also constructions (verb-final sentences), which will prevent a disambiguation by means of syntax. At this point, the phonological information is essential: While the s-forms of the two concepts are identical, the respective p-forms show a difference ((2)). If the word stress is on the first syllable, the associated concept would be ‘to cross over’. If, on the other hand, the word stress is on the third syllable, the concept would be ‘to translate’.

(2) **S-string**: Lass uns übersetzen
Let us translate / cross over

Lexical entry:

concept	s-form	p-form
translate	übersetzen (V)	/yː.bɐ.zˈɛts.n̩/
cross over	übersetzen (Particle-V)	/ˈyː.bɐ.zɛts.n̩/

If this information on word accent is available ‘outside’ of the lexicon (e.g., in form of a speech signal representation), a disambiguation of meaning can take place. Syntactic analysis on its own, on the other hand, cannot differentiate between the two verbs.

2.2 Disambiguating p-forms via spelling and syntax

While the example in (2) shows disambiguation via information from the p-form of an ambiguous s-form, syntax is, on the other hand, often needed to disambiguate ambiguous p-forms. This can be seen in (3), where the p-form of German /ʃpɪnən/ (‘spiders’ / ‘to be crazy’ / ‘to spin’ (with a spinning wheel)) refers to several concepts.

(3)

	concept	s-form	p-form
a.	*spiders*	Spinnen (N)	/ʃpɪnən/
b.	*be crazy*	spinnen (V)	/ʃpɪnən/
c.	*spin*	spinnen (V)	/ʃpɪnən/

On the basis of the different word category, (3a) can be disambiguated from (3b) and (3c) with the help of the s-form in that the position of the word in a sentence gives an indication of its word-class and as a consequence, of its concept. The ambiguous p-form cannot be disambiguated via the lexicon, but relies on the connection of the s-form to the s-string and the c-structure to resolve the ambiguity between the verb and the noun. This is also true for the distinction between (3b) ('to be crazy') and (3c) ('to spin'), which have completely identical lexical entries and cannot be disambiguated on the basis of the lexicon only. In this case, the former is an intransitive verb and the latter a transitive verb; a distinction, which is part of the verb's internal features, but needs to be resolved in the syntax, where the respective arguments are applied. Thus, while the lexicon is essential for the alignment of p-form and s-form and for the resolution of frequently occuring ambiguities on the lexical level, other cases have to be resolved with the help of other levels of grammar.

3 The Strings

The notion of p- and s-string has been discussed before by Dalrymple and Mycock (2011). While I agree with their notion of the s-string, I would like to extend their notion of the p-string.

The s-string on the one hand represents the orthographically spelled out text with the appropriate s-form/concept boundaries, ready for further processing in c-structure. The p-string on the other hand is an *abstract representation* of the speech signal, which is, by definition, a sound wave and thus not visible to the eye. Thus, the p-string as described here is simply a phonological representation for the reader, displaying information on syllable structure, phonemic representations and word stress (in the style of the respective lexical entries).

Like p-form and s-form, p-string and s-string are two sides of the same coin: the information from prosody is minted on one side and the morphosyntactic representation is minted on the other side. However, the coin still has the same value (the concept of the string) and both sides contribute to its expression.

s-string: lass uns übersetzen
"let us cross over"

concept	**lemma**	**form**
translate	übersetzen	/ y:.bɐ.zˈɛts.n̩/
cross over	übersetzen	/ˈy:.bɐ.zɛts.n̩/

p-string: Las.ʊns.ˈy:.bɐ.zɛ.tsn̩

Figure 1: The relationship between s- and p-string

P-string and s-string are aligned with one anoher via the lexicon. That is, the lexicon serves as a look-up instrument for pieces of information. A specific speech signal would thus be tokenized into sets of syllables and aligned with possible lexical entries which are of the form described above in section 2. At this point, it is quite clear that the 'way' from the speech signal to the respective c-structure is not a "pipeline" as proposed by Bögel et al. (2009), but rather a parallel process, where c-structure helps to disambiguate and tokenize the output of the speech signal on the one hand, and p-form entries (and p-structure as we will see in section 5) help to disambiguate syntactically ambiguous concepts/constructions on the other hand.

Although the representation of the p-string provided here already conveys information about the phonological/prosodic side of a string, it must be understood that the p-string is merely a partial visualization of the utterance's sound wave. The speech signal itself carries much more information, e.g., the specific intonation of an utterance, the length of the breaks or the rhythmic grouping of the units. This information is captured in p-structure.

4 P-structure

Prosody is the aspect of grammar which is connected to the speech signal. This includes intonation, rhythmic grouping and stress, and the information connected with each of these subcategories of prosody (e.g., information on intensity and duration of single elements). Each of these aspects has received growing attention among researchers, but the most debated topic within prosodic research is probably the theory of the Prosodic Hierarchy, as proposed by Selkirk (1984, 2011), which separates the elements of the speech signal into (rhythmic) groups according to indications given by phonological processes, intonation patterns or other indications in the speech signal. These rhythmic groups are ordered hierarchically (Figure 2).

IntP	intonational phrase
$\uparrow$	
φ	phonological phrase
$\uparrow$	
ω	prosodic/phonological word
$\uparrow$	
σ	syllable

Figure 2: The Prosodic Hierarchy

As mentioned in the introduction, the exact nature, number and definition of the prosodic units is quite controversial. The problem is that the speech signal mostly consists of a continuous string. Boundaries to indicate prosodic units (be they prosodic words or phrases) might exist, but are not necessarily reliable. If, for example, the sentence in Figure 1 ("lass uns übersetzen") is spoken with a certain speed and intonation, the signal will give no indications of where to put the prosodic boundaries.

The approach pursued in this paper therefore shifts the focus of prosodic grouping to the information received from the speech signal and ties this information to the basic rhythmic unit, the syllable. However, it retains the ability of encoding prosodic grouping if need be, because the relevant information (depending on the theory: the foot, the pauses, the F_0-pattern, or a mixture of the three) is still available and can still be interpreted as grouping information.

4.1 Previous approaches within LFG

Within LFG, the prosodic aspect of grammar has not (yet) received much attention, although some attempts of encoding prosodic information within LFG have been undertaken, the first one being Butt and King (1998), who encoded prosodic structure in an attribute value matrix (AVM) projected from c-structure on the basis of prosodic units.
This was followed by O'Connor (2004), who combines the higher units (`IntP` and φ) of the Prosodic Hierarchy with a rough description of the speech signal's fundamental frequency by means of the ToBI framework.[3] In contrast to the AVM-approach of Butt and King (1998), he chooses a tree-like representation in order to avoid an intermediate projection between the AVM and the string itself on the one hand and to encode hierarchical structures between the two prosodic rhythmic units, based on the annotated high and low tones, on the other hand. While Butt and King (1998) view prosodic structure as being projected from syntactic structure, O'Connor treats syntax and prosody as independent projections. Bögel et al. (2009) follow this view of parallelism and discuss a range of mismatches between syntactic and prosodic grouping. They do not encode a separate prosodic representation per se, but include prosodic bracketing into the syntactic string.

Dalrymple and Mycock (2011) (building on Mycock 2006) develop an elaborate prosodic representation to account for comma intonation and question intonation. They view the string as being at the heart of the projection architecture and use it as the intermediate step between a prosodic tree (based on the prosodic units as described in Figure 2) and the c-structure tree. In addition, they project two further structures, which contain the bracketing information relevant for the alignment of the syntactic and the prosodic string, information on the intonational contour, etc.

All of these approaches encode a specific set of information relevant for a specific phenomenon. While some approaches have a broader potential (e.g., Dalrymple and Mycock 2011), others are too narrow for a wider range of prosodic phenomena (Bögel et al. 2009). Most use either the tree-representation or an AVM-approach to encode prosodic representations. As we will see, both representations are suboptimal when it comes to representing prosodic structure.

4.2 The representational problem

Within LFG, several ways of representing prosody have been proposed. These representations can be divided into two categories: The tree-based representations

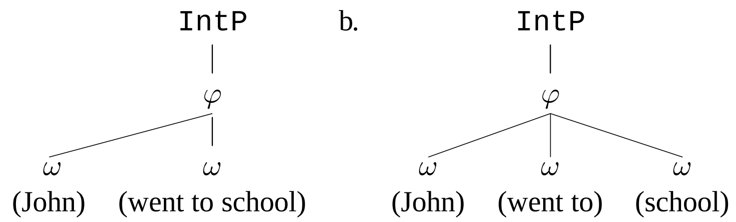
[3] ToBI represents conventions for assigning High and Low tones to the fundamental frequency of a speech signal, thus describing relevant aspects of the intonational contour (Silverman et al., 1992).

(Dalrymple and Mycock 2011, O'Connor 2004) on the one hand, and the AVM-approach proposed by Butt and King (1998) on the other hand.

The tree-based representation relies on hierarchical structures and allows only for a single aspect of the prosodic information to be encoded. O'Connor (2004) uses the tree to encode the intonation of a sentence via the TOBI annotation scheme. On the basis of this annotation, the prosodic grouping is represented in the string via bracketing. Further information given by the speech signal is not encoded within this approach. Dalrymple and Mycock (2011), on the other hand, use the hierarchical structure of the tree to represent the units of the prosodic hierarchy. Further information (e.g., on boundaries, intonation, discourse functions) has to be projected into a separate (AVM) structure, which serves as an intermediary for the remainder of the grammar.

While syntactic structure is hierarchical, the inherent nature of the speech signal is linear; thus, a hierarchically organised representation is not necessarily the right representation. There is no doubt of a certain rhythmic grouping of units, but these mostly apply to more complex constructions and are but one aspect of the speech signal. For the majority of four-word sentences, this grouping is rather irrelevant, as can be seen in (4). The respective tree can look very different depending on either the theory of phrasing (i.e., which elements are phrased together on the basis of which reasons) or the speech tempo and quality (casual, slow, ...). (4a) and (4b) show two possible encodings.

(4) a.



Furthermore, a tree structure is always bound to one possible aspect out of all possible aspects of the speech signal. Additional information on, e.g., intonation has to be stored in an extra structure projecting away from (and thus depending on) the elements of the tree-based structure. Thus, a representation of p-structure should not be built on the less fine-grained and highly variable phrasing of hierarchically organised prosodic units, but should provide a fine-grained approach to the utterance by enabling the description of various aspects and layers of the speech signal in combination with the smaller prosodic units encountered in a sentence (in the p-diagram approach, this would be the syllable).

In contrast to the tree-based representations, the AVM-structure provided by Butt and King (1998) allows for the representation of a broad spectrum of information in that the attribute value pairs can encode several aspects of a speech signal in one structure. However, typical LFG AVMs, for example f-structure, do not represent information in a linear order. If an AVM-approach is pursued, the AVM must include information on precedence relations; otherwise, an extra ordering instance between string and structure is needed. Apart from these ordering issues, the AVM has another drawback: If the speech signal and all its values are to be encoded into an AVM and its inherent attribute-value pairs, the AVM would grow to an enormous size. For two single syllables, the corresponding AVM would already be quite large (Figure 3).

$$\left[\begin{array}{ll} syll_1 & \left[\begin{array}{ll} \textsc{value} & [\text{ra}] \\ \textsc{stress} & + \\ \textsc{tone} & \text{H} \end{array}\right] \\ syll_2 & \left[\begin{array}{ll} \textsc{value} & [\text{vi}] \\ \textsc{stress} & - \\ \textsc{rhbt} & 3 \end{array}\right] \end{array}\right]$$

$\rightarrow$ VALUE: IPA representation of the syllable.

$\rightarrow$ STRESS: present + or absent $-$.

$\rightarrow$ RHBT: *right hand break time* - indicates the length of a break following this syllable in relation to other breaks in the sentence (1= short break, 3= long break).

$\rightarrow$ TONE: H indicates a high tone in the pattern of the fundamental frequency (the "melody").

Figure 3: A possible AVM-representation of */ravi/*

Thus, while the AVM is, in principle, able to encode the relevant information, the representation is not the most desirable one. With more information added (e.g., on syllable length or intensity), the AVM representation would quickly grow in size and, in parallel, become less clear and thus less interpretable.

For this reason, I have developed a new representation, which a) meets the desideratum of a fine-grained representation of the speech signal, b) allows for easy extraction of relevant information and c) provides a compact representation.

4.3 A new approach: The p-diagram

The approach presented in this paper does not view p-structure as an attribute-value matrix or a prosodic tree. Instead, the speech signal (i.e., the sound wave) is transformed into a human-readable way by describing different aspects of the signal in relation to the syllables and the breaks in between these syllables. These calculations depend on the nature of intonation patterns and stress behavior, e.g., the fact that the general level of the fundamental frequency F_0 (the "melody") will decrease towards the end of the sentence. Thus, different layers of information are extracted from the speech signal, e.g.,

- The basic rhythmic unit (the syllable)
- The stress pattern of the syllables (word accent, in combination with lexicon)
- The stress of the overall sentence (sentence accent)
- Possible pauses and their time frames
- The fundamental frequency F_0 indicating the intonation pattern

- ... and other relevant aspects as they are needed for the analysis of a specific prosodic phenomenon

All of these variables are connected with the each syllable via the following function, which decribes the relevant information as an ordered list, i.e., a vector.

(5) **S(u(n), n)**
(Where **S** is the syllable, **u** is the vector and **n** is the index of the syllable)

That is, each syllable of the p-string and each pause between two strings of syllables receives a vector, which includes the relevant information as an ordered list. The result is a set of vectors, which include ordered information of different speech signal dimensions in relation to the syllable of the p-string. These vectors can be generalized as

$$\text{S:} \begin{pmatrix} value \\ stress \\ F_0 \\ \ldots \end{pmatrix} \text{(n)} \quad \text{which would yield, e.g.,} \quad \text{S:} \begin{pmatrix} /y:/ \\ prim \\ H* \\ \ldots \end{pmatrix} \text{(1)}$$

for a specific syllable S_1, which would be the first syllable in a string with the value /y:/, and which carries primary stress and a high tone. With this vector formula, the speech signal is transformed into a bundle of syllable-dependent vectors, encoding the prosodic information as it is related to the specific syllable in the speech signal.

This set of vectors is then read into a representation, which is similar to a diagram, where the syllables and the pauses encountered in the signal are the basic entities of the prosodic representation (the 'x-axis'). On top of these basic entities, layers which include different prosodic information (e.g., stress patterns or intonation) are added (the 'y-axis').

Vector

...	...	...	...	...	...	...
Pause	-	-	-	-	3	...
Fund. Freq.	H*	L	-	-	-	...
Stress	prim	-	sec	-	-	...
Syllables	/y:/	/bɐ/	/zɛ/	/tsn̩/	-	...
Vectorindex	$\mathbf{S_1}$	$\mathbf{S_2}$	$\mathbf{S_3}$	$\mathbf{S_4}$	$\mathbf{S_5}$	... →

Figure 4: The p-diagram of */ˈy:bɐzɛtsn̩/* ('to cross over')

Via the general variable on the y-axis and the respective syllable index on the x-axis, every value of the p-diagram is accessible. For example, the stress value of the third syllable can be extracted directly from the diagram via the relevant function

[S_3, STRESS], which would return STRESS = SEC (for secondary stress). For an extraction of all syllable-values of the utterance [, SYLLABLES] would return a list of all values found at the position of the SYLLABLES of each vector, that is, all x-axis values for this y-axis variable: SYLLABLES = */y:/ /bɐ/ /zɛ/ /tsn̩/*.

A special 'syllable' is S_5 in Figure 4. It encodes a pause in the speech signal with the value 3 (long break, as opposed to the value 1, which encodes a short break (Silverman et al., 1992)). For processing reasons, it is much easier to treat the pauses as special 'syllables' instead of encoding their presence on the syllable to their left or right. Thus, apart from its pause value, a pause vector contains epsilons (-) for the other variables.

Note that the p-diagram could easily include more layers, e.g., information on the length of each syllable. It is not, per se, hierarchically constructed, but depicts the speech signal in a linear way. However, the possibility of encoding prosodic units and the implied hierarchies is given. For example, a notion like [R PHP] as used by Dalrymple and Mycock (2011) to represent the right edge of a phonological phrase can be easily encoded within a vector and retrieved from the p-structure if an analysis of prosodic phrasing is desired. Depending on the theoretical assumptions of the researcher as to which elements of the speech signal encode prosodic units, these prosodic units can be extracted from the p-diagram and its representation of speech signal elements. The p-diagram is thus a 'neutral' representation of a speech event on whose basis the individual theories of prosodic phrasing can be projected.

Figure 5 gives a general overview of the architecture proposed in this paper.

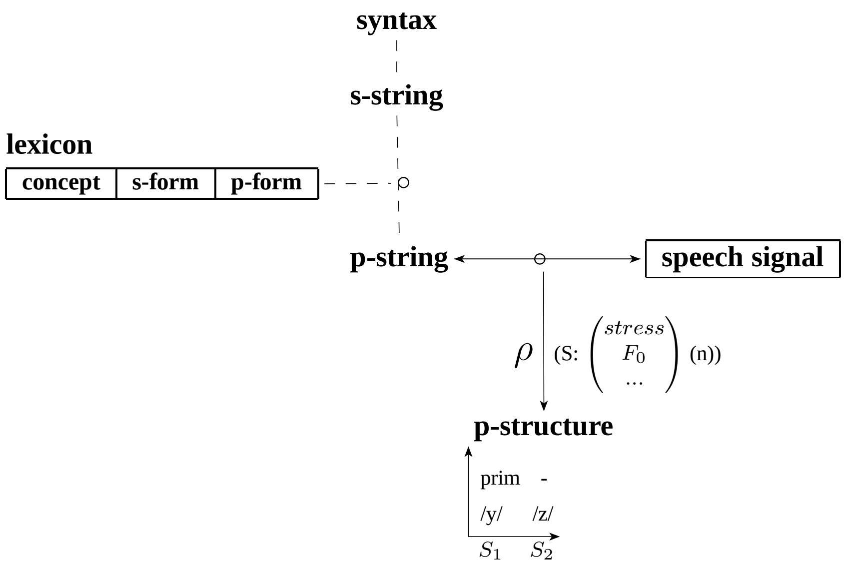


Figure 5: An architectural overview

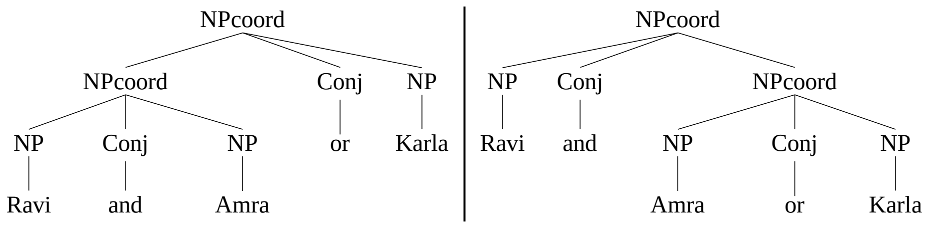
In this architecture, s- and p-string together with the lexicon are at the heart of the grammar. While p-string and speech signal seem to be different structures in this figure, they are, in fact, one unit, where the p-string merely *represents* a sound wave. A bundle of vectors carrying prosodic information in relation to each syllable is extracted from the speech signal and projected to p-structure via the correspondence relation ρ. P-structure displays the content of these vectors in a compact and easily accessible way via a p-diagram.

5 Disambiguating coordination

There are several aspects of the speech signal which are relevant to structures in the grammar. At this point I will show how the information extracted from the speech signal can help to disambiguate syntactically ambiguous sentences. Consider the syntactic phrasing possibilities of the coordination in (6):

(6) Ravi and Amra or Karla

a) b)



(Ravi $\wedge$ Amra) $\vee$ Karla   Ravi $\wedge$ (Amra $\vee$ Karla)

The two possibilities of phrasing for the coordination in (6) correspond to a fundamental difference in interpretation. The syntactic tree in (6a) groups Amra and Ravi and opposes the two to Karla, while the tree in (6b) takes Ravi and groups him with either Amra or Karla. However, the s-string does not allow for a syntactic (and semantic) distinction between the two choices. While the s-string and syntax are thus not able to disambiguate the sentence, p-structure is able to do so. The difference can, for example, be seen in an oscillogram of the above sentence (Figure 6), which represents (in short) the ‘waveform’ of the signal.

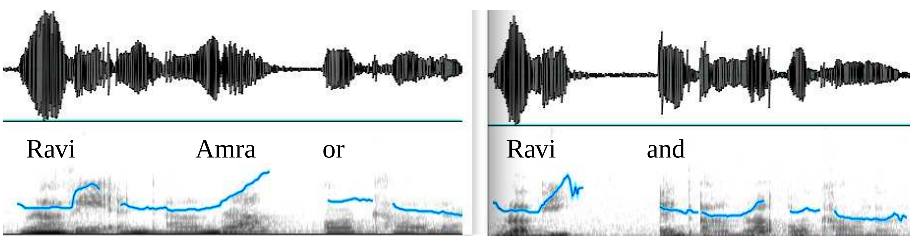


Figure 6: oscillogram for c-str. (6a) on the left and c-str. (6b) on the right

left oscillogram (c-str. (6a)): Clear break after *Amra*
right oscillogram (c-str. (6b)): Clear break after *Ravi*

These pauses in the speech signal give a clear interpretation of the intended grouping.[4] In the specific case of grouping, further indicators would be the pitch in the fundamental frequency (below the waveforms) and a possible lengthening of the last syllable of *Amra* on the left and *Ravi* on the right. For c-structure (5b) and the associated speech signal on the right of Figure 6, the following vectors represent the relevant part of the signal (*Ravi* + *pause*):

$$\text{S: }\begin{pmatrix} pause \\ length \\ F_0 \\ stress \\ value \end{pmatrix}\text{(n)} \quad \Rightarrow \quad \text{S: }\begin{pmatrix} - \\ 20ms \\ L \\ prim \\ /ra/ \end{pmatrix}\text{(1); S: }\begin{pmatrix} - \\ 25ms \\ H* \\ - \\ /vi/ \end{pmatrix}\text{(2); S: }\begin{pmatrix} 3 \\ - \\ - \\ - \\ - \end{pmatrix}\text{(3)}$$

Figure 7: Vectors representing the speech signal (*Ravi* + *pause*)

These vectors are encoded in the p-diagram in Figure 8.

Vector				
Pause	-	-	3	
Length	20ms	25ms	-	
$\mathbf{F_0}$	L	H*	-	
Stress	prim	-	-	
Value	/ra/	/vi/	-	
Vectorindex	$\mathbf{S_1}$	$\mathbf{S_2}$	$\mathbf{S_3}$	... $\rightarrow$

Figure 8: The p-diagram of *Ravi* + *pause*

The relevant information for the syntactic disambiguation can be retrieved from various variables and vectors. The most important factor in this constellation is the break after the string *Ravi* ([S_3, PAUSE]). Further indication comes from the long second syllable ([S_2, LENGTH]) and the high tone on this syllable ([S_2, F_0]). For the specific problem of grouped coordination, I will only refer to the break information.[5]

The information on breaks and the resulting boundaries is not automatically transferred to another structure of the grammar, as it is th case in Bögel et al. (2009),

[4]Such a clear-cut break is not always available, as has been noted by e.g., Allbritton et al. (1996). However, speakers who are aware of the grouping intention produce signals similar to the one in Figure 6.

[5]The tone and length indications have to be calculated in relation to the tones and length of other syllables in the sentence. While the tone information can be interpreted by itself as well, the length information has to be encoded differently for it to be meaningful. I leave this for further research.

and to a certain extent with Dalrymple and Mycock (2011), where the prosodic constituents are matched against and aligned with the syntactic constituents. Instead, the necessary information about possible breaks is *requested* by the respective structure. For the ambiguous noun coordination, this would result in an annotation like the one in example (7).

(7) NPcoord $\rightarrow$ NP Conj NPcoord
($\uparrow_{\pi^{-1}\rho}$ S_{Nmax+1} PAUSE) =c 4

OR NPcoord Conj NP

(7) shows a (simplified) NP-coordination rule, which allows for a choice between the two groupings introduced in (6). The first rule represents [Ravi $\wedge$ (Amra $\vee$ Karla)] (tree 6b) and the second one [(Ravi $\wedge$ Amra) $\vee$ Karla] (tree 6a). Encoded under the first NP node is a restriction, which a) shows the path (the correspondence relation), b) indicates the relevant syllable vector and c) constrains the value of a specific attribute. The path describes the relation between two structures and thus refers to the general idea of correspondence, which allows for the parallel description of different aspects of linguistic information and the resulting dependencies (e.g., Halvorsen and Kaplan, 1995; Kaplan, 1995; Asudeh, 2006). The correspondence relation from this c-structure node to p-structure is described by the composition of the inverse correspondence relation from c-structure to the string (π^{-1}) and the correspondence relation between string and p-structure (ρ).

As discussed in sections 2 and 3, I view the p- and s-string as two parts of the same entity. The two representations are aligned with the help of the lexicon, which encodes both the p-form and the s-form of all lexical entries. This means essentially that any s-string element 'knows' the corresponding fragment of syllables of the p-string. Thus, the s-string element *Ravi* and the p-string fragment /ra.vi/ are interlinked. The projection from p-string to p-structure is then managed via the relation ρ; the vectors and the related p-diagram have been discussed in Figures 7 and 8. The functional correspondence relation between c- and p-structure can thus be described by the following formula:

$$\rho(\pi^{-1}(f))$$

However, in this specific case, it is not the information of one of the corresponding syllables that is of interest to the syntactic rule, but the pause, which follows the last syllable. This fact is captured by the annotation S_{Nmax+1}, which refers to the syllable with the maximum index (the last one in any slice) and adds to this index 1 ($S_2 + 1 = S_3$). If this syllable vector has a pause value of 3, then the first c-structure rule in (7) is parsed. If it is another syllable as would be the case with the reverse grouping in tree (6a), this syntactic rule would not apply and the second one would be parsed.[6]

[6]The second possibility could also carry constraints related to the corresponding speech signal. These have been left out for reasons of simplicity. Furthermore, it would probably be more appropriate to implement an OT-constraint instead of an all-or-nothing condition. I will leave this for further research.

Figure 9 shows the architecture in relation to the (partial) parsing of the string *Ravi and Amra or Karla*, where the associated tree is parsed according to the information on the pause (S_3) in the p-structure.

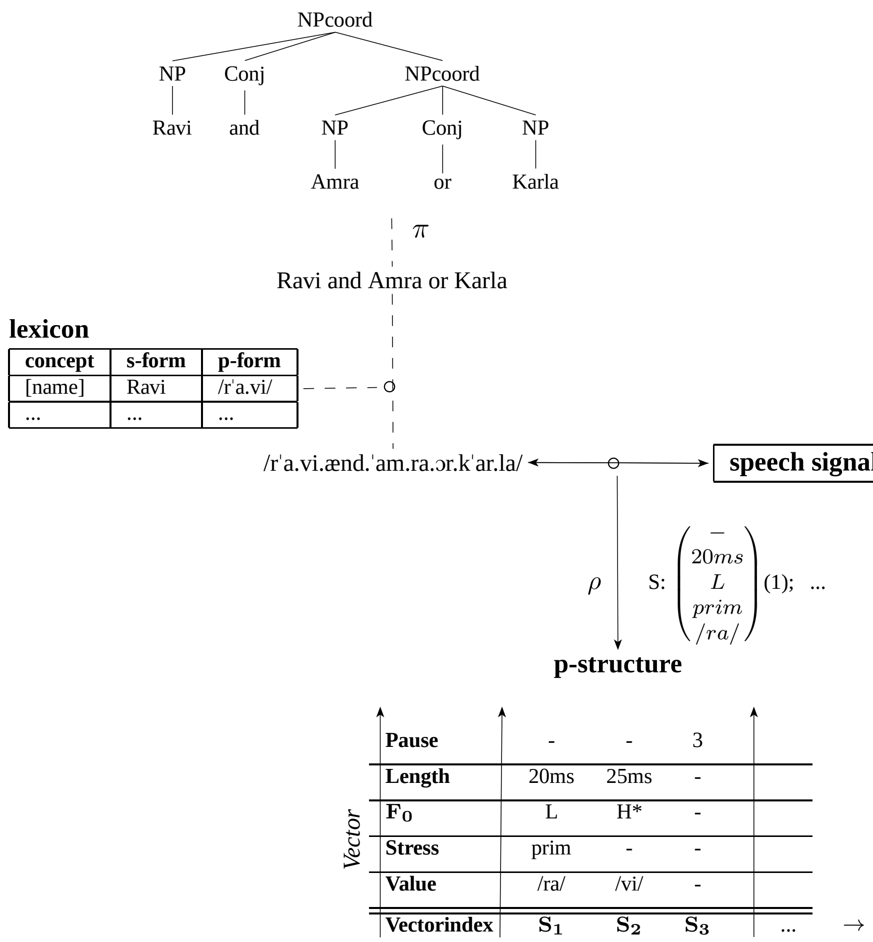


Figure 9: Partial processing of *[Ravi ∧ (Amra ∨ Karla)]*

6 Conclusion

This paper presented a new approach to p-structure in several ways: First, the p-diagram presented in this paper is not based on the units of the Prosodic Hierarchy, as it is the case in Bögel et al. (2009) and Dalrymple and Mycock (2011). The reason for this is that the exact hierarchical structure of the different units, the recognition of these units within spoken language and the phonological theories behind these units are very controversial. Instead, the p-diagram approach bases the representation of the speech signal on the notion of the syllable, because syllables (or at least the nucleus of a syllable) is considered to be a very stable and easily recognized unit in the speech signal and is, in contrast to, e.g., phonological phrases, not as sensitive to external factors like speech tempo. Furthermore, the syllable is ‘small’ enough to allow for a fine-grained analysis of the speech signal, i.e., the syllable allows for a closer look at the phonological events within its range.

Second, the notions of the prosodic tree and the prosodic AVM were replaced by a more compact, linearly structured and easily accessible representation, the *p-diagram*. The p-diagram is composed of syllable-based vectors, which contain different aspects of prosodic information connected to their respective syllable in the speech signal itself. The vectors and thus the resulting p-diagram can be constructed according to the specific needs of the prosodic phenomena to be analysed.

The p-diagram approach is theory-independent, in that it does not assume theories of prosodic grouping according to hierarchical approaches (except for the syllable, which is, however, uncontroversial). Furthermore, it does not make predictions about the alignment of prosodic and syntactic units. It is able to encode all of the above aspects on the basis of the speech signal information encoded in the p-diagram.

It is not the basic intention of this approach to align syntactic and prosodic structures or to project prosodic events per se into other structures; instead, the information is depicted in the p-structure component itself and can be retrieved from the relevant projection in the grammar (e.g., the NP coordination rule in c-structure) via the composition of correspondence relations. Thus, the focus shifts from prosody back to other structures, where specific information from the speech signal is of relevance and can be checked in an easy and compact way via the correspondence relation of any structure to p-structure.

References

Allbritton, David W., McKoon, Gail and Ratcliff, Roger. 1996. Reliability of Prosodic Cues for Resolving Syntactic Ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22(3), 714–735.

Asudeh, Ash. 2006. Direct Compositionality and the Architecture of LFG. In Miriam Butt, Mary Dalrymple and Tracy Holloway King (eds.), *Intelligent Linguistic Architectures: Variations on themes by Ronald M. Kaplan*, CSLI Publications.

Bögel, Tina, Butt, Miriam, Kaplan, Ronald M., King, Tracy Holloway and Maxwell III., John T. 2009. Prosodic Phonology in LFG: A New Proposal. In *Proceedings of LFG09*, CSLI Publications.

Butt, Miriam and King, Tracy Holloway. 1998. Interfacing Phonology with LFG. In *Proceedings of LFG98*, CSLI Publications.

Dalrymple, Mary and Mycock, Louise. 2011. The Prosody-Semantics Interface. In *Proceedings of LFG2011*, CSLI Publications.

Halvorsen, Per-Kristian and Kaplan, Ronald M. 1995. Projections and Semantic Description in Lexical-Functional Grammar. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III and Annie Zaenen (eds.), *Formal Issues in Lexical-Functional Grammar*, pages 279–292, Stanford: CSLI Publications.

Kaplan, Ronald M. 1995. The Formal Architecture of Lexical-Functional Grammar. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III and Annie Zaenen (eds.), *Formal Issues in Lexical-Functional Grammar*, pages 7–27, Stanford: CSLI Publications.

Kleinhenz, Ursula. 1998. *On Words and Phrases in Phonology*. Ph. D.thesis, University of Tübingen.

Lahiri, Aditi and Plank, Frans. 2010. Phonological Phrasing in Germanic: The Judgement of History, Confirmed through Experiment. *Transactions of the Philological Society* 108(3), 370–398.

Lahiri, Aditi and Reetz, Henning. 2002. Underspecified Recognition. In Carlos Gussenhoven, Natasha Werner and Toni Rietveld (eds.), *Labphon 7*, Berlin: Mouton.

Lahiri, Aditi and Reetz, Henning. 2010. Distinctive Features: Phonological Underspecification in Representation and Processing. *Journal of Phonetics* 38, 44–59.

Levelt, Willem J.M., Roelofs, Ardi and Meyer, Antje S. 1999. A Theory of Lexical Access in Speech Production. *Behavioral and Brain Sciences* 22, 1–75.

Mycock, Louise. 2006. *The Typology of Constituent Questions: A Lexical-Functional Grammar Analysis of 'WH'-Questions*. Ph. D.thesis, University of Manchester.

Nespor, Marina and Vogel, Irene. 1986. *Prosodic Phonology*. Dordrecht: Foris.

O'Connor, Rob. 2004. *Information Structure in Lexical-Functional Grammar: The Discourse-Prosody Correspondence in English and Serbo-Croation*. Ph. D.thesis, University of Manchester.

Selkirk, Elisabeth. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. The MIT Press.

Selkirk, Elisabeth O. 2011. The Syntax-Phonology Interface. In John Goldsmith, Jason Riggle and Alan C. L. Yu (eds.), *The Handbook of Phonological Theory*, pages 435–484, Malden, MA: Blackwell.

Silverman, Kim, Beckman, Mary, Pitrelli, John, Ostendorf, Mari, Wightman, Colin, Price, Patti, Pierrehumbert, Janet and Hirschberg, Julia. 1992. TOBI: A Standard for Labeling English Prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff.

ON THE ANALYSIS OF NON-SELECTED DATIVES IN MALTESE

Maris Camilleri and Louisa Sadler
University of Essex University of Essex

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper provides a descriptive overview of extra argumental or non-selected datives in Maltese, poorly described in existing grammars. We outline an LFG approach to the facts we describe bulding on existing LFG work and in particular on Kibort (2008)'s approach to dative arguments, extending her approach to the various subclasses of non-selected dative arguments.

1 Introduction

In this paper we aim to provide the first account of non-selected datives (henceforth NSDs) in Maltese, a Maghrebi/Siculo-Arabic dialect. In presenting Maltese NSDs we add to the growing literature on NSDs in the Semitic languages. A reasonable body of well-described data is available (Al-Zahre, 2003) for Syrian Arabic and we draw some brief comparisons to this data. The Maltese NSDs are described in terms of the typology of NSDs presented in Bosse et al. (2012), which appears to suffice for the Maltese data to be presented.[1]

Before proceeding to a discussion of the distinct types of NSDs in Maltese, we provide some discussion of dative-marked arguments in the language. Section 3 introduces Bosse et al. (2012)'s typology of NSDs (using their German data) and section 4 applies this typology to Maltese. Section 5 provides an LFG analysis for NSDs in Maltese, building on Kibort (2008) and Sadler and Camilleri (2012).

2 Selected Dative Arguments in Maltese

Pronominal accusative (object) and dative arguments are normally expressed affixally, that is, as incorporated pronouns, in Maltese: the relevant paradigms are shown in (1). As is evident, the two sets of forms basically differ in terms of the presence of *-l-* in the dative set, an element which is quite transparently related to the dative marker found with NP arguments, to be illustrated below.

(1)

PNG	OBJ	DATIVE OBJ
1sg	-ni	-lni
2sg	-(V)k	-lVk
3sgm	-u∼h	-lu
3sgf	-ha	-lha
1pl	-na	-lna
2pl	-kom	-lkom
3pl	-hom	-lhom

The accusative forms (i.e. those without *-l-* correspond to the OBJ function: for the moment we will refer to the GF associated with the dative forms as the DAT OBJ

[†]We thank Doug Arnold, Ash Asudeh, Anna Kibort, György Rákosi, participants at LFG 2012 and the editors Miriam Butt and Tracy Holloway King for comments and feedback.

[1]We note however, that this classification omits one less well-described type of NSD, the so-called *subject correferential datives*, which is found in both Syrian Arabic and Hebrew but not in Maltese.

(for further discussion see Sadler and Camilleri (2012)). Dative pronominal affixes and dative NPs occur as the goal or recipient argument in a canonical ditransitive construction as in (2) and (3).

(2) *Bgħat-t-i-l-ha* *l-ittra*
sent.PV-1SG-EP.VWL-DAT-3SGF DEF-letter
I sent the letter to her.

(3) *Bgħat-t* *il-ktieb* *lil* *Marija*
sent.PV-1SG DEF-book.SGM DAT Mary
I sent the book to Mary.

Argumental datives are not restricted to ditransitive predicates: *ċempel* ‘phone’ is a bi-valent verb which takes a dative as its second argument.

(4) *T-i-nsie-x* *iċ-ċempil-l-i*
2-FRM.VWL-forget.IMPER-NEG 2-phone.IMPV-DAT-1SG
Don’t forget to phone me.

While the bound forms that realize the OBJ and DAT OBJ functions are distinct, a slight complication is that the free pronominal non-subject forms are syncretic and derive diachronically from a pronominal inflection attached to *lil*, out of which the contracted form *’l* and the *-l-* marking on the bound dative forms are also derived (Camilleri, 2011). Free pronominal forms are used in a number of specific contexts such as in coordinated constructions as well as contrastively-stressed contexts — see (6).

(5)

PNG	Free pronoun	PNG	Free pronoun
1sg	lili	1pl	lilna
2sg	lilek	2pl	lilkom
3sgm	lilu	3pl	lilhom
3sgf	lilha		

(6) *Raj-t* *lilu, u* *mhux lilek*
saw.PV-1SG him CONJ NEG 2SG
I saw him and not you.

A complicating factor is that the free marker *lil* is also implicated in a form of differential object marking (on accusative objects), operating in accordance with the accessibility hierarchy. With human objects, proper names are obligatorily (and other definites are usually) marked with the ACC *lil* marker, but indefinite human NPs are optionally marked. Non-human NPs are usually not *lil*-marked. Note further that the presence of a dative-marked indirect object inhibits the appearance of *lil* on the direct object, even if human definite, as shown in (7).

(7) a. *Raj-t* *(l)it-tifel*
saw.PV-1SG ACC.DEF-boy
I saw the boy.

b. *Taj-t* **(l)it-tifel* *lil* *omm-u*
gave.PV-1SG ACC.DEF-boy DAT mother-3SGM.ACC
I gave the boy to his mother.

Although dative-marked NPs/pronouns typically realize the goal/recipient argument of a ditransitive verb, in what we have elsewhere called the canonical dative construction (following Kibort (2008)), Maltese also has a (rather restricted) double object or dative-shift construction, found with certain ditransitive verbs, where the goal/recipient is obligatorily expressed as a bound OBJ pronoun. Compare (8), a canonical dative construction, with the double object construction in (9). (10) is a further example of the DOC.

(8) *Wera* *t-triq* *lil* *Pawlu*
showed.PV.3SGM DEF-road DAT Paul
He showed the road to Paul. CDAT

(9) *Wrie-h* *it-triq*
show.PV.3SGM-3SGM.ACC DEF-road
He showed him the road. DOC

(10) *Ma* *n-af-x* *min* *għallm-u* *l-Malti*
NEG 1-know.IMPV.SG-NEG who taught.PV.3SGM-3SGM.ACC DEF-Maltese
I don't know who taught him Maltese. DOC

In other work on the Maltese ditransitive predicates, Sadler and Camilleri (2012) provide a number of arguments showing that the recipient/goal argument corresponds to an OBJ function in the DOC illustrated in (9) and (10), and further that the canonical dative construction (examples (2), (3) and (8)) involves a secondary or restricted OBJ rather than a prepositional OBL.

This section has briefly introduced the use of the dative-marked argument in *selected* contexts, typically where it functions as the third argument of the predicate. We now consider the NSD use of dative pronominal affixes (optionally doubled by a dative-marked NP) in a range of other constructions, but before doing so, provide a brief introduction to the classification of non-selected dative constructions, drawing principally on that proposed by Bosse et al. (2012) (henceforth BBY).

3 Types of Non-selected Datives

On the basis of data from a (relatively modest) spread of languages, BBY identify essentially four distinct types of NSDs; external possessor datives (EP), benefactive datives (BEN), affected experiencer datives (AE) and attitude holder datives (AH). All of the following German examples are due to BBY.[2]

(11) illustrates an external possessor dative, in which a relation of possession exists between the NSD and (typically) the OBJ: in some languages external possession is restricted to cases of inalienable possession. As is frequently the case, an EP interpretation may occur alongside an AE interpretation, in which the dative participant is interpreted as particularly affected by the event (here, by the cleaning of the suit).

(11) *Sie säuberte mir den Anzug.*
she cleaned me.DAT the suit
She cleaned my suit. EP
She (went and) cleaned the suit on me. AE

In the benefactive (BEN) dative construction the argument is not required to be either a possessor or sentient (although it is, in this particular example).

(12) *Dennis installierte seinem Freund das Programm.*
Dennis installed his.DAT friend the program
Dennis installed the program for his friend. BEN

The affected experiencer (AE) construction is illustrated in (13): here the argument is interpreted as an experiencer and must be both sentient and aware.

(13) *Alex zerbrach Chris Bens Vase.*
Alex broke Chris.DAT Ben's vase
Alex broke Ben's vase on Chris.
Alex broke Ben's vase, and this mattered to Chris. AE

The final type, the attitude holder (AH) construction involves an argument that holds an attitude towards the proposition as a whole. The AE construction is often of very restricted distribution — for example, BBY state that it is restricted to first person attitude holders only in German and first and second person in French. Furthermore, this NSD type is widely thought of as entirely non-truth conditional, that is, making no contribution to the at-issue semantics.

(14) *Du sollst mir nicht wieder fernsehen.*
you shall me.DAT not again watch.television
You shall not watch TV again and I want this to come true. AH

For completeness, we can add to this list a further type of NSD, in which the dative pronoun is co-referential with the SUBJ, the so-called **coreferential dative** construction, illustrated in (15) (Al-Zahre and Boneh, 2010). Such examples typically express the **speaker**'s own attitude towards the eventuality. We do not discuss this type further in this paper (they are not found in Maltese).

[2] Bosse et al. (2012) eschew use of the term *ethical dative*, which has been the locus of some terminological confusion, sometimes used in the literature to refer to their (AH) (Rákosi, 2008; Gutzmann, 2007, 2011), and sometimes their (AE) type. Borer and Grodzinsky (1986) use *ethical dative* to cover Hebrew POSS and AE and Al-Zahre and Boneh (2010) to refer to AE in Hebrew and SA. They use "interested hearer datives" to refer to the AH in these languages.

(15) *Salma raʔṣet-l-a* *šway*
Salma dance.PAST.3SGF-to-3SGF a little
Salma (just) danced a little (it's a minor issue). [Syrian Arabic] SA

A central insight of BBY is to establish that these four (EP, BEN, AE, AH) subtypes of NSD construction have distinct properties: these are summarized in (16). One important dimension concerns whether or not the added dative argument makes a contribution to the truth-conditional semantics (TC vs. NTC in (16)). The distribution of these NSD construction types in the languages of their sample is shown in (17).

(16)

	EP	BEN	AH	AE
Semantics	possession	benefit	attitude to Prop	psychological
Poss reqd	yes	no	no	no
Nec. sentient	no	no	yes	yes
TC	yes	yes	no	y (Ger)/n (Heb)
NTC	no	no	yes	yes

(17)

	EP	BEN	AH	AE
Albanian	-	-	-	yes
German	yes	yes	yes	yes
French	-	-	yes	yes
Japanese	yes	yes	-	yes
Korean	yes	yes	-	-
Hebrew	yes	-	-	yes
Micmac	-	yes	-	-

4 Maltese Non-Selected Datives

In this section we consider how Maltese fits within this typology of NSDs: showing how the tests and diagnostics which they provide behave in this language. Unlike argument datives, which may be pronominal (affixal) or lexical — Maltese NSD are only pronominal in form (although the dative affix may be doubled by an external lexical NP associated with a discourse function). With one small exception involving certain cases of inalienable possession in a construct state construction, NSDs are optional.

4.1 Possessor Datives

Maltese involves two distinct means with which to realize possession. The first is a construct state or *iḍāfa* construction. The second involves the possessive prepositional marker *ta'* 'of', as in (18) and (19). (20) illustrates an external possessor dative (EP), while (21) shows that when a prepositional possessive is present, a NSD may not be interpreted as an external possessor EP: this example is ungrammatical *on the* EP *reading* of the NSD.

(18) *Pawlu farrak il-karozza ta' Marija*
Paul broke.up.3.PV.SGM DEF-car of Mary
Paul broke/ruined Mary's car.

(19) *Pawlu farrak l-karozza tagħ-ha*
Paul broke.up.3.PV.SGM DEF-car of-3SGF.ACC
Paul ruined her car.

(20) *Pawlu farrk-i-l-ha l-karozza*
Paul broke.up.3.PV.SGM-EP.VWL-DAT-3SGF DEF-car
Paul ruined her car. EP

(21) **Pawlu farrk-i-l-ha$_i$ l-karozza ta' Marija$_i$*
Paul broke.up.3.PV.SGM-EP.VWL-DAT-3SGF DEF-car of Mary
Paul broke/ruined Mary's car.

(22) shows that the NSD can be optionally doubled by an external topic NP or a strong pronoun (*lil*-marking is obligatory on the human, proper name).

(22) *Pawlu farrk-i-l-ha l-karozza (lil Marija)*
Paul broke.up.3.PV.SGM-EP.VWL-DAT-3SGF DEF-car DAT Mary
Paul ruined Mary's car /Mary, I ruined her car. EP

As noted above, except for such cases of discourse topics, the possessor can be only expressed once in such possessive constructions; either internally as in (18) and (19) or externally through the presence of a NSD in (20). Examples such as (23) and (24) are entirely parallel in interpretation, and both receive affected experiencer (rather than possessor) interpretations.

(23) *Pawlu farrk-i-l-ha$_i$ l-karozza tagħ-ha$_i$*
Paul broke.up.3.PV.SGM-EP.VWL-DAT-3SGF DEF-car of-3SGF.ACC
Paul (went and) ruined her car on her. AE

(24) *Pawlu farrk-i-l-na$_j$ l-karozza tagħ-ha$_i$*
Paul broke.up.3.PV.SGM-EP.VWL-DAT-1PL DEF-car of-3SGF.ACC
Paul went and) ruined her car on us. AE

Although the examples so far have involved an external possessor associated with an OBJ function, it is possible also with other functions. In (25), *t-tifel* 'DEF-boy' is the SUBJ and *-lha* refers to the 'possessor' of *t-tifel*. *Marija* is optional, bears a discourse function and is intonationally offset. (26) and (27) illustrate possessor 'raising' from other grammatical functions.

(25) *Marija n-sterq-i-l-ha t-tifel*
Mary PASS-stole.PV.3SGM-EP.VWL-DAT-3SGF DEF-boy
Mary's boy was stolen. SUBJ

(26) *Għamil-t-l-u servis lill-karozza*
did.PV-1SG-DAT-3SGM service DAT.DEF-car
I serviced his car. DAT OBJ

(27) *Hdim-t-l-u fuq il-karozza illum*
worked.PV-1SG-DAT-3SGM on DEF-car today
I worked on his car today. OBL OBJ

With the external possessor datives, a question such as (28) asks about an event concerning *his (Mario's) car* (so the possession relation is within the questioned event) and it is also possible for the possessor to be the target of a wh-question: these observations support the view that the NSD contributes to the at-issue or truth-conditional semantics, as argued by BBY. Note that (29) corresponds to a question on an EP NSD - the form of question corresponding to a prepositional possessive is shown in (30).

(28) *Fark-u-l-u ('l-)karozza (lil Mario)?*
broke.up.3.PV-PL-DAT-3SGM DEF-car DAT Mario?
Did they ruin Mario's car?

(29) *'L min fark-u-l-u ('l-)karozza?*
DAT who broke.up.3.PV-PL-DAT-3SGM DEF-car?
To whom did they ruin the car = Whose car did they ruin?

(30) *Il-karozza ta' min fark-u?*
DEF-car of who broke.up.3.PV-PL
Lit: The car of whom did they ruin? = Whose car did they ruin?

Consistent with the fact that a NSD interpreted as an EP makes a truth-conditional contribution we see that the presence/absence of the NSD is associated with an interpretational distinction in (31): (31a) involves reference to a car owned by someone in the discourse context while (31b) refers to any car.

(31) a. *Jekk j-fark-u-l-u ('l-)karozza ...*
If 3-bring.to.pieces.IMPV-PL-DAT-3SGM DEF-car ...
If they ruin HIS car ... EP

b. *Jekk j-fark-u ('l-)karozza ...*
If 3-bring.to.pieces.IMPV-PL DEF-car ...
If they ruin the car - ANYONE'S car

4.2 Benefactive Datives

The NSD in Maltese can also have a benefactive interpretation. Under this interpretation the NSD in (33) can express the same meaning (abstracting away from the lexical content of the beneficiary itself) as (32), which involves an OBL with the preposition *għal*.

(32) *Ġab il-ktieb għal Marija biex*
got.PV.3SGM DEF-book for Marija in.order.to
t-a-qra-h
3-FRM.VWL-read.IMPV.SGF-3SGM.ACC
He got the book for Mary to read.

(33) *Ġab-i-l-ha l-ktieb biex*
got.PV.3SGM-EP.VWL-DAT-3SGF DEF-book in.order.to
t-a-qra-h
3-FRM.VWL-read.IMPV.SGF-3SGM.ACC
He got the book for her to read.

A BEN NSD can co-occur with an OBL with the preposition *għal* ‘for’, provided that they do not co-refer. In (34) the intended sense is consistent with a scenario in which the dative ‘she’ had been intending to give the book to Mary to read, but had not been able to because she did not have it herself.

(34) *Ġab-i-l-ha*$_i$ *l-ktieb għal Marija*$_j$ *biex*
got.PV.3SGM-EP-DAT-3SGF DEF-book for Marija in.order.to
t-a-qra-h$_j$
3-FRM.VWL-read.IMPV.SGF-3SGM.ACC
He got her (i.e. for her benefit) the book for Mary, in order for her (Mary) to read it.

An important fact about BEN datives, according to BBY, is that these do not require the BEN argument to be sentient, or alive (unlike the AE datives) and this holds true of NSDs with benefactive interpretations in Maltese:

(35) *Bdej-t t-i-xgħel-l-u xemgħa wara*
started.PV-2SG 2-FRM.VWL-light.up.IMPV-DAT-3SGM candle after
li miet.
COMP died.PV.3SGM
You started lighting a candle for him after he died. BEN

Note that evidence that a NSD can correspond to an argument which may be distinguished from an (external) possessor is provided by the fact that a BEN dative may coexist with an internal possessor (which would itself give rise to an EP in the possessor NSD construction). Example (36) involves both a NSD with a benefactive interpretation and a (distinct) possessor, indicating that a BEN NSD is distinct from an EP one.

(36) *Had-t-l-u t-tfal ta’ Marija l-iskola*
took.PV-1SG-DAT-3SGM DEF-children of Mary DEF-school
I took Mary’s children to school for him (i.e. for his benefit). BEN

In the case of the BEN argument, the event *involving* (interpreted as including) the NSD can be negated (37) and questioned (38), providing evidence that the contribution made by the NSD is part of the truth-conditional or at-issue semantics. Note however that the BEN role cannot be directly negated when it is expressed as an NSD (see (39)) but only when it is expressed as a PP OBL as in (40); a restriction which perhaps follows from the affixal nature of the NSD.

(37) *Ma seraq-hom-l-i-x*
NEG stole.3SGM-3PL.ACC-DAT-1SG-NEG
He didn't steal them for me. BEN

(38) *'L min bdej-t t-i-xgħel-l-u xemgħa*
DAT who started.PV-2SG 2-FRM.VWL-light.up.IMPV-DAT-3SGM candle
wara li miet?
after COMP died.3SGM
Who did you start lighting a candle for after he died? BEN

(39) **Seraq l-affar-ijiet imma ma*
stole.PV.3SGM DEF-thing-PL but NEG
seraq-hom-l-i-x
stole.PV.3SGM-3PL.ACC-DAT-1SG-NEG
He stole the things, but he didn't steal them for me.

(40) *Seraq-ha l-karozza. Biss ma*
stole.PV.3SGM-3SGF.ACC DEF-car. but NEG
seraq-hie-x għal-i-ja
stole.PV.3SGM-3SGF.ACC-NEG for-EP.VWL-1SG.ACC
He stole the car, but not for me. BEN

4.3 Affected Experiencer Datives

A NSD may also be interpreted as an affected experiencer (AE), in which case the referent must be sentient and aware.

(41) *Is-subien ta' Rita żżewġ-u-***l-hom** *kollha (***lil bniet ta'**
DEF-boys of Rita married.PV.3-PL-DAT-3PL all DAT girls of
Carmen*), u issa ma fadal ħadd mir-raħal*
Carmen CONJ now NEG left.3.PV.SGM no.one from.DEF-village
għal-i-hom
for-EP.VWL-3PL.ACC
All of Rita's boys (went and got) married on-them (Carmen's daughters) all, and now there is no one in the village left for them (Carmen's daughters).

(42) *Wasal-l-i tard mill-iskola t-tifel*
arrived.3SGM-DAT-1SG late from.DEF-school DEF-boy
The boy arrived late from school, affecting me by doing so. AE

Because a AE interpretation is only available for alive and sentient participants, *Pawlu* cannot antecede the NSD in (43):

(43) *Meta miet Pawlu$_i$, ftit wara miet-it-l-u$_j$*
when died.PV.3SGM Paul a.little after died-PV.3SGF-DAT-3SGM
omm-u$_j$
mother-ACC.3SGM
When Paul died, his ($\neq$ Paul) mother died soon after.

Just as in the case of the BEN dative, we see that an AE NSD can co-occur with a separate possessor, and hence that AEs are not simply possessors.

(44) *Hbej-t-i-l-ha l-kotba ta' ħi-ja ħalli*
hid.PV-1SG-EP.VWL-DAT-3SGF DEF-book.PL of brother-1SG.ACC so.that
ma t-a-qra-hom-x
NEG 3-EP.VWL-read.IMPV.SGF-3PL.ACC-NEG
I went and hid my brother's books (i.e. adversely affecting her), so that she does not read them. AE

BBY argue that AEs are the locus of parametric variation in a number of respects. In particular, they suggest that AE are wholy non-truth conditional in some languages (contributing conventionally implicated (*ci*) content only), but may also contribute to the truth-conditional (*at issue*) semantics in other languages. In fact a major concern of their paper is to establish that AE NSDs *may* contribute to both *ci* and *at issue* domains and to propose a treatment of such hybrid elements. Detailed discussion of their assumptions, and in particular of their claim that the observed behaviour of German AEs is evidence for a putative dual contribution to both domains is beyond the scope of this paper. Nonetheless it is interesting to note their claim that AE datives are entirely non-truth conditional in Hebrew. The evidence suggests that this is not so in Maltese: (45), which shows that the NSD with an AE interpretation may be within the scope of negation, is just as good as (37). We think, therefore, that in Maltese at least, AE NSD contribute to the *at-issue* semantics. Further evidence comes from the fact that an event involving the AE can be questioned (46), and the affected experiencer can be wh-questioned, as in (47).

(45) *Għad-hom ma żżewġ-u-l-hom-x kollha (lil bniet ta'*
still-3PL.ACC NEG married.PV.3-PL-DAT-3PL-NEG all DAT girls of
Carmen), is-subien ta' Rita, jiġifieri għad-hom fiċ-ċans.
Carmen DEF-boys of Rita, so.this.means still-3PL.ACC in.DEF-chance
Rita's boys have still not all married on them, which means that they (Carmen's girls) still have a chance (i.e. to get married to Rita's remaining boys).

(46) *Żżewġ-u-l-hom kollha (lil bniet ta' Carmen) is-subien ta'*
married.PV.3-PL-DAT-3PL all DAT girls of Carmen DEF-boys of
Rita?
Rita
Did all of Rita's boys get married on them - (Carmen's daughters)? AE

(47) *Min huma dawk li jekk j-iżżewġ-u-l-hom kollha is-subien ta'*
who cop.PL those COMP if 3-marry-PL-DAT-3PL all DEF-boys of
Rita, ma j-i-bqa-x raġel mir-raħal
Rita, NEG 3-EP.VWL-left-NEG man from.DEF-village
għal-i-hom?
for-FRM.VWL-3PL.ACC
Who are the ones who if all of Rita's boys marry on-them, there will be no man left for them from the village? AE

A further relevant observation concerns conditional sentences. If the AE makes a contribution to the (regular) semantics, then the inclusion of an experiencer dative in the clause should make a difference to the interpretation of the antecedent of a conditional clause. The following pair do in fact differ in meaning precisely in terms of whether the speaker is affected by all the boys marrying.

(48) *Jekk j-iżżewġ-u-***l-i** *kollha s-subien ta' Rita, Rina se*
If 3-marry.IMPV-PL-DAT-1SG all DEF-boys of Rita Rina FUT
t-a-għti lil Rita 100 ewro.
3SGF-EP.VWL-give DAT Rita 100 euros
If all of Rita's boys get married on me, then Rina will give Rita $100 euros.

(49) *Jekk jiżżewġu kollha, s-subien ta' Rita, Rina se*
If 3-marry-PL-DAT-1SG all DEF-boys of Rita Rina FUT
t-a-għti lil Rita 100 ewro.
3SGF-EP.VWL-give DAT Rita 100 euros
If all of Rita's boys get married, then Rina will give Rita $100 euros.

4.4 Attitude Holder Datives

We turn now to the fourth type, the AH or attitude holder dative, in (50). Unlike the other types of NSD, the Maltese AH dative cannot be doubled by an external topic, (51), and it cannot be questioned or negated (see (52)).

(50) *Rebħ-i-l-na, lilna, kien*
won.PV.3SGM-EP.VWL-1PL.DAT-1PL, lilna, was.PV.3SGM
He had won on us (ie. affecting us by doing so). AE

(51) *Ejja ħa t-i-rbħ-i-l-na*
come.IMP.2SG so.that 2-FRM.VWL-win.IMPV.SG-EP.VWL-DAT.1PL
**lilna/*aħna*
we.DAT/we.NOM
Come on! Win! AH

(52) **'L min sejjer t-i-rbaħ-l-u?*
ACC who going.SGM 2-FRM.VWL-wins.IMPV.SG-DAT-3SGM
Whom are you going to win on-him?

The construction occurs only with 1st and 2nd person pronouns, and only in imperative and exclamative clause types.[3] Pragmatically, such expressions may serve as a politeness strategy directed towards the addressee as in (55).

(53) *Ara!* *ħa* *t-i-tilq-u-l-i* *mid-dar*
See.IMP.2SG FUT 2-FRM.VWL-leave.IMPV-PL-DAT-1SG from.DEF-house
fl-aħħar *?!*
in.DEF-last
See/Look at this! You are finally leaving the house?! AH

(54) *Itilq-u-l-i* *minn quddiem-i*
leave.IMP-PL-DAT-1SG from in-front-1SG.ACC
Get away from in front of me! AH

(55) *ħa* *n-e-ħod-l-ok* *naqra ilma jekk*
FUT 1-EP.VWL-take.IMPV.SG-DAT-2SG a.little water if
j-o-għġb-ok
3-EP.VWL-like/please.SGM-2SG.ACC
I will take on-you some water please AH

4.5 Summary

This section has applied the classification of non-selected dative types developed in BBY to Maltese. All NSDs in Maltese are expressed as dative verbal affixes. Maltese has all four types proposed in the BBY classification, with a major distinction emerging between the AH datives on the one hand, and the three other types of NSD (EP, BEN and AE) on the other hand. The former make no contribution to the at-issue semantics, and indeed syntactically, the NSD affix is limited to 1st and 2nd person and may not be doubled by a (dative-marked) NP (topic). On the other hand, the other three types show the syntactic behaviour expected of syntactic arguments.

The picture emerging for Maltese is rather different than that BBY claim for Hebrew: that language, they assert, has just two types of NSD, the EP and the AE, with the latter being entirely non-truth conditional. However, examples provided in Al-Zahre and Boneh (2010) indicate that the AE type is probably also found in Hebrew, and as shown below, Syrian Arabic shares at least the EP, AE and AH types (and just as in Maltese, these are expressed by means of a verbal affix).[4]

(56) *Sami kasar-lo* *ən-naDDaar-aat le-ʕali*
Sami break.PV.3MS.-to.3MS the-glass-PL to-Ali
Sami broke Ali's glasses. [Syrian Arabic] EP

[3]With imperatives this is highly colloquial in use and is most likely with the imperative form *ejja* 'come', which functions like 'come on' in English.

[4]We believe it is highly likely that BEN datives also exist in SA and other Arabic dialects. On the other hand, Maltese appears to lack the coreferential dative. Further cross-dialectal work is required to determine whether Maltese is merely exceptional in this regard or whether distributional variables may be detected.

(57) *ʕali ʕam-yətfalsaf-la la-salma*
Ali PROG-philosophize.IMPV.3MS-to.3FS to-Salma
Ali is philosophizing on Salma (this aggravates her). [Syrian Arabic] AE

(58) *šəf-tə-lek šaʔfət šab!*
see.PV-1S-to.2FS piece young man
I saw one of these guys! [Syrian Arabic] AH

5 The analysis

The syntactic analysis we offer makes a fundamental distinction between the AH dative and the remaining three types of NSD. These latter are essentially distinguished from each other in the semantics, in terms of the different entailments over the added participant which they involve: from a morphosyntactic point of view, at least the AE and the BEN NSDs (and we would suggest also the EP datives) are indistinguishable. As we have seen in the discussion of data above, NSD which are interpreted as EP, BEN or AE participants are *syntactically active*, participating in syntactic constructions such as wh-question formation, and also contribute to the at-issue semantics. These properties indicate that these NSDs result from a valency-increasing operation in the morphology which introduces an additional argument. The AH dative is clearly distinct, showing a markedly different behaviour in the syntax (for example, it cannot be the focus of a wh-question, cannot be doubled by a co-referential NP topic or occur as a free pronoun) and does not contribute to the at-issue semantics: this behaviour is fully consistent with BBY's observations concerning AH datives in other languages. We will propose that while both sets of NSDs involve the same morphological realization, they do not share the same morphosemantic operation: the *syntactically active* types of NSD involve the introduction of an additional syntactic argument, but the AH type does not.[5]

A reasonable starting point would seem to be to model the analysis of the syntactically active NSDs (EP, BEN, AE) on that of selected dative arguments in ditransitive constructions such as (59), for they share the syntactic properties of these arguments (that is, they can be doubled by a dative-marked external topic, can be focused, and involve a dative-marked pronoun attached to the verb).

(59) a. *Bgħat-t-i-l-ha l-ittra*
sent.PV-1SG-EPENT.VWL-DAT-3SGF DEF-letter
I sent the letter to her. PRON. CDAT

Bgħat-t il-ktieb lil Marija
end.PV-1SG DEF-book.SGM DAT Mary
I sent the book to Mary. CDAT

[5] A theory-internal consequence which is perhaps of some passing interest is that if we are correct, then one morphological operation (affixation of a dative pronominal marker) can correspond to a multiplicity of different effects (ie is not classified as either morphosemantic or morphosyntactic), as claimed also in Kroeger (2007).

In the canonical ditransitive construction in Maltese the goal/recipient argument is expressed as a dative NP or incorporated dative pronominal. Although the *l*-marking (*lil Marija*) derives diachronically from a preposition, it does not function synchronically as such, and the dative argument does not correspond to a PP in c-structure or an OBL in f-structure. Unlike a primary OBJ, it is not accessible to promotion to SUBJ by passivization, and of course, shows distinct morphological marking when incorporated. Sadler and Camilleri (2012) argue that in the canonical ditransitive construction in Maltese the goal/recipient argument corresponds to a restricted OBJ, in particular a OBJ_{recip} as proposed (for some languages) in Kibort (2008). Kibort argues that standard LMT does not provide an adequate account of the range of syntactic realizations of ditransitive constructions. In standard LMT two surface mappings are provided by associating different intrinsic classification features with the arguments. As a result, in the prepositional construction the theme maps to OBJ and the recipient/goal to OBL while in the dative shift construction the recipient/beneficiary/goal is the OBJ (and accessible to promotion under passivization) and the theme is a restricted object OBJ_{theme}:[6]

(60)	**dative shift**	recip OBJ	theme OBJ_θ
	dative oblique	recip OBL	theme OBJ

Kibort (2008) argues persuasively that dative arguments are distinct from both (first, direct) objects and prepositional obliques, and recognises three mappings for RECIP (and similar) arguments. In her approach, which uses a layer of ordered arguments mediating between semantic roles (or rather, sets of semantic entailments) and intrinsic features (underspecifying grammatical functions), the RECIP argument may map variously to arg2 (when it will surface as OBJ in active clauses), arg3 (when it will surface as a canonical dative in languages permitting this encoding), and arg 4 (when it surfaces as a prepositional oblique). In this version of LMT, then, argument positions (i.e. the valency slots of the predicate) constitute an independent level of representation which mediates the relation between semantic participants and grammatical function assignment.[7]

(61)	<	arg1	arg2	arg3	arg4	...	argn >
		$-$o/$-$r	$-$r	$+$o	$-$o		$-$o

The association of semantic arguments with argument positions is guided by the (relative prominence of the) sets of entailments associated the different arguments, and hence a recipient argument associated with arg3 is associated with more

[6]A different alternative to the standard LMT approach to dative arguments in ditransitive predicates is proposed in Alsina (1996), in which function argument biuniqueness is abandoned and both arguments are treated as (primary) OBJ. However it seems that this approach fails to distinguish adequately between dative objects and 'shifted' goal/recipient arguments, that is, between the canonical dative construction and the shifted construction. This is clearly inadequate for Maltese, where both are found, with different properties associated with the goal/recipient argument. See Sadler and Camilleri (2012) for discussion.

[7]For arguments in favour of the tiered approach using an ordered args list in additional to the semantic roles, see, *inter alia* Ackerman and Moore (2001).

Proto-Benefactive entailments (Primus, 1999) than one associated with arg4, and a recipient argument associated with arg2 bears a significant number of Proto-Patient entailments (and hence outranks the theme argument in dative shift constructions). For clarity, such sets of entailments are abbreviated (by Kibort) in the notation x, y, b, standing for the three participants in a ditransitive event: where x stands for the participant with the most P-A entailments, y for the (Proto-Patient) theme argument and b for the recipient/beneficiary argument. It is important to bear in mind that in the different cases, distinct sets of entailments may be associated with these participants. The point of reference which remains constant in modelling argument structure is the syntactic representation of the predicate's valence and not the ordering of the semantic participants themselves (Ackerman and Moore, 2001, 44ff).

This approach to ditransitive constructions therefore accommodates three distinct mappings for such predicates, as shown in (62):

(62)	**canonical dative**	recip OBJ_{recip}	theme OBJ
	dative shift	recip OBJ	theme OBJ_{theme}
	dative oblique	recip OBL	theme OBJ

Semantic participants should be understood as sets of semantic entailments of the predicate but not as discrete thematic roles which are part of the lexical entry of verbs. In subsection 5.1 we briefly illustrate how this approach may be applied to Maltese ditransitive verbs, before extending it to non-selected datives in 5.2.

5.1 Maltese Ditransitives

In Maltese, the canonical dative mapping is the default realization for ditransitive verbs and is available for all verbs in this class (with semantic arguments x,y,b. An example such as (59) is mapped as in (63).

(63)			x	y	b	
	bagħat	<	arg1	arg2	arg3 >	
			-o	-r	+o	**canonical dative**
			SUBJ	OBj	OBJ_{recip}	

Here the theme (y) argument outranks the b argument: the latter corresponds to a dative-marked (thematically restricted) OBJ_{recip}. The OBJ_{recip} may be a lexical NP, a free pronoun (under certain syntactic conditions) or an incorporated pronominal as in (59a). As noted above, the y argument, but not the b argument is accessible to promotion to SUBJ under passivisation in this construction, which is as predicted by this mapping.

The dative oblique (or prepositional) mapping is also available for verbs with the ditransitive argument frames (x,y,b) where the b argument may be encoded by an appropriate preposition consistent with the semantic interpretation. An example can be provided for the verb *bagħa* 'send', as in (64). Here the b argument corresponds to an arg4, which maps to an OBL.

(64)

Il-kmandant	*bagħat*	*’l*	*kull*	*tifla*	*għand*	*omm-ha*
DEF-commander	sent.PV.3SGM	ACC	every	girl	at	mother-3SGM.ACC

The commander sent every girl to her mother.

(65)

		x	y	b	
bagħat	<	arg1	arg2	arg4 >	
		-o	-r	+r	**dative as oblique**
		SUBJ	OBj	OBL	

Interestingly, there is some evidence that the dative shift construction is also found in Maltese (with the consequence that all three mappings are attested and hence that a theory which accommodates only two is problematic). The dative shift mapping in Maltese is subject to two major restrictions: it is (i) available only with a subset of the ditransitive verbs, and (ii) it is limited to cases where the recipient (b) argument is (an accusative, or OBJ) pronominal (and hence expressed in the verbal morphology). This is somewhat reminiscent of an applicative, although the morphological marker *is* the recipient argument and not simply an (additional) applicative morph. For fuller discussion of this construction, the argumentation underlying this analysis, and further data illustrating the alternations more fully, see Sadler and Camilleri (2012). The dative shift mapping is shown in (67): the diacritic *+OM* on the lexeme should be read as indicating that the verbal morphology includes an object affix.

(66)

wrie-h	*it-triq*
show.PV.3SGM-3SGM.ACC	DEF-road

He showed him the road.

(67)

		x	b	y	
wera$_{+OM}$	<	arg1	arg2	arg3 >	
		-o	-r	+o	**shifted dative**
		SUBJ	OBj	OBJ$_\theta$	

We are now in a position to extend this approach to the set of non-selected dative arguments which are the focus of this paper. We propose that the analysis of what we have called syntactically active non-selected datives should be closely modelled on that of the canonical datives in the ditransitive construction with which they share many significant properties. The difference between SDs and NSDs is that the latter are not included as part of the verb’s basic valence, but are added by a general valency increasing morphosemantic operation which is widely applicable to Maltese verbs, including, for example, intransitive verbs such as *raqad* ’sleep’, as illustrated in (68).

(68)

Raqad-l-i	*l-kelb*
slept.PV.3SGM-DAT-1SG	DEF-dog

The dog slept on me = affected me by sleeping.

5.2 Maltese Non-Selected Datives

Our proposal is that Maltese NSDs with BEN, EP and AE interpretations result from a morphosemantic operation in the lexicon which (i) applies to a base predicate introducing an additional argument associated with a small range of closely related lexical entailments; (ii) introduces a pronominal argument (affix) associated with that additional argument. The output of this morphosemantic process is to increase the valency of the predicate by addition of an argument whose semantic entailments are consistent with the arg3 role. This in turn means that the added argument will be mapped (under Kibort (2008)'s mapping theory) to (one of a small number of) OBJ_θ. Although it is not selected as part of the basic valency of the verb, a NSD in one of these classes is not non-thematic, for it results from a process which extends the predicate's *a*-structure, in much the same way as an applicative construction may extend a predicate's argument structure.[8] This morphosemantic operation adding an arg3 is schematized in (69): **a** stands for a participant associated with entailments consistent with beneficiaries, affected arguments or possessors. In the case of a ditransitive predicate, as discussed in the previous section, a (dative) pronominal affix (DAT.OM) results from a morphosyntactic operation in the sense that it simply realizes an (appropriate) arg3. (70) shows the mapping which results for predicates extended by a non-selected dative (in this case, added to a transitive predicate).

$$\text{(69)}\quad +\text{DAT.OM} \quad < \begin{array}{c} \text{a} \\ \text{arg3} \\ +\text{o} \end{array} > \quad \begin{array}{c} \text{+affected/ben/poss} \\ \\ \\ \end{array}$$

$$\text{(70)}\quad \begin{array}{lccccccc} & & \text{x} & \text{y} & \text{a} & & \text{+affected/ben/poss} \\ \text{V+DAT.OM} & < & \text{arg1} & \text{arg2} & \text{arg3} & > & \\ & & \text{-o} & \text{-r} & \text{+o} & & \\ \hline & & \text{SUBJ} & \text{OBj} & \text{OBJ}_{b/p/ae} & & \end{array}$$

If this approach is along the right lines, it is clear that dative case can signal a range of closely related OBJ_θ roles (a similar point is established, looking at different construction types, in Kibort (2008)). This raises the question of whether multiple dative arguments might co-occur. Given limitations on morphological resources, the addition of two NSD is not expected in Maltese, as such non-selected arguments are necessarily morphological in this language, and the morphology makes available only one 'slot' in the verbal template for such affixes. However one might wonder whether examples might be found in which a non-pronominal CDAT

[8]As Kibort (2008) notes, in symmetrical applicative languages, two alternative mappings are found, so an applied argument in such languages map may to arg3.

$$\text{(i)}\quad < \begin{array}{ccc} \text{x} & \text{y} & \text{b} \\ \text{arg1} & \text{arg2} & \text{arg3} \\ \text{-o} & \text{-r} & \text{+o} \end{array} > \quad \textbf{ben as canonical dative}$$

Further, in languages in which a transitivising applicative can add up to two core arguments, the second applied argument position will also be pre-specified as [+o] and mapped into OBJ_θ, resulting in two secondary objects which "will be distinguished by their subscripts" (Kibort, 2008, 19).

(in a standard ditransitive) and a NSD co-oocur. Consider the following example, which seems to exemplify just this combination. Here the dative affix introduces an argument with an affected experiencer interpretation.

(71) *Bagħat-l-i* *l-ittra* *lil* *Pawlu* *bi* *żball*
sent.PV.3SGM-DAT-1SG DEF-letter DAT Paul with mistake
He sent the letter to Paul by mistake, affecting me in doing so. AE

(72)

		x	y	a	b	
V+DAT.OM	$<$	arg1	arg2	arg3	arg3	$>$
		-o	-r	+o	+o	
		SUBJ	OBj	OBJ$_{recip}$	OBJ$_{ben/poss/ae}$	

Before turning to the analysis of the (syntactically inactive) attitude holder datives, we flag an issue concerning the analysis of EP non-selected datives such as (73). In such cases the external possessor is semantically (also) an argument of one of its co-arguments, here the OBJ *l-pum* 'the handle'. Further, for reasons that we do not fully understand, if the external possessor (EP dative) is such that it would have been expressed *inside* the NP argument by means of the Maltese construct state (which is heavily restricted, mainly to cases of inalienable possession, most usually kinship terms and body parts), then it is often obligatory to double the EP by a pronominal affix on the noun it would be in construct with, as in (74).

(73) *Qsam-t-l-u* *l-pum* *('ill-bieb).*
broke-1SG-DAT-3SGM DEF-handle DAT.DEF-door
I broke the handle of the door (door handle).
The door, I broke its handle. EP

(74) *Marija weġġħ-et-l-i* *id-i* *x'ħin*
Mary hurt.CAUSE.PV-3SGF-DAT-1SG hand-1SG.ACC what.time
qars-it-ni
pinched.PV-3SGF-1SG.ACC
Mary hurt my hand when she pinched me.

We do not have anything to add at this point about cases such as (74) involving the construct state, but the question arises in connection with examples such as (73) as to whether the possessor should be represented syntactically within the f-structure corresponding to the possessum. Such an approach is often adopted in the literature for cases of possessor raising, in which (typically) a possessor 'raises' to (non-thematic) OBJ, 'displacing' the second argument to an OBL, as in *John kissed Mary on the cheek*: for example Lødrup (2009) proposes a functional control equation (↑ OBJ) = (↑ OBL OBJ POSS) in such cases. If cases of dative external possession were similar, they would involve a functional control equation added as a side-effect of the morphosemantic operation in the lexicon. There are, however, a number of differences between possessor raising and the dative external possessor construction - in particular, the possessor is a non-thematic OBJ in the former and hence a syntactic control relation is required for completeness and coherence. Further, the possessum is not restricted to an OBL OBJ function, but can correspond to a range of different GFs, and hence an f-control equation along the lines of (75). We tend to the view that there is no motivation for representing the possessor-possessum relation syntactically by means of a control equation, but leave this question open.[9]

(75) $(\uparrow \{\text{OBJ} \mid \text{SUBJ} \mid \text{OBL OBJ} \mid \text{OBJ}_{recip}\} \text{ POSS}) = (\uparrow \text{OBJ}_{poss})$

Finally, we turn to the treatment of AH non-selected datives: we have shown there is no evidence that they are syntactically active. In common with other subtypes of NSD, attitude holder arguments are expressed by means of a dative affix, but AH datives *cannot* be linked to topicalised NP arguments, unlike other types of NSD. The AH interpretation is also only available for first and second person markers (denoting speaker/hearer participants). There is no evidence that the AH dative contributes to the at-issue semantics. We suggest, therefore, an additional role for the 1person and 2person dative affix: effectively, it may simply realize *pragmatic* information. A possible analysis is that the AH non-selected dative is simply absent from the syntax and the semantics —- the morphology encodes only *ci* meaning. An analysis along these lines is effectively proposed (although in the context of different syntactic assumptions) in Gutzmann (2007) as shown in (76) (for German *mir* 'me.DAT').

(76) $\text{MIR}_{DE}: : \lambda\text{P.MIR}_{DE}(\text{P}) =_{def} \lambda \text{ P. want(Speaker)(P)}: < \text{t}^a, \text{t}^c >$

Within an LFG context, there is no reason, of course, to rule out a morphology-pragmatics correspondence which has no representation on the syntactic levels. This seems to us to be a promising direction in which to develop an analysis of morphologically expressed AH datives.

References

Ackerman, Farrell and Moore, John. 2001. *Proto-Properties and Grammatical Encoding: A Correspondence Theory of Argument Selection*. CSLI Publications.

Al-Zahre, Nisrine. 2003. *La Structure du Groupe Verbal en Arabe*. Ph. D.thesis, Université Paris VIII.

Al-Zahre, Nisrine and Boneh, Nora. 2010. Coreferential Dative Constructions in Syrian Arabic and Modern Hebrew. *Brill's Annual of Afroasiatic Languages and Linguistics* 2, 248–282.

Alsina, Alex. 1996. *The Role of Argument Structure in Grammar: Evidence from Romance*. CSLI Publications.

Borer, Hagit and Grodzinsky, Yosef. 1986. Syntactic cliticization and lexical cliticization: The case of Hebrew dative clitics. In *Syntax and Semantics*, *19*, pages 175–217, Academic Press.

[9]It is possible that the obligatory presence of the pronominal affix within the possessum in (74) might weigh in favour of an f-control/a-control approach to these EP constructions, but do not yet understand the nature of this interaction with the construct state.

Bosse, Solveig, Bruening, Benjamin and Yamada, Masahiro. 2012. Affected Experiencers. *Natural Language & Linguistic Theory* 30/4, 1185–1230.

Camilleri, Maris. 2011. On pronominal verbal enclitics in Maltese. In Sandro Caruna, Ray Fabri and Thomas Stolz (eds.), *Syntax and Semantics; Syntax and the Lexicon*, pages 131–56, Academie Verlag.

Gutzmann, Daniel. 2007. Eine Implikatur konventioneller Art: Der Dativus Ethicus. *Linguistische Berichte* 211, 277–308.

Gutzmann, Daniel. 2011. Expressive Modifiers and Mixed Expressions. In Olivier Bonami and Patricia Cabredo Hofherr (eds.), *Empirical Issues in Syntax and Semantics 8*, pages 123–41, http://www.cssp.cnrs/fr/eiss8.

Kibort, Anna. 2008. On the syntax of Ditransitive Constructions. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG2008 Conference*, CSLI Publications.

Kroeger, Paul. 2007. Morphosyntactic vs. morphosemantic functions of Indonesian -kan. In Annie Zaenen, Jane Simpson, Tracy King, Jane Grimshaw, Joan Maling and Chris Manning (eds.), *Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan*, CSLI Publications.

Lødrup, Helge. 2009. Looking Possessor Raising in the mouth: Norwegian Possessor Raising with Unergatives. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG2009 Conference*.

Primus, Beatrice. 1999. *Cases and Thematic Roles: Ergative, Accusative and Active*. Niemeyer.

Rákosi, György. 2008. Some remarks on Hungarian ethical datives. In József Andor, Béla Hollóst, Tibor Laczkó and Péeter Pelyvás (eds.), *When grammar minds language and literature*, pages 413–22, University of Debrecen.

Sadler, Louisa and Camilleri, Maris. 2012. Ditransitive Predicates and Dative Arguments in Maltese, University of Essex.

NUMBER MARKING: AN LFG OVERVIEW

Mary Dalrymple
University of Oxford

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper presents a general overview of plurality and number marking and its treatment in LFG. It was originally prepared as an introduction to the Workshop on Number and Plurals at LFG12.

Broadly speaking, number marking is morphological marking which indicates the number of verbal dependents or events involved in a situation. Number marking can appear on pronouns or nouns, indicating the number of members in the group referred to, or as agreement marking on determiners, adjectives, verbs, prepositions, and other categories. Verbal number, or pluractionality, indicates the number of events in a complex event description. Here we present a general overview of number marking patterns and their analysis in LFG. For a thorough and in-depth treatment of these issues from a crosslinguistic perspective, see Corbett (2000) and Kibort and Corbett (2008).

1 Number systems

English and many European languages make a two-way distinction in number, contrasting singular and plural:

(1) a. the boy (singular: one boy)

b. the boy**s** (plural: more than one boy)

More complex number systems are common in Austronesian languages. Subject pronouns in Boumaa Fijian (Austronesian, Oceanic; Dixon 1988) distinguish four numbers: singular, dual (two participants), paucal (a small number of participants), and plural (a larger number of participants).

(2)

	singular	dual	paucal	plural
first person inclusive	–	etaru	tou	eta
first person exclusive	au	’eirau	’eitou	’eimami
second person	o	omudrau	omudou	omunuu
third person	e	erau	eratou	era

It is possible for number distinctions to vary across the pronominal paradigm, or to vary according to the type of noun being marked. For example, personal pronouns in Biak (Austronesian, South Halmahera-West New Guinea; Mofu 2009) make a four-way distinction in the third person, but a three-way distinction in first and second person.[1]

[1]Biak has an additional animate/inanimate distinction in the plural only, violating Greenberg’s Universal 45, which states that if a language makes gender distinctions in the plural, it also makes some gender distinctions in the singular; for more discussion, see Steinhauer (1985) and Mofu (2009).

(3)

				plural	
	singular	dual	paucal	animate	inanimate
first person inclusive	–	ku	–	ko	–
first person exclusive	aya	nu	–	inko	–
second person	au	mu	–	mko	–
third person	i	su	sko	si	na

2 Number specification and number agreement

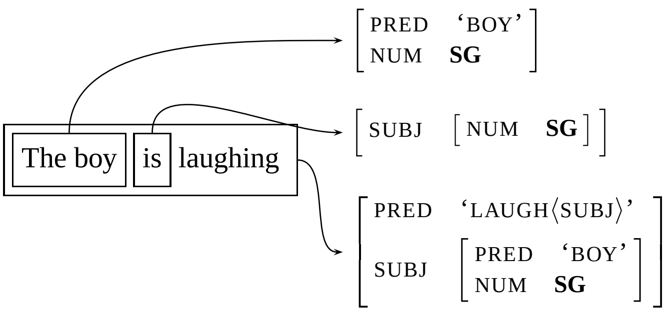
Number can be cospecified by the noun and the verb, as in the English examples in (4), where a singular subject requires a singular verb, and a plural subject requires a plural verb:

(4) a. *The boy is* laughing.
SG SG

b. *The boys are* laughing.
PL PL

In LFG treatments of verb agreement, the features of the dependent must match the features of the agreeing verb. Here the subject *the boy* is singular, and the number specified by the finite verb *is* must be compatible with the number of the subject:[2]

(5)



As usual in agreement relations, number specifications must match:

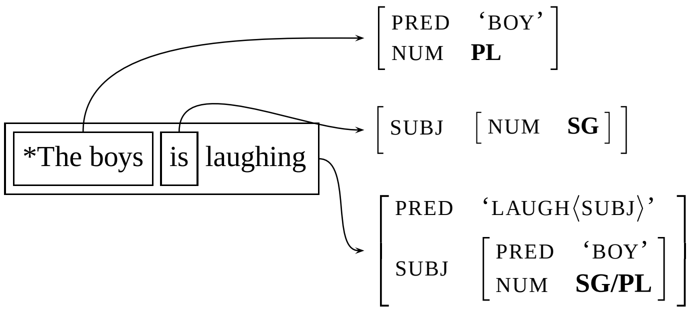
(6) a. **The boy are* laughing.
SG PL

b. **The boys is* laughing.
PL SG

The examples in (6) are ruled out because of a clash between the subject's number and the number required by the verb:

[2]To avoid clutter, we omit most f-structure features other than number, including definiteness marking, person, gender, tense, aspect, and other grammatical features.

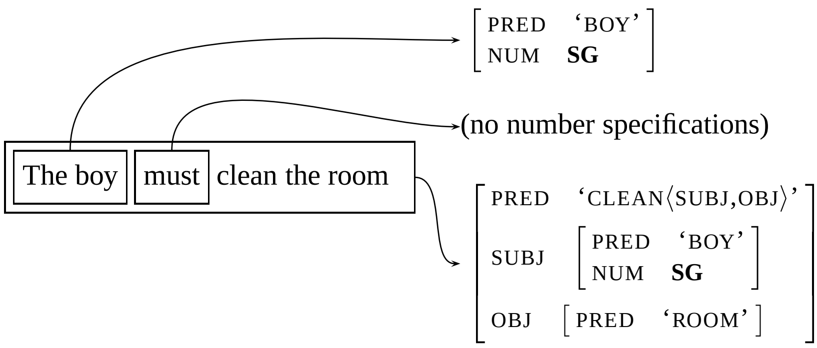
(7) Ill-formed f-structure with clashing number specification:



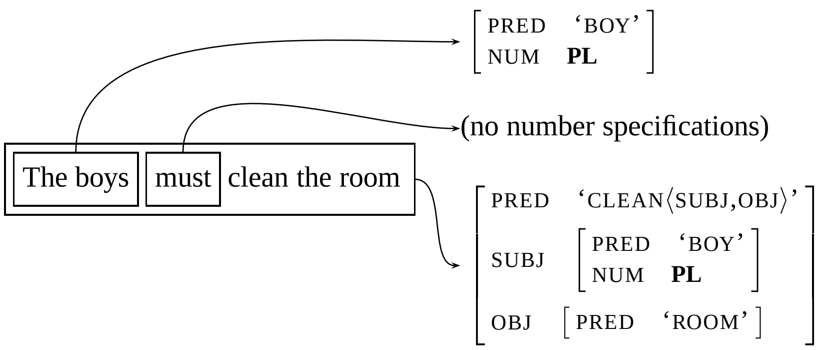
In English and many other languages, number is not always overtly specified. Many English verbs, including modals like *must*, do not impose person or number constraints on their subjects. In such cases, number may be specified only by the noun:

(8) The *boy*/*boys* must clean the room.
SG PL

(9)



(10)



3 Indeterminacy, ambiguity, or underspecification?

For a small number of English nouns, singular and plural forms are not distinguished. When such nouns appear as subjects, the verb may provide the only indication of number:

(11) a. The *sheep/fish/deer* *is* moving quickly.
? SG

b. The *sheep/fish/deer* *are* moving quickly.
? PL

There are in principle several possibilities for the treatment of the number value of these nouns, according to standard LFG analyses of indeterminacy, ambiguity, and underspecification, as follows:

- Ambiguity: there are two homophonous nouns, *sheep*$_{SG}$: [NUM SG], *sheep*$_{PL}$: [NUM PL]
- Underspecification: No NUM value is specified for *sheep*; it can be specified by another component of the sentence, such as the determiner or verb, as either SG or PL.
- Indeterminacy: *sheep* is both singular and plural (in a sense to be defined in the following).

We begin by exploring and dismissing a treatment of nouns like *sheep* as having indeterminate number. A hallmark of indeterminacy is the ability to simultaneously satisfy conflicting requirements on the same feature (Dalrymple and Kaplan, 2000; Dalrymple, King and Sadler, 2006). The case feature for the Polish noun *kogo* 'who' is indeterminate, as shown by Dyła (1984), since it can simultaneously satisfy an ACC requirement and a GEN requirement:

(12) Kogo Janek lubi a Jerzy nienawidzi (Polish)
who Janek likes and Jerzy hates
? OBJ CASE = ACC OBJ CASE = GEN
'Who does Janek like and Jerzy hate?' (Dyła, 1984)

This is also true for the case feature in German; as shown by Groos and van Reimsdijk (1979), *was* 'what' can simultaneously satisfy an ACC requirement and a NOM requirement:

(13) Ich habe gegessen was übrig war (German)
I have eaten what was left
OBJ CASE=ACC NOM/ACC SUBJ CASE=NOM
'I ate what was left.' (Groos and van Reimsdijk, 1979)

Formal analyses of indeterminacy have been proposed by Dalrymple and Kaplan (2000) and Dalrymple, King and Sadler (2006); though there are important differences between these analyses, they share the property that indeterminate features have complex values which can simultaneously satisfy multiple conflicting constraints.

(14) Specification of NOM/ACC case indeterminacy:

a. *was* according to Dalrymple and Kaplan (2000): $\begin{bmatrix} \text{CASE} & \{\text{NOM, ACC}\} \end{bmatrix}$

b. *was* according to Dalrymple, King and Sadler (2006): $\begin{bmatrix} \text{CASE} & \begin{bmatrix} \text{NOM} & + \\ \text{ACC} & + \end{bmatrix} \end{bmatrix}$

However, such analyses are inappropriate for the number feature, since the same form cannot satisfy singular and plural requirements at the same time:

(15) *The sheep *is* here and *are* happy.
SG PL

Hence, the correct treatment for nouns like *sheep* must involve ambiguity or underspecification, not indeterminacy.

Although it is difficult to provide clear evidence as to whether English nouns like *sheep* are ambiguous or underspecified, in other languages it is clear that underspecification is the right treatment. Biak nouns are never marked for number, though number distinctions are obligatory elsewhere in the language; in the examples in (16), the demonstrative determiners and verbs show number agreement (Mofu, 2009).

(16) a. $\boxed{\text{rum}}$ ine iwawa
house this.SG 3SG.shake
'This house is shaking.'

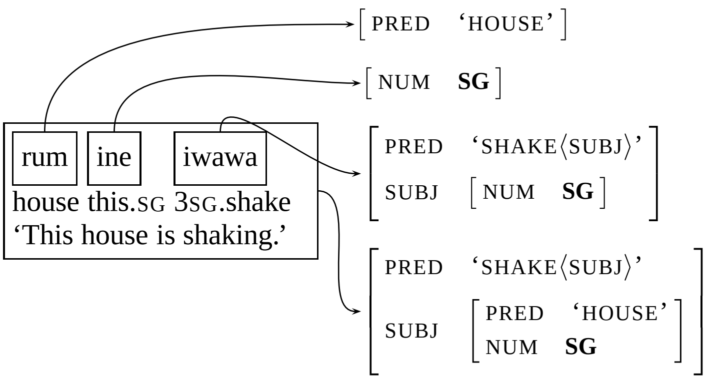
b. $\boxed{\text{rum}}$ suine suwawa
house this.DUAL 3DUAL.shake
'These two houses are shaking.'

c. $\boxed{\text{rum}}$ skoine skowawa
house this.PAUCAL 3PAUCAL.shake
'These (several) houses are shaking.'

d. $\boxed{\text{rum}}$ nane nawawa
house this.PL.INANIM 3PL.INANIM.shake
'These houses are shaking.'

Occam's razor precludes an analysis of all nouns in Biak as four ways ambiguous, with no morphological evidence for the ambiguity: Biak nouns are underspecified for NUM. As shown in (17), number specifications imposed by the determiner and the verb constrain the same feature, and these features must be compatible.

(17)



4 Double indexing

In the examples we have seen so far, the verb specifies the required features of one of its dependents, its subject. In more complex cases, the same word can specify number constraints for more than one dependent. The Biak possessive construction illustrates this pattern (Mofu, 2009): the possessive determiner specifies the person and number of the possessor, and additionally specifies the number of the possessee head noun. An overt possessor may appear before the head noun, as in (18e).

(18) a. roma bye-di
son POSS.3SG-DET.SG
'his son'

b. roma bye-suya
son POSS.3SG-DET.DUAL
'his two sons'

c. roma aye-di
son POSS.1SG-DET.SG
'my son'

d. roma aye-skoi
son POSS.1SG-DET.PAUCAL
'my (several) sons'

e. Yohanes rum bye-dya
Yohanes house POSS.3SG-DET.SG
'Yohanes's house'

Such patterns are sometimes referred to as "double indexing".[3] The LFG analysis is straightforward; the doubly-agreeing word specifies the number of the possessor as well as the possessee:

[3]This use of the term "double indexing" is different from its use in formal semantics, which refers to the representation of different kinds of referential dependencies by different kinds of indices that can appear on the same phrase (e.g. Heim 1993).

(19)

$\boxed{\text{Yohanes}}\ \boxed{\text{rum}}\ \boxed{\text{bye-dya}}$

Yohanes house POSS.3SG-DET.SG
'Yohanes's house'

$\left[\text{PRED}\quad \text{`YOHANES'}\right]$

$\left[\text{PRED}\quad \text{`HOUSE'}\right]$

$\begin{bmatrix} \text{NUM} & \textbf{SG} \\ \text{POSS} & \left[\text{NUM}\quad \textbf{SG}\right] \end{bmatrix}$

$\begin{bmatrix} \text{PRED} & \text{`HOUSE'} \\ \text{NUM} & \textbf{SG} \\ \text{POSS} & \begin{bmatrix} \text{PRED} & \text{`YOHANES'} \\ \text{NUM} & \textbf{SG} \end{bmatrix} \end{bmatrix}$

Palmer (2012, this volume) provides more discussion of double indexing in Oceanic languages.

5 Representing the number feature

In much LFG work, the number feature is assumed to have atomic values such as SG, DUAL, and PL, as in the examples above. More recent work on *constructed number* assumes that the value of the NUM feature is not atomic, but a complex value whose form may be constrained in different ways by different parts of the sentence. Sadler (2011) provides an analysis of the constructed dual in Hopi which assumes a complex value for the NUM feature.

Sadler (2011) presents the following data from Hopi, taken from Corbett (2000, 169), and notes that "dual number is expressed constructively through the combination of a plural pronoun and a verb showing SG agreement", as shown in (20c):

(20) a. Pam wari
that.SG run.PERFECT.SG
'S/he ran.'

b. Puma yùutu
that.PL run.PERFECT.PL
'They ran.'

c. Puma wari
that.PL run.PERFECT.SG
'They (two) ran.'

According to Sadler's analysis, number may be partially specified by different components of the sentence: singular verbs in Hopi contribute the specification [SG $+$] for the NUM feature, and plural pronouns contribute [PL $+$]. Dual number is defined as [SG $+$, PL $+$]:

(21) | *puma* that.PL | *wari* run.SG |

$$\begin{bmatrix} \textsc{pred} & \text{‘}\textsc{pro}\text{’} \\ \textsc{num} & \begin{bmatrix} \textbf{PL} & + \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \textsc{pred} & \text{‘}\textsc{run}\langle\textsc{subj}\rangle\text{’} \\ \textsc{subj} & \begin{bmatrix} \textsc{num} & \begin{bmatrix} \textbf{SG} & + \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \textsc{pred} & \text{‘}\textsc{run}\langle\textsc{subj}\rangle\text{’} \\ \textsc{subj} & \begin{bmatrix} \textsc{pred} & \text{‘}\textsc{pro}\text{’} \\ \textsc{num} & \begin{bmatrix} \textbf{SG} & + \\ \textbf{PL} & + \end{bmatrix} = \textsc{dual} \end{bmatrix} \end{bmatrix}$$

A similar analysis for Marori, involving a nonatomic value for the NUM feature, is discussed below and by Arka (2012, this volume).

This analysis of constructed number may appear similar to Dalrymple, King and Sadler's treatment of feature indeterminacy, discussed above: in both analyses, a complex f-structure appears as the value of a feature (the value of the feature CASE for Dalrymple, King and Sadler 2006, NUM for Sadler 2011), with the possibility for more than one component of the complex structure to have the value '$+$' ([NOM $+$, ACC $+$] for Dalrymple, King and Sadler, [SG $+$, PL $+$] for Sadler). Importantly, however, the two analyses are in fact very different. For indeterminate features like CASE, a complex value such as [NOM $+$, ACC $+$] allows an indeterminate form to simultaneously fulfil **conflicting** case requirements. In contrast, Sadler's analysis does not entail that dual nouns are in any sense simultaneously SG and PL; instead, complete patterns of feature values holistically represent the kinds of values that are expressed as atoms in other analyses (SG, DUAL, PL). This shows that structures that are similar in appearance can be used to express very different linguistic intuitions.

6 Syncretism and number

Recent LFG work has explored patterns of syncretism and underspecification in the morphology of number. We review this work here as an illustration of the general issue of the tradeoff in complexity between grammatical constraints and grammatical structures: often, similar phenomena can be analysed either in terms of relatively simple constraints on complex structures, or by complex statements of constraints on simple structures.

6.1 Murrinh-Patha

Nordlinger (2011) presents an analysis of the morphology of number agreement on verbs in Murrinh-Patha (Australian), which we will use as an example of complex constraints on simple structures. Nordlinger's analysis has been considerably simplified for the purposes of this discussion; for the complete analysis, see Nordlinger (2011) and, for more discussion of number marking in Murrinh-Patha, Nordlinger (2012, this volume).

A partial paradigm for the Murrinh-Patha verb meaning 'see' is shown in (22):

(22) a. bam-ngkardu
3SG-see
'He/she saw him/her.'

b. bam-ngintha-ngkardu
3SG-FEM.DUAL-see
'They two (female non-siblings) saw him/her.'

c. pubamka-ngkardu
3DUAL-see
'They two (siblings) saw him/her.'

d. pubamka-ngkardu-ngime
3DUAL-see-FEM.PAUCAL
'They (paucal female nonsiblings) saw him/her.'

e. pubam-ngkardu
3PL-see
'They (paucal siblings/plural) saw him/her.'

A notable feature of this analysis is the reuse of forms in different and seemingly incompatible parts of the paradigm. As Nordlinger (2011) notes, an analysis involving accidental coincidence of form is unsatisfying; the patterns shown here are systematic.

(23)

bam	singular
bam + ngintha/nintha	dual non-sibling
pubamka	dual sibling
pubamka + ngime/neme	paucal non-sibling
pubam	paucal sibling
pubam	plural

In her analysis of these forms, Nordlinger (2011) proposes to use the standard atomic values SG, DUAL, PAUCAL, and PL for the NUM feature. Crucially, the constraints associated with each form involve disjunction and the use of constraining equations to control the contribution of the morpheme combinations, as follows:

(24) bam: $\{(\uparrow \textsc{subj num}) = \textsc{sg}$
$\mid (\uparrow \textsc{subj num}) =_c \textsc{dual}\}$

pubamka: $\{(\uparrow \textsc{subj num}) = \textsc{dual}$
$(\uparrow \textsc{subj sib}) = +$
$\mid (\uparrow \textsc{subj num}) =_c \textsc{paucal}\}$

ngintha: $(\uparrow \textsc{subj num}) = \textsc{dual}$
$(\uparrow \textsc{subj sib}) = -$

ngime: $(\uparrow \textsc{subj num}) = \textsc{paucal}$
$(\uparrow \textsc{subj sib}) = -$

Consider, for example, the form *bam-ngkardu*:

(25) a. bam-ngkardu
3SG-see
'He/she saw him/her.'

b. bam: $\{(\uparrow \textsc{subj num}) = \textsc{sg}$
$\mid (\uparrow \textsc{subj num}) =_c \textsc{dual}\}$

The constraints associated with *bam* can be paraphrased as follows: *bam* contributes the value SG for the number feature, or it appears in a context in which the value DUAL for the number feature is provided by another form. In (26) there is no other form to provide the value DUAL, so SG is correctly chosen:

(26)
$$\begin{bmatrix} \textsc{pred} & \text{`}\textsc{see}\langle\textsc{subj,obj}\rangle\text{'} \\ \textsc{subj} & \begin{bmatrix} \textsc{num} & \textbf{SG} \end{bmatrix} \end{bmatrix}$$

If the form *ngintha* is present, it contributes DUAL number, which is incompatible with a SG value:

(27) a. bam-ngintha-ngkardu
3SG-FEM.DUAL-see
'They two (female non-siblings) saw him/her.'

b. bam: $\{(\uparrow \textsc{subj num}) = \textsc{sg}$
$\mid (\uparrow \textsc{subj num}) =_c \textsc{dual}\}$

ngintha: $(\uparrow \textsc{subj num}) = \textsc{dual}$
$(\uparrow \textsc{subj sib}) = -$

Since the SG specification for *bam* cannot be satisfied, the DUAL constraining equation for *bam* must be satisfied. The DUAL value contributed by *ngintha* satisfies the constraining equation, and the result is as in (28):

(28) $\begin{bmatrix} \text{PRED} & \text{`SEE}\langle\text{SUBJ,OBJ}\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{NUM} & \text{DUAL} \\ \text{SIB} & - \end{bmatrix} \end{bmatrix}$

Nordlinger's analysis uses disjunctive constraints over atomic values such as SG, DUAL and PL to achieve simple and familiar f-structures for Murrinh-Patha verbs. For further discussion of Murrinh-Patha verb morphology, see Nordlinger (2012, this volume).

6.2 Marori

Arka (2011) provides a discussion of verb morphology in Marori (isolate, Trans New Guinea) which is similar to Sadler's (2011) in using complex values for the NUM feature; the use of complex values allows a very simple statement of the contribution of different forms in the Marori verbal paradigm. Arka's analysis distinguishes singular, dual, paucal, and plural agreement through a verb-internal constructive strategy. Like the preceding analysis, the discussion of Marori presented in the following has been considerably simplified; see Arka (2011) for the full analysis, and Arka (2012, this volume) for more discussion of number marking in Marori.

Like Hopi, the Marori dual is formed as a combination of nonsingular and nonplural. Subject agreement in example (29a) is singular (singular and nonplural), (29b) is dual (nonsingular and nonplural), and (29c) is plural (nonsingular and plural):

(29) a. keswemeb
ksw=∅-∅-me-∅
hit=3-$\boxed{\text{2SG}}$-3MASC-$\boxed{\text{2NONPLURAL}}$
'You (SG) will hit him.'

b. kesneme
ksw=∅-n-me-∅
hit=3-$\boxed{\text{2NONSING}}$-3MASC-$\boxed{\text{2NONPLURAL}}$
'You (2) will hit him.'

c. kesnemem
ksw=∅-n-me-∅
hit=3-$\boxed{\text{2NONSING}}$-3MASC-$\boxed{\text{2PL}}$
'You (more than 2) will hit him.'

Constructed dual can also be expressed by a nonsingular subject with a nonplural verb:

(30) emnde (yanadu) na=n bosik eyew ∅-nda-m
$\boxed{\text{3NONSING}}$ two 1SG=for pig see 3-3FEM-$\boxed{\text{2/3NONPLURAL}}$
'They (2) hunted a pig for me.'

Singular subjects take nonplural agreement:

(31) efi yewrifam na=n bosik eyew ∅-nda-m
$\boxed{\text{3SG}}$ female 1SG=for pig see 3-3FEM-$\boxed{\text{2/3NONPLURAL}}$
'She/the woman hunted a pig for me.'

A nonsingular pronominal form in combination with a plural verb means 'three or more' (nonsingular, nondual):

(32) emnde (usindu) fis na=n bosik eyew ∅-ndi-m
$\boxed{\text{3NONSING}}$ all yesterday 1SG=for pig see 3-3MASC-$\boxed{\text{3PL}}$
'They (3 or more) hunted a pig for me.'

Like Sadler, Arka (2011) assumes that the value of the NUM feature is nonatomic; unlike Sadler, dual is treated as [SG −, PL −] rather than [SG +, PL +]. Arka's feature treatment for Marori is as in (33):

(33) Singular: $\begin{bmatrix} \text{SG} & + \\ \text{PL} & - \end{bmatrix}$

Dual: $\begin{bmatrix} \text{SG} & - \\ \text{PL} & - \end{bmatrix}$

Plural: $\begin{bmatrix} \text{SG} & - \\ \text{PL} & + \end{bmatrix}$

Crucial to Arka's analysis is that nonatomic features give rise to natural classes of features: singular and dual are [PL −], dual and plural are [SG −]. The form glossed 'singular' contributes [SG +], the nonsingular form is [SG −], plural is [PL +], and nonplural is [PL −]. This allows the following analyses:

(34) 'You (SG) will hit him':
keswemeb
ksw=∅-∅-me-∅
hit=3-2SG-3MASC-2NONPLURAL

$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{SG} & + \end{bmatrix} \end{bmatrix}$ $\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PL} & - \end{bmatrix} \end{bmatrix}$ = singular

(35) 'You (2) will hit him':
kesneme
ksw=∅-n-me-∅
hit=3-2NONSING-3MASC-2NONPLURAL

$\begin{bmatrix}\text{SUBJ} & \begin{bmatrix}\text{SG} & -\end{bmatrix}\end{bmatrix}$ $\begin{bmatrix}\text{SUBJ} & \begin{bmatrix}\text{PL} & -\end{bmatrix}\end{bmatrix}$ = dual

Rather than using disjunctive constraints on atomic features, Arka's analysis makes use of simple underspecified constraints on complex features: complex **structures** and simple **constraints**. It is not clear whether such a strictly compositional approach to constructed number is possible for all languages: Nordlinger (2012, this volume) presents data from Murrinh-Patha that are difficult to analyse on an approach involving complex values and underspecification. Arka (2012, this volume) provides additional discussion of number marking and verbal number in Marori.

7 Inverse number

We conclude with a brief description of two phenomena for which there are no standard or established analyses within LFG, but which must be addressed in a complete analysis of the morphology, syntax, and semantics of number.

In languages with **inverse number**, different classes of nouns have different 'unmarked' or default values for the number feature, and inverse morphological marking changes the number value to a 'marked' value; see Corbett (2000) for more discussion. A notorious case of inverse number is found in Kiowa (Kiowa-Tanoan; Watkins 1984; Harbour 2008), and shown in (36), excerpted from Harbour (2008, Chapter 2). Nouns in Kiowa fall into a number of classes, three of which are illustrated here. In the class represented by the noun 'young man', the unsuffixed noun is singular or dual, and nouns with the inverse number suffix *dɔ́* are plural. In the class represented by 'tomato', the unsuffixed noun is dual, and the suffixed form is singular or plural; in the class represented by 'tree', the unsuffixed noun is dual or plural, and the suffixed noun is singular. Such patterns pose interesting questions for the morphology-syntax-semantics interface.

(36)

	singular	dual	plural	
young man	tógúl	tógúl	tógúú-$\boxed{\text{dɔ́}}$	(suffixed: plural)
tomato	k!ɔ̂ɔ̨-$\boxed{\text{dɔ}}$	k!ɔ̂n	k!ɔ̂ɔ̨-$\boxed{\text{dɔ}}$	(suffixed: non-dual)
tree	áá-$\boxed{\text{dɔ}}$	áá	áá	(suffixed: singular)

More discussion of inverse number in Oceanic is provided by Palmer (2012, this volume).

8 Verbal number

Verbal number marking indicates that an event took place a number of times or that an event had a number of (usually theme) participants. Veselinova (2006) illustrates the phenomenon with examples from Mupun (Chadic; Frajzyngier 1993; Veselinova 2006). The Mupun verb meaning ‘kill’ has two forms: one for a single event, and one for multiple killing events:

(37) a. tù: ‘kill, singular action’

b. tù-é: ‘kill, plural action’

The form in (37a) is used for one killing event, and the form in (37b) is used for several events. Since a rat can be killed only once, in (37b) several rats must have been involved; plural marking on the noun meaning ‘rat’ is possible but not necessary.

(38) a. n-tu joos
1SG-kill.SG rat
‘I killed a rat.’

b. n-tue joos (mo)
1SG-kill.PL rat (PL)
‘I killed rats.’

Suppletive forms are also found:

(39) a. cīt: ‘beat, singular action’

b. nás: ‘beat, plural action’

The singular-action form cannot be used with a plural object, as shown in example (40a). Example (40b) shows that it is possible to use the plural-action form with a singular object to describe multiple beating events involving the same individual:

(40) a. *wu cit mo
3SG hit.SG 3PL
‘He hit them.’ (with a plural object, singular ‘hit’ cannot be used)

b. wu nás war
3SG hit.PL 3SG
‘He hit her many times.’ (multiple events with singular object)

Arka (2012, this volume) discusses verbal number in Marori, which involves multiple exponence of the kind discussed in Section 6.

9 Conclusion

LFG provides a solid theoretical basis for work exploring the syntax and semantics of number, but a good deal of work remains to be done. Besides a complete treatment of inverse number and verbal number, described in the final two sections of this overview paper, unexplored or underexplored areas include the following areas:

- the determination and specification of number for coordinate structures, including structures with singular number, such as “my friend and colleague”, as well as structures with plural number, like “George and Fred” and “the dog and cat”
- the analysis of nouns which seem to exhibit both singular and plural properties, such as British English “company”, which require singular determiners but can appear with plural verbs (“this company are well managed”): see Hristov (2012) for interesting and illuminating discussion of these examples, and proposals for their analysis from an LFG perspective
- the formal representation of the number feature, particularly for languages whose number systems include dual, trial, or paucal in addition to singular and plural
- related to the general issue of the representation of the number feature, the question of whether the number feature should be treated as privative, with one of the values of the feature represented as the absence of a value for the feature

The papers presented in the Workshop provide a firm basis for the exploration of these and other issues.

Acknowledgments

Thanks to I Wayan Arka for proposing and organising the Workshop on Number at LFG12, to the workshop participants I Wayan Arka, Rachel Nordlinger, and Bill Palmer, and to the audience at LFG12 for stimulating discussion of these issues. And as always, thanks to Miriam Butt and Tracy King for helpful comments on an earlier draft of this paper, and for their usual exemplary efficiency in preparing the conference proceedings for publication.

References

Arka, I Wayan. 2012. Verbal Number, Argument Number and Plural Events in Marori. In *On-line Proceedings of the LFG2012 Conference*.

Arka, Wayan. 2011. Constructive Number Systems in Marori and Beyond. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG2011 Conference*.

Corbett, Greville G. 2000. *Number*. Cambridge, UK: Cambridge University Press.

Dalrymple, Mary and Kaplan, Ronald M. 2000. Feature Indeterminacy and Feature Resolution. *Language* 76(4), 759–798.

Dalrymple, Mary, King, Tracy Holloway and Sadler, Louisa. 2006. Indeterminacy by Underspecification. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG2006 Conference*.

Dixon, Robert M. W. 1988. *A Grammar of Boumaa Fijian*. Chicago: University of Chicago Press.

Dyła, Stefan. 1984. Across-The-Board Dependencies and Case in Polish. *Linguistic Inquiry* 15(4), 701–705.

Frajzyngier, Zygmunt. 1993. *A Grammar of Mupun*. Berlin: Dietrich Reimer Verlag.

Groos, Anneke and van Reimsdijk, Henk. 1979. Matching Effects in Free Relatives: A Parameter of Core Grammar. In Adriana Belletti, Luciana Brandi and Luigi Rizzi (eds.), *Theory of Markedness in Generative Grammar: Proceedings of the 1979 GLOW Conference*, pages 171–216, Pisa: Scuola Normale Superiore di Pisa.

Harbour, Daniel. 2008. *Morphosemantic Number: From Kiowa Noun Classes to UG Number Features*. Studies in Natural Language and Linguistic Theory, Springer.

Heim, Irene. 1993. Anaphora and Semantic Interpretation: A Reinterpretation of Reinhart's Approach. Technical Report, University of Tübingen, reprinted in Sauerland and Percus (1998, 205–246).

Hristov, Bozhil. 2012. *Agreement, Case Assignment, and Nominal Coordination*. D.Phil. dissertation, University of Oxford.

Kibort, Anna and Corbett, Greville G. 2008. Number. In *Grammatical Features*, University of Surrey.

Mofu, Suriel. 2009. *Biak Morphosyntax*. D.Phil. dissertation, Oxford University.

Nordlinger, Rachel. 2011. LFG and Language Documentation. Abstract in *On-line Proceedings of the LFG2011 Conference*, ed. Miriam Butt and Tracy Holloway King.

Nordlinger, Rachel. 2012. Number Marking in the Daly River Languages. In *On-line Proceedings of the LFG2012 Conference*.

Palmer, Bill. 2012. Nominal Number in Meso-Melanesian. In *On-line Proceedings of the LFG2012 Conference*.

Sadler, Louisa. 2011. Indeterminacy, Complex Features and Underspecification. *Morphology* 21, 379–417.

Sauerland, Uli and Percus, Orin (eds.). 1998. *The Interpretive Tract*. MITWPL, Department of Linguistics and Philosophy, MIT.

Steinhauer, Hein. 1985. Number in Biak: Counterevidence to Two Alleged Language Universals. *Bijdragen tot de Taal-, Land- en Volkenkunde* 141, 462–485.

Veselinova, Ljuba N. 2006. Verbal Number and Suppletion. In Martin Haspelmath, Matthew S. Dryer, David Gil and Bernard Comrie (eds.), *The World Atlas of Language Structures*, Oxford: Oxford University Press.

Watkins, Laurel. 1984. *A Grammar of Kiowa*. Lincoln: University of Nebraska Press.

AN LFG APPROACH TO WOLOF CLEFT CONSTRUCTIONS

Cheikh Bamba Dione
University of Bergen

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper investigates the syntax of clefts in Wolof and proposes an analysis based on the Lexical Functional Grammar (LFG) formalism. Wolof clefts illustrate an interaction between morphology, syntax and information structure. In particular, they vary morphosyntactically depending on what item is clefted. Structurally, the clefts lack the cleft pronoun, are mono-clausal at the phrasal level, however, bi-clausal at the functional level. Furthermore, they relate to copular constructions in that both instantiate the same form. Thus, an understanding of these constructions is a prerequisite for understanding how clefting works.
In this paper, I review different approaches towards copula predication within LFG and present my analysis of Wolof data. I propose a parallel syntactic approach that assumes a close-complement (PREDLINK) for copular and cleft clauses. In addition, I posit an i(nformation)-structure projection to allow for extra-syntactic analysis.

1 Introduction

This paper explores the copula-cleft connection in Wolof, which has basically three types of clefts (Torrence, 2005), as given in (1)[1].

(1) a. (*xale yi*) ***ñu-a*** *lekk jën wi*. subject cleft
child the 3pl-COP eat fish the
'It's the children who ate the fish.'

b. i. (*xale yi*) *jën wi* ***la-ñu*** *lekk*. non-subject cleft
child the fish the COP-3pl eat
'It's the fish that the children ate.'

ii. *jën wi* ***la*** *xale yi lekk*. non-subject cleft
fish the COP.3 child the eat
'It's the fish that the children ate.'

c. (*xale yi*) ***da-ñu*** *lekk jën wi*. verb cleft
child the COP-3pl eat fish the
'What the children did is eat the fish.'

The cleft sentences in Examples (1a-1c) vary morphosyntactically depending on what item is clefted. Such an item is determined by means of special morphemes (e.g. *ñu-a*, *la-ñu* and *da-ñu*) which put the discourse function (DF) focus on the subject (1a), non-subject (1b-i-1b-ii) and verb constituent (1c), respectively. Morphophonologically, the discourse markers can be decomposed into a pronominal base (e.g. *ñu*) combined with a copula (e.g. *–a*, *la-*, *da-*). In subject and non-subject

[†]I thank Miriam Butt for kindly providing me sample data on discourse structure analysis. Also, I thank my advisor, Koenraad De Smedt, for valuable comments on different versions of this paper.

[1]The material in parenthesis is a non-obligatory subject. Wolof permits an independent clause to lack an explicit subject (see Torrence (2003) among others).

focus clauses, the clefted material immediately precedes the discourse marker. In verb clefts, however, it follows this marker. Thus, the initial subject in (1a) (i.e. *xale yi* "the children") appears in the standard subject position which is also a focus position. In contrast, in the non-subject cleft in (1b-i), the same position is occupied by the verbal object which typically bears a complement function (CF). Accordingly, this sentence has a completely different structure with the object in focus and the initial subject in topic position. Furthermore, Example (1b-i) shows that in case of a topicalization of the subject, this constituent must be resumed by a subject marker (i.e. *ñu*). In (1b-ii) however, no topicalization holds and hence no need for resumption, which otherwise would lead to ungrammaticality.[2]

The clefts in (1a-1c) are related to copular constructions in that both construction types instantiate the same form. Like clefts, each of the clauses in (2) exhibits a distinct morphology. Futhermore, sentences (2a) and (2c) may have a cleft reading, as can be seen from their translations. Such constructions basically contrast with ordinary clefts in that they contain a nominal predicate and often instantiate the imperfective (IPFV) aspect marker.

(2) Wolof copular constructions

a. *xale yi* ***ñu-a-y*** *baykat*. Subject copula
child the 3pl-COP-IPFV farmer
'The children are farmers.' / 'It's the children who are farmers.'

b. *xale yi baykat* ***la-ñu***. Non-subject copula
child the farmer COP.3pl
'The children are farmers.'

c. *xale yi* ***da-ñu-y*** *baykat*. Predicate copula
child the COP-3pl-IPFV farmer
'The children are farmers.' / 'It's because the children are farmers.'

This paper is organized as follows. Section 2 will review the different approaches to copular constructions within LFG and then present a new analysis of Wolof predication. Section 3 is devoted to an examination of clefting and its relationship to the information structure. In Section 4, I will present the LFG-based analysis proposed for Wolof clefts. In Section 4.3 I deal with a few complex problems of the syntax-information-structure interface raised by Wolof clefts and give arguments for the appropriateness of an i-structure projection. Section 5 will briefly discuss the analysis of Wolof adjectival constructions. I conclude with Section 6.

2 Copula Constructions in LFG

The literature on copular analyses in the LFG framework can basically be split into two main types: a single-tier and a double-tier analysis (Nordlinger and Sadler,

[2]Note that in structures like (1b-ii), the copular form *la* expresses only the person feature and remains the same for singular and plural subjects.

2007). The second analysis type has been investigated in more details by Dalrymple et al. (2004) who divided it into two variants that differ in a significant way: (i) an open-complement double-tier (XCOMP) and (i) a closed complement double-tier analysis PREDLINK (Butt et al., 1999). These approaches are extensively discussed by Attia (2008) and Sulger (2009). I will only briefly review them in Sections 2.1 and 2.2.

2.1 Single Tier Analysis

The single-tier analysis is one possible way of dealing with copula constructions in LFG (Nordlinger and Sadler, 2007). Nordlinger and Sadler's (2007) assumption about the single-tier analysis is drawn from the structure of verbless copula constructions. As noted by Rosén (1996), the analysis of 'verbless syntactic constructions' is appealing for the Lexical Functional Grammar (LFG) architecture which "does not stipulate that syntactic functions must be expressed configurationally at some level of the grammar" (Rosén, 1996). The single-tier analysis type stipulates that in copula constructions the predicate bears the sentential head function by selecting for a subject. This is illustrated by the clause in (3) from Russian and its associated functional representation in (4) (Nordlinger and Sadler, 2007, p. 141).

(3) Ona vrač.
3sg.fem.nom doctor.sg.nom
'She is a doctor.'

(4)
$$\begin{bmatrix} \text{PRED} & \text{`DOCTOR}\left\langle\text{SUBJ}\right\rangle\text{'} \\ \text{CASE} & \text{nom} \\ \text{NUM} & \text{sg} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{fem} \\ \text{PERS} & 3 \\ \text{CASE} & \text{nom} \end{bmatrix} \end{bmatrix}$$

Beside verbless structures, Nordlinger and Sadler (2007) argued that this analysis type can also handle copular constructions for languages which have overt copulas such as English. Hence, for this approach it does not matter whether the copula is present or absent. Furthermore, Dalrymple et al. (2004) consider it adequate for cases in which the copula is optional, such as with Japanese predicative adjectives.

However, the single-tier analysis is troublesome because it has to provide evidence that the predicate can subcategorize for a subject, as the Russian noun *vrač* in (3) does. This is particularly problematic because a separate analysis must be posited depending on whether the category of the predicate constituent (e.g. adjective vs. noun) can license a subject or not, although the predication is the same. As pointed out by Attia (2008) and Sulger (2009), the presence or absence or the copula is not enough motivation for postulating two separate analyses.

2.2 Double-Tier Analysis

The LFG formalism provides another possibility for handling predicative constructions: the double-tier analysis (Nordlinger and Sadler, 2007). This analysis type involves a copular verb which has two arguments: a subject and a predicate. The LFG literature distinguishes between two variants of this type, which differ in a significant way. In the first variant the postverbal phrase fills an open complement XCOMP function. In the second variant it bears a closed complement function PREDLINK. Both approaches are briefly reviewed in Sections 2.2.1 and 2.2.2.

2.2.1 Open-complement Double-tier

Under the open complement double-tier approach, the postverbal phrase is not considered as the sentential head. It is rather interpreted as an XCOMP whose subject is controlled by the copular verb's subject (Butt et al., 1999). In turn, either the copular verb (if present) or a null element (e.g. in verbless structures) is assumed to act as the sentential head (Dalrymple et al., 2004).

The main advantage with this approach is that it can easily capture phenomena like agreement between the post-copular complement and the subject of the copula via functional control. This makes it appropriate for copular constructions found in languages such as French and Norwegian which exhibit this agreement type. The French example in (5) and its related representation in (6), both from Dalrymple et al. (2004), illustrate how the open-complement double-tier analysis works for this language.

(5) Elle est petite.
she.F.SG is small.F.SG
'She is small.' (French)

(6) XCOMP analysis of French copula

$$\begin{bmatrix} \text{PRED} & \text{`be}\left\langle \text{XCOMP} \right\rangle \text{SUBJ'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`she'} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{fem} \end{bmatrix}_1 \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{`small}\left\langle \text{SUBJ} \right\rangle\text{'} \\ \text{SUBJ} & [\,]_1 \end{bmatrix} \end{bmatrix}$$

Under the double-tier analysis, Example (5) consists of a subject *elle*, a copular verb *est* and a predicate which surfaces as an adjective *petite*. The adjective is assumed to have a subject which is identified with the matrix subject. As can be seen from the f-structure in (6), the control from the matrix subject through the adjective can be captured by control equations, e.g. (↑SUBJ)=(↑XCOMP SUBJ).

However, such an approach is problematic for several reasons. First of all, the assumption that the post-copular element is open implies a constraint for it to have a subject. The subject argument is required in order to satisfy the completeness criterion (i.e. a subject is needed in order to fill the control equation of the verb). This is naturally troublesome for phrasal constituents (e.g. NPs, PPs) which do not have an overt subject and, hence, would require two different subcategorization frames: one without a SUBJ argument and one for the predicative use (Butt et al., 1999). Secondly, the argument for the open complement analysis is weakened by the status of agreement across languages. The most important objection to this approach regarding agreement is that in "languages like Norwegian, for example, there is no subject-verb agreement, so that subject-adjective agreement must be treated differently from subject-verb agreement in any case" (Dalrymple et al., 2004, p. 196). Furthermore, Attia (2008) makes the counterargument that even for languages with subject-verb agreement like French, this agreement form is not the same as the subject-predicate agreement found in copular constructions. Third, the open complement analysis does not bring enough arguments to represent predicative constructions in a way that makes them distinct from normal subject raising verbs (Attia, 2008; Sulger, 2009). Finally, the main drawback of such an approach is that it results in a clash of PRED values if the post-copular complement has a subject (Dalrymple et al., 2004). This is for instance the case in closed complement clauses in English headed by an overt complementizer, as shown in (7) and (8) from Dalrymple et al. (2004, p. 194). The feature clash comes from the fact that two elements of the sentence in (7) are associated with the SUBJ function in the embedded f-structure in (8).

(7) The problem is that they appear.

(8)
$$\begin{bmatrix} \text{PRED} & \text{`be}\left\langle\text{XCOMP}\right\rangle\text{SUBJ'} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{`problem'}\end{bmatrix} \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{`appear}\left\langle\text{SUBJ}\right\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED} & *\text{`they/problem'}\end{bmatrix} \end{bmatrix} \end{bmatrix}$$

2.2.2 Closed Complement Double-Tier

The PREDLINK analysis is the second variant of the double-tier analysis. This approach assumes that the predicate is a closed complement and that no control equation between the subject and the predicate is needed. This avoids the difficulty encountered with the XCOMP analysis regarding subcategorization. Categories which do not have an overt subject do not need to subcategorize for it, since this is not required for completeness. More precisely, the PREDLINK approach "models the fact that a particular property is predicated of the subject in a syntactically reasonable way and provides enough information for subsequent semantic analysis" (Butt et al., 1999, p.70). Examples (9) and (10) from Butt et al. (1999, p.70)[3], illustrate a typical PREDLINK analysis.

(9) The tractor is red.

(10)
$$\begin{bmatrix} \text{PRED} & \text{`be}\left\langle\text{SUBJ, PREDLINK}\right\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{`tractor'}\end{bmatrix} \\ \text{PREDLINK} & \begin{bmatrix}\text{PRED} & \text{`red'} \\ \text{ATYPE} & \text{predicative}\end{bmatrix} \end{bmatrix}$$

The f-structure for (9) using a PREDLINK analysis is given (10). It is interpreted as there being a copular verb *be* subcategorizing for a subject and a PREDLINK identified with the adjective *red*. This adjective, in turn, is interpreted as predicating a property of the subject and, contrary to the open-complement XCOMP approach, does not need to subcategorize for it.

As outlined in Butt et al. (1999), Attia (2008) and Sulger (2009), the close-complement analysis presents several advantages. First, it is a universal LFG treatment for predicative constructions. To this extent, it provides a deeper representation of these constructions abstracting away the several behaviours of the different constituents occupying the predicate position. As Attia (2008) pointed out, under this approach, predicative structures "receive a default f-structure analysis that expresses the existence of subject (SUBJ) and predicate (PREDLINK) as primitive grammatical functions and to consider the use of a copula as a parameter of variation across languages" (Attia, 2008, p. 148). This analysis type is independent of whether the copula is present or absent, obligatory or optional or whether agreement features hold between the subject and predicate. These parameters vary according to the specific language's properties. Second, as noted by Sulger (2009), the PREDLINK analysis is not affected by the constituent type of the copula complement. Hence, it can handle any constituent types with different semantic roles. Finally, the close-complement analysis can capture all the representations that can be modeled using the XCOMP analysis, even if encoding long-distance agreement using XCOMP may look more intuitive (Dalrymple et al., 2004, p. 196).

[3]For the f-structure in (10) I only give the parts relevant for this discussion. For the complete structure see Butt et al. (1999).

2.3 Towards an Analysis of Wolof Copula in LFG

In this section, I present my analysis of Wolof copula predication in LFG. Wolof copular constructions are similar to the those found in Maltese and Hebrew in that they are 'verbless' (Nordlinger and Sadler, 2007). 'Technically speaking' they do not contain a verbal copula element. More precisely, they derive from morphologically complex markers which are not properly lexical words, and hence, do not project the level of lexical category. The copular constructions given in (2) have a structure which consists of (i) a subject, (ii) a complex word incorporating the copular morpheme and (iii) a predicate. The copula may surface in different forms according to the construction type. As shown in Example (2a), this complex word is the result of an incorporation process of the copula with the subject agreement marker. Morphologically, the whole complex is an agreement marker that consists of a person marker (e.g. *mu*), a copula (e.g. *-a*) and an imperfective aspect marker. It, additionally, expresses focus features. I argue that these markers bear the functional position *I* (originally for INFL) (Falk, 1984) and acts in the clause like a head. More precisely, they belong to the functional category *Icop*, which, in turn, shares the categorial features of the lexical category V.

The LFG annotation in (11) illustrates a possible lexical entry for *ñu-a-y* in (2a). Recall, however that Example (2a) is two-way ambiguous between a purely predicative and a cleft construction. Therefore, the same item *ñu-a-y* needs to be annotated in two different ways. The lexical annotation in (11) illustrates a possible analysis of this item as a copular element. The alternative analysis of the same item as cleft will be given in Section 3.4.

(11) *ñu-a-y* Icop (↑PRED)='a⟨ (↑SUBJ)(↑PREDLINK) ⟩'
(↑VTYPE)=copular
(↑SUBJ NUM)=$_c$ pl
(↑SUBJ PERS)=$_c$ 3
(↑FOCUS)=(↑SUBJ).

Unlike French and Norwegian, post-copular complements in Wolof do not show any agreement with the subject. Hence, for Wolof, agreement is not enough reason for postulating an XCOMP analysis. With regard to these Wolof data, the copula is always present and obligatory, except for stative "adjectival" constructions discussed in Section 5. The occurrence of the copula is often accompanied with subject and focus marking, which are both optional. Thus, for Wolof the 'parameters of variation' (Attia, 2008) seem to play a minor role.

3 Clefts

3.1 Clefting and Information Structure

The concept of information structure relates to the type of information encoded in a particular utterance, denoted as discourse functions. In the LFG framework, DFs are commonly classified into one of the three categories (King and Zaenen, 2004).

1. Topic/Theme/Given
2. Focus/Rheme/New
3. Contrastiveness

This traditional division assigns each of these three a particular function. Focus usually encodes new information; e.g. something that the speaker or writer expects their hearer or reader might not already know. In contrast, Topic is assumed to be given information, i.e. information that the speaker or writer expects the hearer or reader may be familiar with or that has been introduced in the discourse.

Cleft constructions are typical examples to illustrate how discourse functions can be encoded at a syntactic level. The organization of information structure is tightly linked to the clefts' function as focusing tools used by the speaker/writer when it comes to draw attention to salient parts of their message (Hasselgård, 2002). The term focus will refer in this article only to contrastive focus, so that, the clefted constituent always conveys new information in the context, which is explicitly contrasted with something in the preceding context, as Example (12) from Kihm (1999, p. 245) (emphasis, parentheses and labeling mine) illustrates. In (12) the subordinate clause contains given or known information (i.e. that someone has written *Ulysses*), while the clefted constituent introduces new information (i.e. the author of *Ulysses*, which is contrasted with other possible authors).

(12) It is [Joyce]$_{Focus}$ [who]$_{Topic}$ wrote Ulysses.

King and Zaenen (2004) highlighted three different ways of encoding discourse functions: encoding via (i) a privileged structural position, (ii) discourse markers or particles, and (iii) a specific intentional pattern. In the first possibility, termed 'structural encoding', the particular discourse function is expected to surface in a particular phrase structure position. It has been argued, for instance, that topic and background traditionally bear an initial and postverbal position respectively, while focus often appears in both positions (i.e. pre- or post-verbal). Another possibility to encode DFs consists of using discourse markers. This possibility is used by a wide range of languages, e.g. Wolof, Japanese and Hindi. Japanese, for instance, has a topic marker *wa* while Hindi makes use of different markers to encode (exclusive or inclusive) contrastive focus (King and Zaenen, 2004). Finally, some languages such as English use intonation to signal the focused element in a sentence. Examples (13) and (14) from King and Zaenen (2004) illustrate focused constructions using discourse markers and focus stress, respectively.

(13) [rAdha=ne=hI]$_{Focus}$ baccho=kO kahAnI sunAyI
Radha=erg=Foc children=ACC story hear
'It was (only) Radha who told the children a story' (Hindi)

(14) a. Did you see Mary or John?
b. I saw [JOHN]$_{Focus}$.

3.2 Cleft structure

As discussed in the previous section, clefting is essential to spread information of a single proposition over two clauses, hence two information units (Hasselgård, 2002). Accordingly, in many languages, including English, typical cleft sentences are overtly bi-clausal: i.e. they consist of a main clause and a subordinate clause. In turn, the subordinate clause may contain a copula and the focused element, as given in (15) (Ebert, 2011).

(15) it COPULA X [RELATIVE [$_s$ ....]] (Cleft)

There are, however, many other languages, including Wolof and some Bantu languages, which do not follow this typical structure. Kikuyu (Bantu), for instance, uses the discourse marker *ne* to put an element into focus via clefting (Ebert, 2011). As can be seen in (16), the cleft construction does not contain a relative clause. Hence, this sentence does not exhibit a bi-clausal structure, but rather seems to be mono-clausal.

(16) ne mae Abdul a-ra-nyu-ir-ɛ
FOC water Abdul 3SG-PRT-drink-ASP-FV
'It is water that Abdul drank.'

3.3 Analysis of DFs in LFG

LFG offers different possibilities for analyzing discourse functions (King and Zaenen, 2004). DFs can be captured structurally or functionally (e.g. at f- or i-structure level). LFG also proposes optimality theoretic approaches for this issue.

An LFG analysis of DFs via structural encoding involves basically two possibilities. One assumption is that predicates subcategorize for DFs. Following Alsagoff (1992), King and Zaenen (2004) argued that verbs in Malay subcategorize for topic. Accordingly, in this language, topic arguments can be identified with a particular grammatical function and annotated on the corresponding affixes. Alternatively, DFs can be assigned via functional annotations on c-structure nodes. According to Bresnan and Mchombo (1987) topics in Chicheŵa are associated with a privileged c-structure position and are therefore identified with some grammatical function via anaphoric binding.

An LFG approach to discourse markers encoding DFs involves case / morphology; i.e. the link between the DF and the discourse marker is captured at the morphological level. For instance, the Hindi focus marker *hI* specifies its relation to the DF focus it is associated with via the inside out equation (FOCUS ↑) in its lexical entry, as shown in Figure 1.

X(P)
X(P) Cl-disc
(FOCUS ↑)
hI

Figure 1: LFG analysis of the Hindi DF marker *hI* (King and Zaenen, 2004).

Even if the different ways of encoding DFs in LFG are well described, their adequate representation level remains controversial. DFs can potentially be represented at the f-structure or the i-structure. These two possibilities reflect the distinction between syntacticized and real discourse functions[4](Bresnan, 2001). A new analysis for Wolof will be discussed in Section 4.

3.4 Wolof Cleft

The interest of exploring the morphosyntactic properties of Wolof clefts is twofold. First of all, compared to their English counterparts, Wolof clefts look as if they have a completely different structure consisting of one clause. They also crucially differ from clefts in languages like English in that they lack the cleft pronoun, e.g. 'it'.

Secondly, in Wolof cleft clauses, the information structure is not merely a secondary component superimposed on a core syntactic one, but it rather organizes the verbal system and even conditions the choice of the inflectional markers for each cleft type. So, as Robert (2000) pointed out, Wolof represents an extreme case of grammaticalization of focus. According to Kihm (1999, p. 246) clefting is "the only means in the language to put an element into focus" and consequently, Wolof lacks the other focus expression forms found in many languages such as "focus stress as in English or focusing through position as in Hungarian". The Wolof data presented in this work, however, will show that, in fact, Wolof makes a combination of structural and morphological encoding to mark the discourse function focus. Let us first discuss some of the properties mentioned for Wolof clefts.

The morphosyntactic structure of Wolof clefts appears to be similar to those found in Hindi (see Section 3.1) in that the focus argument is encoded morphologically. Furthermore, the language exhibits a case of multiple encoding in the sense that it uses more than one strategy at the same time to put a specific element into focus. Unlike Kihm (1999), I argue that structural encoding is still available in Wolof, but seems to be deficient just because it is combined with discourse marking (i.e. morphology). In this regard, Wolof behaves like Tagalog which uses the same multiple encoding mechanism which combines position and marker (King and Zaenen, 2004). Hence, in the subject and non-subject clefts constructions given in (1), *ñu-a* and *la-ñu* are discourse function markers and the head of *I*; the constituent in SpecIP is always a focus and maps to subject function if the marker is *ñu-a* and a non-subject function if it is *la-ñu*.

[4]For more details see Dalrymple (2001); King and Zaenen (2004) among other authors.

Abstracting away from the individual constructions given in (1), I assume for subject, verb and non-subject clefts in Wolof a unique structure consisting of: (i) an optional constituent (*XP*) which can be of different categories, (ii) a multicategorial item *SMCOP* incorporating the copula and the optional subject marker and (iii) a sentential clause *S*. This multicategorial item has two morphological structures: the subject marker can precede the copula as in subject cleft (i.e. *(SM-)COP*) or follow it as in verb and non-subject clefts (i.e. *COP(-SM)*). The Wolof cleft structure I propose is given in (17). It has a linear ordering common for the three cleft types.

(17) Wolof cleft structure:

a. XP (SM-)COP / COP(-SM) $[_S\ [_{VP}$ ....V....]] (Cleft)

The structure in (17) is similar to the one for Kikuyu given in Section 3.2. Due to an opacifying effect of morphology, the cleft structure appears mono-clausal. At the surface level, there is no subordinate clause which is overtly introduced as some kind of relative clause. I follow Kihm (1999) in claiming that, clefts in Wolof are, in fact, bi-clausal just like their English counterparts (Kihm, 1999, 246). However, unlike Kihm (1999), I argue that this bi-clausal structure only holds at the functional level, and is not overtly expressed at the c-structure level. Hence, I rather assume that they are mono-clausal at the phrasal level.

4 Analyzing Wolof clefts in the LFG Framework

In this section I present my ideas on how Wolof clefts can be analyzed using the LFG mechanism. The next two following sections present the analysis of clefts at the c-structure and the representation of the involved DFs at the f-structure. These DFs are interpreted as the syntacticized discourse functions FOCUS and given topic (GVN-TOP), which is a special type of the topic argument. I will then examine the problems relative to the representation of these DFs in the f-structure and outline possible solutions in Section 4.2.

4.1 Constituent structure

As mentioned in Section 3.4, Wolof combines two encoding possibilities: structural and morphological encoding. DFs are assigned using functional annotations on special c-structure nodes and, at the same time, using morphology, i.e. focus markers. As Figures 2-3 show, the DF foci in Wolof are associated with *Spec IP-cop* and identified with the grammatical functions subject and complement, respectively. The assignment of focus is regulated by the morphology: the focus type varies depending on the morpheme used (e.g. *moo* vs. *la*). In addition, these figures are evidence for assuming that topics and foci are distinct in Wolof (Russell, 2006).

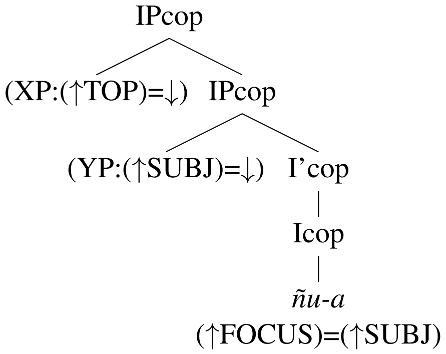


Figure 2: C-structure for subj. cleft

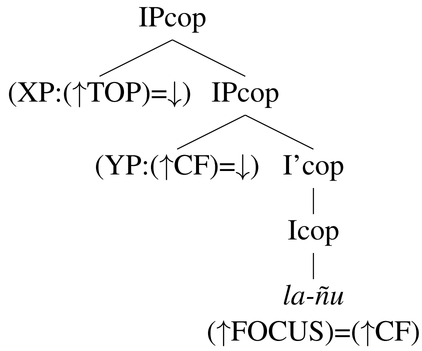


Figure 3: C-structure for non-subj. cleft

4.2 Functional Representation

In the standard assumption of LFG, discourse function information has traditionally been encoded in the f-structure via annotations on the c-structure. Structurally encoded topic and focus arguments are considered syntacticized[5] (Bresnan and Mchombo, 1987) and placed in the f-structure alongside grammatical functions. This approach works well for languages which encode DFs using subcategorization or exhibit phenomena such as pronoun incorporation (King and Zaenen, 2004).

Concerning the Wolof data, let us first assume that the DFs found in clefts have a syntactic role similar to those discussed in Bresnan and Mchombo (1987). According to Bresnan and Mchombo (1987), in a cleft construction, the clefted constituent typically bears both functions FOCUS and TOPIC, as can be seen in (12). It is FOCUS in the main clause and TOPIC in the embedded one. However, the same constituent cannot bear both functions at the same level. Thus, one can argue that the DFs found in subject and non-subject clefts have a clear syntactic role, hence grammaticalized, and do not exhibit a mismatch regarding the associated grammatical functions. These DFs would, therefore, be represented in the f-structure.

(18) *ñu-a* Icop (↑PRED)='be⟨ (↑SUBJ)(↑PREDLINK) ⟩'
(↑PREDLINK)=(↑FOCUS)
(↑FOCUS-TYPE)=contrastive
(↑SUBJ SUBJ NUM)=$_c$ pl
(↑SUBJ SUBJ PERS)=$_c$ 3.

The LFG annotation in (18) illustrates a possible lexical entry for *ñu-a* found in subject clefts as in (1a). At the functional level, the copular inflectional element *Icop* in (18) is analyzed with two arguments SUBJ and PREDLINK. As the c-structures in Figures 2 and 3 show, the content of the S-clause in (1a) *lekk jën*

[5]However, as Bresnan (2001, p. 97) noted, grammaticalised discourse functions like TOPIC and FOCUS should be distinguished from real discourse functions which are a part of discourse in a sense of communicative functions like information packaging.

wi "eat the fish" is presupposed to be already known: it is a 'given topic' (GVN-TOP, in addition to being a SUBJ). In other words, the fact that someone ate the fish is assumed to be known and is, therefore, an old information; the new information is that the one doing it was *xale yi* "the children". This can be captured by saying that 'the children' in (1a) is predicated of the property of having eaten the fish. This reversal of the roles of logical subject and predicate is what is achieved by clefts. Hence, the lexical DP (e.g. *xale yi*) which fills the specifier position of the clause (see the c-structures in Figures 2 and 3) bears both the grammatical function associated with PREDLINK and the syntacticized FOCUS function, as indicated by the equation (↑PREDLINK)=(↑FOCUS). Likewise, the subject of the cleft clause also has a subject which is here identified with FOCUS. This functional relation is captured by the equation (↑FOCUS)=(↑SUBJ SUBJ) while (↑FOCUS-TYPE)=contrastive specifies the type of the linked discourse function. The constraining equations (e.g. (↑SUBJ SUBJ NUM) =c sg) describe the subject-verb agreement required for the subject of the embedded clause.

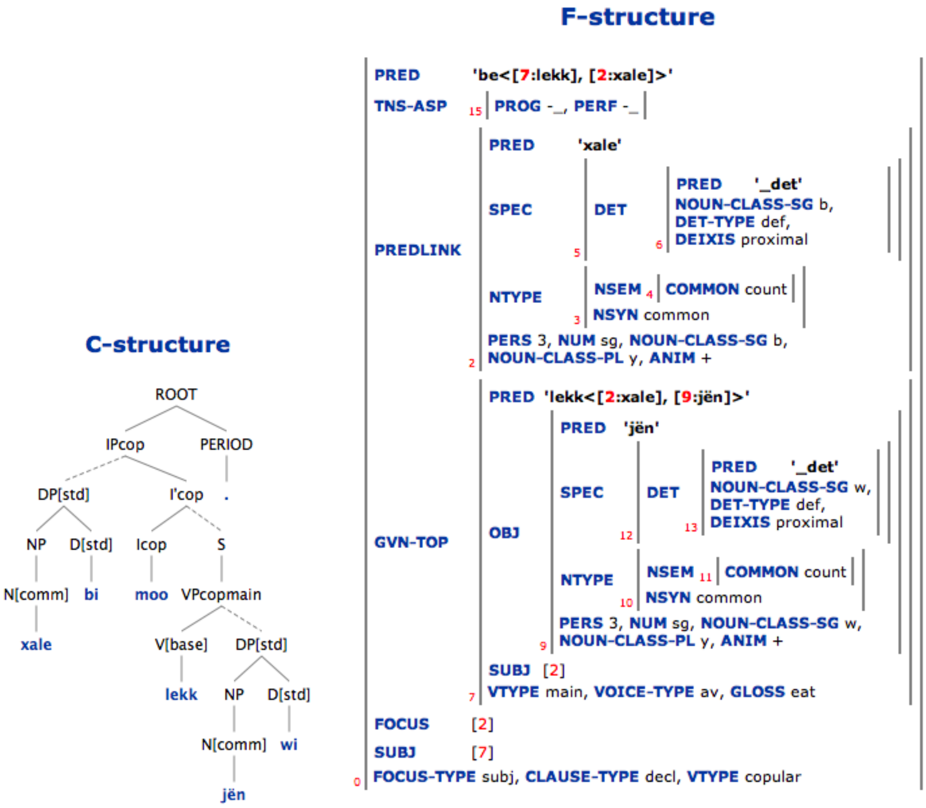


Figure 4: C- and F-structure for subject cleft

In non-subject cleft clauses, I assume that the given topic is the same, i.e. the old information. Unlike the subject cleft, however, here the non-subject constituent bears both the grammatical function PREDLINK and the focus function since it contains the new information. This is easily captured by the parallel approach.

Figures 4 and 5 illustrate the proposed c- and f-structure for Wolof subject and non-subject clefts respectively. As the parsed samples show, the representation of DFs via lexical annotations and annotation at the c-structure tree works well for subject and non-subject clefts, assigning them both a grammatical function and a discourse function.

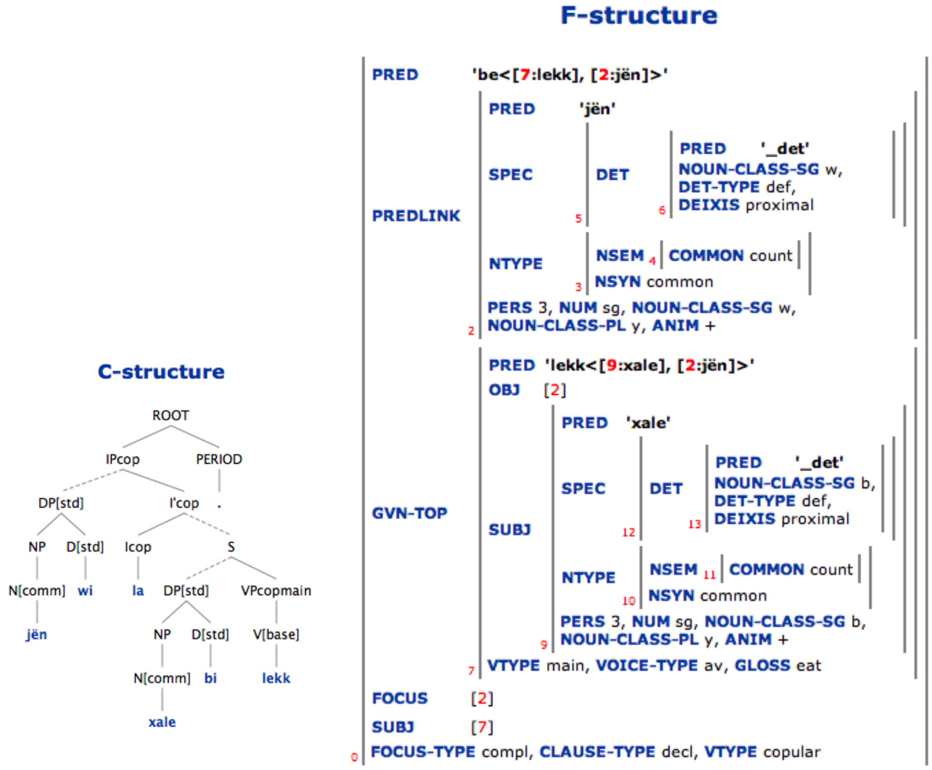


Figure 5: C- and F-structure for non-subject cleft

4.3 Focusing F-structure Heads

However, as King and Zaenen (2004) reported, the representation at f-structure turns out to be ultimately problematic due to the mismatches between DFs and grammatical functions. For instance, in cases of focusing f-structure heads, "the standard annotations result in the incorrect scoping of the discourse functions in that more material is focused or topicalized than intended" (King, 1997, p. 2). This is exactly the problem encountered for the analysis of Wolof verb cleft in that focusing the f-structure head results in wider scope than desired. By assigning to the verb the discourse function contrastive focus, all the arguments included in the sub-f-structure containing the head are also contained within the discourse function. Hence, not only the verb PRED is focused, but so are its arguments (SUBJ and OBJ). This is linguistically incorrect since contrastive focus on the verb excludes focus on any material but the verb itself (King, 1997). Hence, the analysis of the verb cleft cannot be correctly predicted by this approach, and therefore needs an alternative approach.

In Example (1c), contrastive focus picks out the verb as prominent information. However, trying to capture this by lexically annotating the DF marker *da-ñu* with (↑PREDLINK)=(↑FOCUS) would result in a too wide scope. As already discussed, some DF material will not always overlap with f-structure elements (e.g. phrases which are part of the f-structure, but not of the i-structure), yielding mismatches between both structures. In order to overcome such divergences, many of the recent works on DFs within LFG proposed an independent component called i-structure for representing the information structure of a sentence instead of analyzing it within the f-structure (Butt and King, 1996; King, 1997).

To tackle this issue in Wolof, I follow King (1997) in proposing that the DF information found in the clefts be captured in this independent projection, i.e. i-structure, and that i-structure be related to c-structure through a delta projection *d* for a discourse structure. This projection is assumed to be accessible to the s(emantic)-structure, and relates to the argument structure as well. Hence, the information relevant to the i-structure is assumed to be the core predicate value without its associated argument structure.

More precisely, I combine two solutions proposed in Kaplan and Maxwell (1996) and King (1997). First, I posit an i-structure projection distinct from the f-structure, which can easily capture these mismatches. Secondly, I remove arguments of the verb retaining only the core PRED in the i-structure. This approach yields the desired partial f- and i-structures given in (19) and (20).

(19) F-structure

$$\left[\begin{array}{ll} \text{PRED} & \text{'be}\left\langle\text{SUBJ, PREDLINK}\right\rangle\text{'} \\ \text{SUBJ} & \left[\begin{array}{ll}\text{PRED} & \text{'xale'}\\ \text{...} & \end{array}\right] \\ \text{PREDLINK} & \left[\begin{array}{ll} \text{PRED} & \text{'lekk}\left\langle\text{xale, jën}\right\rangle\text{'}\\ \text{SUBJ} & \left[\begin{array}{ll}\text{PRED} & \text{'xale'}\\ \text{...} & \end{array}\right] \\ \text{OBJ} & \left[\begin{array}{ll}\text{PRED} & \text{'jën'}\\ \text{...} & \end{array}\right]\end{array}\right]\end{array}\right]$$

As (20) shows, only the core meaning of the PRED is focused using the PRED FN value (Kaplan and Maxwell, 1996), which remains when the arguments of PRED are removed, avoiding projecting the argument structure into the i-structure.

(20) I-structure

$$\left[\begin{array}{ll} \text{FOCUS} & \left[\begin{array}{ll} \text{D-PRED} & \text{lekk} \\ \text{FOCUS-ON} & \text{verb} \\ \text{FOCUS-TYPE} & \text{contrast} \end{array}\right] \\ & \\ \text{GVN-TOP} & \left[\begin{array}{ll} \text{D-PRED} & \text{'xale'} \\ \text{TOPIC-TYPE} & \text{given} \end{array}\right] \end{array}\right]$$

For Wolof, the lexical item associated with verb cleft is annotated as in (21).

(21) *da-ñu* Icop (↑PRED)='be⟨ (↑SUBJ)(↑PREDLINK) ⟩'
d::* = (d::M* FOCUS)
(↑PREDLINK PRED FN)=(d::* D-PRED)
(d::* FOCUS-TYPE)=contrastive
(d::* FOCUS-ON)=verb
...

This analysis places the relevant core PRED of the PREDLINK in the FOCUS of the i-structure. The projection ***d::*** indicates the DF projection. Additionally, I use the * and *M** notations as proposed by Kaplan (1987). The annotation * and *M** refer to the node (i.e. ↓) and its mother (↑), respectively. I further use additional annotations to specify the focus type and the focused constituent.

5 Discussion: Adjectives as a Missed Category

Another issue that is crucially raised by this proposal concerns the lack of adjectives in Wolof, where the adjective's role is taken over by stative verbs (McLaughlin, 2004). Adopting the single-tier analysis, I argue that, Wolof stative (adjectival) verbs behave like Japanese adjectives (Dalrymple et al., 2004) in that: (i) they provide the main PRED for the clause, i.e licensing their own subject and (ii) they do not require the copula, as seen in (22b). Also like Japanese, Wolof adjectival constructions can take an overt copula, as in (22a). However, in Wolof, this may result in a focused construction, as the English translation of Example (22a) shows.

(22) a. sa bët bi da-fa xonq.
Poss2sg eye the COP.3sg red
'Your eye is red.' / 'Your eye is RED.'

b. sa bët bi xonq na.
Poss2sg eye the red 3sg
'Your eye is red.'

Following Dalrymple et al. (2004), I assume that the stative (adjectival) verb is an open function and subcategorizes for a SUBJ. Examples (23-24) propose a possible analysis which illustrates the contrast between the neutral reading in (23) and the focus one in (24).

(23) $\begin{bmatrix} \text{PRED} & \text{`xonq}\left\langle\text{SUBJ}\right\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{`bët'}\end{bmatrix} \end{bmatrix}$

(24) $\begin{bmatrix} \text{PRED} & \text{`xonq}\left\langle\text{SUBJ}\right\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{`bët'}\end{bmatrix} \\ \text{FOCUS} & \begin{bmatrix}\text{xonq}\end{bmatrix} \end{bmatrix}$

6 Conclusion

In this paper I have proposed a parallel syntactic approach for Wolof cleft and copular constructions using the LFG architecture. In this proposal, both the simple copula and clefts in Wolof share an uniform phrase structure and make use of the close-complement double-tier (PREDLINK) at the functional level. The PREDLINK analysis is appropriate for these constructions in that it provides a universal LFG treatment for predicative constructions and is not affected by divergent analyses of copula constructions within this language. This paper has also investigated different possibilities for capturing the discourse functions related to the Wolof clefts at the adequate representation level. On the one hand, contrastive focus in subject and non-subject clefts is considered as syntacticized as is the case in a wide range of languages which show agreement between discourse function and f-structure grammatical functions. For such languages, it has been argued that there is a syntactic topic and focus which therefore should be placed in the f-structure alongside grammatical functions. On the other hand, however, this approach cannot account for an appropriate representation of the discourse function found in Wolof verb clefts in that it includes more material in the i-structure than intended. For this purpose, this paper has postulated an independent projection i-structure to correctly account for focusing f-structure heads. Such a projection has been modeled as a projection of the c-structure, which can be accessed by the semantic structure. Furthermore, the information relevant to the i-structure has been extracted as the core predicate value without the argument structure. The proposed analysis has been implemented in a computational LFG grammar using the XLE software (Crouch et al., 2012). In the current development of the grammar, however, the encoding of DFs within the i-structure is still experimental while the internal organization of this additional projection requires further research.

7 Acknowledgements

This research has received support from the EC under FP7, Marie Curie Actions SP3-People-ITN 238405 (CLARA).

References

Alsagoff, Lubna. 1992. *Topic in Malay*. Ph. D.thesis, Stanford University.

Attia, Mohammed. 2008. *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation*. Ph. D.thesis, University of Manchester.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Bresnan, Joan and Mchombo, Sam A. 1987. Topic, pronoun, and agreement in Chicheŵa. *Language* 63(4), 741–782, reprinted in Masayo Iida, Steven Wechsler, and Draga Zec (eds.), Working Papers in Grammatical Theory and Discourse Structure: Interactions of Morphology, Syntax, and Discourse, pp. 1–59. Stanford: CSLI Publications.

Butt, Miriam and King, Tracy Holloway. 1996. Structural topic and focus without movement. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG '96 Conference*, Rank Xerox, Grenoble.

Butt, Miriam, King, Tracy Holloway, Niño, María-Eugenia and Segond, Frédérique. 1999. *A Grammar Writer's Cookbook*. Center for the Study of Language and Information, Stanford, California.

Crouch, Dick, Dalrymple, Mary, Kaplan, Ron, King, Tracy, Maxwell, John and Newman, Paula. 2012. *XLE Documentation*. On-line documentation, Palo Alto Research Center (PARC).

Dalrymple, Mary. 2001. *Lexical-Functional Grammar (Syntax and Semantics, Volume 34) (Syntax and Semantics)*. Academic Press.

Dalrymple, Mary, Dyvik, Helge and King, Tracy Holloway. 2004. Copular Complements: Closed or Open? In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG '04 Conference*, University of Canterbury.

Ebert, Christian. 2011. Focus and Syntax. Course on information structure [Class Handout]. SoSe 2011. Universität Tübingen.

Falk, Yehuda N. 1984. The English Auxiliary System: A Lexical-Functional Analysis. *Language* 60(3), 483–509.

Hasselgård, Hilde. 2002. Adverbials in *IT*-cleft constructions. In Karin Aijmer and Bengt Altenberg (eds.), *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, pages 195–211, Amsterdam: Rodopi.

Kaplan, Ronald M. 1987. Three seductions of computational psycholinguistics. In Peter Whitelock, Mary McGee Wood, Harold L. Somers, Rod Johnson and Paul

Bennett (eds.), *Linguistic Theory and Computer Applications*, pages 149–181, London: Academic Press.

Kaplan, Ronald M. and Maxwell, John T. 1996. LFG Grammar Writer's Workbench. Technical Report, Xerox PARC.

Kihm, Alain. 1999. Focus in Wolof: a study of what morphology may do to syntax. In G. Rebuschi & L. Tuller (ed.), *The Grammar of Focus*, pages 245–273, Amsterdam: John Benjamins.

King, Tracy Holloway. 1997. Focus Domains and Information-Structure. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG '97 Conference*, University of California, San Diego.

King, Tracy Holloway and Zaenen, Annie. 2004. F-structures, Information Structure, and Discourse Structure. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG '04 Conference*.

McLaughlin, Fiona. 2004. Is there an adjective class in Wolof? In R.M.W. Dixon and Alexandra Y. Aikhenvald (eds.), *Adjective classes. A crosslinguistic typology.*, pages 242–262, Oxford University Press.

Nordlinger, Rachel and Sadler, Louisa. 2007. Verbless Clauses: Revealing the Structure within. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling and Chris Manning (eds.), *Architectures, rules, and preferences: variations on themes by Joan W. Bresnan*, pages 139–160, CSLI Publications.

Robert, Stéphane. 2000. Le verbe wolof ou la grammaticalisation du focus. Louvain: Peeters, Coll. Afrique et Langage, 229-267. Version non corrigée.

Rosén, Victoria. 1996. The LFG Architecture and "Verbless" Syntactic Constructions. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG '96 Conference*, Grenoble, France.

Russell, Margaret A. 2006. *The Syntax and Placement of Wolof Clitics*. Ph. D.thesis, University of Illinois at Urbana-Champaign.

Sulger, Sebastian. 2009. Irish Clefting and Information-Structure. In Miriam Butt and Tracy Holloway King (eds.), *Online Proceedings of the LFG09 Conference*, CSLI Publications.

Torrence, Harold. 2003. Verb Movement in Wolof. *UCLA Working Papers in Linguistics* pages 1–31.

Torrence, William Harold. 2005. *On the Distribution of Complementizers in Wolof*. Ph. D.thesis, University of California Los Angeles.

COMPLEX PREDICATES IN ARRERNTE

Mark Dras[1] François Lareau[1]
Benjamin Börschinger[1,2] Robert Dale[1]
Yasaman Motazedi[1] Owen Rambow[3]
Myfany Turpin[4] Morgan Ulinski[3]

[1]Macquarie University [2]Universität Heidelberg
[3]Columbia University [4]University of Queensland

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

Using the example of Murrinh-Patha, Seiss (2011) illustrates how Australian Aboriginal languages can shed light on the morphology-syntax interface: one aspect of their polysynthetic nature is that information often encoded in phrases and clauses in other languages is instead found in a single morphological word. In this paper, we look at another instance, the Australian Aboriginal language Arrernte, and in particular at complex predicates within the language, to examine the implications for the morphology-syntax interface. Following from this, we show how a glue semantics-based approach can be applied to Arrernte complex predicates, in a way that fits neatly with the use of glue semantics to model lexical functions in LFG in a multilingual natural language generation environment.

1 Introduction

Using the example of Murrinh-Patha, Seiss (2011) illustrates how Australian Aboriginal languages can shed light on the morphology-syntax interface: one aspect of their polysynthetic nature is that information often encoded in phrases and clauses in other languages is instead found in a single morphological word, one manifestation of the morphology-competes-with-syntax idea discussed in Bresnan (2001). In this paper, we look at another instance, the Australian Aboriginal language Arrernte, and in particular at complex predicates within the language, to examine the implications for the morphology-syntax interface.

The context for this work is a data-to-text multilingual natural language generation (MNLG) system, where one of the languages to be generated is Eastern/Central Arrernte. For the language realisation component, the grammar is developed in XLE and the morphology in XFST. Some aspects of the language can be handled quite straightforwardly using standard LFG mechanisms. Complex predicates, however, require more consideration.

Complex predicates in LFG have most often been handled using some kind of predicate composition, sometimes at the level of argument structure elaborated via Jackendoff's Lexical Conceptual Structures (LCS) (Jackendoff, 1990), and in some cases through the use of the restriction operation (Butt, 1993; Alsina, 1997; Andrews and Manning, 1999; Butt and King, 2006; Nordlinger, 2010, exemplify some approaches). An alternative proposed in an appendix of Andrews and Manning (1999) and elaborated in Andrews (2007) is to use glue semantics, which is outlined there with a sketch of Romance causatives.

Following from our consideration of the morphology-syntax interface, we show how a glue semantics-based approach can be applied to Arrernte complex predicates, in a way that fits neatly with the use of glue semantics to model lexical collocations in LFG in an MNLG environment (Lareau et al., 2011).

[†]We acknowledge the support of ARC Discovery grant DP1095443. We thank MK Turner for help with Arrernte; and Miriam Butt, Rachel Nordlinger and Meladel Mistica for useful comments.

2 Arrernte: The Language

2.1 Structure of the Grammar

Eastern/Central Arrernte is a language of the Arandic group of the Pama-Nyungan family of Australian languages. It is one of the larger Australian languages, with perhaps 1500 to 2000 speakers, who mostly reside around Alice Springs in Australia's Northern Territory. It is also a language of regular communication, which children still speak as a first language. In addition to the early work of Strehlow (1944), there are two major written descriptions of aspects of Arrernte grammar, PhD theses by Wilkins (1989) and Henderson (1998); in addition, there is a large dictionary (Henderson and Dobson, 1994).

Henderson (1998, Sec 1.4) gives an overview of the structure of the language:

> E/C Arrernte is suffixing and agglutinative and is rich in compounding morphology of various types. The principal parts of speech are nominals, verbs, adverbs and particles/clitics. ... The core syntactic cases are Ergative (ERG), Nominative (NOM), Accusative (ACC) and Dative (DAT). In nominals other than pronouns Nominative and Accusative are syncretised, while in pronouns Ergative and Nominative are syncretised except for the first person singular which distinguishes all four cases. Other cases include Locative/Instrumental (LOC/INST), Ablative (ABL), Allative (ALL), Possessive (POSS) and Proprietive (PROP).[1] Pronouns distinguish singular, dual and plural .... All verbs must bear one of a set of suffixes referred to as the obligatory morphology. These indicate tense, mood and clausal status including switch reference marked dependent clauses.

There are also non-obligatory suffixes that precede obligatory suffixes; these may include aspect and subject number. There is a special kind of verbal category among the non-obligatory suffixes called Associated Motion, also found in other Australian languages, which indicates that "a verb-stem action happens against the background of a motion event with a specific orientation in space" (Wilkins, 2006). The (somewhat simplified) morphological structure of verbs is schematically depicted in Figure 1.

There is also reduplication, which when applied to verbs does not have a straightforward relationship to the verb structure of Figure 1. There are many varieties of reduplication: one that we refer to later in the paper is the Attenuative. The Attenuative form of some verb X is often glossed as 'start to X', and is

[1]Authors' note: Other abbreviations we use in examples are ASSOC MOTION = Associated Motion, ATTEN = Attenuative reduplicant, DO.COMING = type of Associated Motion marker, EMPH = emphasis particle, FOC = focus particle, INTENS = intensifier, ITER = iterative, IV = intransitive verbaliser, NOMLSR = nominaliser, NUM = number, PRES = present tense, PRIOR.MOTION = type of Associated Motion marker, PST = past tense, PV = preverb, QUICK:DO&GO.BACK = type of Associated Motion marker, RECIP = reciprocal, REFL = reflexive, SR = switch reference, SS = same subject, TV = transitive verbaliser, V = verb.

V root (TV) Base + IV (TV) PV … Root Non-V + TV	REFL / RECIP	number	ASSOC MOTION	aspect	ITER	tense (SR) mood (SR) SR NOMLSR

Figure 1: Simplified Eastern/Central Arrernte Verb Structure, adapted from Henderson (1998, p.276).

formed by taking the first syllable of the verb stem, adding *elpe* or *erlpe*, and then repeating the (inflected) verb; an example, giving the reduplicated form of the verb *ampangkeme*[2] in (1a), is in (1b), with the reduplicant and the reduplicated part of the base underlined.

(1) a. *ampangk-eme*
groan-PRES
(he) is groaning

b. *ampelpe-ampangk-eme*
ATTEN-groan-PRES
(he) is starting to groan (Henderson, 1998, (4.60a))

Henderson (1998, Section 5.3.1) contends that "a large proportion of verb components, and therefore verb forms, involve more than one stem or morphological word": these are referred to as complex, as opposed to simple, verbs. These complex verbs include lexical compounds consisting of preverbs (defined below) in combination with (following) inflecting verbs. Complex verbs and the phenomenon of Associated Motion will be central to this paper; we discuss them in more detail in Section 4 on complex predicates.

In contrast to the rigid morphology, Arrernte clausal syntax is relatively free, with no obvious constraints on the order of phrases, although it may be the case, as in some other Australian languages (Simpson, 2007, for example), that it is pragmatically constrained. Word order within the Noun Phrase, however, is much more restricted, and case is marked exclusively on the final element of the NP.

2.2 An LFG Analysis

Following Nordlinger and Bresnan (2011), we capture Arrernte's free word order at the sentence level by assuming a flat exocentric c-structure rooted in S. We do not model in the grammar the potential pragmatic factors that control linearization; instead, for the actual system we have a separate reranking post-process to handle this. There is no explicit copula in the present tense; like Nordlinger and Sadler (2007), we allow all nominals to act predicatively. The head of a sentence can be

[2] In this paper we follow the standard practice of referring to verbs by their 'dictionary form', the stem combined with the present tense ending *-eme*. Also, in glossing we use the morpheme boundaries of Henderson (1998) and Henderson (2002). In full sentences, we follow the usual orthographic convention of starting with an uppercase letter.

a verb or a nominal, but only a finite verb can carry tense; with other heads, the auxiliary *aneme* (lit., 'sit') can be added to carry such information if desired. Grammatical functions are specified with dependent-marking (Bresnan, 2001, p111), and NPs with a semantic case such as LOC function as modifiers.

NPs, in contrast, have relatively fixed internal word order, with the possible exception of some 'floating' of demonstratives and counting terms; we model this with the separation of immediate dominance and linear precedence constraints (Falk, 2001). NP case (ergative/absolutive, as well as the numerous other cases such as locative, ablative, etc.) is handled via 'particles' in syntax, which always appear at the end of the whole NP. We model this by adding a projection level over the core NP.

For the most part, verb morphology is handled in the morphological component; this includes some relatively complicated cases of prefixing reduplication including the Attenuative mentioned in Section 2.1, which we handle in XFST using compile-replace rules (Beesley and Karttunen, 2003; Bögel et al., 2007). In between these clear-cut cases of aspects handled by the grammar versus those handled by the morphology, however, there is the grey area of complex predicates.

3 Definitions of Complex Predicates

According to an influential definition by Butt (1993), a complex predicate has to satisfy three conditions: (a) the argument structure is complex (two or more semantic heads contribute arguments); (b) the grammatical functional structure is that of a simple predicate — it is flat, and there is only a single predicate (paraphrased by Nordlinger (2010) as "monoclausal"); and (c) the phrase structure may be either simple or complex — it does not necessarily determine the status of the complex predicate.

While this definition is widely accepted within the LFG community, it ought to be noted that there also are more general definitions, e.g. Amberber et al. (2010) in their book on complex predicates, where they acknowledge that there is no agreed set of criteria for defining a complex predicate; Butt (1993) also gives the same caveat. Further, there are interesting cases that are still naturally analyzed within LFG as compex predicates although they do not meet Butt's criteria, such as the Associated Motion construction in Wambaya as described by Nordlinger (2010). We review this briefly here, along with work by Wilson (1999) on the Australian language Wagiman, both for what they have to say about the definitions of complex predicates, and for the characteristics that are similar to the complex predicates that we discuss in this paper.

Wambaya is a non-Pama-Nyungan language, and therefore relatively distant from Arrernte. Its word order is free, but there is an auxiliary obligatorily in second position (Nordlinger, 1998b). An Associated Motion marker is optionally attached to the auxiliary; there must also be a main verb. Nordlinger (2010) notes: "When combined with a motion verb, the Associated Motion marker adds the direction of the motion event, since motion verbs in Wambaya are direction-oriented. When the main verb is a non-motion verb, however, the Associated Motion affix encodes a sequential event 'go/come and VERB'." Following Broadwell (2000) on Choctaw, Nordlinger (2010) characterises these as single events and dual events respectively.

She asserts that the Associated Motion construction is monoclausal, as it contains only a single main verb and a single subject, and the Associated Motion marker cannot constitute a clausal predicate on its own. The satisfaction of condition (a) above, however, is less clear, in terms of the extent to which the Associated Motion marker can be considered a semantic predicate in the absence of syntactic predication tests. Nordlinger (2010) argues that it does on the grounds that it adds a motion predication, as well as sensitivity to the semantics of its lexical verbs.

Wagiman is also a non-Pama-Nyungan language, aspects of which are described by Wilson (1999) and reanalysed in Andrews and Manning (1999). It contains coverbs analogous to the preverbs of Arrernte (see Section 4), which are analysed by Wilson (1999) as complex predicates. They differ from other instances of complex predicates in that both components can occur independently and act as full predicates, in contrast to earlier examples of complex predicates where one component was essentially a light verb (e.g. the Romance causatives of Alsina (1997) or the Urdu permissive of Butt (1993)).

In this paper we take the definition of what counts as a complex predicate from Henderson (2002). Our analysis based on this definition is broadly in the spirit of Butt (1993); we will draw attention to the situations where it is not.

4 Complex Predicates in Arrernte

4.1 Three Types of Complex Predicate

Henderson (2002), our source of the definition of Arrernte complex predicates, more generally investigates the problematic nature of a word in Arrernte: in some contexts an element of the language will appear to be, say, a derivational morpheme, and in others a separate word. Henderson (1998) claims that "a large proportion of verb components, and therefore verb forms, involve more than one stem or morphological word, [and that this notion of complex predicates] provides a coherent account of a number of phenomena"; he then gives phonological, morphological (e.g. the possibility of reduplication) and grammatical criteria for determining wordhood in Arrernte. These broadly agree, although not always. The following are two of the phonological criteria from Henderson (2002).

Prosodically conditioned allomorphy The forms of the Reciprocal, Dual and Plural verb suffixes depend on the number of syllables between the beginning of the phonological word and later verb suffixes. For Dual and Reciprocal, the morphemes *err* and *irr* are used if the stem has an odd or even number of syllables, respectively; for Plural these morphemes are *errirr* and *irrer*, with an additional alternative *ewarr* that may apply to stems of more than one syllable.

Stress Each word bears a primary stress on the first syllable beginning with a consonant. In (relatively uncommon) words of four or more syllables, there may be stress on alternating syllables after the primary stress.

Henderson (2002) notes that there is no simple definition of a word in Arrernte in terms of grammatical criteria. Nominal morphology is limited to compounding. Verbs take suffixes as described in Section 2.1; the order of these morphemes is largely fixed, as in Figure 1. There are a number of other factors that can be taken as indicators of word status, however, including the two following:

Intervening material It is possible for some non-verbal morphemes to intervene at specific points within the verb. For example, in (2), the particle *akwele* 'supposedly' (which can appear on its own, outside of a verb or an NP) appears inside the verb, between the stem *arrerne* and its suffixes.

(2) *arrerne akwele lh-eme*
place SUPPO REFL-PRES
supposedly sit down (Henderson, 2002, (9))

Reduplication In addition, the manner and location of reduplication, which applies to verbs, gives an indication of the boundaries of a verbform.

Given the definition of complex predicate for Arrernte based on the criteria of Henderson (2002), we look at three particular types: the intransitive verbaliser (IV), lexical compounds, and Associated Motion. In the following subsections, we then consider, in light of the problematic nature of the notion of word, whether each of these three should be handled in the syntax or the morphology of our overall LFG grammar; and if in the syntax, what kind of verbs — e.g. full verbs, light verbs or auxiliaries — are involved.

Intransitive Verbaliser The IV *irreme*, in (3), is a highly productive element of the language that follows and combines with a base that can be a nominal, adverb, NP, or sometimes a clause: in (3a), it combines with the nominal *mwerre* 'good'. Its basic sense is inchoative, although it can function as a copula; the derived form of base + IV functions as an intransitive verb. The IV has often been treated as a derivational morpheme, as in Figure 1.

(3) a. *Utyene tiwelhe-me mwerre-irr-eme-le*
sore fall.off-PRES good-IV-PRES-SS
The scab falls off and the sore gets better.
(Henderson and Dobson, 1994, entry for *tiweme*)

b. ***Alakenhe*** *re ampe akweke mpwe ulk-etyenh-ele*
thus 3sg.NOM child small urine excrete-FUT-SS
irr-entye.akngerre
IV-NOMLSR
Little kids behave that way when they need to have a leak.
(Henderson, 2002, (20))

Lexical compounds Lexical compounds as in (4) consist of a preverb followed by and combined with an inflecting verb: in (4a), *lthere* is the preverb, and *iweme* the inflecting verb. In some ways, then, it is similar to the IV, although it is much less productive and more lexically idiosyncratic; the fixed order and lack of productivity also distinguish them from adverb-verb combinations. Also as with the IV, lexical compounds have often been treated as the result of derivational morphology, as in Figure 1.

(4) a. *Arelhe-le ampe lthere iw-eme*
woman-ERG child pinch$_1$ pinch$_2$-PRES
The woman is pinching the child. (Henderson, 1998, (5.28), modified)

b. *Angeme the pelhe-iw-eke*
fly 1sg.ERG spit$_1$-spit$_2$-PST
I spat the fly out. (Henderson and Dobson, 1994, entry for *pelhe-iweme*)

c. *Ampe yanhe-le-ame apmere irnterre anthurre*
child that-ERG-EMPH place INTENS INTENS
akerre-iw-eme
scatter$_1$-scatter$_2$-PRES
That child is scattering things all over the camp.
(Henderson and Dobson, 1994, entry for *akerre-iweme*)

d. ***Ikerrke** anthurre akwele re **iw-elh-eke***
stick$_1$ INTENS SUPPO 3sg.NOM stick$_2$-REFL-PST
He supposedly got himself really stuck. (Henderson, 1998, (5.5))

Associated Motion In Associated Motion constructions as in (5), the Associated Motion marker occurs between the verb stem and the obligatory morphology: in (5a), the Associated Motion marker is *artn.alp*,[3] which adds the meaning of quickly going and returning while performing some other action.

(5) a. *Artwe angk-artn.alp-eke*
man speak-QUICK:DO&GO.BACK-PST
The man quickly spoke and then went back.
(Wilkins, 2006, (15c), modified)

b. *Ar-ety-arle akwele alh-err-eme*
see-PRIOR.MOTION-FOC SUPPO GO-DUAL-PRES
Two supposedly go and then see. (Henderson, 2002, (29))

c. *Artwe angk-inty-eke*
man speak-DO.COMING-PST
The man spoke while coming this way. (Wilkins, 2006, (15a), modified)

[3]The period in the middle is conventionally used in a gloss of a separable Associated Motion marker, discussed below.

Associated Motion can occur with almost all verbs; the exceptions are what Wilkins (2006) characterises as “deictic” motion verbs (e.g. *alheme* ‘go’, *alpeme* ‘go (and come) back’). This incompatibility accords with Associated Motion having much the same semantics as the deictic motion verbs. Apart from this, the semantic contribution of the Associated Motion marker is broadly the same for motion and non-motion verbs. The Associated Motion marker in Arrernte, then, is not sensitive to the verb to which it is linked in the manner of Wambaya, but the grounds of semantic predication for complex predicate status still hold. Wilkins (1989) sees the Associated Motion marker as a morpheme for which there is a specific slot in the verb stem.

4.2 Syntax or Morphology?

The default position, then, might be to handle all three types in the morphology, as suggested by Figure 1. However, in light of Henderson (2002), we note the following points and then make proposals about where to handle each type.

For the IV construction, Attenuative reduplication — with its two possible positions as given in (6) — in conjunction with the other criteria mentioned above, indicates that the combined form is not a single simple verb. If the verb were a simple one, and the IV consequently an unequivocal derivational morpheme, only (6a) would be valid; (6b) indicates that *irreme* has at least a quasi-independent status.

(6) a. *mwelpe-mwerre-irr-eme*
ATTEN-good-IV-PRES
start to get better (Henderson, 1998, (4.66))

b. *mwerre-irrerlpe-irr-eme*
good-ATTEN-IV-PRES
start to get better (Henderson, 1998, (4.66))

The same argument can be made for lexical compounds, to which the Attenuative applies in a similar fashion. Each inflecting verb used in a lexical compound is in all cases homophonous with a free verb, whose meaning is sometimes obviously related but sometimes not. In (4a) and (4c), the inflecting verb is homophonous with *iweme* ‘throw (away)’. Henderson (1998) notes that there is a “continuum of semantic compositionality” ranging from cases where there is almost no sense of the free verb, as in (4d), to ones with a more transparent sense, as in (4c).

For associated motion, Henderson (2002), based on a range of further data than Wilkins (1989), notes that the construction e.g. in (5a) could instead be glossed as an Associated Motion particle *artn* and the full verb of motion *alpeme* ‘to go and come back’; that there are phonological grounds for considering them separate words; and that some intervening material is possible. (5b) contains the Associated Motion marker *ty.alh* (potentially an Associated Motion particle *ty* and the full verb of motion *alheme* 'go'), which is separated by the focus particle *arle* and *akwele* 'supposedly' (and is consequently written as two words in the example).

We note that all of these constructions in fact permit intervening material. In addition to (5b) just described, in (4d), the lexical compound *ikerrke-iweme* (in bold) also has *akwele* intervening, as well as the intensifier *anthurre* and the pronoun *re*; in (3b), between the base *alakenhe* 'thus' and the IV (also in bold) there is a dependent clause. This brings into focus the question of whether all of these should be considered separate words for an LFG analysis, and so, for an XLE implementation, perhaps more naturally handled in the grammar rather than the morphology.

In coming to a view about this, it is useful to consider the separability cline of Henderson (2002), reproduced here. This cline groups into classes the kinds of intervening material permitted in complex predicates:

1. certain particles and clitics: e.g. *anthurre* Intensifier, *akwele* 'supposedly', *arle* FOC;
2. *akwete* 'still';
3. third person singular pronoun functioning non-referentially as an emphatic;
4. other pronominal NPs;
5. simple non-pronominal NPs, most likely being a single nominal;
6. other adverbs, complex NPs;
7. dependent clauses.

In terms of applicability to the various kinds of complex verbs, the cline ranges from most to least widespread: that is, intensifiers such as *anthurre* in class 1 are applicable to the widest range of complex verbs, while the dependent clauses of class 7 are the most restricted. Furthermore, as can be seen from the cline above, the size of the units of intervening material broadly increases from class 1 to class 7. In our examples, (5b) illustrates intervening material from classes 1 and 2; (4d) from classes 1 and 3; and (3b) from class 7, the most extensive attested type of intervening material. Henderson (2002) notes that the cline is roughly implicational in a number of ways, in particular that if a type of complex verb allows intervening material of class n, it also allows intervening material of classes $1 \ldots n-1$.

As illustrated by (3b) for the IV construction, the potentially unbounded amount of intervening material from all separability classes between the base and the IV suggests handling these within the grammar as separate words.

Verbs with Associated Motion, by contrast, are attested as permitting intervening material only from classes 1 and 2, like the class 1 element *akwele* 'supposedly' in (5b); these are finite and quite small in extent. In addition, not all Associated Motion morphemes can be decomposed into smaller components that correspond to some motion verb: (5c) has the morpheme *intye*, typically glossed 'do X while coming this way', which has no obvious free verb counterpart.[4] The nature of the

[4]Wilkins (1989, p277) states that *intye* "is itself likely to have originated from a former motion

Associated Motion construction is then similar to those of Wambaya, as described in Nordlinger (2010) (and briefly in Section 3), although there is not the same freedom of movement of the component parts. An appropriate place to handle Associated Motion would then be in the morphology, with a slot inside the verb structure as in Figure 1, and with additional internal slots for morphemes of classes 1 and 2 of the separability cline.

Lexical compounds fit somewhere in the middle. There are attested utterances with intervening material of classes 1–5, but none with the extent of intervening material of the IV construction. The situation here then is more ambiguous between syntax and morphology than the other two cases. We propose to handle this in the syntax, given its similarity to the IV construction; there are also further reasons, which will become clear in Section 5.1, where we discuss our glue semantics-based approach and the incorporation into our representation of the notion of lexical functions from Meaning-Text Theory (MTT).

4.3 Full Verb, Light Verb or Auxiliary?

We now consider the two constructions to be handled in the grammar, the IV construction and lexical compounds. Our position is that neither the IV nor the inflecting verb of lexical compounds should be considered a full verb, which would imply that the base or preverb respectively would consequently be some kind of argument. Attenuative reduplication applies only to verbs, so the base or preverb, which can validly be reduplicated as in (6a), cannot be a plain nominal (or adverb, etc).

In addition, case marking supports this. If *mwerre* 'good' in (3a) were the object of a full verb *irreme*, then the derived compound form would be transitive, and the subject *utyene* 'sore' would be marked with the ergative marker *-le*; and this is not the case. And for lexical compounds, case marking (the ergative *-le* on the NP *ampe yanhe* 'that child') in (4c) indicates that the verb is transitive.

Regarding the status of the IV and inflecting verbs as light verbs versus auxiliaries, we draw on the helpful synthesis of Seiss (2009), which aims to bring together various definitions that have been used in the field for auxiliaries, serial verbs and light verbs, in order to work towards a common cross-linguistic usage.

Our lexical compounds consisting of preverb + inflecting verb are in fact quite similar to the inflecting verb + coverb that is part of the case study of the Australian language Ngan'gityemerri that Seiss (2009) uses to illustrate definitions of inflecting elements as auxiliaries, serial verbs or light verbs. Given this parallel, and drawing on the characterisation of light verbs in Butt (2010), the inflecting verb would be a light verb: a key characteristic is that "light verbs exhibit subtle lexical semantic differences in terms of combinatorial possibilities", and lexical compounds as we have already noted, have a high degree of lexical idiosyncrasy in terms of inflecting verbs.

---

verb meaning 'come' ", and that this has been argued for related languages by Koch (1984). However, this has no bearing on a synchronic analysis as in this paper.

The IV is less straightforward. It is not lexically idiosyncratic in the manner of the lexical compound's inflecting verb. However, it would be odd to characterise it as an auxiliary. Seiss (2009) notes that auxiliaries typically do not contribute semantic information about the type of event, whereas light verbs can: the IV by its nature typically adds an inchoative meaning to the whole complex verb. In addition, light verbs and not auxiliaries can change the valency of a construction. In the IV construction, the base is fundamentally a nominal (or adverb, etc), which can perhaps be considered to be acting as a verb in this context, based on the Attenuative as discussed above. It is unclear what the valency of the base by itself would be, but the IV definitely enforces intransitivity as its fundamental function.

We then treat the IV and the lexical compound's inflecting verb as light verbs. The discussion following links this to our existing treatment of lexical resources that have much in common with light verbs, and hence to our handling of complex predicates at the level of f-structure. In light of the complexities of the Arrernte system of spatial and movement relations (Wilkins, 2006), we leave open exactly what the semantics for Associated Motion should be, and consequently what the resulting complex predicate would look like; we thus only discuss the IV and lexical compounds in the remainder of the paper.

5 A Representation of Arrernte Complex Predicates

5.1 Meaning-Text Theory's Lexical Functions

It has been the experience of large-scale MNLG systems (Wanner et al., 2010, for example) that as much of the system as possible should be language-independent, a position advocated from the early days of the field by Bateman et al. (1991) and Cahill et al. (2000) among others. Among the mechanisms for enhancing language independence are the so-called lexical functions from MTT (Mel'cuk, 1996; Kahane and Polguère, 2001), which embody recurrent patterns of collocations. These abstract away from language-dependent collocations, such as the English *outright lie* versus French *mensonge éhonté* 'shameless lie', as well as language-internal collocational variation, such as *heavy rain*, *strong wind* or *intense bombardment* which all refer to the intensification of some phenomenon. This particular semantic notion of intensification or strength is represented by `Magn(L)`; another lexical function of interest is $\texttt{Oper}_1\texttt{(L)}$, where a semantically (mostly) empty verb serves as syntactic support to link a predicative noun to its most prominent semantic argument, for example $\texttt{Oper}_1$(TALK)=GIVE, $\texttt{Oper}_1$(ATTENTION)=PAY. Lexical functions provide an efficient mechanism for describing a wide range of collocations. In Lareau et al. (2011) we showed how these can be incorporated into LFG using glue semantics; see also the companion paper in this volume.

Among the types of collocations described in MTT are support verbs; one such is $\texttt{Oper}_1\texttt{(L)}$ above. While there is no universally agreed definition of support verbs, we follow Fillmore et al. (2003) who, in discussing the nature of "semantically transparent" lexical elements as part of the FrameNet formalisation, charac-

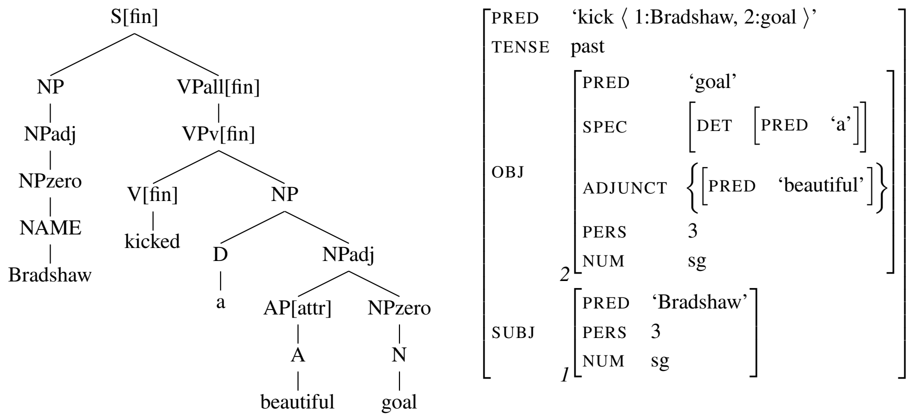


Figure 2: c-structure (left) and f-structure (right) for *Bradshaw kicked a beautiful goal*

terise support verbs as "broader than the traditional notion of light verb". We apply our mechanism for this broader class of support verbs to the light verbs of Section 4.3. We thus review briefly here our mechanism for dealing with lexical functions, which then leads in to our glue-semantics-based approach to handling Arrernte complex predicates.

The domain of our MNLG system is Australian Football League (AFL) football. Consider sentence (7), with c-structure and f-structure as in Figure 2.[5]

(7) Bradshaw kicked a beautiful goal.

In MNLG, the system starts with some representation of the input, and generates text from that, generally passing through a number of stages; in an LFG context, this first-stage input (after selection of the relevant content) might be first-order predicate logic, or perhaps more expressive representations such as intensional logic or Discourse Representation Theory (DRT) (Kamp and Reyle, 1993), with this semantic representation then mapping to f-structure and then c-structure, and at the end producing the generated text. If the meaning representation for generating sentence (7) were as in (8a), a more literal meaning representation where there is a bijection between words and terms in the meaning representation, the mapping to f-structure would be quite straightforward; if, however, the meaning representation were as in (8b), where the semantically empty element is omitted and the attribute generalised, the mapping would need some more powerful mechanism. This second meaning representation is in fact the relevant one for our system: it abstracts away from the collocationally determined use of *beautiful* to describe a good goal (which in the language of football commentary is largely interchangeable with *magnificent*, *superb*, ...) and *kick* to describe the scoring of the goal. A

[5]The c-structure here broadly follows the PARC Starter English Grammar: `http://www2.parc.com/isl/groups/nltt/xle/doc/PargramStarterGrammar/starternotes.html`.

Bradshaw	N	$(\uparrow$PRED$)=$'Bradshaw'
		Bradshaw : $\uparrow_\sigma$
goal	N	$(\uparrow$PRED$)=$'goal'
		goal : $\uparrow_\sigma$
beautiful	A	$(\uparrow$PRED$)=$'beautiful'
		$\lambda X.\text{beautiful}(X)$: $(\text{ADJ} \in \uparrow)_\sigma \multimap (\text{ADJ} \in \uparrow)_\sigma$
kicked	V	$(\uparrow$PRED$)=$'kick$\langle(\uparrow$SUBJ$),(\uparrow$OBJ$)\rangle$'
		$(\uparrow$TENSE$)=$past
		$\lambda X.\lambda Y.\text{kick}(X,Y)$: $(\uparrow\text{SUBJ})_\sigma \multimap [(\uparrow\text{OBJ})_\sigma \multimap \uparrow_\sigma]$
a	D	$(\uparrow$PRED$)=$'a'
		$\lambda X.X$: $(\text{DET} \uparrow)_\sigma \multimap (\text{DET} \uparrow)_\sigma$

Figure 3: Lexical entries with meaning constructors for mapping between the f-structure of Figure 2 and the literal meaning of (8a).

goal, worth six points, can only be scored in AFL by kicking; touching with any other body part results in a 'behind', worth one point. *Kick* therefore is semantically empty.

(8) a. kick(bradshaw, beautiful(goal))
 b. good(goal(bradshaw))

To describe the mapping between our desired semantics in (8b) and our f-structure in Figure 2, we use glue semantics as described in Dalrymple (2001). Briefly, in a glue semantics approach a lexical entry contains a meaning constructor made up of two parts: the lefthand (meaning) side represents the meaning, and the righthand (glue) side represents a logical formula over semantic structures corresponding to those meanings. We first give in Figure 3 the lexical entries that would be required for mapping the literal semantics of (8a) to the f-structure. The entry for *kick*, for example, is just the standard one for a transitive verb.[6]

To handle the mapping between the semantics of (8b) and the f-structure, we would add the entries of Figure 4. Here GOAL, by contrast, is a unary predicate: $\lambda X.\text{goal}(X)$, i.e. 'X goals', so to speak. However, in the construction under consideration here, its semantic predicativity is not echoed in syntax: there is no verb *to goal* in standard English, which is why a support verb is needed in the first place. *Kick* is this support verb, and so adds nothing to the final semantic form. *Kick* is only a support verb, however, in the context of *goal*, which is enforced by the constraining equation. *Beautiful* is similar to before, but has the more generic semantics of positive appreciation: $\lambda X.\text{good}(X)$.

Many of these more complex syntax-semantics mappings are in fact fairly regular for MTT lexical functions, and can be captured using templates, an XLE mechanism that can be used to implement LFG's Lexical Rules. For example, for the

[6]There are various ways of handling the determiner *a*, e.g. as a quantifier. We give only a simple treatment for illustrative purposes, where *a* contributes nothing to the semantics.

goal	N	(↑PRED)=‘goal’
		$\lambda X.\text{goal}(X)$: ((OBJ↑) SUBJ)$_\sigma \multimap \uparrow_\sigma$
kicked	V	(↑PRED)=‘kick⟨(↑SUBJ),(↑OBJ)⟩’
		(↑OBJ PRED)=$_c$‘goal’
		(↑TENSE)=past
		$\lambda X.X$: (↑OBJ)$_\sigma \multimap \uparrow_\sigma$
beautiful	A	(↑PRED)=‘beautiful’
		((ADJ ∈ ↑) PRED)=$_c$‘goal’
		$\lambda X.\text{good}(X)$: (ADJ ∈ ↑)$_\sigma \multimap$ (ADJ ∈ ↑)$_\sigma$

Figure 4: Additional lexical entries with meaning constructors for mapping between the f-structure of Figure 2 and the desired meaning of (8b).

lexical function `Oper`$_1$`(L)`, which represents the use of support verbs in contexts such as that of *kick* in our examples, the following template in (9) could be defined.

(9) @OPER1(L)=(↑PRED)=‘%stem⟨(↑SUBJ),(↑OBJ)⟩’
(↑OBJ PRED)=$_c$‘L’
$\lambda X.X$: (↑OBJ)$_\sigma \multimap \uparrow_\sigma$

5.2 Handling Complex Predicates via Glue

As may already be apparent from the analysis of the IV and lexical compound complex predicates as containing light verbs, we can use this exact same mechanism to handle them in the grammar component: we can take the elements of the f-structure corresponding to complex predicates and use glue semantics to combine them together. This is quite different from the more common approach in LFG, which has generally used the LCS of Jackendoff (1990). In an early version this approach, exemplified by Butt (1995), complex predicates are formed at a-structure by combining an LCS containing a ‘transparent event’ position (for example, an LCS corresponding to a light verb) with a fully specified LCS representing its argument. Andrews and Manning (1999) demonstrate some problems with this approach, including that it does not handle the combination of two full predicates, such as in Wagiman. They propose instead an approach to complex predicates using restriction projections; their approach can also use the mechanism of LCS, but they also put forward (in Appendix A) glue semantics as an alternative mechanism for the complex predicate combination. Andrews (2007) notes, albeit with a different formulation of glue semantics in that paper, that in some ways glue semantics is mimicking the effect of the alternative LCS approach at a-structure. In fact, it does away with a-structure, enforcing in the appropriate way the combination of complex predicate components at s-structure. Our approach in this section has some similarities to those of Andrews and Manning (1999) and Andrews (2007), but is implemented using the lexical functions we have adopted from MTT, as described in Section 5.1. We illustrate it using the IV construction and lexical compounds.

For the IV, the appropriate lexical function is `IncepOper`$_1$`(L)`, similar to `Oper`$_1$`(L)` but referring to a support-like verb indicating the start of something (e.g. *contract a disease*). We define the template INCEPOPER1 at the top of Figure 5. To illustrate its use, we take a slightly simpler version of sentence (3a) above, in (10). *irreme* would instantiate this template; *mwerre* 'good' is a nominal that is verbalised by *irreme*. For the generalised f-structure and lexical items with semantics (Figure 5 centre left), we obtain via the glue semantics of the lexical entries the desired overall semantics for the complex predicate (Figure 5, bottom and centre right, respectively).[7]

(10) *Artwe mwerre-irr-eme*
man good-IV-PRES
The man is starting to get better.

Note that in contrast with the LCS approach to complex predicates, *mwerre* 'good' appears as a separate, embedded f-structure, not yet combined with the IV, as this happens at s-structure via glue semantics. We noted in Section 4.3 that the base *mwerre* would not be an object; in fact it is not obvious what the most appropriate category might be. For this example we use PREDLINK (Butt et al., 1999): it is used standardly for copula or copula-like constructions (e.g. by Sulger (2009) for a closed complement double-tier analysis of Irish copulas). We also note that the lexical function `IncepOper1` is contributing the inchoative aspect of the semantics, which is somewhat different from the earlier LCS approaches.

For lexical compounds, we consider examples (4a) and (4b). One question is whether there are two full predicates here that are able to operate independently, as in Wagiman (Section 3), or whether a support verb analysis would be suitable. While for (4a) there is no attested independent use of the nominal *lthere*, in (4b) *pelhe* is an attested independent nominal meaning 'spit', and as noted earlier there is the full verb *iweme* 'throw' which in this context is perhaps indicating the motion of the spittle. However, although *pelhe* as a preverb can occur with the Attenuative (suggesting some characteristics of a verb) it cannot occur with any other verbal morphology, and it is not attested as acting as a full predicate independently, so for the construction as a whole we adopt the support verb analysis. Consequently, we use the lexical function `Oper`$_{12}$ to represent a plain support verb that takes as its subject the first semantic argument of the base of the collocation (for (4a), *arelhe*), the second argument as its direct object (*ampe*), and the base itself (*lthere*); the structure is parallel for (4b).

The question then is what grammatical function is appropriate for the base. In the case of (4b), it may again be PREDLINK: this is a resultative construction which is quite similar to a construction of the Urdu/Hindi reference dependency bank for complex predicates (Ahmed et al., 2012), there termed the resultative complex predicate with noun. There, PREDLINK represents what is predicated of a certain

[7]We note that there is nothing tying us to this particular semantics. Neo-Davidsonian semantics would work equally well, or the DRT or intensional semantics mentioned earlier.

@INCEPOPER1(L)=
(↑PRED)='IncepOper1⟨(↑SUBJ),(↑PREDLINK)⟩'
(↑PREDLINK PRED)$=_c$'L'
$\lambda X.\text{start}(X) : (\uparrow \text{PREDLINK})_\sigma \multimap \uparrow_\sigma$

$$
{}_{1}\!\left[\begin{array}{ll}
\text{PRED} & \text{`}\mathtt{IncepOper}_1\langle \textit{2}\text{:artwe}, \textit{3}\text{:mwerre}\rangle\text{'} \\
\text{TENSE} & \text{pres} \\
\text{SUBJ} & {}_{2}\!\left[\begin{array}{ll} \text{PRED} & \text{`artwe'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \end{array}\right] \\
\text{PREDLINK} & {}_{3}\!\left[\begin{array}{ll} \text{PRED} & \text{`mwerre'} \end{array}\right]
\end{array}\right]
\qquad \text{start(well(man))}
$$

[artwe] man : $\uparrow_\sigma$
[mwerre] $\lambda X.\text{well}(X) : ((\text{PREDLINK}\uparrow)\ \text{SUBJ})_\sigma \multimap \uparrow_\sigma$
[irreme] $\lambda X.\text{start}(X) : (\uparrow\text{PREDLINK})_\sigma \multimap \uparrow_\sigma$

Figure 5: Analysis for (10): lexical function template for INCEPOPER1 (top left), f-structure for IV (centre left), semantic representation for IV (centre right), lexical entries with glue semantics (lower left)

entity (in their case and ours, the syntactic object; for us, what is predicated is that it has been spat out). Ahmed et al. (2012) comment that the resultative nature of the construction is not overt in the f-structure, and must be inferred from the existence of PREDLINK. This is somewhat different from the canonical use of PREDLINK, and for us does not seem applicable to (4a).

An alternative could be COMP, assuming as does Lødrup (2012) that nominal COMPs are possible. This analysis is perhaps more controversial, and we do not have space to present a full justification of it here, but we note that the essential point is that what we are looking for is a fairly neutral grammatical function whose role is just to give the base a place in the f-structure. This is indeed how Lødrup (2012) describes nominal COMP: "The intuition behind the COMP function could be verbalized this way: COMP differs from the other complement functions by not having their properties; it is a complement that just 'is there', and does not take part in grammatical processes."[8] Our analysis is then as in Figure 6. COMP would

[8]Rachel Nordlinger (personal communication) has considered nominal COMP to be potentially applicable to Australian Aboriginal languages such as Wambaya, for instance in (11).

(11) *Ngawu ngu bungmanya mirra*
1sg.NOM 1sg.S-FUT old.woman be
I will live to be an old woman. (Nordlinger, 1998a)

She notes that NPs such as *bungmanya*, often called (subject) complements in the literature, are a kind of nominal which relates in some way to another nominal without it clearly being a modifica-

$$
{}_{\mathit{1}}\left[\begin{array}{ll}
\textsc{pred} & \text{‘}\texttt{Oper}_{12}\langle \mathit{2}\text{:arelhe, } \mathit{3}\text{:ampe, } \mathit{4}\text{:lthere}\rangle\text{’} \\
\textsc{tense} & \text{pres} \\
\textsc{subj} & {}_{\mathit{2}}\left[\begin{array}{ll} \textsc{pred} & \text{‘arelhe’} \\ \textsc{pers} & 3 \\ \textsc{num} & \text{sg} \end{array}\right] \\
\textsc{obj} & {}_{\mathit{3}}\left[\begin{array}{ll} \textsc{pred} & \text{‘ampe’} \\ \textsc{pers} & 3 \\ \textsc{num} & \text{sg} \end{array}\right] \\
\textsc{comp} & {}_{\mathit{4}}\left[\begin{array}{ll} \textsc{pred} & \text{‘lthere’} \end{array}\right]
\end{array}\right]
\qquad \text{pinch(woman, child)}
$$

[arelhe] woman : $\uparrow_\sigma$
[ampe] child : $\uparrow_\sigma$
[lthere] $\lambda X.\lambda Y.\text{pinch}(X, Y)$:
$((\textsc{comp}\uparrow)\ \textsc{subj})_\sigma \multimap [((\textsc{comp}\uparrow)\ \textsc{obj})_\sigma \multimap \uparrow_\sigma]$
[iweme] $\lambda X.X$: $(\uparrow\textsc{comp})_\sigma \multimap \uparrow_\sigma$

Figure 6: Analysis for (4a): f-structure for lexical compound (upper left), semantic representation (upper right), lexical entries with glue semantics (lower left)

also be applicable to the IV construction from earlier.

From these examples it can be seen that the fundamental indicator of complex predication in our approach is the occurrence of lexical functions as the PRED value, in conjunction with the result at s-structure: we thus also arguably diverge from the LFG-specific definition of complex predicate by Butt (1993), in that in our representation they do not constitute a single predicate at f-structure.[9]

6 Conclusion

In this paper we have looked at three kinds of complex predicate in the Australian Aboriginal language Eastern/Central Arrernte: the intransitive verbaliser construction, lexical compounds, and Associated Motion. Based on the criteria of Henderson (2002) for characterising a word in Arrernte, we have argued that the first two types of complex predicate are more naturally handled in an LFG grammar, and the last in the morphology. We have then shown how a mechanism for incorporating the lexical functions of Meaning-Text Theory into LFG via glue semantics, developed as part of a multilingual natural language generation system, extends

tional structure.

[9]Note that whereas our semantic structure in Figure 5 happened to have two separate elements at s-structure corresponding to the two parts of the complex predicate, because of choices made in the semantic representation, in Figure 6 the two parts of the complex predicate combine to give a single element at s-structure, as in more traditional approaches to complex predicate combination in LFG.

naturally to form a mechanism for formation of complex predicates of at least the first two types in the semantics. In terms of future work, a full treatment of Associated Motion remains to be carried out. A potential direction here is the approach detailed in recently published monograph of Mani and Pustejovsky (2012), which describes itself as "analyz[ing] the semantics of motion expressions in terms of the formalisms of qualitative spatial reasoning".

References

Ahmed, Tafseer, Butt, Miriam, Hautli, Annette and Sulger, Sebastian. 2012. A Reference Dependency Bank for Analyzing Complex Predicates. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Alsina, Alex. 1997. A Theory of Complex Predicates: Evidence from Causatives in Bantu and Romance. In Alex Alsina, Joan Bresnan and Peter Sells (eds.), *Complex Predicates*, pages 203–246, Stanford, CA, US: CSLI Publications.

Amberber, Mengistu, Baker, Brett and Harvey, Mark (eds.). 2010. *Complex Predicates*. Cambridge, UK: Cambridge University Press.

Andrews, Avery. 2007. Projections and Glue for Clause-Union Complex Predicates. In *Proceedings of the LFG07 Conference*, pages 44–65.

Andrews, Avery and Manning, Christopher. 1999. *Complex Predicates and Information Spreading in LFG*. Stanford, CA, US: CSLI Publications.

Bateman, John, Matthiessen, Christian, Nanri, Keizo and Zeng, Licheng. 1991. The re-use of linguistic resources across languages in multilingual generation components. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, volume 2, pages 966–971, Sydney.

Beesley, Kenneth R. and Karttunen, Lauri. 2003. *Finite State Morphology*. Stanford, CA, US: CSLI.

Bögel, Tina, Butt, Miriam, Hautli, Annette and Sulger, Sebastian. 2007. Developing a Finite-State Morphological Analyzer for Urdu and Hindi. In T. Hanneforth and K. M. Würzner (eds.), *Finite-State Methods and Natural Language Processing, Revised Papers of the Sixth International Workshop on Finite-State Methods and Natural Language Processing*, pages 86–96, Potsdam, The Netherlands: Potsdam University.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford, UK: Blackwell.

Broadwell, George Aaron. 2000. Choctaw directionals and the syntax of complex predication. In Miriam Butt and Tracy King (eds.), *Argument Realization*, pages 111–134, Stanford, CA, US: CSLI Publications.

Butt, Miriam. 1993. *The Structure of Complex Predicates in Urdu*. Ph. D.thesis, Stanford University.

Butt, Miriam. 1995. *The Structure of Complex Predicates in Urdu*. Stanford, CA, US: CSLI Publications.

Butt, Miriam. 2010. The Light Verb Jungle: Still Hacking Away. In Mengistu Amberber, Brett Baker and Mark Harvey (eds.), *Complex Predicates*, pages 48–78, Cambridge, UK: Cambridge University Press.

Butt, Miriam and King, Tracy Holloway. 2006. Restriction for Morphological Valency Alternations: The Urdu Causative. In Miriam Butt, Mary Dalrymple and Tracy Holloway King (eds.), *Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan*, pages 235–258, Stanford, CA, US: CSLI Publications.

Butt, Miriam, King, Tracy Holloway, Niño, Maria-Eugenia and Segond, Frédérique. 1999. *A Grammar Writer's Cookbook*. Stanford, CA, US: CSLI Publications.

Cahill, Lynn, Doran, Christie, Evans, Roger, Kibble, Rodger, Mellish, Chris, Paiva, Daniel, Reape, Mike, Scott, Donia and Tipper, Neil. 2000. Enabling resource sharing in language generation: an abstract reference architecture. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*, volume 42 of *Syntax and Semantics Series*. New York, NY, US: Academic Press.

Falk, Yehuda. 2001. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Stanford, CA, US: CSLI Publications.

Fillmore, Charles, Johnson, Christopher and Petruck, Miriam. 2003. Background to Framenet. *International Journal of Lexicography* 16(3), 235–250.

Henderson, John. 1998. *Topics in Eastern and Central Arrernte grammar*. Ph. D.thesis, University of Western Australia.

Henderson, John. 2002. The Word in Eastern/Central Arrernte. In R. M. W. Dixon and Alexandra Aikhenvald (eds.), *Word: A Cross-Linguistic Typology*, pages 100–124, Cambridge, UK: Cambridge University Press.

Henderson, John and Dobson, Veronica. 1994. *Eastern and Central Arrernte to English Dictionary*. Alice Springs, Australia: IAD Press.

Jackendoff, Ray. 1990. *Semantic Structures*. Cambridge, MA, US: MIT Press.

Kahane, Sylvain and Polguère, Alain. 2001. Formal foundation of lexical functions. In *Proceedings of ACL 2001*, Toulouse, France.

Kamp, Hans and Reyle, Uwe. 1993. *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42 of *Studies in Linguistics and Philosophy*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Koch, Harold. 1984. The Category of 'Associated Motion' in Kaytej. *Language in Central Australia* 1, 23–34.

Lareau, Francois, Dras, Mark, Börschinger, Benjamin and Dale, Robert. 2011. Collocations in Multilingual Natural Language Generation: Lexical Functions meet Lexical Functional Grammar. In *Proceedings of the Ninth Australasian Language Technology Workshop (ALTA'11)*, pages 95–104, Canberra, Australia.

Lødrup, Helge. 2012. Is there a 'third object'? Searching a nominal counterpart to COMP. In *Proceedings of the LFG12 Conference*.

Mani, Inderjeet and Pustejovsky, James. 2012. *Interpreting Motion: Grounded Rep-*

resentations for Spatial Language. Oxford, UK: Oxford University Press.
Mel'cuk, Igor. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In Leo Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102, Amsterdam, The Netherlands: Benjamins.
Nordlinger, Rachel. 1998a. *A Grammar of Wambaya*. Canberra, Australia: Pacific Linguistics.
Nordlinger, Rachel. 1998b. *Constructive Case: Evidence from Australian Languages*. Stanford, CA, US: CSLI Publications.
Nordlinger, Rachel. 2010. Complex predicates in Wambaya: detaching predicate composition from syntactic structure. In Mengistu Amberer, Brett Baker and Mark Harvey (eds.), *Complex Predicates*, pages 237–258, Cambridge, UK: Cambridge University Press.
Nordlinger, Rachel and Bresnan, Joan. 2011. Lexical-Functional Grammar: interactions between morphology and syntax. In Robert Borsley and Kersti Börjars (eds.), *Non-Transformational Syntax: Formal and Explicit Models of Grammar*, West Sussex, UK: Wiley-Blackwell.
Nordlinger, Rachel and Sadler, Louisa. 2007. Verbless clauses: revealing the structure within. In J. Grimshaw, J. Maling, C. Manning, J. Simpson and A. Zaenen (eds.), *Architectures, Rules and Preferences: A Festschrift for Joan Bresnan*, pages 139–160, Stanford, CA, USA: CSLI Publications.
Seiss, Melanie. 2009. On the Difference Between Auxiliaries, Serial Verbs and Light Verbs. In *Proceedings of the LFG09 Conference*, pages 501–519.
Seiss, Melanie. 2011. Implementing the Morphology-Syntax Interface: Challenges from Murrinh-Patha Verbs. In *Proceedings of the LFG11 Conference*, pages 482–502.
Simpson, Jane. 2007. Expressing pragmatic constraints on word order in Warlpiri. In Annie Zaenen, Jane Simpson, Chris Manning and Jane Grimshaw (eds.), *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*, Stanford, CA, US: CSLI Publications.
Strehlow, T. G. H. 1944. *Aranda phonetics and grammar*. Sydney, Australia: Oceania Monographs.
Sulger, Sebastian. 2009. Irish clefting and information-structure. In *Proceedings of the LFG09 Conference*, pages 562–582.
Wanner, Leo, Bohnet, Bernd, Bouayad-Agha, Nadjet, Lareau, François and Nicklaß, Daniel. 2010. MARQUIS: Generation of user-tailored multilingual air quality bulletins. *Applied Artificial Intelligence* 24(10), 914–952.
Wilkins, David. 1989. *Mparntwe Arrernte (Aranda): studies in the structure and semantics of grammar*. Ph. D.thesis, Australian National University.
Wilkins, David. 2006. Towards an Arrernte grammar of space. In Stephen Levinson and David Wilkins (eds.), *Grammars of space: explorations in cognitive diversity*, pages 24–62, Cambridge, UK: Cambdridge University Press.
Wilson, Stephen. 1999. *Coverbs and complex predicates in Wagiman*. Stanford, CA, US: CSLI Publications.

NON-SUBJECT PARTICIPANTS IN TOLAKI

Owen Edwards
Australian National University

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

I present data from Tolaki, an Austronesian language of Central Indonesia, which challenges the notion that grammatical functions form discrete categories. I argue that current models of grammatical functions within Lexical Functional Grammar cannot account for the data we find. If we were to posit discrete categories for grammatical functions on the basis of different behaviour under different morpho-syntactic tests, we would be forced to posit a minimum of nine categories in order to account for the results; nearly double the number of categories currently provided for by LFG. A better way of analysing the data we find in Tolaki is to posit a continuum of grammatical functions between the most and least privileged grammatical functions, subject and adjunct. Participants are located along this continuum and are either more subject-like or more adjunct-like.

1 Introduction

In this paper I will investigate grammatical functions (GFs) in the Austronesian language Tolaki, with a particular emphasis on non-subject participants.

LFG posits a universally available inventory of eight GFs (Dalrymple, 2001; Bresnan, 2001). In this paper I only discuss nominal GFs, and will not be further concerned with the clausal GFs COMP, XCOMP or XADJ. The five nominal GFs are:

(1) SUBJ, OBJ, OBJ_θ, OBL_θ, ADJCT

In this paper I argue that if we were to posit discrete categories for Tolaki GFs we would be forced to posit a minimum of nine categories in order to account for the results; nearly double the number of categories allowed by LFG. I propose that a better way of analysing the data is to posit a continuum of GFs between the most and least privileged GFs, SUBJ and ADJCT.

In the first section of my paper I provide evidence for the most privileged GF, SUBJ, and the least privileged GF, ADJCT. We will see that the three tests, relativi-

[†]Unless otherwise cited, data is drawn from my own fieldwork conducted at the beginning of 2012. I would like to thank my main informants, Darmin, Untung, Sukur Tabara and Omar Pidani.

Glosses follow the list of standard abbreviations that can be found in the Leipzig Glossing Rules, with the following exceptions: ACCID ‘accidental passive, CN ‘common noun’, eSi ‘older sibling’, EX ‘exclusive’, IN ‘inclusive’, NFIN ‘non-finite’, NSG ‘non-singular’, PN ‘proper noun’ and ySi ‘younger sibling’.

Example sentences are given in standard Tolaki orthography, with dashes (-) added to indicate morpheme breaks. Tolaki letters have the same values as Indonesian equivalents, with the exception of the apostrophe <’> which represents the glottal stop /ʔ/. Full sentences receive appropriate capitalisation and punctuation, with the exception of ungrammatical sentences. A capital ‘N’ (i.e. *poN-*) indicates a morpheme after which the morpho-phonemic process of prenasalisation occurs. Under this process the voiceless stops /p, t, k/ become the prenasalised stops /mb, nd, ŋg/. Two lines of Tolaki are given in sentences when this morphophonemic process operates. The top line shows the standard orthography, the second line the morpheme breaks.

sation, nominative agreement and plural agreement allow us to identify SUBJ. We will also identify ADJCT as a participant that cannot appear as a bare noun phrase.

The bulk of my paper is an investigation of the behaviour of the remaining non-subject participants; participants which cannot be identified as either SUBJ or ADJCT. After identifying the coding strategies used for eight such participants, I will show their behaviour under the syntactic tests of external possession, secondary predication and passivisation.

I conclude my paper with a proposal for explaining the continuum-like nature of grammatical functions in Tolaki.

2 The Limits: Subject and Adjunct

I begin my investigation by defining the limits among GFs; the most privileged GF, SUBJ, and the least privileged GF, ADJCT.

2.1 Subject

One syntactic test and two morphological tests allow us to reliably identify the SUBJ in Tolaki. The tests of relativisation, plural agreement and indexation with nominative prefixes refer uniquely to the group of roles comprised of S, A and the derived S in a passive sentence.

2.1.1 Relativisation

The criteria by which we can identify a relative clause in Tolaki are: (a.) the relativised noun phrase occurs before the verb, (b.) the verb is in the non-finite form[1], and (c.) no affixes occur on the verb indexing the relativised argument.

Examples (2)-(4) show the successful relativisation of an S, A and derived S respectively. In these examples the relative clause is enclosed within square brackets.

(2) *Ingoni laa* [$_{\text{NP}}$ *toono* [$_{\text{RC}}$ *i-luara* *⟨m⟩o-susua.*]]
earlier EXIST person LOC-outside ⟨NFIN⟩INDF.P-sing.
There was someone earlier today who sang [something] outside.

(3) *Ku-kokolea-'i* [$_{\text{NP}}$ *hai-nggu* [$_{\text{RC}}$ *t⟨um⟩idu-'aku.*]]
1NOM-annoy-3ABS ySi-1GEN ⟨NFIN⟩punch-1ABS
I annoy my younger brother who punched me.

[1] The non-finite form is coded with the infix *⟨um⟩*. When the indefinite P prefix *poN-* occurs, the non-finite form is encoded by the process /p/ → /m/; i.e. *moN-*. Passive verbs do not have a non-finite form. For a fuller discussion of the form and use of the non-finite verbal form in Tolaki see Edwards (2012, 56-9) and Mead (1998, 291-4).

(4) *Ku-tidu-'i* [$_{\text{NP}}$ *hai-nggu* [$_{\text{RC}}$ *k⟨in⟩okolea-nggu.*]]
1NOM-punch-3ABS ySi-1GEN ⟨PASS⟩annoy-1GEN
I punched my younger brother whom I annoy.
lit. I punched my younger brother who is annoyed by me.

Non-subjects cannot be relativised in Tolaki. This is shown in (5), which would otherwise fulfil the criteria for the successful relativisation of a P.

(5) * *Ku-tidu-'i* [$_{\text{NP}}$ *hai-nggu* [$_{\text{RC}}$ *ku-k⟨um⟩okolea.*]]
1NOM-punch-3ABS ySi-1GEN 1NOM-⟨NFIN⟩annoy
I punched my younger brother whom I annoy.

2.1.2 Plural Agreement

When the SUBJ of a clause is plural, the verb can optionally take the plural prefix *mbeN-*. Plural in Tolaki consists of a group of three or more. Verbal indexation of plural participants is necessarily non-singular and the participant can take the optional NSG suffix *-Cako*.

An example of the verb agreeing with each of a plural S, A and derived S, is given in sentences (6)-(8) respectively.

(6) *Lako-ro-to* *mbe-lako* *hada* *dadio.*
go-3NSG.GEN-PRF PL-go monkey many
Then the many monkeys left. (Untung, 2009, 31)

(7) *Rombenggii'ito* *kolopua.*
ro-mbeN-kii-'i-to kolopua
3NSG.NOM-PL-see-3ABS-PRF tortoise
They [the monkeys] saw the tortoise. (Untung, 2009, 32)

(8) *Rombinendopaki* *poteha'akonggu.*
ro-mb⟨in⟩eN-topaki poteha-'ako-nggu
3NSG.NOM-⟨PASS⟩PL-slap cousin-NSG-1GEN
My cousins were slapped.

Finally, the ungrammatical sentence (9), in which the only non-singular participant is the P, shows that the prefix can only agree with the SUBJ.

(9) * *ku-mbe-langgu-'iro* *banggona-hako-nggu*
1NOM-PL-hit-3NSG.ABS friend-NSG-1GEN
I hit my friends.

2.1.3 Nominative Agreement

Only an S, A or derived S can be indexed with nominative prefixes. Examples (10), (11) and (12) illustrate the indexation of each of these roles with a nominative prefix

(10) *Ku-lako i'-aa-homa.*
1NOM-go LOC-area-forest
I went to the forest.

(11) *Ku-soro-'i oto-nggu.*
1NOM-push-3ABS car-1GEN
I pushed my car.

(12) *I-Bio no-k⟨in⟩ii.*
PN-Bio 3NOM-⟨PASS⟩see.
Bio [is the one who] was seen.

Absolutive and genitive suffixes are also used to index the SUBJ under certain circumstances[2]. Thus, while nominative prefix indexation is not the only coding strategy that can be used to index the SUBJ, SUBJs are the only participants that can be coded in this way.

2.2 Adjuncts

SUBJ forms the upper limit among grammatical fuctions; the most privileged, while ADJCT forms the lower limit, the least privileged GF.

Adjuncts in Tolaki fail all of the tests listed listed above for subjects, in addition to all other syntactic tests discussed in this paper. Furthermore, adjuncts cannot appear as a bare noun phrase and must be marked with either of the prefixes *i-* 'locative' or *kei/ine-* 'adjunct'[3], and/or are introduced by one of the prepositions *ari* 'from', *ronga* 'with' and *sambe* 'until'.

Sentence (15) shows a locative adjunct marked with *i-*, sentence (13) an adjunct marked with both the preposition *ari* and the prefix *ine-* and sentence (14) shows an adjunct marked with the preposition *ronga* 'with'.

(13) *A-no te-bua pele-hada* **ari** **ine**-*kowuna.*
and-3NOM ACCID-fall palm-monkey from ADJCT.CN-bamboo
And a monkey's hand fell out of the bamboo. (Untung, 2009, 31)

(14) *Ku-laa ⟨m⟩e-tulura* **ronga** *hai-nggu.*
1NOM-PROG ⟨NFIN⟩INTR-speak with ySi-1GEN
I'm speaking with my younger sibling.

Furthermore these participants can be multiply specified, a test for adjuncthood (Dalrymple, 2001, 12). This is shown in sentence (15).

(15) *Ki-laa mbe-lako i'-aa-homa i-kambo*
1EX.NOM-PROG PL-go LOC-area-forest LOC-village
⟨m⟩e-tamo-'ako Okonda.
⟨NFIN⟩INTR-name-APPL *Okonda*.
We were walking in a forest in a village which was called *Okonda*.

[2]See Edwards (2012, 46-56) and Mead (1998, 300-343) for a full discussion of the circumstances under which different affixes are used.

[3]*kei* is used with pronouns and proper nouns, *ine* is used with all other nouns. These prefixes introduce roles with a wide variety of semantic roles including (but not limited to) GOAL, SOURCE, LOCATION and ACCOMPANIMENT.

We can thus define the limits of GFs in Tolaki: the least and most privileged GFs. This is shown in table 1

	Rel.	Pl.	NOM	Bare NP
SUBJ	✓	✓	✓	✓
ADJCT	-	-	-	-

Table 1: The Limits

3 The Middle: Objects and Obliques

In this section I will investigate GFs which fall between the two extremes of SUBJ and ADJCT. I begin by discussing the way in which these participants are coded. I will discuss a total of eight participants.

3.1 The Coding of Non-subject Participants.

3.1.1 Definite P and Idenfinite P

The first two non-subject participants I will investigate are the 'Definite P' and the 'Indefinite P'.

Definite P's and Indefinite P's are in complementary distribution with one another. Definite P's are indexed on the verb with absolutive suffixes, as in sentence (16), while indefinite P's are unindexed and co-occur with the INDF.P prefix *poN-*, as in sentence (17).

(16) *Ku-soro-***'i** **oto-nggu**.
1NOM-push-3ABS car-1GEN
I pushed my car.

(17) *Ano po'alo o'aso boto, ano ponggaa*
a-no **po**-alo **o'aso boto** a-no poN-kaa
and-3NOM INDF.P-take one CLF and-3NOM INDF.P-eat
And he takes a single one [banana] and eats [it]. (Untung, 2009, 30)

That this is always the pattern, is shown in circumstances in which the P is inherently definite, such as with a pronominal P. Thus sentence (18) with a pronominal referent indexed with the absolutive suffix is grammatical, while the equivalent sentence (19) with the prefix *poN-* and an unindexed P is ungrammatical.

(18) *Ku-langgu-ko.*
1NOM-hit-2ABS
I hit you.

(19) * *ku-po-langgu inggo'o*
1NOM-INDF.P-hit 2SG
I hit you.

Furthermore, an unindexed P does not usually occur with demonstratives or possessive suffixes, as these usually indicate definite referents. Thus, when talking about a grub which I had photographed, my informants found (20) with an absolutive P acceptable, while sentence (21) was judged strange.

(20) *Laa-nggu k⟨um⟩ii-kii-'i inono uwato ...*
PROG-1GEN ⟨NFIN⟩REDUP-see-3ABS this grub
While I was looking at this grub ...

(21) ? *Laanggu monggii-kii inono uwato ...*
laa-nggu ⟨m⟩oN-kii-kii inono uwato
PROG-1GEN ⟨NFIN⟩INDF.P-REDUP-see this grub
While I was looking at this grub ...

When unindexed P's do occur with a demonstrative or possessor, they indicate that the P is an uncertain member of a group. Thus, when asked about sentence (22) with an unindexed, but possessed, P one informant explained that "we can't know yet who is hit".

(22) *No-po-langgu hai-nggu.*
3NOM-INDF.P-hit ySi-1GEN
He hit one of my younger siblings.

Thus, absolutive indexed P's are definite, while unindexed P's are indefinite, even when there is no other indication of this in the clause.

We can represent the mapping of the argument structure to morphological categories of sentence (16) in (23), and that of (17) in (24).

(23) **Definite P:**
'PRED ⟨ __, __ ⟩'
NOM $\underline{\text{ABS}}_{\text{definite}}$

(24) **Indefinite P:**
'*poN-* PRED ⟨ __, __ ⟩'
NOM $\underline{\text{Ø}}_{\text{indefinite}}$

3.1.2 Dative P and Applicative P

Another two participants which are in complementary distribution with one another are the 'Dative P' and the 'Applicative P'. While the usual strategy for indexing a P with definite reference is with absolutive suffixes, a small subset of verbs in Tolaki indexes such P's with a dative suffix. A simple example is given in sentence (25)

(25) *Ano tealonggee kolopua.*
a-no tealoN-**kee** kolopua
and-3NOM fetch-3DAT tortoise
And he fetched the tortoise. (Untung, 2009, 28)

Such verbs have a less affected P with semantic roles such as THEME OR STIMULUS, rather than PATIENT. Nonetheless, among those verbs which have non-PATIENT P's, it must be stipulated at the lexical level which take a Dative P. Thus, for instance, while both *to'ori* 'know' and *kolupe* 'forget' both have a P with the semantic role STIMULUS, *to'ori* indexes P's with a dative suffix while *kolupe* indexes them absolutively. A sample of verbs which take a dative P include: *watu* 'join with, go along with', *to'ori* 'know', *te'eni*[4] 'say, tell' and *teposua* 'meet'.

Partially, pre-empting my conclusions in section 4, I propose that the difference in coding (and other morpho-syntactic behaviour) between predicates which take a Dative P and those that take an absolutive Definite P arises from each type of predicate having a different argument structure. While predicates with a Dative P are monovalent, predicates with an absolutive Definite P are bivalent.

Concerning verbs which take a Dative P, there is thus as a mismatch between the lexico-conceptual structure, which contains two semantic roles, and the argument structure of the relevant verb, which subcategorises for only one.

Because the argument structure of such verbs only contains one argument, when the second semantic role in the lexico-conceptual structure is included, it is encoded in the same way as a participant external to the argument structure, such as a beneficiary (see section 3.1.3 below). The mapping of participants between the lexico-conceptual structure and argument structure of the verb *to'ori* is shown in (26), with the lexico-conceptual structure on top and the argument structure below.

(26)
to'ori (EXP , STIM)
| |
'PRED⟨ __ ⟩' __

Historically, the aberrent indexation pattern of many of these verbs can be explained as resulting from the accidental passive prefix *te-* becoming fossilised onto the verb. The argument structure of these verbs originally contained two arguments, however, with the fossilisation of this prefix their historic transitivity was lost.

Explanations for forms without initial /t/ (such as *watu*) are not so clear. One likely source is that at some point in the history of the language, the applicative suffix *-Cako* become an obligatory part of the verb stem, to which the absolutive suffixes later fused.

Historically, the dative suffix arose through a combination of applicative + absolutive suffix (Mead, 1998, 207-12).[5] It would appear then, that these verbs were originally monovalent and that the 'extra' participant was originally included through applicativisation with regular absolutive agreement for definite participants.

[4]In the case of *te'eni*, absolutive suffixes can be optionally used to index the message, that is, what was said. Dative suffixes index the addressee.

[5]A *synchronic* analysis of the dative suffixes as applicative + absolutive is problematic, given verbs with two dative suffixes such as those in sentences (33) and (46).

We therefore expect that when this participant is indefinite, it will be unindexed and the applicative suffix *-Cako* will appear on the verb. This is indeed what we find. An example is given in (27).

(27) *Nopondeposuangako kadue.*
no-poN-teposua-**ngako** kadue
3NOM-INDF.P-meet-APPL dwarf.buffalo
He met/came across some dwarf buffalo.

Synchronically, this is best analysed as a method of including the participant which exists at the lexico-conceptual structure but is absent from the argument structure of the base verb, without agreeing with it.

The argument structure of these verbs, along with the mapping of participants to morphological categories, is shown for Dative P's in (28) and Applicative P's in (29).

Dative P:

(28) 'PRED ⟨__, __⟩'
NOM <u>DAT</u>$_{\text{definite}}$

Applicative P:

(29) 'APPL⟨__, __⟩ 'PRED⟨__⟩'' __
NOM <u>Ø</u>$_{\text{indefinite}}$

3.1.3 Other Participants

Beneficiary A participant with the semantic role BENEFICIARY can be indexed on the verb with dative suffixes. An example of such a beneficiary with a monovalent verb is given in (30) and an example with a bivalent verb in (31).

(30) *No-laa ⟨m⟩e'-indio-***kee** *ama-nggu.*
3NOM-PROG ⟨NFIN⟩INTR-work-3DAT father-1GEN
He's working for my father.

(31) *Ku-tidu-'i-***ko'o**.
1NOM-punch-3ABS-2DAT.
I'll punch him for you./I'll get him (back) for you.

The mapping of participants to morphological categories for sentence (30) is shown in (32) below.

(32) **Beneficiary:**
'PRED ⟨ __ ⟩' BEN
NOM <u>DAT</u>

Note that when both a BENEFICIARY *and* a Dative P are found, we find two dative suffixes on the verb:

(33) *Okino ehe teposuanggeekona nggiro'o otina.*
oki-no ehe teposuaN-**kee-kona** nggiro'o o-tina
NEG-3NOM like meet-3DAT-1DAT that CN-woman
He doesn't want to meet that woman for me.

The Transitive Instrument and Theme One strategy for including an instrument in Tolaki is to index it with an absolutive suffix. When this occurs, the PATIENT/THEME is indexed with a dative suffix. An example is given in (34)

(34) *No-langgu-'i-kona o-kasu.*
3NOM-hit-3ABS-1DAT CN-wood
He hit me with a piece of wood.
lit. He hit a piece of wood to/at me.

I will refer to the instrument of such constructions as the 'Transitive Instrument' and the P as the 'Transitive Instrument Theme'. An equivalent meaning can also be expressed with various periphrastic constructions, as in (35), as well as with the applicative *-Cako* or with a prepositional phrase headed by *ronga* 'with'.

(35) *No-langgu-'aku mombake o-kasu.*
3NOM-hit-1ABS using CN-wood
He hit me using a piece of wood.

When both the PATIENT and INSTRUMENT are realised by a full NP, the PATIENT usually precedes the INSTRUMENT in word order. This is illustrated in sentence (36) below, which would be interpreted as indicated unless uttered in a context that would force the asterisked reading.

(36) *No-langgu-'i-***kee** **o-watu** *o-kasu.*
3NOM-hit-3ABS-3DAT CN-stone CN-wood
He hit the stone with a piece of wood.
*He hit the wood with a stone.

A non-3rd person instrument cannot be used in the Transitive Instrument construction, as shown by the ungrammatical (37a). In such circumstances a periphrastic construction, such as in sentence (37b), or the Intransitive Instrument construction (see section 3.1.3 below) must be employed instead. The pragmatic situation motivating the sentences in (37) is a dream in which an anthropomorphic piece of wood with the power of speech questions the motives behind the actions of the dreamer.

(37) a. * *mbaako'i u-langgu-'aku-kee toono*
why 3NOM-hit-2NOM-1DAT person

b. *Mbaako'i u-langgu-'i toono mombake inaku?*
why 2NOM-hit-3ABS person using 1SG
Why did you hit the person (by) using me?

The mapping of the Transitive Instrument and Theme to morphological categories is given in (38).

(38) **Transitive Instrument and Theme:**

‘PRED ⟨ __, __ ⟩’ INST
| | |
NOM DAT ABS

The Intransitive Instrument The final participant I will discuss I call the ‘Intransitive Instrument’. Another strategy for including an instrument is to index it with a dative suffix. When this occurs, the verb takes the intransitive prefix *pe-* and the THEME/PATIENT is introduced with the *kei/ine-* prefix (see section 2.2). An example is given in sentence (39) below.

(39) *Nopedondonggee okasu ine banggonano.*
no-pe-dondoN-kee o-kasu ine-banggona-no
3NOM-INTR-hit-3DAT CN-wood ADJCT.CN-friend-3GEN
He hit his friend with a piece of wood.

Note that such constructions appear marginal and not all speakers allow them. A different informant from the one who provided sentence (39) found the analogous sentence (40) to be unacceptable, though he still said it could be understood.

(40) ? *Kupehotonggee opade ine banggonanggu.*
ku-pe-hotoN-**kee** **o-pade** ine-banggona-nggu
3NOM-INTR-cut-3DAT CN-machete ADJCT.CN-friend-1GEN
I cut my friend with a machete.

The mapping of the participants to morphological categories in this Intransitive Instrument construction is shown in (41).

(41) **Intransitive Instrument:**

‘INTR⟨__⟩‘PRED ⟨__, __ ⟩” INST
| | |
NOM ADJCT DAT

3.2 Coding Summary

We have thus identified a total of nine non-subject participants in Tolaki, including adjuncts. The coding of each of these participants is listed in table 2

While this is not an exhaustive list of all non-subject participants found in Tolaki, it does provide a representative sample and includes all dative non-subject participants known to the author.

	Indexation					
	ABS	DAT	APPL	INDF.P	INTR	Bare NP
1. Definite P	✓	-	-	-	-	✓
2. Indefinite P	-	-	-	✓	-	✓
3. Dative P	-	✓	-	-	-	✓
4. Applicative P	-	-	✓	✓	-	✓
5. Trans Inst	✓	-	-	-	-	✓
6. Trans Inst Thm	-	✓	-	-	-	✓
7. Intrans Inst	-	✓	-	-	✓	✓
8. Beneficiary	-	✓	-	-	-	✓
9. Adjunct	-	-	-	-	-	-

Table 2: Coding of Tolaki Non-Subject Participants

3.3 The Syntactic Behaviour of Non-subject Participants.

In this section I will investigate the behaviour of these non-subject participants under different syntactic tests. We will find that while each morpho-syntactic test is sensitive to a restricted set of non-subject participants, no test is sensitive to exactly the same set of participants as another test. Calling the most privileged non-subject participant OBJ and the least privileged ADJCT, we can observe a continuum-like scale of non-subject participants in which some participants have more behaviour in common with OBJ and some have more behaviour in common with ADJCT.

I will discuss three syntactic tests that have been found to consistently discriminate among non-subject participants in Tolaki. These tests are external possession, secondary predication and passivisation. The results of each of these tests will be discussed in turn.

For each of these tests, only a subset of the data will be presented in the following sections. The remainder of the data showing the behaviour of each non-subject under each test can be found in Edwards (2012, 94-9).

3.3.1 External Possession

External possession[6] is a test in which the possessor of a participant is indexed on the verb with dative suffixes under certain semantic and pragmatic conditions. An example is given in sentence (42).

(42) *No-langgu-'i-***kona** *hai-***nggu**
3NOM-hit-3ABS-1DAT ySi-1GEN
He hit my younger brother.

[6]What is here called 'external possession' can be correlated with what is also termed 'possessor raising' or 'possessor ascension' in the literature. For a general discussion of this phenomenon see Payne and Barshi (1999). For a specifically LFG treatment of such a phenomenon see Lødrup (2009).

In such sentences the possessor is indexed twice in the sentence, once with a dative suffix on the verb and once with a genitive suffix in the possessed NP. While sentence (42) is also grammatical without a dative suffix indexing the possessor, external possession is in general preferred.[7]

In situations where the action performed is beneficial for the possessor of the P, it is unclear whether the dative is a simple benefactive or whether it is due to external possession. An example is sentence (43) below, from Mead (1998, 238).

(43) *Oheo, pe'eka kabusa-'i-***keito** *ana-***ndo** *tewuta-'i-to.*
Oheo, ascend clean-3ABS-1IN.DAT child-1IN.GEN defecate-3ABS-PRF
Oheo, come up and clean our child (for us), he's become dirty.

This sentence is ambiguous between the external possession structure shown in (44b), in which the possessor of the NP is encoded twice, once with a genitive suffix on the NP and once with a dative suffix on the verb, and the structure in (44a) in which the dative encodes a beneficiary which happens to be coreferential with the possessor of the NP.

(44) a. CLEAN (SUBJ, OBJ:[(POSS) THM]) BEN
imp GEN$_i$ ABS DAT$_i$

b. CLEAN (SUBJ, OBJ:[(POSS) THM])
imp DAT ABS
GEN

While sentence (42) is also similarly structurally ambiguous, the beneficiary reading in (44a) is pragmatically highly unlikely, and speakers can distinguish between the two meanings.

When we turn to those participants which are eligible to be externally possessed, we find that all non-subject participants except for an Indefinite P, an Applicative P and an Adjunct are eligible to have their possessor indexed on the verb with dative suffixes.

While Definite P's can be externally possessed, as in sentence (42), Indefinite P's cannot. This is shown in sentence (45) which is pragmatically odd as the dative can only be interpreted as a beneficiary, as in the structure in (44a).

(45) ? *Nopolanggunggona hainggu.*
no-po-langguN-kona hai-nggu
3NOM-INDF.P-hit-1DAT ySi-1GEN
He hit one of my younger siblings for me.

Finally, the data for beneficiaries on external possession, in sentences such as (46), is inherently ambiguous between the structures shown in (44).

[7] When asked to explain the difference between sentence (42) and the equivalent without external possession, one native speaker felt as though the sentence without external possession described an accident, while the one with external possession was a deliberate attempt to harm.

(46) *Ku-po-wai-keero-ko'o banggona-mu o-tee.*
1NOM-INDF.P-make-3NSG.DAT-2DAT friend-2GEN CN-tea
I made tea for your friends [for you].

3.3.2 Secondary Predication

Secondary predication in Tolaki was tested using the depictive secondary predicate *molangu* 'drunk'.

The adjective *molangu* 'drunk' can be included in a sentence in several ways. The first is in a separate verb phrase headed by the auxiliary *laa*. When this is the case nominative prefixes can optionally occur indexing the SUBJ:

(47) *Kuteposuanggee banggonanggu, (no)laa molangu.*
ku-teposuaN-kee banggona-nggu (no)-laa molangu
1NOM-meet-3DAT friend-1GEN (3NOM)-PROG drunk
I met my friend, he was drunk.

Secondly, the adjective can be included internally in the NP it modifies, either before the noun it modifies, as in sentence (48), or after the noun it modifies as in sentence (49). In the case of sentence (49) the adjective can be shown to be internal to the noun phrase it modifies because the genitive affix occurs after it.

(48) *Ihawi ku-kii-'i [molangu banggona-nggu. $_{NP}$]*
yesterday 1NOM-see-3ABS drunk friend-1GEN
Yesterday I saw my drunk friend.

(49) *Ihawi ku-langgu-'i [banggona molangu-nggu. $_{NP}$]*
yesterday 1NOM-hit-3ABS friend drunk-1GEN
Yesterday I hit my drunk friend

Finally, the adjective can occur clause finally, but external to the noun phrase it modifies. A simple example is shown by sentence (50). In this sentence the boundary of the noun phrase is indicated by the position of the genitive affix.

(50) *No-leu [banggona-nggu$_{i}$ $_{NP}$] molangu$_{i}$*
3NOM-come friend-1GEN drunk
My friend arrived drunk.

When the secondary predicate occurs external to the noun phrase it modifies, only certain participants are eligible to launch it. The participants which cannot launch it are a Beneficiary, Transitive Instrument, Theme and Adjunct.

Sentence (51), shows that an Adjunct cannot launch a secondary predicate. When presented to informants, this sentence was accompanied by laughter, as the only grammatical interpretation is one in which the dative instrument 'water' launches the secondary predicate; rather than the pragmatically more likely prepositional adjunct.

(51) *Kupebahongee* *iwoi* *kei i Bio* *molangu.*
ku-pe-bahoN-kee $iwoi_i$ $kei\text{-}i\text{-}Bio_j$ $molangu_{i,*j}$
1NOM-INTR-wash-3DAT water ADJCT.PN-PN-Bio drunk
I washed Bio with drunk water.
*I washed Bio with water [while he was] drunk.

3.3.3 Passivisation

The final syntactic test I will discuss is passivisation. In section 2.1 we saw that in Tolaki the A/S role maps onto the GF SUBJ. However, when a verb is passivised, the P is assigned the GF SUBJ.

The passive in Tolaki is marked by the infix *⟨in⟩* for stems beginning with a voiceless consonant and the prefix *ni-* for stems beginning with a voiced consonant. Either form can be used for vowel initial stems. When the passive occurs in a relative clause, the A can be indexed with genitive suffixes, otherwise the A can be included as an Adjunct with the prefix *kei/ine-*.

Only a Definite P, an Indefinite P or an Applicative P are eligible to be the input of a passive. A typical example is given in sentence (52) below, in which the passive has been used in order to allow the PATIENT of the verb to be relativised.

(52) *hai-nggu* *ni-langgu-mu*
ySi-1GEN PASS-hit-2GEN
my younger sister who you hit

When an Indefinite P is passivised, the verb does not retain the INDF.P prefix. However, we find that indefinite participants can still be the input to the passive, as in sentence (53).

(53) *Ohawo laa* *ni-lolaha-mu?*
what PROG PASS-search-2GEN
What are you looking for?

In order to passivise the Dative P of a verb like *to'ori*, the participant must be first applicativised. Thus, sentence (54) is ungrammatical, while sentence (55) with an applicative suffix is grammatical.

(54) * *kaaka-mu t⟨in⟩o'ori-nggu*
eSi-2GEN ⟨PASS⟩know-1GEN

(55) *kaaka-mu t⟨in⟩o'ori-***'ako***-nggu*
eSi-2GEN ⟨PASS⟩know-APPL-1GEN
Your older sister who I know

Transitive Instruments, which are indexed in the same way as definite P's cannot, however, be an input for the passive. Thus, sentence (57), the passive version of (56), is ungrammatical.

(56) *No-langgu-'i-kona o-kasu.*
3NOM-hit-3ABS-1DAT CN-wood
He hit me with a [piece of] wood.

(57) * *o-kasu ni-langgu-kona*
CN-wood PASS-hit-1DAT

However, this is not a restriction blocking participants with the semantic role INSTRUMENT from undergoing passivisation, as instruments included in a sentence with the applicative suffix are an eligible input to the passive, as in sentence (58).

(58) *o-kasu ni-langgu-***'ako** *kei-inaku*
CN-wood PASS-hit-APPL ADJCT.PN-1SG
the piece of wood with which I was hit

3.4 Summary of Results

A summary of the results of each syntactic test can be found in Table 3 over the page. This table also summarises the morphology used to code each participant. Subjects are also included in this table. A tick (✓) indicates that a participant 'passes' the test, a dash (-) indicates that it does not while neither ( ) indicates that the data is currently lacking (for the Transitive Instrument under External Possession), or inherently ambiguous (for the Beneficiary under External Possession).

The results of those tests that distinguish between non-subject participants are presented in Figure 1. A score of 0.1 indicates the participant fails the test, 0.5 insufficient data and 1 that a participant passes the test. In Figure 1 the results for participants with identical syntactic behaviour are combined. Such participants include the Applicative P and Indefinite P, as well as the Dative P and Intransitive Instrument.

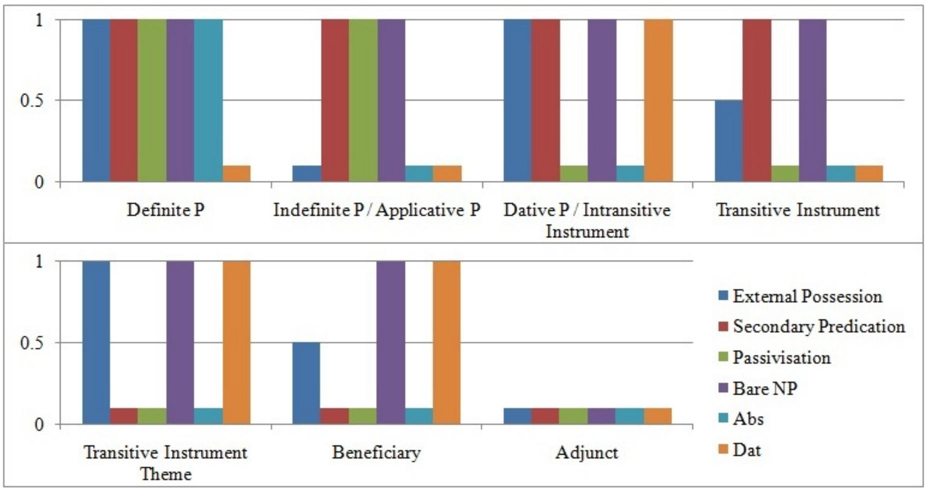


Figure 1: Non-Subject Participants

	Syntactic Criteria					Morphological Criteria						
			Ext	2nd		Agreement						Bare
	Rel	Pl.	Poss	Pred	PASS	NOM	ABS	DAT	APPL	INDF.P	INTR	NP
Subject S	✓	✓	-	✓	-	✓	✓	-	-	-	-	✓
Subject A	✓	✓	-	-	-	✓	✓	-	-	-	-	✓
Definite P	-	-	✓	✓	✓	-	✓	-	-	-	-	✓
Indefinite P	-	-	-	✓	✓	-	-	-	-	✓	-	✓
Dative P	-	-	✓	✓	-	-	-	✓	-	-	-	✓
Applicative P	-	-	-	✓	✓	-	-	-	✓	✓	-	✓
Trans Inst	-	-		✓	-	-	✓	-	-	-	-	✓
Trans Inst Thm	-	-	✓	-	-	-	-	✓	-	-	-	✓
Intrans Inst	-	-	✓	✓	-	-	-	✓	-	-	✓	✓
Beneficiary	-	-		-	-	-	-	✓	-	-	-	✓
Adjunct	-	-	-	-	-	-	-	-	-	-	-	-

Table 3: Morpho-Syntactic Tests for Grammatical Functions

4 Conclusions

These results show that if we were to posit discrete categories for Tolaki non-subject participants we would be forced to posit a minimum of six categories; two more than are provided for by current models of LFG.

However, even this characterisation is a best case scenario and assumes that further testing with additional morpho-syntactic tests and non-subject participants will not reveal yet more distinctions.

In fact initial results from quantifier float suggest that this is indeed the case. While it has not yet been tested thoroughly for every participant, a Beneficiary can launch a floating quantifier while a Transitive Instrument Theme cannot; this would force us to identify seven non-subject categories.

Furthermore, these tests cannot all be described as unidirectional. The idea of unidirectionality is best explained by reference to a subset of English data. In English, only the SUBJ is eligible to be controlled, only a SUBJ and OBJ are expressed as a bare noun phrase and SUBJ, OBJ and OBL_θ (but not ADJCT) can launch a secondary predicate. This data is summarised in Table 4.

	Cont.	Bare NP	2^{nd} Pred.
A. SUBJ	✓	✓	✓
B. OBJ	-	✓	✓
C. OBL	-	-	✓
D. ADJCT	-	-	-
	x	*y*	*z*

Table 4: English Syntactic Tests

Each test is sensitive to a progressively more restricted set of GFs, and the scope of each test can be partially predicted on the basis of others. If a GF is eligible to be controlled it is eligible to be expressed in a bare noun phrase, likewise if a GF is eligible to be expressed in a bare noun phrase it is eligible to launch a secondary predicate. Maximal unidirectionality is defined formally in (59).

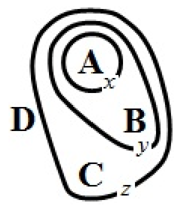


Figure 2: Unidirectional Tests

Maximal Unidirectionality:

(59) Given a scale of GFs ranging from SUBJ(1) through OBJ(2) to ADJCT(n), there is no syntactic privilege that applies to GF(x) such that GF($x-1$) is not eligible to participate in the same privileges.

Unidirectionality is represented in Figure 2, where uppercase letters represent GFs and the lines represent groupings of GFs to which a test is sensitive.

However, Tolaki morpho-syntactic tests are not unidirectional. This is shown in figure 3, in which nine of the tests are represented. While each test is sensitive to a unique set of GFs, they are not progressively more restricted.

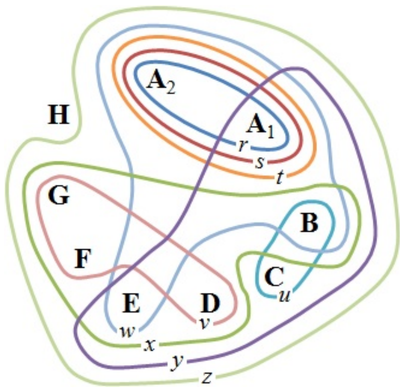


Figure 3: Non-Unidirectional Tests

A_1 = Subject S	r = Relativisation
A_2 = Subject A	s = Plural Agreement
B = Definite P	t = NOM
C = Indefinite P	u = Passivisation
D = Dative P	v = DAT
E = Transitive Instrument	w = ABS
F = Transitive Instrument Theme	x = External Possession
G = Beneficiary	y = 2^{nd} Predication
H = Adjunct	z = Bare NP

Furthermore, these complications do not exist only among non-subject participants. While the grammatical relations A and S have been grouped together in Tolaki as the SUBJ, these two roles do not have exactly the same set of behaviour. While, an S can launch a secondary predicate, an A cannot.

Instead of positing discrete categories of GFs for Tolaki, a better way to model the data is to posit a continuum of functions. Within this continuum 'SUBJ' defines the upper limit of GFs, the maximum amount of behaviour associated with a GF, while 'ADJCT' defines the lower limit, the minimum amount of behaviour associated with a GF.

Other participants exist on this continuum somewhere between these two limits, some are more privileged than others and are thus more or less 'subject-like'.

One way of representing this continuum would be to simply sum the number of tests which a participant is sensitive to. However, such an approach would not capture the non-unidirectionality of the tests and if such an approach were taken, the best analysis of the Tolaki data would appear be to posit 9 GF categories. A graph which captures the continuum-like nature of GFs as well as their non-unidirectionality is given in Figure 4

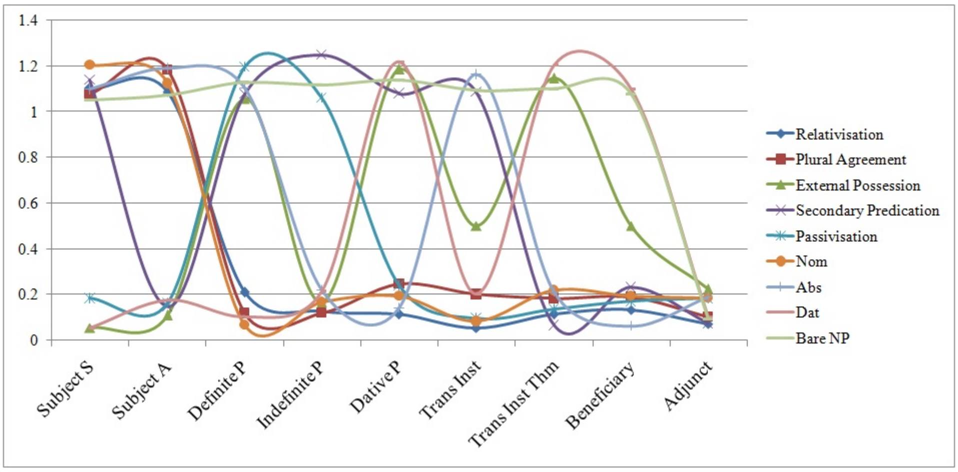


Figure 4: Continuum of Grammatical Functions Test by Test[8]

However, none of these models explains *why* Tolaki participants display this continuum-like behaviour. It is possible to reach an explanation by proposing that the morphology that would typically treated as inflectional is, in fact, derivational.

Tolaki transitive verb roots are defective, thus while it is possible to identify a disyllabic root $\sqrt{kaa}$ meaning 'eat', this form never surfaces. As noted by Mead (1998, 156) the transitive verb in Bungku-Tolaki languages is always accompanied by additional morphology. In Tolaki, even imperatives must be accompanied minimally by the indefinite P prefix or an absolutive suffix.

I propose that each affix is derivational and alters the argument structure of the predicate. As the argument structure of a predicate is slightly altered, so too is the morpho-syntactic behaviour of the participants slightly altered.

One such example, that we have already briefly touched upon, is the difference between verbs which take a definite P, indexed absolutively, and verbs which take a Dative P. Thus, we saw in section 3.1.2 that the argument structure of Dative P verbs differs from that of canonical transitive verbs in that Dative P verbs subcategorise for one argument, while canonical transitive verbs subcategorise for two. The linking between the lexico-conceptual structure and argument structure of the canonical transitive verb *kaa* 'eat' and the Dative P verb *to'ori* 'know' are given below:

(60) *kaa* (AGT, THM)
$\quad\quad\quad\quad$ | $\quad$ |
'PRED⟨ __ , __ ⟩'

(61) *to'ori* (EXP , STIM)
$\quad\quad\quad\quad$ | $\quad$ |
'PRED⟨ __ ⟩' __

I propose that the difference in morpho-syntactic behaviour between these two non-subject participants can be explained by their different argument structure.

[8] A score of 0.3 or lower indicates that a participant is not sensitive to a test, a score of 1 or higher indicates a participant is sensitive to a test. To make all lines visible, scores are randomised by ±0.3.

Extending this idea further, I propose that the difference in behaviour between other participants is likewise explained in this way. Observe the two sentences below:

(62) *Noponggaa* *o'ika.*
no-poN-kaa o-ika
3NOM-INDF.P-eat CN-fish
He eats some fish.

(63) *Nokaa'i* *o'ika.*
no-kaa-'i o-ika
3NOM-eat-3ABS CN-fish
He eats the fish.

We have seen that the Indefinite P in (62) and the Definite P in (63) do not have the same morpho-syntactic behaviour. Therefore, I propose that each involves a different argument structure derivation. In the case of (62) the new derivation prespecifies that the P is indefinite in reference, in the case of (63) the new derivation prespecifies that the P is 3[rd] person and definite:

(64) ‘*poN-*⟨__, INDF⟩ ‘*kaa*⟨__, __⟩”

(65) ‘*-'i*⟨__, 3PRS:DEF⟩ ‘*kaa*⟨__, __⟩”

References

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Blackwell.

Dalrymple, Mary. 2001. *Syntax and Semantics 34: Lexical Functional Grammar*. Academic Press.

Edwards, Owen. 2012. Grammatical Functions in Tolaki. Honours Thesis, Australian National University.

Lødrup, Helge. 2009. Looking Possessor Raising in the Mouth: Norwegian Possessor Raising with Unergatives. In *Proceedings of the LFG09 Conference*, CSLI Publications.

Mead, David. 1998. *Proto-Bungku-Tolaki: Reconstruction of its Phonology and Aspects of its Morphosyntax*. Ph. D.thesis, Rice University.

Payne, Doris and Barshi, Immanuel. 1999. External Possession: What, Where, How and Why. In Doris Payne and Immanuel Barshi (eds.), *External Possession*, pages 3–29, Benjamins.

Untung. 2009. *Cerita Rakyat Tolaki* [Tolaki Folklore]. Bintang.

MISSING RESOURCES IN A RESOURCE-SENSITIVE SEMANTICS

Gianluca Giorgolo and Ash Asudeh

King's College, London & University of Oxford

Carleton University & University of Oxford

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

In this paper, we present an investigation of the argument/adjunct distinction in the context of LFG. We focus on those cases where certain grammatical functions that qualify as arguments according to all standard tests (Needham and Toivonen, 2011) are only optionally realized. We argue for an analysis first proposed by Blom et al. (2012), and we show how we can make it work within the machinery of LFG. Our second contribution regards how we propose to interpret a specific case of optional arguments, optional objects. In this case we propose to generalize the distinction between transitive and intransitive verbs to a continuum. Purely transitive and intransitive verbs represent the extremes of the continuum. Other verbs, while leaning towards one or the other end of this spectrum, show an alternating behavior between the two extremes. We show how our first contribution is capable of accounting for these cases in terms of exceptional behavior. The key insight we present is that the verbs that exhibit the alternating behavior can best be understood as being capable of dealing with an exceptional context. In other words they display some sort of control on the way they compose with their context. This will prompt us also to rethink the place of the notion of subcategorization in the LFG architecture

1 Introduction

The distinction between *arguments* and *adjuncts* is central for the LFG architecture as it influences the way in which representations of linguistic expressions are generated both at the functional and the semantic level. At the functional level the distinction between arguments and adjuncts is crucial for the definition of the notion of *completeness* and *coherence* of an f-structure, which is in turn one of the parameters that determines the grammaticality of an expression. Similarly, at the level of semantics, the distinction between arguments and adjuncts has important consequences on the semantic representations we choose and on the way we control the composition of these meanings. Arguments are in fact usually represented as resources that are consumed by predicates, while adjuncts tend to be represented as functions that consume predicates to generate modified versions of them. The choice of whether a particular grammatical function is an argument or an adjunct requires particular attention from a semantic perspective, as it determines important properties such as the scopal relations between quantified expressions and the way in which these relations are captured by our semantic theory. While the importance of this distinction for the f-structural level has been recognized by many in the LFG literature (Bresnan, 1978; Dalrymple, 2001), we think that its effects

[†]This research was supported by an Early Researcher Award from the Ontario Ministry of Research and Innovation, NSERC Discovery Grant #371969 and by a grant from the John Fell Oxford University Press Research Fund. The authors thank Avery Andrews, Doug Arnold, Boban Arsenijević, Miriam Butt, Mary Dalrymple, Ron Kaplan, Dag Haug, Stefano Quaglia, Louisa Sadler, Sebastian Sulger and the audience at LFG12 for their comments and questions. Any remaining errors are our own.

at the compositional semantics level are understudied, and an overview of these effects can help clarify this important grammatical issue.

Traditionally the distinction between arguments and adjuncts has been made on the basis of a mixture of ontological and syntactic tests. Needham and Toivonen (2011) provide an overview of these tests in an LFG perspective. At the same time, they point out that there are cases where the tests fall short of providing a clear distinction between the two classes. In this paper we will analyze some of these examples under the perspective of compositional semantics. We are particularly interested in cases where grammatical functions that are usually considered arguments become *optional*, a property typical of adjuncts. We motivate the discussion on the basis of the equivalences in (1-6).

(1) Alice ate yesterday afternoon. ⇔ Alice ate something yesterday afternoon.

(2) Bob drank last night. ⇔ Bob drank something last night / Bob drank something alcoholic last night.

(3) (?) Bob loves drinking. ⇔ Bob loves drinking alcohol.

(4) Yesterday, Alice debugged for three hours. ⇔ Yesterday, Alice debugged some code/some programs for three hours. *(in a context in which it is known that Alice finds debugging annoying)*

(5) Silvio was accused of tax fraud. ⇔ Silvio was accused of tax fraud by someone. / Someone accused Silvio of tax fraud.

(6) Silvio was accused. ⇔ Silvio was accused of something by someone.

The lefthand side of the equivalences we have in (1), (2), (3) and (4) show that we can omit the object with certain transitive verbs. However the righthand side of the same equivalences show that the argument is not deleted from the semantic representation of the verb, but rather it is filled by some default value. In most cases the argument slot is bound by an existentially bound variable, but this is not always true. For example, in (8) the omitted object is interpreted as a universally bound variable, while in (9) the intuitive reading for the omitted destination of the arriving event is a deictic or indexical one (the origin seems to be interpreted as an existentially bound variable).[1]

(8) W.H.O. warns against homeopathy use. ⇔ W.H.O. warns everyone against homeopathy use.

[1] However notice that Stanley (2000) proposes an analysis of cases like (9) in which the unexpressed arguments are considered bound by a linguistic operator. Stanley bases his analysis on examples like (7) where the raining event location co-varies with the locations quantified over by "everywhere".

(7) Everywhere Bob goes, it rains.

(9) Bob arrived yesterday. ⇔ Bob arrived from somewhere yesterday to the contextually relevant location.

Another important aspect of this phenomenon is that it seems to be lexically specified. Not all transitive verbs can in fact be constructed with an implicit object, and what is even more interesting is that related verbs may present opposite behaviors. For instance the verb *eat* can be constructed without an explicit object but the intensified form *devour* can not.

The equivalences in (5) and (6) show a similar situation for agents in passive constructions. The by-phrase is always optional, but the described events are always understood as requiring an agent. Notice that in this case there is no lexically specified preference for this construction. All transitive verbs allow for an implicit agentive by-phrase.

These examples challenge a resource sensitive semantics, such as Glue Semantics, in two ways. First of all we have to clarify whether these optionally realized semantic roles should be considered arguments or adjuncts. This decision will determine how we represent them in terms of semantic resources. Given that they seem to contribute to the semantic content of an utterance even when they are not present, we are inclined to consider them core arguments of their predicates. This choice motivates the second challenge to a resource sensitive semantics. We need in fact to clarify how these default resources are introduced in the semantic derivation despite the fact that they are not apparently introduced by any linguistic item.

In what follows we present our solution. In a nutshell we propose to consider verbs that support implicit objects and constructions like passive as being capable of actively operating on their context during the semantic derivation. We will reject the hypothesis that these verbs and constructions are in some way ambiguous. Instead we will associate a single core meaning to them, but give them the power to operate on their context in *exceptional* cases. For the case of verbs that allow implicit object, this notion will prompt us to reconsider the standard distinction between transitive and intransitive verbs. We will suggest that the distinction between transitive and intransitive verbs is not binary. Instead we propose a continuum of verbal behaviors with certain verbs leaning more clearly towards the transitive end of the continuum, other more towards the intransitive end, and still others presenting less marked uses. This shift in perspective has the effect of changing the way in which we look at the issue of how to distinguish arguments from adjuncts. By showing that the richer categorization we propose helps to clarify these notions in the context of the transitive / intransitive divide, we show that looking at compositional semantics may be crucial to better understanding the notion of *argument*.

The paper is structured as follows. Section 2 discusses prior analyses of the kind of data we are interested in. We will survey a number of proposals and identify a recent one by Blom et al. (2012) as the most promising. In section 3 we show how this proposal can be adapted to work in the context of LFG, and how it relates to a previous extension of Glue Semantics that we have proposed in Giorgolo and Asudeh (2011). In section 4 we provide a detailed analysis of some examples. In section 5 we discuss the consequences of our proposal for the distinction between transitive and intransitive verbs and, more in general, for the notion of subcategorization. We conclude in section 6 with some final remarks.

2 Prior work

We start our review of the literature on the topic with Bresnan's (1978) analysis of transitive verbs constructed without an explicit object. Bresnan proposes a solution based on *lexical ambiguity*. The verbs that can be constructed without an explicit object therefore have two entries in the lexicon: the first one correspond to the standard transitive construction whose meaning is represented by a binary function, while the second specifies an intransitive syntactic structure coupled with a unary predicate, constructed from the original binary relation by *existentially* binding the object argument. For example, Bresnan (1978) gives the following lexical entries for the verb *eat*:

eat:	V,	[___ NP],	NP_1 $\mathbf{eat}$ NP_2
		[___],	$(\exists\ y)$ NP_1 $\mathbf{eat}$ y

The second entry for *eat* is obtained from the first one by removing the syntactic requirement of an object, and by existentially binding the object position of the predicate $\mathbf{eat}$. We will see that the idea that treating verbs like *eat* as ambiguous is a common analysis in the literature.

Dowty (1982) proposes a similar analysis in the context of Montague Grammar by introducing so-called *Relation-Reducing Rules*. One such rule transforms a transitive verb like *eat* into an intransitive one and at the same time changes the semantics of the verb by binding its second argument to an existential quantifier. This approach is completely equivalent to the one of Bresnan (1978), with the additional complication that we have to introduce a device in the lexicon that controls the applicability of the Relation-Reducing Rules. Without such a device the grammar would over-generate as it would allow us to derive ungrammatical sentences such as **Yesterday, John devoured*.

The proposal in Bresnan (1978) does not specify how the implicit quantifier should behave with respect to other quantificational elements in the sentence. Fodor and Fodor (1980) tackled this question by first noticing that, in the case of a quantifier in subject position, the existential binding the second argument of $\mathbf{eat}$ must take narrow scope:

(10) Every boy ate.

The intuitive reading for (10) is the one were there are (possibly) different entities that are eaten by the boys.[2]

To capture this generalization, and to explain it in terms of general logical properties, Fodor and Fodor (1980) resort to an approach based on *meaning postulates*. Their first assumption is that the ambiguity that in Bresnan's proposal

[2] The strongest reading were a single entity is shared by the boys is of course available, as it is entailed by the weaker reading.

$$
\begin{tikzcd}
\forall x(P_1(x)) \arrow[rr, leftrightarrow] \arrow[d, leftrightarrow] & & \forall x \exists y(P_2(x,y)) \arrow[d, leftrightarrow] \\
P_1(c_1) \wedge \ldots \wedge P_1(c_n) \wedge \top \arrow[rr, leftrightarrow] & & \exists y(P_2(c_1,y)) \wedge \ldots \wedge \exists y(P_2(c_n,y)) \wedge \top
\end{tikzcd}
$$

Figure 1: Equivalences between first-order formulae justifying the narrow scope of implicitly introduced existential quantifiers.

was restricted to the syntactic component is extended to the semantics. They assume that at the semantic representation level there are two predicates for transitive verbs that allow for implicit objects: a standard binary version, and a unary one. For instance, for a verb like *eat* we have a binary predicate $\mathbf{eat}_2$ that represents its meaning when used with an overt object, and a unary $\mathbf{eat}_1$ that corresponds to the meaning of a use like the one in (10). The two versions of the predicates are then related through meaning postulates like those in (11) and (12).

(11) $\mathbf{eat_1}(c) \leftrightarrow \exists y(\mathbf{eat_2}(c, y))$ with c a constant term

(12) $Qx(\mathbf{eat_1}(x)) \leftrightarrow Qx\exists y(\mathbf{eat_2}(x, y))$ with Q a quantifier

In words, the unary version of the predicate is required to be equivalent to the binary one, where the second argument is bound to a an existential quantifier. The relative order between the implicit existential quantifier and other quantificational operators in postulates like (12) is fixed, with the implicit quantifier having narrow scope. This order is not arbitrary but is explained by Fodor and Fodor on the basis of postulates of the kind in (11) and general logical equivalences. For example, in the case of a universal quantifier the relative order of the two quantifiers is determined by the equivalences illustrated in the diagram in Figure 1. The assumption here is that there is at least one constant naming each element in the domain of quantification. The equivalence on the top row is justified by the equivalence in the bottom row, which results from the repeated application of a postulate of the form of (11), and the two "vertical" equivalences which are general logical equivalences.

This solution is therefore a mixture of lexical ambiguity, also extended to the semantic representation language, and general logical axioms. This approach requires in any case a lexical specification that controls when the unary predicates are available (for example restricting them to passive constructions). Notice that Fodor and Fodor (1980)'s observation is not in contrast with the solution of Bresnan (1978): in Asudeh and Giorgolo (2012) we show how we can control the relative scope of explicitly and implicitly quantified arguments in an LFG setting without resorting to meaning postulates.

However this solution presents some drawbacks, mainly connected to the fact that we may have to list a large number of postulates in cases of verbs constructed with more than two arguments. Consider (13), for example.

(13) Most politicians were accused of at least two crimes.

In this case we would have to control the meaning of the predicate $\mathbf{accuse}$ by listing postulates covering all the possible allowed combinations of implicit and explicit quantifiers and constant terms (in the case of a ternary predicate like $\mathbf{accuse}$ this would amount to 12 (non-equivalent) postulates).

The analyses we have considered so far are based on the idea that the alternation between constructions where all the arguments of predicates are expressed and the cases where some specific arguments are left implicit can be best captured in terms of lexical ambiguity. Carlson (1984) and Lasersohn (1993) depart from this assumption and instead propose two similar analyses that explain the alternation on ontological grounds. Here we focus on the analysis of Lasersohn (1993) as it is motivated on an interesting problem that solutions based on lexical ambiguity encounter when dealing with distributive readings of certain predicates. Lasersohn considers sentences like the one in (14) which is usually interpreted to mean something along the lines of (15). The core intuition here is that the unexpressed agent is not necessarily the same for all atomic grading events. Therefore the existential quantifier binding the agent variable in the predicate needs to have narrow scope with respect to the universal quantifier that enumerates the atomic grading events.

(14) The papers were graded.

(15) $\forall y \exists x (y \in \mathbf{paper}^* \rightarrow \mathbf{grade}(x, y))$

While this seems in line with the observation of Fodor and Fodor (1980), this is not the case. The problem is in the way in which the universal quantifier is introduced in the semantic representation. The lexical solution of Fodor and Fodor (1980) can in fact control the scope of the quantifiers associated with implicit arguments only with respect to quantifiers that are introduced by other lexical resources. However, the standard assumption is that the universal quantifier in (15) is part of the lexical entry of the verb *grade*, specifically of its distributive reading that we can represent as (18).[3]

(18) $\lambda y \lambda x \forall z (z \in y \rightarrow \mathbf{grade}(x, z))$

It is easy to see that an approach based on existentially binding the variable corre-

[3]Lasersohn (1993) uses a meaning postulate similar to the one in (16) to construct the distributive reading of a verb.

(16) $\alpha(X) \leftrightarrow \forall y (y \in X \rightarrow \alpha(y))$

This approach makes his argument less compelling as it would be easy to modify this postulate in the case of a binary (or in general n-ary predicate) predicate to obtain the desired result:

(17) $Qx(R(x, Y)) \leftrightarrow \forall y Qx (y \in Y \rightarrow R(x, y))$ where Q is a quantifier and Y a plural entity.

This approach would however suffer the same drawbacks described for Fodor and Fodor (1980).

sponding to the agent role would lead to the wrong interpretation:

$$(\lambda y \lambda x \forall z(z \in y \to \mathbf{grade}(x, z)))\ \mathbf{paper}^* \leadsto$$
(19) $$\exists x \forall z(z \in \mathbf{paper}^* \to \mathbf{grade}(x, z))$$

The solution proposed by Lasersohn (1993) is radically different from those we have seen so far and is based on a flat semantic representation centered around *events*. According to Lasersohn (1993), a sentence like (14) is interpreted as (20). In this representation the event is considered central and the various roles are introduced by specific predicates that link the individuals involved with the event. Most importantly, in a case like (14) we are not required to specify an agent.

(20) $\exists e(\mathbf{grade}(e) \wedge \textsc{patient}(\mathbf{paper}^*, e))$

To explain the equivalences of the kind we saw in (1-6) Lasersohn (1993) introduces ontological postulates that require that every *atomic* event (at least those showing the behavior under discussion) must have an agent. For instance, to capture the intuition behind the interpretation of (14) we can introduce the postulate in (22).[4] In words, postulate (22) says that every atomic grading event must have at least one agent. The restriction to atomic events is crucial to obtain the correct reading. In fact, by restricting the applicability of the postulates to atomic events, Lasersohn (1993) is able to obtain the correct relative scope between the universal quantifier introduced by the distributive reading of the verb (which is nothing but a collection of atomic events), and the existential quantifier introduced by the ontological rules that Lasersohn assumes govern the meaning of verbs.

(22) $\forall e(\textsc{atom}(\mathbf{grade}, e) \to \exists x(\textsc{agent}(x, e)))$

While this analysis has some attractive features, like the ability to explain in a simple and general way the interactions between implicit arguments and distributive readings, it also has some major drawbacks. The first problem is connected to the fact that implicit arguments are analyzed as being introduced at a very general level. This gives the wrong predictions for certain cases. Consider for example the verb *break*. This verb supports distributive readings, as attested by (23), which is usually interpreted as equivalent to *each mirror was broken by someone/something*.

(23) The mirrors were broken

However, the ontological postulate that would explain this equivalence, would also give us the wrong prediction that in the case of the atomic event introduced by the unaccusative use of *break* in (24) there is an unexpressed agent. This interpretation does not seem to be supported by intuition, as sentence (24) is not necessarily equivalent to *Someone/something broke my TV*.

[4] ATOM is defined by Lasersohn (1993) as follows:

(21) $\textsc{atom}(\alpha, e) \leftrightarrow (\alpha(e) \wedge \neg\exists e'(e' < e \wedge \alpha(e')))$

where $<$ is an ordering relation on events (e.g. a mereological one).

(24) My TV broke.

In general, unaccusative uses seem to contradict the ontological requirement that atomic events must have an agent.[5] Ontological postulates are applied to all cases that involve the predicates they control, and there is no way to limit their application to predicates introduced by specific linguistic expressions without shifting to a lexical perspective.[6] This suggests that the ontological requirement is too strong. Natural language has the expressive means (through specific constructions or specific lexical items) to refer to atomic events that may lack an agent, so we should assume a more flexible semantic model that allows both cases.[7]

Another problem with the analysis of Lasersohn (1993) is in the assumption that the existential quantifier that binds an implicit agent in a passive construction with a distributive reading must always have narrowest scope. There are cases like (26) where this generalization does not seem to hold.

(26) The numbers were summed.

The preferred reading for sentence (26) is one where a single entity sums the numbers (or where at least this entity performs the final addition that gives the final result). In this case, the problem again seems to be the strength of the ontological postulates, more specifically in the fact that they focus on atomic events.

In our opinion, the drawbacks of Lasersohn's event-based approach far outweigh its benefits. At the same time we recognize that in the case of lexical approaches to implicit arguments, the problem of how to derive the correct scope between quantifiers binding implicit arguments and quantifiers controlling distributive readings remains open. A naive lexical solution to this problem would amount to hard coding the relative order of the quantifiers in the lexical entries. (27) would give us the correct interpretation in a case of a sentence like (14).

(27) $\lambda y \forall z \exists x (z \in y \rightarrow \mathbf{grade}(x, z))$

The price to pay to get the correct reading would be to say that the meaning of a passive construction is not derived by general principles from the meaning we associate with the active form of the verb, but rather that it is separately specified in the lexicon. In turn the lexicon should be enriched with postulates that govern the relation between the meaning of active and passive forms. We believe that this

[5]Some languages use a reflexive-like construction in place of unaccusative constructions. This could suggest an implicit agent co-referring with the patient. However this is does not seem to be the case as the following Italian example shows:

(25) Ieri, alle tre, la porta si è chiusa. È stato Marco. (*Yesterday, at three, the door closed. Marco did it.*)

[6]A variant of this approach is implemented in the event-based lexical solution we discuss in Asudeh and Giorgolo (2012).

[7]To be fair, any lexical solution that does not employ a flat semantics like the one used by Lasersohn (1993) will have difficulties in explaining unaccusative constructions.

is a rather high price to pay. A more promising approach would probably first tackle the question of how distributive readings are represented in the lexicon. We leave this line of research for future work.

We conclude this review of the literature with a recent proposal by Blom et al. (2012), that approaches the problem from a lexicalist perspective but avoids the problematic assumption that implicit arguments are evidence of an underlying ambiguity. In a nutshell, Blom et al. (2012) assume that the interpretation of predicates that allow for implicit arguments corresponds to functions that can take a variable number of arguments. The implicit arguments are marked at the type level as belonging to special *option* or *sum types*, which are derived from standard types by the addition to the domain of a distinguished element that is meant to represent the absence of any other usable value. The introduction of sum types is then coupled with an extension of the meaning language, the simply-typed λ-calculus, that includes equality predicates for the sum types and a *choice* term constructor, e.g. the familiar `if-then-else` construction used in many programming languages. Blom et al. (2012) then represent the semantics of verb like *eat* as a function that takes arguments both of standard types and of sum types. The function performs checks on the sum type argument and in case it does not find a usable value (i.e. the function has been passed the distinguished additional element) it proceeds to bind the missing argument with an existential quantifier. The result is a meaning capable of adapting to the context in which it is used.

We believe that this solution is superior to the other presented so far for a number of reasons. First of all, being a pure lexical solution, it avoids the difficulties encountered by the approaches based on meaning postulates (i.e. the explosion of the number of postulates) and those based on ontological restrictions. Lexical solutions are capable of a much finer grained control on the distribution of implicit arguments, also capturing the intuitive idea that these constructions are somehow tied to lexical knowledge, as the opposite behavior of two related verbs like *eat* and *devour* so clearly suggests. At the same time, this solution avoids one of the weak points of lexical solutions: the idea that constructions with implicit arguments are in some way connected to ambiguous lexical entries.

We see two problems with analyses based on the use of ambiguous lexical entries. First of all the two entries are not effectively expressing two different meanings, but rather two different patterns of composition. Our intuition is that the difference between *Yesterday, John ate* and *Yesterday, John ate something* is not related to the meaning of the verb *ate*, but rather to pragmatic considerations, such as the availability of certain information or the focus on certain aspects of the message transmitted. The second problem, which is connected to the first one, is that the use of ambiguous lexical entries introduces a form of non-determinism that, again, is not reflected in our linguistic intuitions.

These considerations lead us to think that solutions based on lexical ambiguity are at best approximations of the phenomenon we want to model. The use of ambiguous lexical entries seems justified only because of the limitations of traditional formal semantic tools. Instead, the data we consider in this paper suggest the need of more refined formal tools, capable of dealing in this case with flexible forms of composition. The analysis of Blom et al. (2012) gives us exactly this form of flexibility.

In what follows we will specify how this solution can be integrated in the LFG architecture. We will show that the analysis of Blom et al. (2012) is actually an instance of a more general framework for an enriched form of composition that we have already presented in Giorgolo and Asudeh (2011), where it was justified on the basis of the unrelated phenomenon of conventional implicatures.

3 Adapting Blom et al. (2012) to the LFG framework

Blom et al. (2012) develop their proposal in the context of Abstract Categorial Grammar, a framework that shares many similarities with LFG-Glue but that also has some differences. The most important difference is the higher importance that Glue Semantics assigns to the resource sensitive nature of semantics. We show here how Blom et al. (2012)'s ideas can be integrated in Glue Semantics and more generally in the LFG framework.

3.1 Monads for implicit arguments

We present here the essential technical details of the analysis of Blom et al. (2012) and how they are related to the extension of Glue Semantics we presented in Giorgolo and Asudeh (2011).

We have already briefly said that Blom et al.'s (2012) solution is based on two parallel extensions. At the level of semantic types, Blom et al. introduce option or sum types, constructed by generating (inductively) for each type τ a new type τ^o, whose domain is the domain of τ with the addition of a distinguished element $*_\tau$. At the level of the meaning language, they add an equality relation $=$ for each new optional type and a special polymorphic function `option`, shown in (28), and defined on the basis of a choice construction, here represented by the familiar mathematical notation for functions defined by cases.

(28) $$\texttt{option}(x, f, d) = \begin{cases} d & \text{if } x = *_\tau \\ f(x) & \text{otherwise} \end{cases} : \tau^o \to (\tau \to \sigma) \to \sigma$$

`option` works by inspecting its first argument: if it is the distinguished element of the optional type τ^o the function returns its third argument, otherwise the function returns the application of its second argument to its first argument.[8] These extensions are put to use in the entries of lexical items that allow implicit arguments. For instance, (29) is the interpretation proposed by Blom et al. (2012) for the verb *eat* (where o is the object argument and s is the subject argument). The verb denotes a

[8]The assumption here is that every value in τ^o that is not $*_\tau$ is also in τ. The fact that τ is a subtype of τ^o would not be a sufficient condition in this case, given that the second argument of `option` has a negative occurrence of τ.

function of two arguments, the first of type "optional" e and the second a pure e. In case the first argument is a regular object its value is used by applying the function $\lambda u(\mathbf{eat}(s,u))$ to it, otherwise the object argument is bound in the predicate $\mathbf{eat}$ by an existential quantifier.

(29) $\lambda o \lambda s(\mathtt{option}(o, \lambda u(\mathbf{eat}(s,u)), \exists x(\mathbf{eat}(s,x)))) : e^o \to e \to t$

Given that the type of the denotation of *eat* is $e^o \to e \to t$, the framework requires also an *optionalization* operation of type $\tau \to \tau^o$, that embeds any type into its optional extension.[9] In this way the entry for a verb like *eat* can be combined with an object of type e by lifting the last one to the type e^o. At this point we cannot combine the verb with a quantifier object. In his master's thesis, Blom shows that, in a system that includes functional abstraction and application (as even the simply-typed λ-calculus does) and the optionalization operator, a type like $e^o \to e \to t$ can be "lowered" to $e \to e \to t$, the standard type for a transitive verb. We now show that the system described by Blom et al. (2012) is (almost) a monad, the core mathematical object of the extension of Glue Semantics we proposed in Giorgolo and Asudeh (2011).

We start with a brief introduction to monads (the interested reader may find a more thorough introduction aimed at linguists in Shan (2001) and Giorgolo and Asudeh (2012)). A monad can be defined as a triple $\langle M, \eta, \star \rangle$. M is a type constructor that maps any type τ to a new type $M\ \tau$. η is a function of type $\tau \to M\ \tau$ that lifts values of type τ into values of type $M\ \tau$. η must satisfy certain rules with respect to the third object of the triple, $\star$, so that it functions as a sort of "identity" function that simply embeds values of τ in the new type $M\,\tau$. $\star$ is a binary function of type $M\ \alpha \to (\alpha \to M\ \beta) \to M\ \beta$ that allows us to bind a value contained in its first monadic argument to a standard name to be used in the body of a function generating a new monadic object. Intuitively $\star$ plays the role of a special functional application that mediates between monadic and non-monadic values.

There are different ways in which we can intuitively understand how monadic values are different from traditional ones. The metaphor we will use in this paper sees monadic values as *computations* that produce values. The idea is that lexical resources that have a monadic type require some effort to be unpacked. What is most important is that these computations may have side effects besides producing a value, much like computations run by a computer may have other effects besides returning a value (e.g. writing some intermediate results to a log file, access a database or printing warnings to a console). In this paper we will be using monads to model computations that possibly fail without returning a value.

In Giorgolo and Asudeh (2011), we presented a system to integrate monads in Glue Semantics. Here we present a superior system suggested to us by Avery Andrews (p.c.) and based on the logic presented in Benton et al. (1998). The present system is at the same time simpler and more elegant and all the analyses we discussed in Giorgolo and Asudeh (2011) can be easily translated in the new

[9] The operator is explicitly introduced and discussed in Blom (2012).

formalism. The idea is to extend the set of linear connectives with a unary connective $\Diamond$ — used to mark monadic resources. In natural deduction format, the proof theory of this new connective is captured by the usual *introduction* and *elimination* pair of rules, shown here respectively in (30) and (31) with the corresponding Curry-Howard-like correspondence.

(30) $$\frac{x : a}{\eta(x) : \Diamond a}\,\Diamond I$$

(31) $$\frac{m : \Diamond a \qquad \begin{array}{c}[x : a]_i \\ \vdots \\ n : \Diamond b\end{array}}{m \star \lambda x.n : \Diamond b}\,\Diamond E_i$$

The introduction rule states that when we have a resource we are always free to lift it to a monadic level. This is reflected in the λ-calculus side by embedding the value in the monadic type by using the "innocuous" η map. The elimination rule is better understood if we consider first the λ-terms that encode the proof step. What rule (31) says is that if we are able to produce a monadic value n by assuming some value that we call x associated with a resource a and we have a proof of a computation that generates such a resource, then we can use the $\star$ operator to extract the value from the computation and plug it into the body of n. At the level of Glue Logic, we go from a situation with two monadic resources and an open hypothesis, to one without the hypothesis and with only one monadic connective.

We can now show that the system of Blom et al. (2012) is a monad, more specifically what is known in the functional programming tradition as the $Option$ or $Maybe$ monad. The type constructor M is represented in this case by the operation $\cdot^o$ which generates a new type τ^o for each type τ by adding a distinguished element $*_\tau$ to its domain. The *optionalization* operator corresponds to η, as it maps values of type τ into values of type τ^o. η defined in this way can be proven to satisfy the rules we mentioned above with respect to $\star$, which is not used by Blom et al. (2012), but which would be defined as in (32).

(32) $$m \star k = \begin{cases} * & m = * \\ k(m) & \text{otherwise} \end{cases}$$

Despite the similarity between (32) and the definition for the `option` operation given in (28), the two functions are quite different and operate at different levels, as $\star$ is used to combine different monadic resources, while `option` is used only internally in lexical entries. This means that we will keep this last operation as a primitive addition to the language used to specify lexical entries. Finally, the *de-optionalization* operation that is needed for the analysis of the composition of verbs taking monadic arguments with quantified arguments can also be shown to be derivable in this system. We give the proof in (33), where we show that a resource $\Diamond\alpha \multimap \beta$ can be "lowered" to a resource $\alpha \multimap \beta$ without monadic subformulae in negative contexts.

$$\text{(33)} \qquad \dfrac{\dfrac{\dfrac{[x:\alpha]^1}{\eta(x):\Diamond\alpha}\,\Diamond I \qquad f:\Diamond\alpha\multimap\beta}{f(\eta(x)):\beta}\multimap E}{\lambda x.f(\eta(x)):\alpha\multimap\beta}\multimap I_1$$

In this framework verbs like *eat*, *read* or *drink* subcategorize for an object but they consume it only when wrapped in a monad. For example, (34) would be the lexical entry for the verb *eat*.[10]

(34) eat V $(\uparrow\ \textsc{pred}) = \text{‘eat’}$
$\lambda o\lambda s(\texttt{option}(o, \lambda u(\mathbf{eat}(s,u)), \exists x(\mathbf{eat}(s,x))))$
$\Diamond(\uparrow\ \textsc{obj})_\sigma \multimap (\uparrow\ \textsc{subj})_\sigma \multimap \uparrow_\sigma$

The λ-term in (34) represents the semantic contribution of *eat*. It is a function of two arguments, o and s, the former a *computation* returning a value of type e while the latter is a pure value of type e (possibly produced by a computation at a different level). The body of the function uses the `option` procedure to test the result of o: if it is a value of type e then the term $\lambda u(\mathbf{eat}(s,u))$ is applied to the result and the result is used as the second argument of the relation $\mathbf{eat}$, otherwise `option` returns its third argument $\exists x(\mathbf{eat}(s,x))$ were the second argument of $\mathbf{eat}$ is bound by the existential quantifier.

In the case of a passive construction we can derive its denotation from the active form using the function `passivize` defined in (36) that takes as argument a function of type $e \to e \to t$ and returns a new function of type $Option\ e \to e \to t$.[11]

$$\text{(36)} \qquad \texttt{passivize}(f) = \lambda a\lambda p(\texttt{option}(a, \lambda a(f(a,p)), \exists x(f(x,p))))$$

At the level of Glue terms, this corresponds to remapping the template on the left in (37) to the one on the right.

$$\text{(37)} \qquad (\uparrow\ \textsc{obj})_\sigma \multimap (\uparrow\ \textsc{subj})_\sigma \multimap \uparrow_\sigma \qquad \Diamond(\uparrow\ \textsc{obl})_\sigma \multimap (\uparrow\ \textsc{subj})_\sigma \multimap \uparrow_\sigma$$

[10] For the sake of simplicity we do not require anything of the implicit object but to exist. A more realistic lexical entry would require the bound variable to be something that is food for the referent. For example if Dr. McCoy from Star Trek utters “Every subject ate” referring to a group of alien beings in his lab, we expect that each subject ate something compatible with its biology. Notice that the lexical entry in (34) allows us to make the value of the bound variable x dependent on the value of the variable s as required. Also see Asudeh and Giorgolo (2012).

[11] Alternatively we could move the `option` outside the second lambda abstraction:

$$\text{(35)} \qquad \texttt{passivize}(f) = \lambda a(\texttt{option}(a, \lambda a\lambda p(f(a,p)), \lambda p\exists x(f(x,p))))$$

The two definitions are equivalent.

3.2 Implicit arguments and projections

The analysis based on the idea of optional arguments can be easily integrated in Glue Semantics as we just described. However, the projective nature of LFG forces us to be more precise about the way in which the absence of an argument is specified in the derivation. In fact, although Blom et al. (2012) speak about optional arguments, the absence of an argument is only implicitly signalled by the use of the distinguished value $*$ which still counts as a resource that is regularly consumed. The fact that LFG allows us to have access to all the linguistic structures computed at earlier stages gives us a way to actually indetify the contexts in which the introduction in the semantic derivation of the special value $*$ is necessary. In this way we are able to eliminate any source of non-determinism that may stem from the uncertainty connected to introduction of the "silent" $*$ value.

The solution we propose is based on the idea that the $Option$ monad can be interpreted as representing a computation that may fail. In the case of implicit arguments the computation that may fail is the one that constructs the semantic resources out of the actual linguistic elements of the sentence. In LFG the f-structure projects a semantic structure (s-structure) that is used to construct the premises for the glue proof. In Glue Semantics the s-structure is then used together with the lexicon as the input for the procedure that generates *resources* (i.e. premises) for the semantic derivation. This procedure is normally understood as producing a set of resources/premises. What we make explicit is the possibility that this procedure encounters an exceptional situation, such as when attempting to instantiate the linear formula template for the verb *eat* (cf. (34)). In that case, there is no linking with the s-structure projected by the OBJ feature, as no such feature is present. We assume that the procedure signals this error and links it to the rest of the template formula which can instead be instantiated. The error therefore becomes a (faulty) premise for the semantics derivation.[12]

Alternatively we can reuse some of the intuitions of the second analysis we presented in Asudeh and Giorgolo (2012). If we posit that the lexical entry of a verb like *eat* introduces in the s-structure both an AGENT and a PATIENT feature whose values are determined on the basis of the f-structure by the σ projection we can understand the presence in our derivations of an error premise in two (roughly equivalent) ways:[13]

1. the values of the features of the s-structure may all be initialized to $*$ signaling by default that no resource, corresponding to that semantic feature, has been explicitly introduced yet. The σ projection fills the values of the features that have a corresponding f-structural counterpart. In the case of an

[12] A fundamental assumption of this analysis is that grammatical functions are subctegorized in the semantic representation of lexical resources rather than at the syntactico-functional level. We elaborate more on this idea in section 5.

[13] In Asudeh and Giorgolo (2012), we use features like ARG_1 and ARG_2 instead, since AGENT and PATIENT are redundant with predicates in the event semantics in the meaning language, but here we do not assume an event semantics.

Word	Category	Constraints
John	N	$(\uparrow \text{ PRED}) = \text{'John'}$ $\mathbf{john} : \uparrow_\sigma$
ate	V	$(\uparrow \text{ PRED}) = \text{'eat'}$ $\lambda o \lambda s(\mathtt{option}(o, \lambda u(\mathbf{eat}(s,u)), \exists x(\mathbf{eat}(s,x))))$ $\Diamond(\uparrow \text{ OBJ})_\sigma \multimap (\uparrow \text{ SUBJ})_\sigma \multimap \uparrow_\sigma$
something	N	$(\uparrow \text{ PRED}) = \text{'some'}$ $\lambda P \exists x(P(x))$ $(\uparrow_\sigma \multimap X) \multimap X$
kiss	V	$(\uparrow \text{ PRED}) = \text{'kiss'}$ $\lambda o \lambda s(\mathbf{kiss}(s,o))$ $(\uparrow \text{ OBJ})_\sigma \multimap (\uparrow \text{ SUBJ})_\sigma \multimap \uparrow_\sigma$
$\text{kissed}_{\text{pass}}$	V	$(\uparrow \text{ PRED}) = \text{'kiss'}$ $\mathtt{passivize}(\lambda o \lambda s(\mathbf{kiss}(s,o))) \rightsquigarrow$ $\lambda a \lambda p(\mathtt{option}(a, \lambda a(\mathbf{kiss}(a,p)), \exists x(\mathbf{kiss}(x,p))))$ $\Diamond(\uparrow \text{ OBL})_\sigma \multimap (\uparrow \text{ SUBJ})_\sigma \multimap \uparrow_\sigma$

Table 1: Toy lexicon

implicit object the PATIENT feature receives no value.

2. σ by default attempts to fill the values of all s-structural features. If a feature cannot be assigned a value an error is raised and registered in the s-structure using the special value $*$.

If we choose this second approach we have to change the lexicon accordingly to make direct reference to the s-structural features. The changes are straightforward. However, in this paper we choose the first implementation of the monadic approach.

To make our proposal clearer, we now present a detailed analysis of some interesting cases.

4 Analysis

In all the analyses, we assume the toy lexicon in Table 1.

4.1 Implicit objects

The first example we analyze is the case of an implicit object (38).

(38) John ate

The simplified f-structure associated with (38) is shown in (39).

(39)
$$e\begin{bmatrix}\textsc{pred} & \text{`eat'} \\ \textsc{subj} & j\begin{bmatrix}\textsc{pred} & \text{`John'}\end{bmatrix}\end{bmatrix}$$

When instantiating the Glue Term for the verb *ate*, the parser / interpreter tries to access the OBJ function in (39). This leads to an error given that no such function is represented in (39). The error is signaled and propagated to the rest of the interpretation process. The error is explicitly introduced in the semantic derivation by the premise $\Diamond n$, marked by $\Diamond$ as it is not a pure value but a computational object, and associated with the special value $*$. The resulting proof is shown below and consists of two simple functional applications/$\multimap$-eliminations.

$$\dfrac{\begin{array}{c}\text{John}\\ [\![\text{John}]\!] : j\end{array} \qquad \dfrac{\begin{array}{c}\text{ate}\\ [\![\text{ate}]\!] : \Diamond n \multimap j \multimap e\end{array} \qquad \begin{array}{c}\texttt{error}\\ * : \Diamond n\end{array}}{\lambda s \exists x(\mathbf{eat}(s,x)) : j \multimap e}\multimap E}{\exists x(\mathbf{eat}(\mathbf{john},x)) : e}\multimap E$$

The error is detected by the meaning component of *ate* and a default interpretation for the object (the existentially bound variable) is used. In this way, the error is neutralized and the process is successful, leading to the expected interpretation.

4.2 Explicit objects

The second example we consider shows how the same lexical entry for the verb *ate* generates the correct interpretation when the object is explicitly realized as in sentence (40).

(40) John ate something

Based on the f-structure in (41) we associate with the sentence the semantic derivation in Figure 2.

(41)
$$e\begin{bmatrix}\textsc{pred} & \text{`eat'} \\ \textsc{subj} & j\begin{bmatrix}\textsc{pred} & \text{`John'}\end{bmatrix} \\ \textsc{obj} & st\begin{bmatrix}\textsc{pred} & \text{`some'}\end{bmatrix}\end{bmatrix}$$

The crucial step in the proof is the "lowering" of the type of the denotation of *ate* from the type $Option\ e \to e \to t$ to the type $e \to e \to t$. This corresponds to the de-optionalization proof we presented in (33). At the level of meaning terms, we simply create a new function that wraps its first argument in a monad using η, therefore generating a computation that does nothing besides returning the value passed as an argument.

$$
\dfrac{\dfrac{\text{John}}{[\![\text{John}]\!] : j} \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\text{ate}}{[\![\text{ate}]\!] : \Diamond t \multimap j \multimap e} \qquad \dfrac{[z : t]^1}{\eta(z) : \Diamond t}\,\Diamond I}{\lambda s(\mathbf{eat}(s,z)) : j \multimap e}\multimap E}{\lambda z \lambda s(\mathbf{eat}(s,z)) : t \multimap j \multimap e}\multimap I_1 \qquad [w : t]^2}{\lambda s(\mathbf{eat}(s,w)) : j \multimap e}\multimap E}{\dfrac{\mathbf{eat}(John, w) : e}{\boxed{1} = \lambda w(\mathbf{eat}(John, w)) : t \multimap e}\multimap I_2}\multimap E
$$

$$
\dfrac{\boxed{1} \qquad \dfrac{\text{something}}{[\![\text{something}]\!] : (t \multimap X) \multimap X}}{\exists y(\mathbf{eat}(John, y)) : e}\multimap E
$$

Figure 2: Proof for *John ate something*

4.3 Passives

Finally, we show how a passive construction without a *by*-phrase gets an existential interpretation. The example sentence and the associated f-structure are shown respectively in (42) and (43). The proof has exactly the same shape as the one for the case of an implicit object. What is interesting is how the Glue and meaning terms for the passive form of *kiss* are constructed on the basis of their active counterparts (see Table 1). The resulting denotation corresponds to a function that is capable of providing an existential closure in case the *agent* is not expressed phonologically.

As it was the case for the analysis of implicit objects, the procedure that instantiates the linear formula governing the compositional behavior of *kissed$_{pass}$* fails as there is no projection of an OBL feature in the s-structure. The error is added to the premises that guide the semantic composition reasoning and is linked to the resource corresponding to the passive verb.

(42) John was kissed

(43)
$$
k\begin{bmatrix} \text{PRED} & \text{‘kiss’} \\ \text{SUBJ} & j\begin{bmatrix} \text{PRED} & \text{‘John’} \end{bmatrix} \end{bmatrix}
$$

$$
\dfrac{\dfrac{\text{John}}{[\![\text{John}]\!] : j} \qquad \dfrac{\dfrac{\text{kissed}_{\text{pass}}}{[\![\text{kissed}_{\text{pass}}]\!] : \Diamond n \multimap j \multimap k} \qquad \dfrac{\texttt{error}}{* : \Diamond n}}{\lambda p \exists x(\mathbf{kiss}(x, p)) : j \multimap k}\multimap E}{\exists x(\mathbf{kiss}(x, \mathbf{john})) : e}\multimap E
$$

5 The transitive/intransitive continuum and subcategorization

Before concluding, we would like to briefly discuss some ideas that emerged in the analysis of the data discussed in this paper. We first focus on the specific phenomenon of implicit objects and see how it may be related to other related phenomena and what it can tell us about the traditional transitive/intransitive distinction. We then extend the discussion to the more general notion of subcategorization and its position in LFG.

The fact that verbs like *eat* and *warn* can be constructed either with an explicit or an implicit object blurs the standard distinction between transitivity and intransitivity. This distinction is usually considered to cut across levels of analysis, as transitivity is normally explained as both a syntactic and a semantic property. Yet in the cases we discussed there seem to be a misalignment between syntax and semantics, given that *Yesterday, John ate* represents a syntactically intransitive structure that still retains a semantic interpretation constructed around a binary predicate. These verbs therefore seem to be a sort of in between case, showing a preference for being constructed transitively, but also allowing an intransitive use while retaining their transitive meaning.

On the flip side, in the case of cognate objects and similar constructions, we observe that intransitive verbs are used in a transitive way. However, at the level of semantics, their interpretation remains that of unary predicates. For example, sentence (44) is equivalent to *John died horribly* and does not involve a second entity. Similarly in example (45) we are informed of the specific way in which John was dancing, not he was engaging in a certain relation with or he was performing a certain action on a second entity.

(44) John died a horrible death

(45) Yesterday, John danced the waltz all night long

Here we have verbs that are normally used in intransitive constructions but that in some cases allow for a transitive use, but at the semantic level their meaning remains that of unary predicates.

The generalization that emerges from this considerations is that, from the syntactic perspective, the distinction between transitive and intransitive verbs may be too coarse, and instead a sort of graded continuum between verbs that are always constructed as transitive and others that are always intransitive with in between cases in the middle may better capture the reality of things. At the same time, the data suggests that the distinction between transitive and intransitive meanings is maintained in all the in between cases.

More generally this leads us to reconsider the notion of subcategorization in LFG. It has already been proposed that the best place to capture the idea that predicates require certain arguments is not at the level of f-structure (which is too close to the surface syntactic level) but rather at the level of the linear term that control semantic composition (Kuhn, 2001; Asudeh, 2012). In current LFG practice, there is in fact some duplication of information, given that Glue terms also encode the information about subcategorization, and possibly do so in a more refined and controlled way. But most importantly, the considerations about implicit arguments that we developed in this paper suggest that indeed the requirements about arguments are best expressed at the level of semantics, where we can better observe their effects in problematic cases. Therefore, we propose a revision of the standard practice in LFG of checking for predicate arguments early in the interpretation process, a check that instead should take place in the last phase of the process. At the same time, this idea suggests that the kind of syntactic requirements that we impose on the surface form of linguistic expressions should be more flexible and permit us to account also, for example, for the in between cases we just discussed.

6 Conclusion

In this paper we discussed the case of implicit arguments in a number of constructions that range from optional objects to optional by-phrases in passive constructions. We reviewed a number of proposals in the literature, and identified those based on lexical solutions as the best candidates to properly treat the phenomenon under discussion. We focused in particular on the solution of Blom et al. (2012) as the best one, given that it maintains the benefits of lexical solutions without resorting to the idea that these expressions are in any sense ambiguous, a common assumption in the literature that is not supported by our linguistic intuitions. We adapted this solution to the LFG framework. The adaptation makes use of the extension of Glue Semantics with monads, an addition we introduced Giorgolo and Asudeh (2011), to analyse the unrelated phenomenon of conventional implicatures. In this way we managed to extend the analytic capabilities of LFG at no cost, and we have obtained further evidence that supports the idea that we need to extend the traditional semantic toolkit with more powerful mathematical structures. Finally we have discussed how this data prompts us to rethink the place of subcategorization in the LFG architecture. We proposed a view where subcategorization is considered a semantic/compositional property of linguistic expressions, rather than cutting across multiple levels of analysis.

References

Asudeh, Ash. 2012. *The Logic of Pronominal Resumption*. Oxford: Oxford University Press.

Asudeh, Ash and Giorgolo, Gianluca. 2012. Flexible Composition for Optional and Derived Arguments. In *Proceedings of the LFG12 Conference*, Stanford: CSLI Publications, this volume.

Benton, Nick, Bierman, G. M. and de Paiva, Valeria. 1998. Computational Type from a Logical Perspective. *Journal of Functional Programming* 8(2), 177–193.

Blom, Chris. 2012. *Optional Arguments in Abstract Categorial Grammar*. Masters Thesis, Utrecht University.

Blom, Chris, de Groote, Philippe, Winter, Yoad and Zwarts, Joost. 2012. Implicit Arguments: Event Modification or Option Type Categories? In Maria Aloni, Vadim Kimmelman, Floris Roelofsen, Galit Sassoon, Katrin Schulz and Matthijs Westera (eds.), *Logic, Language and Meaning, 18th Amsterdam Colloquium, Amsterdam , The Netherlands, December 19-21*, volume 7218 of *Lecture Notes in Computer Science*, pages 240–250, Springer Berlin / Heidelberg.

Bresnan, Joan. 1978. A Realistic Transformational Grammar. In Morris Halle, Jan Bresnan and George A. Miller (eds.), *Linguistic Theory and Psychological Reality*, MIT Press.

Carlson, Greg N. 1984. Thematic Roles and Their Role in Semantic Interpretation. *Linguistics* 22(3), 259–280.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. San Diego, CA: Academic Press.

Dowty, David. 1982. Grammatical Relations and Montague Grammar. In Pauline I. Jacobson and Geoffrey K. Pullum (eds.), *The Nature of Syntactic Representation*, pages 79–130, D. Reidel.

Fodor, Jerry A. and Fodor, Janet Dean. 1980. Functional Structure, Quantifiers, and Meaning Postulates. *Linguistic Inquiry* 11(4), 759–770.

Giorgolo, Gianluca and Asudeh, Ash. 2011. Multidimensional Semantics with Unidimensional Glue Logic. In Miriam Butt and Tracy Halloway King (eds.), *Proceedings of the LFG11 Conference*, pages 236–256, CSLI Publications.

Giorgolo, Gianluca and Asudeh, Ash. 2012. $\langle M, \eta, \star \rangle$ Monads for Conventional Implicatures. In Ana Aguilar Guevara, Anna Chernilovskaya and Rick Nouwen (eds.), *Proceedings of Sinn und Bedeutung 16*, volume 1, pages 265–278, MIT Working Papers in Linguistics.

Kuhn, Jonas. 2001. Resource Sensitivity in the Syntax-Semantics Interface and the German Split NP Construction. In W. Detmar Meurers and Tibor Kiss (eds.), *Constraint-Based Approaches to Germanic Syntax*, Stanford, CA: CSLI Publications.

Lasersohn, Peter Nathan. 1993. Lexical Distributivity and Implicit Arguments. In *Proceedings from Semantics and Linguistic Theory III*, pages 145–161.

Needham, Stephanie and Toivonen, Ida. 2011. Derived Arguments. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG11 Conference*.

Shan, Chung-chieh. 2001. Monads for Natural Language Semantics. In Kristina Striegnitz (ed.), *Proceedings of the ESSLLI-2001 Student Session*, pages 285–298, 13th European Summer School in Logic, Language and Information.

Stanley, Jason. 2000. Context and Logical Form. *Linguistics and Philosophy* 23, 391–434.

DEEP SYNTAX IN STATISTICAL MACHINE TRANSLATION

Yvette Graham
Dublin City University

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This work investigates the current performance capabilities of LFG f-structure based transfer machine translation. Our empirical evaluation compares transfer-based machine translation performance to state of the art machine translation. Our investigation reveals that although the LFG-based approach under-performs compared to state of the art method in general, when the evaluation is restricted to translations where the target language f-structure falls within coverage of the generation grammar, the LFG-based system can in fact achieve higher coverage of unseen data in addition to improvements in translation quality.

1 Introduction

Essentially, machine translation (MT) systems need to accomplish two things: translate the source language (SL) word into the target language (TL) and produce these words in the correct order for the TL (Koehn, 2009). Approaches to MT use different levels of linguistic analysis for translation and divide the tasks involved in the translation of words and word order between analysis and generation components and a transfer component. The shallowest approach translates a SL surface form sentence directly into the TL, assigning the tasks of translating both words and word order to the transfer component, as in Phrase-Based Statistical Machine Translation (PB-SMT) (Koehn et al., 2003) for example. At a slightly deeper level of analysis, such as Phrase-Based Factored Models (Koehn and Hoang, 2007), transfer involves translating the lemma form, morpho-syntactic information and word order to the TL. Deep syntactic analysis goes a level deeper and transfer now involves translating SL syntactic representations such as dependency relations, lemmas and morpho-syntactic information to the TL. Even deeper again we have semantic analysis, with transfer translating between SL and TL context and meaning representations, relations, roles and (possibly) morpho-syntactic information. Finally, an interlingual analysis assigns the entire translation task to the analysis and generation components, with no transfer required, since the representation itself is entirely language independent.

Although increasing the depth of analysis can potentially *decrease the difficulty of translation*, there is the inevitable trade-off as a deeper analysis *increases the difficulty of analysis and generation*. In addition, when we divide the task of translation into separate components in a pipeline architecture, we need to consider how well each step in the pipeline fits together. The output of the parser used for analysis must be the input expected by the transfer decoder, and likewise the transfer decoder output must provide good input for generation. In addition, the use of parsers and generators to a deep level of analysis can also restrict the number of translation hypotheses reached by the search. For example, if generation is only

†This work was partly funded by a Science Foundation Ireland PhD studentship P07077-60101.

possible on the sentence level, as opposed to the word level, significantly more pruning of translation options may be necessary.

2 Deep Syntax for Transfer

Deep syntax, such as the Lexical Functional Grammar f-structure (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001), has been used in transfer-based machine translation (Riezler and Maxwell, 2006; Bojar and Hajič, 2008; Graham et al., 2009) as it provides a good level of linguistic analysis for machine translation, for several reasons:

- The reordering model, required by PB-SMT and shallow-syntax approaches is one of the most challenging models to devise (Koehn, 2009; Chen et al., 2006) and is not required for deep syntactic transfer. Source language (SL) word order is eliminated from the translation process since translation happens at the deep syntax level, abstracting away from surface form word order differences.

- The number of nodes in a deep syntactic representation is linear in sentence length, avoiding complexity problems encountered with shallow syntax based approaches (Deneefe et al., 2007; Deneefe and Knight., 2009).

- Non-terminals are allowed in transfer rules to map pieces of SL structure to the correct position in the TL but in a much more constrained way than in, for example, Hierarchical Models (Chiang, 2007) avoiding the severe pruning necessary for decoding in other parsing-based approaches (Li et al., 2009).

- Decoding can be carried out via a top-down application of contiguous transfer rules, so there are no gaps between TL words, eliminating the need for sophisticated heuristic language modeling techniques such as cube-pruning (Chiang, 2007), for example.

- Morpho-syntactic information for source and target sentences is present in deep syntactic representations, enabling the use of statistically richer Factored Models (Koehn and Hoang, 2007) also increasing coverage of inflections of lemmas not observed in bilingual training data.

This work focuses on investigating the current feasibility of deep syntactic transfer by comparing performance of two publicly available machine translation systems and to provide as meaningful a comparison as possible, we use two systems that are trained fully automatically. As such, we compare English translations of German text using the publicly available state-of-the-art phrase-based statistical machine translation (PB-SMT) system, Moses (Koehn et al., 2007) with translations produced by Sulis (Graham, 2010), a (also publicly available) transfer-based SMT system that uses the LFG f-structure as the intermediate representation for transfer. Although Sulis is in fact linguistic theory independent, with the only restriction being that input and output structures are deep syntax, the system was initially developed for LFG f-structure transfer, and therefore is fit for the purpose of our empirical comparison. In addition to investigating just how far off state-of-the-art performance the LFG f-structure transfer system currently is, we are also interested to know if the deep syntax SMT system produces the same kinds of translations as a PB-SMT system, examining one syntactic construction in particular, compound nouns. We investigate if for this particular syntactic construction, if the LFG-based system achieves state-of-the-art performance by providing a human evaluation of translations of a sample of compound nouns occurring in the test data.

3 LFG-based Transfer and PB-SMT Comparison

3.1 Experimental Set-up

German and English Europarl (Koehn, 2005) and Newswire sentences length 5-15 words were parsed using using LFG Grammars (Kaplan et al., 2004; Riezler et al., 2002), resulting in approx. 360K parsed sentence pairs, applying a disambiguation model to select the single best parse for each input. A trigram deep syntax language model was trained on the LFG-parsed English side of the Europarl corpus, with approximately 1.26M English f-structures (again using only the single-best parse) by extracting all unigram, bigram and trigrams from the f-structures before running SRILM (Stolcke, 2002). The surface-form language model, used after generation, consisted of the English side of the Europarl, also computed using SRILM. Word alignment was run on the training data yielding an alignment between local f-structures for each f-structure pair in the bilingual training data. All transfer rules consistent with this alignment were extracted. Minimum Error Rate Training (MERT) (Och, 2003) was carried out on 1000 development sentences for each configuration using Zmert (Zaidan, 2009).[1] Parsing and generation were carried out using XLE (Kaplan et al., 2002) and LFG Grammars (Kaplan et al., 2004; Riezler et al., 2002). We restrict our evaluation to short sentences and use the test set of Koehn et al. (2003), which includes 1755 German-English translations.

We compare the performance of a state-of-the-art PB-SMT system, Moses (Koehn et al., 2007), with the LFG f-structure transfer-based system (Graham, 2010). In our investigation, we examine if the LFG-based system produces the same kinds of translations as the Phrase-Based system, focusing on one specific syntactic construction, the compound noun (CN), to observe if, for this particular syntactic construction, the f-structure system can achieve state-of-the-art performance in a human evaluation of the first 100 CNs in the test data. The same data

[1]Settings for MERT training were as follows: beam=20, m=100, k=1, k-option=shortest. MERT was carried out separately for each method of word alignment. In all other experiments weights for the LFG-INT configuration were used.

	Bleu	Correct CNs	Fuzzy CNs	Precision Grammar Coverage
LFG	17.29 %	56 %	25 %	38%
PB	30.70 %	54 %	22 %	n/a

Table 1: LFG f-structure transfer and PB-SMT comparison

as in previous experiments was used for training and testing of both systems. For training the LFG-based system, we use technologies described in (Graham and van Genabith, 2008; Graham et al., 2009; Graham and van Genabith, 2009, 2010a,b). Configuration settings for the LFG-based system were as follows: word alignment – deep syntax intersection, no rule size limit , beam size of 100, m-best list size of 100 and non-deterministic generation (*allstrings* XLE option).[2]

3.1.1 Results

Table 1 contains automatic evaluation results for the f-structure transfer (LFG) system (17.29 Bleu) compared to the Phrase-Based (PB) system (30.7 Bleu) showing the degree to which the LFG-based system currently under-performs compared to state-of-the-art.[3] For CNs, however, the LFG-based system performs at least as well as the PB system by translating 56% CNs correctly and 25% in a way that contributes at least some correct meaning to the translation (labeled *fuzzy correct*), while the PB system translates 54% correctly and 22% as a fuzzy translation, in our human evaluation.[4]

Table 2 contains results for the 38% of translations that were within coverage of the precision grammar used for generation, showing the PB system (32.69% Bleu) outperforming the LFG-based system (27.85% Bleu), by almost 5 Bleu points absolute. Due to the possibility of (ngram-based) Bleu unfairly biasing in favor of the PB system, we include results for human-targeted Bleu, NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006, 2005) automatic evaluation metrics using reference translations produced by post-editing the first 150 translations from each MT system (Snover et al., 2006). Results for this evaluation show that the LFG system (73.12% Bleu) in fact outperforms the PB system (70.8%) by a little over 2 Bleu points absolute for translations within coverage of the precision grammar used for generation. We also include the number of untranslated words for the LFG-based system (2 words) and the PB system (34 words), showing that for translations in-coverage (by in-coverage we mean the input sentence achieves a full parse by the source language precision grammar)

[2]See http://www2.parc.com/isl/groups/nltt/xle/doc/xle.html for further details of available options with XLE.

[3]The unfair bias of ngram-based Bleu metric in favor of Moses should be noted, and is discussed later.

[4]It is worth noting that it is highly likely that the LFG-based system would not perform as well on a test set of unrestricted sentences length due lower parser coverage of long sentences.

	Bleu	HBleu	HNIST	HTER	HMETEOR	Untrans. Words
LFG	27.85	73.12	8.3602	20.74	82.80	2
PB	32.69	70.80	8.1710	23.63	86.00	34

Table 2: Precision grammar in-coverage comparison with state-of-the-art. Note: H-Bleu = human targeted Bleu for 150 post-edited reference translations (similar to HTER (Snover et al., 2006))

of the precision grammar, the LFG-based system also achieves higher coverage of unseen data.

3.1.2 Discussion

Automatic evaluation results for the entire test set suggest that the LFG-based system under-performs significantly in comparison with state-of-the-art (Table 1). However, the results are unfairly biased in favor of the PB system, due to a combination of the Bleu evaluation metric being ngram-based with legitimate syntactic variations in the LFG system output. The difference in results is, however, too large to claim that this is entirely due to this bias. Table 4 shows a random selection of translations produced by the LFG-based system from the entire test set.

Human evaluation of 100 CNs shows that the LFG system does in fact achieve state-of-the-art performance for this particular syntactic construction, however. Interestingly, the intersection of the CNs that both the LFG and PB systems translate correctly is quite small, with the LFG-based system correctly translating 30% of those not translated correctly by Moses, and Moses correctly translating 23% of those not translated correctly by the LFG-based system, suggesting the possibility of a hybrid MT system (similar to (Eisele et al., 2008; Chen et al., 2007; Eisele, 2005)) or that deep syntax parsing could be used to improve translation of CNs for PB-SMT. Table 3 shows a selection of CNs taken from the entire test set for the PB and LFG systems. The LFG system achieves coverage of CNs not observed in training data where component nouns were observed in training. For example, the CN, *Hafenpolitik*, was not observed in the German training data, but *Hafen* appears combined with other nouns a total of approximately 80 times and *politik* also appears in the German training data approximately 3,400 times combined with another noun. This CN is translated correctly by the LFG-based system but not the PB system.

For translations within coverage of the precision grammar, i.e. where the transfer decoder manages to produce a combination of lemmas, dependency relations and morpho-syntactic information in TL structures that do not clash with constraints during TL generation, human-targeted evaluation results show the LFG system achieves state-of-the-art performance for these translations, in addition to achieving higher translation coverage of unseen data, mainly due to its ability to learn how to translate new unseen CNs from CNs in the training data that con-

CN	PB Translation	LFG Translation
Wiederaufnahme	Resumption	
Tagesordnung	agenda	
Rechnungsführung	accounts	
Unternehmensneugründungen		company start-ups
Vorsichtsmassnahmen		measures precautionary*
Asien-Europa-Stiftung		Asia Europe Foundation
Osttimors		East Timor
ASEM-Gesprächen		ASEM talks
Hafenpolitik		port policy
Schwerpunkt	Emphasis	
Hauptsache	reason*	
Eigenkapital	capital*	invested capital*
Arbeitsrecht	labour law*	employment legislation*
Küstenstaaten		coastal states
Subsidiaritätsprinzip	principle of subsidiarity*	
Bewerberländer	candidate countries	applicant countries*
Parlamentswahlen	parliamentary elections*	general elections
Standpunkt	position*	question*
Ostsee	Baltic	
Änderungsantrag	Amendment	
Dioxinskandal	dioxin scare*	dioxin scandal
Einteilung	classification*	division
Futtermittelsicherheit		feed safety
Futtermittelkette		feed chain
Futtermitteln	feed*	means of feed*
Gemeinschaftsebene	Community level	Community scale*
Weltanschauung	World view*	world like mindedness*
weltweit	in the world*	worldwide*
Gemeinderatswahlen		elections local*
Richtlinien	directives*	directive*
Kernstück	heart*	lifeblood*
Ausnahmemöglichkeiten		opportunity for exceptions*
Änderungsanträgen	amendments	
Änderungsanträge		amendments
Vertragseinhaltung		Treaty compliance*
Entschliessungsantrags	resolution*	
Forschungsraum	research area	period of Research*
Endkontrolle		final*
Gegenprüfung		counter examination

Table 3: German Compound Noun translations for the Phrase-Based SMT system and the LFG-based system, translations evaluated as a fuzzy translation are marked with an asterisk

SRC:	Dies kann nicht hingenommen werden.
REF:	This is an unacceptable situation.
LFG:	Not one that can allow continue
SRC:	Herr Präsident! Die Sicherheit verschiedener Verkehrsarten steht ernsthaft auf dem Spiel.
REF:	Mr President, safety is a serious issue for various forms of transport.
LFG:	Mr President. Die of different forms of transport safety is at stake seriously.
SRC:	Das ist die politische Position.
REF:	That is the political position.
LFG:	That is the political position.
SRC:	Natürlich ist sich auch die türkische Gesellschaft dieses Gegensatzes bewusst.
REF:	Turkish society obviously perceives this contradictory attitude.
LFG:	Of course ist sich the Turkish society also of this contradiction bewusst
SRC:	Solche Gewalttätigkeit potenziert die Hassgefühle nur noch weiter.
REF:	That sort of violence only stirs up feelings of hatred.
LFG:	This violation potenzieren only hate emotions further

Table 4: Randomly selected translations, original reference translations provided (not human-targeted)

tain component nouns, in addition to achieving coverage of inflections of words not seen in bilingual training, since we use Factored Models (Koehn and Hoang, 2007). Table 5 shows a random selection of translations for the PB and LFG systems for translations in coverage of the precision generation grammar and Table 6 shows German words that were not translated by the LFG and PB systems for translations in coverage of the precision grammar.

4 Summary

Compared to state-of-the-art PB-SMT the LFG-based system under-performs, but for sentences in-coverage of the precision grammar used for generation, state-of-the-art performance and higher coverage of unseen data is achieved. Some practical challenges still need to be overcome before reaching state-of-the-art performance for all input. One challenge is parser coverage: depending on the parsing technologies used, coverage of long sentences can be low, resulting in a much smaller sized bilingual corpus used for training in comparison to a phrase-based system. A similar challenge occurs for generator coverage: technologies for generation from deep syntactic structures are usually tested on gold-standard input, and even with adaptation to allow more robust generation, generator coverage can still be low. In addition, even when generation succeeds, a fluent sentence of the target language is not guaranteed. LFG f-structures contain a large amount of information, such as dependency relations between words and morpho-syntactic features and in order for TL generation to produce good quality output, the particular combination of lemmas, dependency relations and morpho-syntactic information in the TL structure must comply with constraints within the generation grammar. If the TL structure contains morpho-syntactic and dependency information that clash when constraints are solved during generation, a fragment grammar can be used, but the quality of output severely deteriorates. Constructing TL deep syntactic structures that do not cause clashes in generation constraints remains a major challenge for f-structure transfer.

References

Banerjee, Satanjeev and Lavie, Alon. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In *Proceedings of Workshop on "Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, at the 43rd Annual Meeting of the Association for Computational Linguistics"*, pages 65–73, Ann Arbor, Michigan.

Bojar, Ondřej and Hajič, Jan. 2008. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation at the 46th Annual Meeting of the Association for Computational Linguistics*, pages 143–146, Columbus, OH.

SRC:	Auf Gesetzesebene gibt es allgemeine Texte, in denen Diskriminierung weltweit verurteilt wird.
REF:	**Legally speaking, there are general texts condemning discrimination everywhere.**
LFG:	General texts condemning worldwide discrimination have been given to any legislative level.
PB:	There is general provisions on gesetzesebene where discrimination is condemned in the world.
SRC:	Das soll sich hier hoffentlich nicht wiederholen!
REF:	**I hope we will not see a repeat performance here!**
LFG:	hopefully that should not be repeated.
PB:	I hope it will not repeat here!
SRC:	Der BSE-Skandal war das schlechteste, bekannteste Beispiel.
REF:	**The BSE scandal was the worst and most notorious example.**
LFG:	The BSE scandal is the worst and most known case.
PB:	The BSE scandal was the worst and most well-known example.
SRC:	In Erwartung von mehr Klarheit haben wir uns deshalb der Stimme enthalten.
REF:	**Pending further clarification, we therefore abstain from the vote.**
LFG:	Therefore I abstained in expectation of greater clarity for.
PB:	In expectation of greater clarity, we have therefore abstain from voting.
SRC:	Der Wiederaufbau Osttimors ist noch im Gange.
REF:	**The rebuilding of East Timor is still an ongoing process.**
LFG:	The reconstruction of East Timor is still taking place.
PB:	The reconstruction osttimors is still in progress.
SRC:	Möchte sich jemand für diesen Antrag aussprechen?
REF:	**Is there a speaker to support this request?**
LFG:	Does anyone wish to speak in support of this motion?
PB:	Does anyone wish to speak in favor of this request?
SRC:	Vielen Dank für diese Klarstellung, Herr Kommissar.
REF:	**Thank you very much for that clarification, Commissioner.**
LFG:	I would like to thank the Commissioner for that clarification.
PB:	Thank you for that clarification, Commissioner.
SRC:	In diesem Punkt sind wir einer Meinung.
REF:	**On this point we are in agreement.**
LFG:	We will be agreement on point about this.
PB:	In this regard, we are in agreement.
SRC:	Gibt es Einwände?
REF:	**Are there any comments?**
LFG:	Are there any objections?
PB:	Are there any comments?
SRC:	Verhaltenskodex für Waffenausfuhren
REF:	**Arms trade code of conduct**
LFG:	Code of Conduct on Arms Exports
PB:	Code of conduct on arms exports

Table 5: Randomly selected sample of translations in-coverage of precision grammar, original reference translations provided.

Phrase-Based System	LFG-based system
interparlamentarischer	liegen
asien-europa-stiftung	vorsichtshalber
osttimors	
interparlamentarischer	
europäers	
zu	
lehnten	
spielzeugbomben	
erfahrenen	
kompetenzverteilung	
marktposition	
enttäuschte	
selbstbewertung	
gegenwert	
kostengünstiges	
bleibenden	
geldverkehrs	
reformpläne	
eindämmungsmassnahmen	
regem	
auslanLFGdiplomatie	
kompetenzabgrenzung	
planungssicherheit	
papua-führer	
dominiert	
ersuchten	
neuzuteilung	
eu-lärmindizes	
zusatzstoffes	
klimafrage	
vorsichtshalber	
sicherheitsspielraum	
un-flüchtlingshilfswerk	
gesamtgesellschaftlichen	

Table 6: German words not translated in translations within coverage of the TL generation precision grammar for the Phrase-Based and LFG-based systems

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Chen, Boxing, Cettolo, Mauro and Federico, Marcello. 2006. Reordering rules for phrase-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 53–58, Kyoto, Japan.

Chen, Yu, Eisele, Andreas, Federmann, Christian, Hassler, Eva, Jellinghaus, Michael and Theison, Silke. 2007. Multi-engine machine translation with an open-source SMT deocder. In *Proceedings of the Second Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association for Computational Linguistics*, pages 193–196, Prague, Czech Republic.

Chiang, David. 2007. Hierarchical Phrase-based Models of Translation. *Computational Linguistics* 2(33), 201âĂŞ228.

Dalrymple, Mary. 2001. *Lexical-Functional Grammar*. Academic Press.

Deneefe, Steve and Knight., Kevin. 2009. Synchronous Tree Adjoining Machine Translation. In *Proceedings of Empirical Methods in Natural Language Processing Conference*, pages 727–736, Edinburgh, United Kingdom.

Deneefe, Steve, Knight, Kevin, Wang, Wei and Marcu, Daniel. 2007. What can Syntax-based MT learn from Phrase-based MT? In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processsing and Computational Natural Language Learning*, pages 755–763, Prague, Czech Republic.

Doddington, George. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. In *Proceedings of Human Languages Technologies Conference*, San Diego, California.

Eisele, Andreas. 2005. First steps towards multi-engine machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Compuational Linguistics Workshop on Building and Using Parallel Texts*, pages 155–158, Ann Arbor, Michigan.

Eisele, Andreas, Federmann, Christian, Saint-Amand, Herve, Jellinghaus, Mochael, Herrmann, Teresa and Chen, Yu. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation at the 46th Annual Meeting of the Association for Compuational Linguistics*, pages 179–182, Columbus, Ohio.

Graham, Yvette. 2010. Sulis: An Open Source Transfer Decoder for Deep Syntactic Statistical Machine Translation. *In The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation* .

Graham, Yvette and van Genabith, Josef. 2008. Packed Rules for Automatic Transfer Rule Induction. In *Proceedings of the European Association of Machine Translation Conference 2008*, Hamburg, Germany.

Graham, Yvette and van Genabith, Josef. 2009. An Open Source Rule Induction Tool for Transfer-Based SMT. *The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation* pages 37–46.

Graham, Yvette and van Genabith, Josef. 2010a. Deep Syntax Language Models and Statistical Machine Translation. In *Proceedings of the Fourth International Workshop on Syntax and Structure in Statistical Translation at The 23rd International Conference on Computational Linguistics*, Beijing, China.

Graham, Yvette and van Genabith, Josef. 2010b. Factor Templates for Factored Machine Translation Models. In *Proceedings of the International Workshop on Spoken Language Translation*, Paris, France.

Graham, Yvette, van Genabith, Josef and Bryl, Anton. 2009. F-structure Transfer-Based Statistical Machine Translation. In *Proceedings of the 14th International Lexical Functional Grammar Conference*, Cambridge: CSLI Publications.

Kaplan, Ronald M. and Bresnan, Joan. 1982. Lexical Functional Grammar, a Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 173–281.

Kaplan, Ronald M., King, Tracy Holloway and Maxwell, John T. 2002. Adapting existing grammars: the XLE experience. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Kaplan, Ronald M., Riezler, Stefan, King, Tracy Holloway, Maxwell, John T. and Vasserman, Alexander. 2004. Speed and accuracy in shallow and deep stochastic parsing. In *Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Meeting*, pages 97–104, Boston, MA.

Koehn, Philip, Och, Franz Josef and Marcu, Daniel. 2003. Statistical phrase-based translation. In *Proceedings of Human Language Technology - North American Chapter of the Association for Computational Linguistics Conference*, pages 48–54, Alberta, Canada.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand.

Koehn, Philipp. 2009. *Statistical Machine Translation*. Cambridge University Press.

Koehn, Philipp and Hoang, Hieu. 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language*

Processing and Computational Natural Language Learning, pages 868–876, Prague, Czech Republic.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondrej, Constantin, Alexandra and Hoang, Evan HerbstHieu. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Li, Zhifei, Callison-Burch, Chris, Khundanpur, Sanjeev and Thorton, Wren. 2009. Decoding in Joshua, Open Source Parsing-Based Machine Translation. In *The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation*, volume 91, pages 47–56.

Och, Franz Josef. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Riezler, Stefan, King, Tracy H., Kaplan, Ronald M., Crouch, Richard, Maxwell, John T. and Johnson, Mark. 2002. Parsing the Wall Street Journal using Lexical Functional Grammar and discriminitive estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Philadelphia, PA.

Riezler, Stefan and Maxwell, John. 2006. Grammatical Machine Translation. In *Proceedings of Human Language Technologies and the 44th Annual Meeting of the Assciation for Computational Linguistics*, pages 248–255, New York City, NY.

Snover, Mathew, Dorr, Bonnie, Scwartz, Richard, Makhoul, John and Micciula, Linnea. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the 7th biennial Conference of the Association for Machine Translaiton in the Americas*, pages 223–231, Boston, MA.

Snover, Mathew, Dorr, Bonnie, Scwartz, Richard, Makhoul, John, Micciula, Linnea and Weischeidel, Ralphe. 2005. A Study of Translation Error Rate with Targeted Human Annotation. Technical Report, University of Maryland, College Park, MD.

Stolcke, Andreas. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Zaidan, Omar. 2009. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation* 91, 79–88.

EXPLORING THE PARAMETER SPACE IN STATISTICAL MACHINE TRANSLATION VIA F-STRUCTURE TRANSFER

Yvette Graham and Josef van Genabith
Centre for Next Generation Localisation

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

Machine translation can be carried out via transfer between source and target language deep syntactic structures. In this paper, we examine core parameters of such a system in the context of a statistical approach where the translation model, based on deep syntax, is automatically learned from parsed bilingual corpora. We provide a detailed empirical investigation into the effects of core parameters on translation quality for the German-English translation pair, such as methods of word alignment, limits on the size of transfer rules, transfer decoder beam size, n-best target input representations for generation, as well as deterministic versus non-deterministic generation. Results highlight just how vital employing a suitable method of word alignment is for this approach as well as the significant trade-off between gains in Bleu score and increase in overall translation time that exists when n-best structures are generated.

1 Introduction

Statistical Machine Translation via deep syntactic transfer is carried out in three steps: (i) parsing the source language (SL) input to SL deep syntactic representation, (ii) transfer from SL deep syntactic representation to target language (TL) deep syntactic representation, (iii) generation of TL string. Figure 1 shows how an example German sentence is translated into English. Bojar and Hajič (2008) present an English to Czech SMT system that uses the Functional Generative Description (FGD) (Sgall et al., 1986) Tectogrammatical Layer (T-layer), i.e. labeled ordered dependency trees, as intermediate representation for transfer, and integrate a bigram dependency-based language model into decoding. Riezler and Maxwell (Riezler and Maxwell, 2006) use the Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001) functional structure (f-structure) for transfer, an attribute-value structure encoding of bilexical labeled dependencies and atomic value features, and extract transfer rules semi-automatically from the training data, by automatically word aligning surface-form sentences using Giza++ (Och et al., 1999) before manually detecting and automatically correcting systematic errors. Most of the transfer rules are automatically extracted from the parsed training data with some transfer rules manually written and deep syntax language modeling is carried out after decoding, on the n-best output structures.[1]

Like Riezler and Maxwell (2006), we use the LFG f-structure as the intermediate representation for transfer, but in contrast we investigate the feasibility of deep syntactic transfer when translation models are learned fully automatically. In addition, we integrate a deep syntax language model to decoder search, similar to Bojar and Hajič (2008) but increase to a tri-gram model. Again in contrast to Riezler and Maxwell (2006) where language modeling is applied to the n-best structures output

[†]This work was partly funded by a Science Foundation Ireland PhD studentship P07077-60101.

[1]Personal communication with authors.

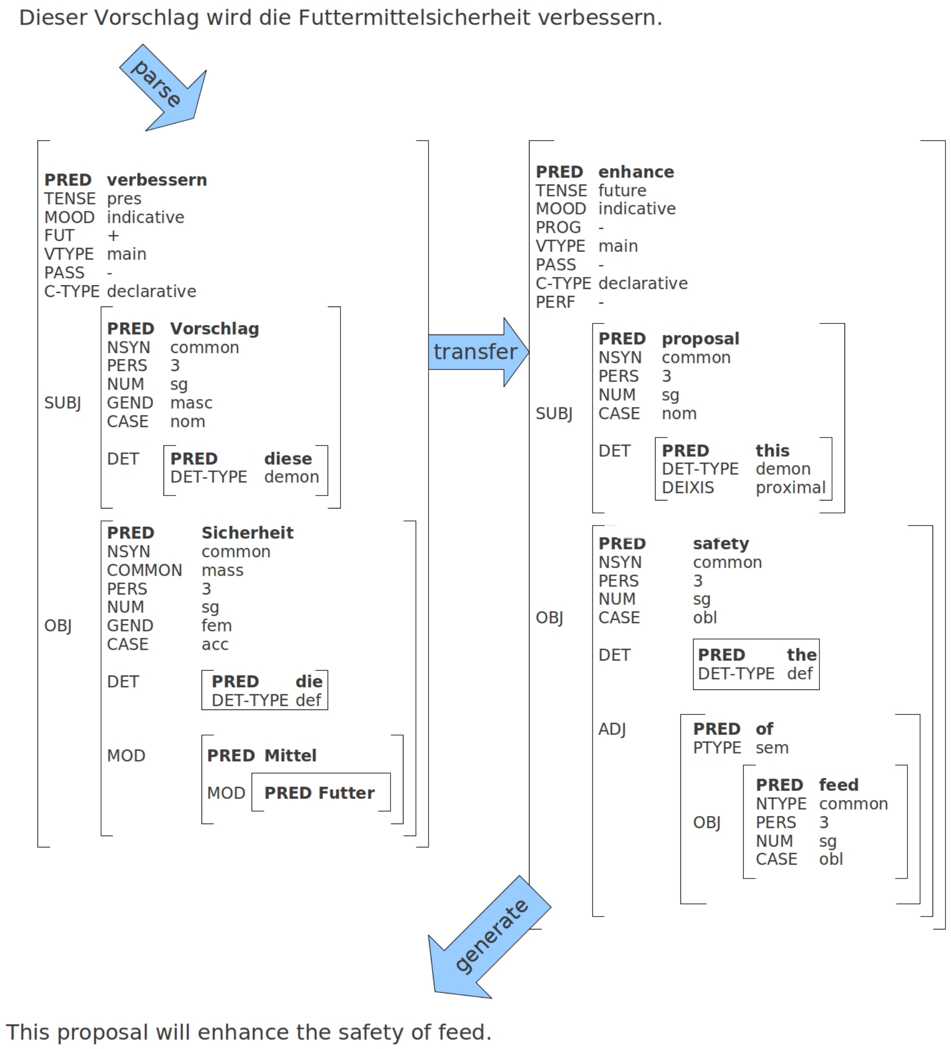


Figure 1: Deep syntactic transfer example via LFG f-structures

after decoding, we integrate language modeling to decoder search. Our empirical evaluation highlights the importance of selecting methods of word alignment most suitable for deep syntax, as well as notable trade-offs that exist between currently achievable translation speed and the quality of translations produced.

Ding and Palmer (2006) use dependency structures for translation, but the approach they take is not strictly deep syntactic transfer, as they use dependency relations between surface form words as opposed to lemmas and morpho-syntactic information, and additionally they use information about source language word order during translation, arguably losing the high level of language pair independence afforded by fully deep syntactic transfer.

2 Translation Model

Similar to PB-SMT (Koehn et al., 2003), our translation model is a log-linear combination of several feature functions:

$$p(e|f) = exp \sum_{i=1}^{n} \lambda_i h_i(e, f) \quad (1)$$

2.1 Word Alignment

An alignment between the nodes of the SL and TL deep syntactic training structures is required in order to automatically extract transfer rules. In our evaluation, we investigate the following three methods of word (or node) alignment, all using Giza++ (Och et al., 1999) for alignment and Moses (Koehn et al., 2007) for symmetrization:

- SF-GDF: input the surface-form bitext corpus to Giza++ and symmetrize with grow-diag-final algorithm.[2] Map many-to-many word alignment from each surface-form word to its corresponding local f-structure. This yields a many-to-many alignment between local f-structures and was used in Riezler and Maxwell (2006).[3]
- DS-INT: reconstruct a bitext corpus by extracting predicates from each local f-structure, input the reconstructed bitext to Giza++, then use the intersection of the bidirectional word alignment for symmetrization. This yields a one-to-one alignment between local f-structures. This method takes advantage of the predicate values of f-structures being in the more general lemma form, and should suffer less from data sparseness problems.

[2]Grow-diag-final works as follows: Word alignment is run in both language directions, for example, German-to-English (f2e) and English-to-German (e2f). For any given training sentence pair, each run (e2f and f2e) can yield a different set of alignment points between the words of the training sentence pair. There are many ways to combine these two sets, grow-diag-final begins with the intersection, then adds unaligned words.

[3]It should be noted that we use a different method of transfer rule extraction, we do not correct word alignment and do not include hand-crafted transfer rules.

- DS-GDF: reconstruct a bitext corpus by extracting predicates from each local f-structure, input the reconstructed bitext to Giza++ (as in DS-INT), but use grow-diag-final for symmetrization yielding up to many-to-many alignments between local f-structures.

2.2 Transfer Rule Extraction

Similar to PB-SMT, the transfer of a SL deep syntactic structure **f** into a TL deep syntactic structure **e** can be broken down into the transfer of a set of rules $\{\bar{f}, \bar{e}\}$:

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) \tag{2}$$

In PB-SMT, all phrases consistent with the word alignment are extracted, with shorter phrases needed for high coverage of unseen data and larger phrases improving TL fluency (Koehn et al., 2003). With the same motivation, we extract all transfer rules consistent with the node alignment. Figure 2 shows a subset of the transfer rules extracted from the f-structure pair in Figure 1.[4] We estimate the translation probability distribution using relative frequencies of transfer rules:

$$\phi(\bar{f}, \bar{e}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} count(\bar{e}, \bar{f}_i)} \tag{3}$$

This is carried out in both the source-to-target and target-to-source directions.[5]

3 Deep Syntax Language Model

In deep syntactic transfer, the output of the decoder is a TL deep syntactic structure with words organized in the form of a graph (as opposed to a linear sequence of words in PB-SMT). A standard surface-form language model cannot be used during transfer decoding because no surface-form representation of the TL deep syntactic structure is available. It is still important for the model to take TL fluency into account so that the structures it outputs contain fluent combinations of words.

A standard language model estimates the probability of a sequence of English words by combining the probability of each word, w_i, in the sequence given the preceding sequence of $i-1$ words. In a similar way, we estimate the probability of a deep syntactic structure d, with root node w_r consisting of l nodes, by combining the probability of each node, w_i, in the structure given the sequence of nodes linked to it via dependency relations that terminates at the node's head. We use the

[4]Morphosyntactic information is left out.

[5]Since we use Factored Models for translating morpho-syntactic information, when computing the translation model we ignore differences in morpho-syntactic information.

$\begin{bmatrix} \text{PRED} & \text{Sicherheit} \\ \text{DET} & \begin{bmatrix} \text{PRED} & \text{die} \end{bmatrix} \\ \text{MOD} & \begin{bmatrix} \text{PRED} & \text{Mittel} \\ \text{MOD} & \begin{bmatrix} \text{PRED} & \text{Futter} \end{bmatrix} \end{bmatrix} \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} \text{PRED} & \text{safety} \\ \text{DET} & \begin{bmatrix} \text{PRED} & \text{the} \end{bmatrix} \\ \text{ADJ} & \begin{bmatrix} \text{PRED} & \text{of} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{feed} \end{bmatrix} \end{bmatrix} \end{bmatrix}$
$\begin{bmatrix} \text{PRED} & \text{Sicherheit} \\ \text{DET} & \begin{bmatrix} \text{PRED} & \text{die} \end{bmatrix} \\ \text{MOD} & X_0 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} \text{PRED} & \text{safety} \\ \text{DET} & \begin{bmatrix} \text{PRED} & \text{the} \end{bmatrix} \\ \text{ADJ} & \begin{bmatrix} \text{PRED} & \text{of} \\ \text{OBJ} & X_0 \end{bmatrix} \end{bmatrix}$
$\begin{bmatrix} \text{PRED} & \text{Sicherheit} \\ \text{DET} & X_0 \\ \text{MOD} & \begin{bmatrix} \text{PRED} & \text{Mittel} \\ \text{MOD} & \begin{bmatrix} \text{PRED} & \text{Futter} \end{bmatrix} \end{bmatrix} \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} \text{PRED} & \text{safety} \\ \text{DET} & X_0 \\ \text{ADJ} & \begin{bmatrix} \text{PRED} & \text{of} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{feed} \end{bmatrix} \end{bmatrix} \end{bmatrix}$
$\begin{bmatrix} \text{PRED} & \text{Sicherheit} \\ \text{DET} & X_0 \\ \text{MOD} & X_1 \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} \text{PRED} & \text{safety} \\ \text{DET} & X_0 \\ \text{ADJ} & \begin{bmatrix} \text{PRED} & \text{of} \\ \text{OBJ} & X_1 \end{bmatrix} \end{bmatrix}$
$\begin{bmatrix} \text{PRED} & \text{die} \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} \text{PRED} & \text{the} \end{bmatrix}$
$\begin{bmatrix} \text{PRED} & \text{Futter} \end{bmatrix}$	$\rightarrow$	$\begin{bmatrix} \text{PRED} & \text{feed} \end{bmatrix}$

Figure 2: Extracted LFG F-structure transfer rule

function m, to map the index of a node to the index of its head node within the structure.

$$p(d) = \prod_{i=1}^{l} p(w_i | w_r, ..., w_{m(m(i))} w_{m(i)}) \tag{4}$$

In order to combat data sparseness, we apply the Markov assumption, as is done in standard language modeling, and simplify the probability of a deep syntactic structure by only including a limited length of history when estimating the probability of each node in the structure. A trigram deep syntax language model estimates the probability of each node in the structure given the sequence of nodes consisting of *the head of the head of the node* followed by *the head of the node* as follows:

$$p(d) \approx \prod_{i=1}^{l} p(w_i | w_{m(m(i))}, w_{m(i)}) \tag{5}$$

Figures 3(a) and 3(b) show how the trigram deep syntax language model probability is estimated for the English f-structure in Figure 1.[6]

4 Decoding

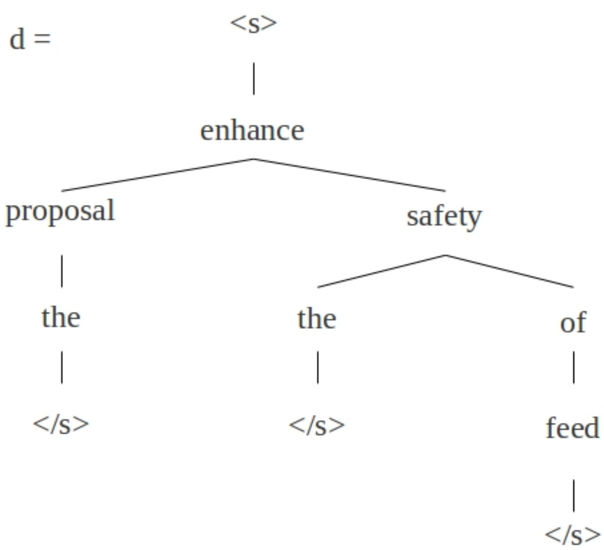
In the (i) parse, (ii) transfer, and (iii) generate architecture of the system, decoding carries out step (ii), the transfer of a SL deep syntactic structure to the target language. Decoding of the SL structure is top-down starting at the root of the structure (usually the main verb of the sentence). Similar to PB-SMT, where decoding search space is exponential in sentence length, our search space is exponential in the number of SL nodes, and we use beam search to manage its size. We use an adaptation of Factored Models (Koehn and Hoang, 2007) to translate morpho-syntactic information.

5 Generation

Generation of the TL output is carried out using XLE rule-based generator (Kaplan et al., 2002), using an English precision grammar (Kaplan et al., 2004; Riezler et al., 2002), designed only to generate fluent sentences of English. When the precision grammar alone is used for generation it often fails due to imperfect input resulting from the transfer step of our system. A fragment grammar is used as a back-off in such cases, to increase the coverage. For some TL structures however, even when the fragment grammar is used the generator can still fail due to ill-formed input structures. The decoder outputs an m-best list of TL structures, the

[6]Argument sharing can occur within deep syntactic structures and in such cases we use a simplification of the actual deep syntax graph structure by introducing the restriction that each node in the structure may only have a single mother node (with the exception of the root node which has no mother node), as this is required for the m function.

(a)



(b)

$$
\begin{aligned}
p(\text{d}) \approx \quad & p(\text{ enhance } | <\text{s}>) \\
& p(\text{ proposal } | <\text{s}> \text{ enhance }) \\
& p(\text{ the } |\text{enhance proposal }) \\
& p(</\text{s}> | \text{ proposal the }) \\
& p(\text{ safety } | <\text{s}> \text{ enhance }) \\
& p(\text{ the } | \text{ enhance safety }) \\
& p(</\text{s}> | \text{ safety the }) \\
& p(\text{ of } | \text{ enhance safety }) \\
& p(\text{ feed } | \text{ safety of }) \\
& p(</\text{s}> | \text{ of feed })
\end{aligned}
$$

Figure 3: Deep Syntax Language Model Example

content of which tends to vary a lot with respect to lexical choice. By increasing the number of structures input to the generator we can improve overall MT system coverage.

The generator is also non-deterministic, generating a k-best list of output sentences for each input TL structure. For (English) grammatical structures, the value of k is usually low, with the list containing a small number of legitimate variations in word order, and for ungrammatical or ill-formed input structures, k is usually very large, with the lists consisting of many permutations of the same words. Since the transfer decoder outputs an m-best list of structures and for each of those structures we generate k strings, the size of the n-best list for the overall MT system is therefore $m * k$.

Besides increasing coverage, by increasing the value of m, increasing either m or k (or both) also has the potential to reduce search error and result in improved MT system performance. Although the size of m can easily be changed to any desired value for the decoder (by simply changing a value in the configuration file), the generator only allows three options for deterministic versus non-deterministic generation: shortest and longest, generating either only the *shortest* or *longest* sentence with respect to number of words or *allstrings* generating all possible strings given an input structure according to the generation grammar. We refer to the three available generation options as k-options.

In the overall translation model, we include some features that are applied to the TL surface-form sentence after generation.[7] To stay true to the deep syntax approach, we do not use features that use information about the source language surface form word order. We compute a standard language model probability for the generated string and a grammaticality feature function, using information output by the generator about the grammaticality of the string. In addition, we omit scope features from f-structures for rule extraction, transfer and generation.

6 Other Features

In addition to feature functions we described thus far, we include the following additional features:[8]

- lexical translation model for source to target and target to source directions
- transfer rule size penalty (phrase penalty)
- TL node penalty (word penalty)
- fragment penalty

[7]Note that if we did not do this then many the n-best translations would be given the same score, because generation is non-deterministic.

[8]Equivalent features used in PB-SMT are in brackets.

- default transfer rule penalty[9]
- morpho-syntactic rule match feature[10]

7 Evaluation

We provide a detailed evaluation of the system to investigate effects on MT performance of using (i) different methods of word alignment, (ii) restricting the size of transfer rules by imposing different limits on the number of nodes in the LHS and RHS of transfer rules used for transferring SL structures to the TL,[11] (iii) different beam sizes during decoding, (iv) generating different sized m-best TL decoder output structure lists, and (v) different k-options for deterministic versus non-deterministic generation.

German and English Europarl (Koehn, 2005) and Newswire sentences length 5-15 words were parsed using using LFG Grammars (Kaplan et al., 2004; Riezler et al., 2002), resulting in approx. 360K parsed sentences pairs with a disambiguation model used to select the single best parse. A trigram deep syntax language model was trained on the LFG-parsed English side of the Europarl corpus, with approximately 1.26M English f-structures (again using only the single-best parse) by extracting all unigram, bigram and trigrams from the f-structures before running SRILM (Stolcke, 2002). The surface-form language model, used after generation, consisted of the English side of the Europarl, also computed using SRILM. Word alignment was run on the training data yielding an alignment between local f-structures for each f-structure pair in the bilingual training data. All transfer rules consistent with this alignment were extracted. Minimum Error Rate Training (MERT) (Och, 2003) was carried out on 1000 development sentences for each configuration using Z-MERT (Zaidan, 2009).[12]

We restrict our evaluation to short sentences (5-15 words) and use the test set of Koehn et al. (2003), which includes 1755 German-English translations.[13] We carry out automatic evaluation using the standard MT evaluation metric, Bleu (Papineni et al., 2002), in addition to a method of evaluation used to evaluate LFG

[9]When a SL word is outside the coverage of the transfer rules, it gets translated using a default rule that translates any SL word as itself (Riezler and Maxwell, 2006).

[10]For high coverage of transfer rules we allow a fuzzy match between morpho-syntactic information in the SL input structure and those of transfer rules. This feature allows the system to prefer translations constructed from transfer rules that matched the SL structure for a higher number of morpho-syntactic factors.

[11]For example, if the limit is 2, only rules with a maximum of 2 nodes in the LHS and a maximum of 2 nodes in the RHS are used for transfer.

[12]Settings for MERT training were as follows: beam=20, m=100, k=1, k-option=shortest. MERT was carried out separately for each method of word alignment. In all other experiments weights for the DS-INT configuration were used.

[13]The test set was selected on the basis that it is a commonly available test set of short sentences of German to English. Another option would have been to use short sentences from one of the WMT test sets. However, the WMT test sets only contain a relatively low number of short sentences, so instead we revert to the 2003 test set, though a little outdated, is the current best option available.

	Align. Pts.		Rules					
	Total	Ave.	Total	Ave.	Bleu	Prec.	Rec.	F sc.
SF-GDF	4.5M	12.5	2.9M	8.1	1.61	15.83	5.46	8.12
DS-GDF	4.1M	11.5	9.7M	27.1	6.04	29.13	28.17	28.64
DS-INT	2.5M	6.9	13.9M	38.8	16.18	40.31	41.25	40.78

Table 1: Effects of using different methods of word alignment. Note: rule size limit = none, beam = 100, m = 100, k = 1, k-option = shortest

parsers comparing parser-produced f-structures against gold-standard f-structures. The method extracts triples that encode labeled dependency relations, such as *subject(enhance,proposal)* and *object(enhance,safety)* for example, and triples encoding morpho-syntactic information, for example *case(proposal,nominative)* or *tense(enhance,future)*, from each parser produced f-structure and corresponding gold-standard f-structure, counting matching triples to finally compute a single precision, recall and f-score computed over the triples of the entire test set.

We evaluate the highest ranking TL decoder output f-structure with an adaptation of this method since we do not have access to gold-standard f-structures for the test set. Instead we use the next best thing, the parsed reference translations. This provides an evaluation that eliminates generator performance. Note, however, that this method of evaluation is somewhat harsh when used for the purpose of MT evaluation. Since it was designed to evaluate parser output, it does not take differences in lexical choice into account, so, for example, if the MT system produces the correct tense but a different lexical item for *enhance*, such as *tense(improve,future)*, the triple is counted as incorrect ignoring the fact that tense was in fact correct. Correct triples, in the evaluation, are those where the correct lexical choice was made by the system *and* the correct dependency relation (or morpho-syntactic information) was produced.

7.1 Results

Table 1 shows statistics and results for each word alignment method. The deep syntax intersection method of word alignment by far achieves the best result with a Bleu score of 16.18. Results drop sharply when the grow-diag-final algorithm is applied to deep syntax word alignment, with scores of 6.04 Bleu. The method of word alignment that uses the surface-form bitext corpus for word alignment achieves an extremely low score of only 1.61 Bleu.

Table 2 shows automatic evaluation results when different limits on rule size are imposed (all for the best performing alignment method DS-INT). As the limit is increased from 1 node per LHS and RHS to 7 nodes, so does the Bleu score, from 10.09 to 16.55, with a slight decrease, to 16.18, when no limit is put on the size of transfer rules. The biggest increase is seen when we compare the results when the limit is increased from 1 node (10.09 Bleu) to 2 nodes (14.94 Bleu), an increase of almost 5 percentage points absolute. In general, precision, recall and

Limit	Bleu	Prec.	Recall	F-score
1	10.09	38.67	33.89	36.12
2	14.94	41.55	39.09	40.28
3	15.85	41.50	39.93	40.70
4	16.31	41.03	40.25	40.63
5	16.14	40.75	40.50	40.62
6	15.52	40.31	40.71	40.51
7	16.55	40.46	41.03	40.74
none	16.18	40.31	41.25	40.78

Table 2: Effects of limiting transfer rule size. Note: word alignment = DS-INT, beam = 100, m = 100, k = 1, k-option = shortest

f-score also increase, as we increase the limit on transfer rule size, for example, from an f-score of 36.12 when the limit is 1 to 40.74 for a limit of 7.

Results for the system for different decoder beam sizes are shown in Table 3.[14] Results show that changing the beam size does not have a dramatic effect on the system performance. However, the difference between the highest and lowest scores is approximately half a Bleu point, which is a notable decrease in translation quality when the beam is increased from size 10 to 400. This is counter to our expectations, since with an increase in beam size we expect to observe an improvement in Bleu score since more target language f-structures are reached by the decoder search. This indicates that the model used to rank target language solutions is introducing error as some target language f-structures reached when the beam size is 400 are incorrectly ranked higher than other solutions reached when the beam size is 10. In addition, due to the extensive resources and time required to carry out minimum error rate training for the system, the same weights were used for all beam sizes (via optimization with a beam size of 100), and the particular weights may by chance be more suited to solutions reached by a beam size of 10. Further investigation is required before we can make any more general statement about what beam size might be best for f-structure transfer.

Table 4 shows automatic evaluation results for different m-best list sizes.[15] Results show that increasing the size of the m-best list of TL structures produced by the decoder, has a dramatic effect on system performance, with the largest increase in results when we increase the size of m from 1 (12.67 Bleu) to 10 (15.34 Bleu), an increase of almost 3 Bleu points absolute. Results increase again when we increase m to 100 (16.18 Bleu) and again for 1000 (16.57). We include Bleu scores for when true casing is used, and, as expected, for all configurations the Bleu score

[14]Note in this experiment that results are lower relative to other experiments because m=1, as when m is larger than the specified beam size, the decoder can increase the beam size in order to ensure enough solutions.

[15]Precision, recall and f-scores are the same for each configuration, since scores are computed on the highest ranking TL structure, which is the same in each configuration. Bleu-tc scores are for Bleu evaluation with true casing.

Beam	Bleu	Prec.	Recall	F-score
1	12.76	40.61	41.19	40.90
5	12.84	40.70	41.54	41.11
10	13.03	40.79	41.43	41.11
20	12.83	40.69	41.31	41.00
50	12.69	40.35	41.18	41.00
100	12.67	40.31	41.25	40.78
200	12.67	40.24	40.99	40.61
400	12.52	40.06	40.78	40.78

Table 3: Effects of increasing the decoder beam size. Note: word alignment = DS-INT, rule size limit = none, *m* = 1, k = 1, k-option = shortest

m-best list size	Bleu
1	12.67
10	15.24
100	16.18
1000	16.57

Table 4: Effect of increasing the size of the m-best decoder output lists. Note: word alignment = DS-INT, rule size limit = none, beam = 100, k = 1, k-option = shortest. Precision = 40.31%, recall = 41.25%, f-score = 40.78%

drops when casing is taken into account, by approximately 1 Bleu point absolute.

Table 5 shows automatic evaluation results for different generation configurations.[16] The lowest result is seen for deterministic generation with k-option *longest* (15.55), where the generator outputs the longest result, while selecting the shortest generator output string for each TL structure results in an increase to 16.18 Bleu, an increase of almost 1 Bleu point. When non-deterministic generation is used and the generator produces all TL strings for the TL input structure the score increases again to 17.29 Bleu.

[16] Precision, recall and f-scores are the same for each method, since scores are computed on the highest ranking TL structure before generation is carried out.

k-option list size	Bleu
longest	15.55
shortest	16.18
allstrings	17.29

Table 5: Deterministic versus non-deterministic generation. Note: word alignment = DS-INT, rule size limit = none, beam = 100, m = 100. Precision = 40.31, recall = 41.25 and f-score = 40.78 for three configurations.

7.2 Discussion

In the sections that follow, we provide some discussion of results observed.

7.2.1 Word Alignment

Results show that system performance varies dramatically depending on how word alignment is carried out and this is caused by each word alignment method producing different quality alignment points and constraining transfer rule extraction differently (Table 1). The best performing method, DS-INT, produces the fewest and highest quality alignment points and subsequently the best MT performance.

7.2.2 Limiting Transfer Rule Size

In general, as we increase the limit on transfer rule size (Table 2), results improve as more fluent combinations of words in TL structures are produced. Larger snippets of TL structure are also less likely to cause clashes with generation constraints. The minor decrease observed when we change from a limit of 7 to no limit on transfer rule size is probably due to a small number of erroneous transfer rules being eliminated when transfer rule size is limited.

7.2.3 Decoder Beam Size

Increasing the beam size of the heuristic search does not dramatically increase MT system performance (Table 3), with a beam size of 10 being sufficient and this is probably due to the search being highly focused on lexical choice, as it is carried out on lemmatized dependency structures with the translation of morpho-syntactic information carried out independently of decoding, using an adaptation of Factored Models.

7.2.4 M-best Decoder Output

Increasing the number of structures generated (Table 4) has a more dramatic effect. When m is increased from 1 to 10, an increase of almost 3 Bleu points absolute is observed and scores increase again when we move to 100 structures by almost 1 Bleu point. Increasing the size of m to 1000 results in an additional increase of 0.39 Bleu points absolute, but a trade-off exists as the increase in computation time required for generation by increasing m from 100 to 1000 is significant, from approximately 2.33 to 26.75 cpu minutes per test sentence.

7.2.5 Deterministic vs. Non-deterministic Generation

Allowing non-deterministic generation (Table 5) results in a significant increase in Bleu score. With respect to the trade-off in additional computation time required by non-deterministic generation, non-deterministic generation indeed is worthwhile, since the average time for generation is only increased by half a cpu minute per test sentence, from 2.33 (shortest) to 2.83 (allstrings) cpu minutes.

8 Summary

A detailed evaluation of a German-to-English SMT via deep syntactic transfer system was presented in which values of core parameters were varied to investigate effects on MT output. Experimental results show that the deep syntax intersection word alignment method achieves by far the best results for the system, with larger rule size limits also improving translation quality as estimated by Bleu. Varying the beam size does not show dramatic effects on MT performance, with a beam size of only 10 being sufficient for the transfer-based system. In addition, significant gains can be made by increasing the size of the m-best decoder output list to 100 and non-deterministic generation, however with the significant trade-off in overall translation time introduced by generating from multiple target language structures. In future work, we would like to investigate to what degree the same effects are observed when the language direction is changed to English-to-German. Translation into German would be interesting for this approach, since German has more free word order and richer morphology compared to English. However, significant adaptation of existing generation technologies for German would be required before this is possible, since generation from imperfect German f-structures is required.

References

Bojar, Ondřej and Hajič, Jan. 2008. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation at the 46th Annual Meeting of the Association for Computational Linguistics*, pages 143–146, Columbus, OH.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Dalrymple, Mary. 2001. *Lexical-Functional Grammar*. Academic Press.

Ding, Yuan and Palmer, Martha. 2006. Better Learning and Decoding for Syntax Based SMT Using PSDIG. In *Proceedings of the Association for Machine Translation in the Americas Conference 2006*.

Kaplan, Ronald M. and Bresnan, Joan. 1982. Lexical Functional Grammar, a Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 173–281.

Kaplan, Ronald M., King, Tracy Holloway and Maxwell, John T. 2002. Adapting existing grammars: the XLE experience. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Kaplan, Ronald M., Riezler, Stefan, King, Tracy Holloway, Maxwell, John T. and Vasserman, Alexander. 2004. Speed and accuracy in shallow and deep stochastic parsing. In *Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Meeting*, pages 97–104, Boston, MA.

Koehn, Philip, Och, Franz Josef and Marcu, Daniel. 2003. Statistical Phrase-based Translation. In *Proceedings of Human Language Technology - North American Chapter of the Association for Computational Linguistics Conference*, pages 48–54, Alberta, Canada.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand.

Koehn, Philipp and Hoang, Hieu. 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondrej, Constantin, Alexandra and Hoang, Evan HerbstHieu. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Och, Franz Josef. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Och, Franz Josef, Tillmann, Christoph and Ney, Hermann. 1999. Improved Alignment Models for Statistical Machine Translation. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, College Park, MD.

Papineni, Kishore, Roukos, Salim, Ward, Todd and Zhu, Wei-Jing. 2002. A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Riezler, Stefan, King, Tracy H., Kaplan, Ronald M., Crouch, Richard, Maxwell, John T. and Johnson, Mark. 2002. Parsing the Wall Street Journal Using Lexical Functional Grammar and Discriminitive Estimation Techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Philadelphia, PA.

Riezler, Stefan and Maxwell, John. 2006. Grammatical Machine Translation. In *Proceedings of Human Language Technologies and the 44th Annual Meeting of the Assciation for Computational Linguistics*, pages 248–255, New York City, NY.

Sgall, Petr, Hajicova, Eva and Panevova, Jarmilla. 1986. *The Meaning of the Sentence and its Semantic and Pragmatic Aspects*. Dordrecht: Reidel and Prague: Academia.

Stolcke, Andreas. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Zaidan, Omar. 2009. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation* 91, 79–88.

FROM DEPENDENCY STRUCTURES TO LFG REPRESENTATIONS

Dag Haug
University of Oslo

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

In this paper, we present the conversion of the PROIEL dependency treebank into LFG representations and the algorithms that were used in converting dependency structures to f- and c-structures. The source corpus has a large amount of non-projective edges, and the conversion to c-structure goes beyond previous work in providing principled representations of non-projective structures.

1 Introduction

When creating a treebank, it is necessary to select an appropriate formalism to express the annotation in. While it might seem obvious to simply pick your favoured linguistic theory, this is in practice rarely what happens. Instead, treebanks are usually expressed in either dependency grammar (DG), which despite some important work (Tesnière, 1959; Sgall et al., 1986; Mel'čuk, 1988; Hudson, 2007), has never really been developed as a unified linguistic theory; or in some phrase structure-based formalism that typically uses much flatter phrase structures than those assumed by linguists who use phrase structure-based paradigms as their theoretical framework.

There are several good reasons why this is so. One practical concern may be that it is hard to get annotators with the appropriate training for performing more theoretically motivated annotation. Second, corpus annotation inevitably needs to deal with constructions that linguistic theory has not developed analyses of, whether because they are thought linguistically uninteresting (e.g. calendar expressions) or simply because they have gone unnoticed in the literature (especially if one is dealing with a less-studied language). So a theoretically motivated corpus will require much theoretical work before the annotation can start. Third, creating a treebank is a very time-consuming task, and it is desirable that the result should be accessible to as many users as possible. The more theoretically motivated the treebank is, the less it is likely to be accessible to 'outsiders' who might, for example, have difficulties in navigating the AVMs of pure HPSG or LFG-based corpora.

But while there are good reasons why treebanks avoid theory-driven representations, this can lead to a gap between corpus linguistics and linguistic theory (Frank, 2001). One of the main uses that a treebank can offer would seem to be the testing of linguistic theories and analyses against 'real data', not just intuitions. But if the raw data that a corpus search yields is simply not compatible with the theory to be tested, hypothesis testing can be more difficult.

The solution to this dilemma is to annotate and store corpora using simplified representations, but to take care that it is possible to enrich these representations to proper, theoretically motivated structures – ideally, to different structures motivated by several different theoretical frameworks.

[†]We thank Mary Dalrymple for hints about the earlier literature, the audience at LFG12 for useful comments and Tracy Holloway King for editorial suggestions.

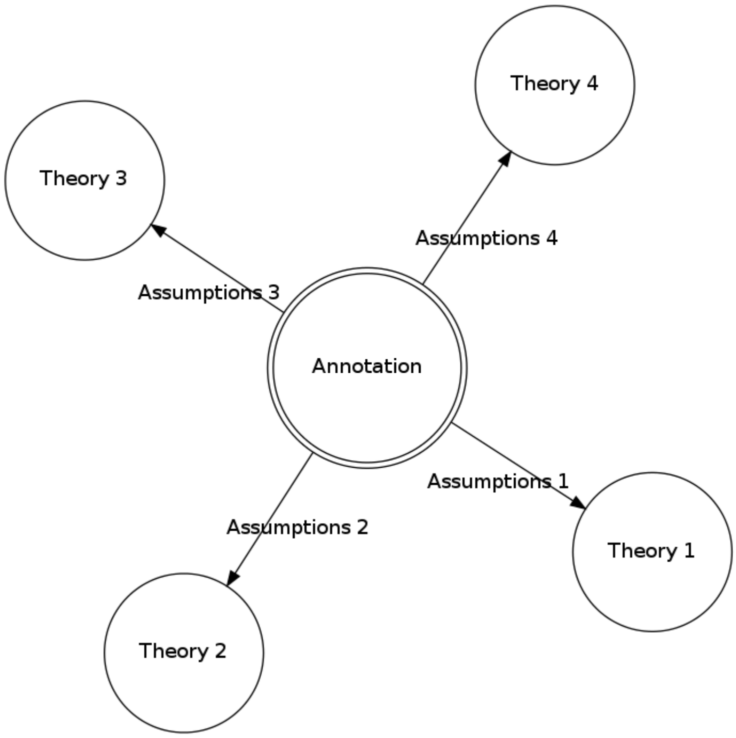


Figure 1: Linguistic annotation and theories

1.1 Theory-neutral representation

We are therefore aiming at a ‘theory-neutral’ representation. By theory-neutral annotation, we understand annotation that respects at least the first, and possibly the second of the following two constraints.

1. Encode enough structure to allow the reconstruction of theoretically motivated structures in the target frameworks

2. Encode no more structure than is common to the target frameworks (including structure that in some frameworks are seen as derived/secondary)

Of course, there are limits to such theory-neutrality: some theories might simply be too different. But in the ideal situation it will be possible to reconstruct full theoretical representations by supplementing the information in the corpus with the specific assumptions of each target theory. If constraint 2 is not violated, it will be possible to do this by monotonically adding information. The situation is summed up in figure 1.

When generating phrase structures from a corpus that does not have them, which is the most challenging problem in creating LFG representations out of a dependency treebank, the added theoretical assumptions will typically concern such

language	source	nonprojective	projective	% nonproj
Latin	Gallic War	1547	19086	8.1%
	Letters to Atticus	2269	22693	10.0%
	Vulgate NT	2721	65671	4.1%
	Per. Aeth.	1279	14890	8.6%
Greek	Herodotus	4137	33522	12.3%
	NT	3997	94028	4.3%
OCS	Marianus NT	1828	47719	3.8%
	Zographensis NT	26	701	3.7%
	Suprasliensis	327	6068	5.4%
Gothic	NT	1886	46884	4.0%
Armenian	NT	409	18063	2.3%
	Koriwn	48	1539	3.1%

Table 1: Projectivity in the PROIEL corpus

things as categories and X′-theory. In fact, converting a corpus can be seen as hypothesis testing: the conversion will only succeed if it is in fact possible to convert the dependency structures into phrase structures that accord with our assumptions, and failure will indicate that the data falsifies our assumptions about phrase structure.[1]

Our approach to annotation entails a different strategy for conversion than what is found in much other work. The information added during conversion is intended to embody assumptions of linguistic theories, not to be guesses about information that is lacking in the source data. With few exceptions, we are only enriching the annotation to the extent that this can be done in a deterministic way; we do not attempt to make up for missing information in the source through heuristics.

1.2 The source corpus

The corpus to be converted in this experiment is the PROIEL corpus, which consists of 447,008 words in Greek, Latin, Gothic, Armenian and Old Church Slavic (OCS). The core of the corpus is made up by translations of the New Testament in these languages (or the original, in the case of Greek), and there are also some other Greek and Latin texts. All these languages are morphologically rich, dependent-marking, non-configurational languages with a high-degree of non-projectivity, although this varies a lot between texts. As can be seen in table 1, the non-biblical texts have very high non-projectivity rates, in one case exceeding 10%. This makes the creation of motivated c-structures an interesting challenge.

The corpus is annotated with dependency structures, i.e. labelled, asymmetric relations between words, but with some changes from what is normal in DG. There

[1] It should be noted that manually annotated data often has errors and inconsitencies in it, which can also create problems for the conversion. Such annotation mistakes can and should be corrected on the source side rather than being dealt with in the conversion.

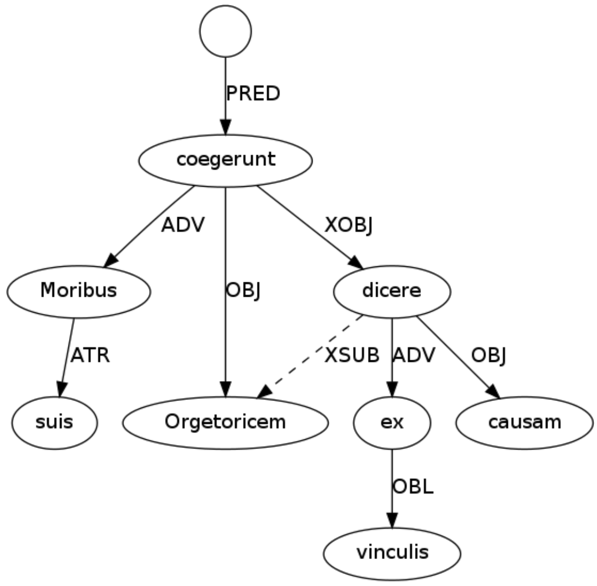


Figure 2: Object control (1)

are two problems in particular that led us to deviate from standard DG, namely structure-sharing and ellipsis.

Structure-sharing phenomena are problematic for DG because of the unique head principle which says that each word has exactly one head, ensuring that the dependency graph is in fact a tree. In the PROIEL corpus, this constraint is respected in the primary dependency graph, but structure-shared elements are related to their second head via secondary edges, as in the object control example in (1) with the associated dependency graph in figure 2.

(1) Moribus suis Orgetoricem ex
custom.PL.M.ABL own.3.PL.M.ABL Orgetorix.SG.M.ACC from
vinculis causam dicere coegerunt
chain.PL.N.ABL cause.SG.F.ACC say.PRES.INF.ACT force.3.PL.PRF.ACT
'Following their customs, they forced Orgetorix to speak his cause from chains.' (Caes. Gal. 1.4.1)

Orgetoricem is both the object of the matrix verb *coegerunt* and the subject of the embedded infinitive *dicere*. Only the first function is captured in the primary dependency graph whereas the second function is represented via the dotted secondary edge labelled XSUB (external subject). Notice that the same annotation is used for functional and anaphoric control. Also no distinction is made between raising and control, i.e. whether the higher position is thematic or not. The reason is that there is simply no reliable answers to such questions for our languages at the moment. On the other hand, the annotation does distinguish between cases like (1) and those where the accusative does not have a grammatical function in the matrix clause, i.e. accusatives with infinitives. In these cases, the accusative is made the SUBJ of the infinitive (which is COMP), without any structure-sharing.

Ellipsis is problematic for DG because DG normally relies on the words of the sentence to make up the nodes of the dependency tree. But when there is ellipsis, there is structure without any word. Consider (2).

(2) partes tres quarum unam
parts.PL.F.ACC three.F.ACC who.PL.F.GEN one.SG.F.ACC
incolunt Belgae aliam
inhabit.3.PL.PRES.ACT Belgian.PL.M.NOM other.SG.F.ACC
Aquitani tertiam qui ipsorum
Aquitani.PL.M.NOM third.SG.F.ACC who.PL.M.NOM himself.3.PL.M.GEN
lingua Celtae nostra Galli
language.SG.F.ABL Celt.PL.M.NOM our.SG.F.ABL Gauls.PL.M.NOM
appellantur
call.3.PL.PRES.PAS
'three parts, of which the Belgians inhabit one, the Aquitani another, and those called Celts in their own language and Gauls in ours the third.' (Caes. Gal. 1.1.1)

There are three coordinated clauses sharing the relative pronoun *quarum*, but only the first has an overt verb, *incolunt*. The others two clauses contain ellipted instances of the same verb, but this is not easy to capture in DG – there is no node of which *Aquitani* can be the subject. And there is no overt conjunction coordinating the three clauses. Our solution is to use empty nodes in such cases, as shown in figure 3. The empty nodes are typed (V for ellipted verb and C for ellipted conjunction) and bear an arbitrary, unique identifier. Via secondary edges labeled PID (predicate identity), we can also capture the fact that the empty verbal node instantiates *incolunt*.

Notice finally that this solution is in principle equivalent to that adopted in many DG corpora (notably the Prague Dependency Treebanks and corpora inspired by these) where 'invisible structure' is captured in the labels instead. We can remove empty nodes and instead make them dependents of their empty head's own head and label the resulting dependencies with the concatenation of the original label, a unique ID for the empty node, and the label that the empty node bears to its own head in the origin structure. This yields the structure in figure 4. But such structures are more difficult to work with for annotators.

The labels used in the corpus and their LFG equivalents are shown in table 2. As we can see, the labels have been chosen to match those of LFG. Still, there are a couple of differences. First, the annotation scheme does not attempt to distinguish between OBJ_θ/OBL, as that distinction is hard to draw for annotators.[2] Second,

[2]Unless it were reduced to a category difference, where all PPs are OBL and all non-accusative argument NPs are OBJ_θ – but that distinction is in any case retrievable from the morphological anno-

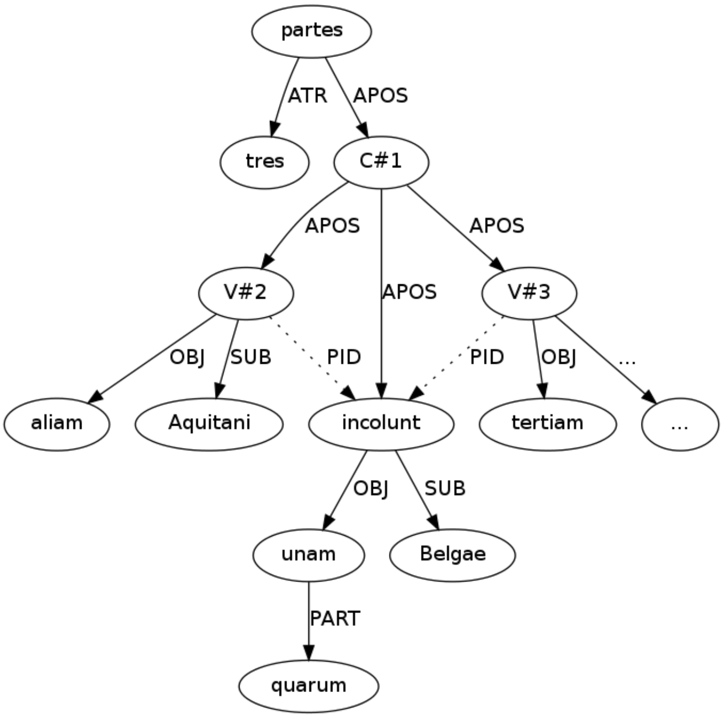


Figure 3: Ellipsis (2)

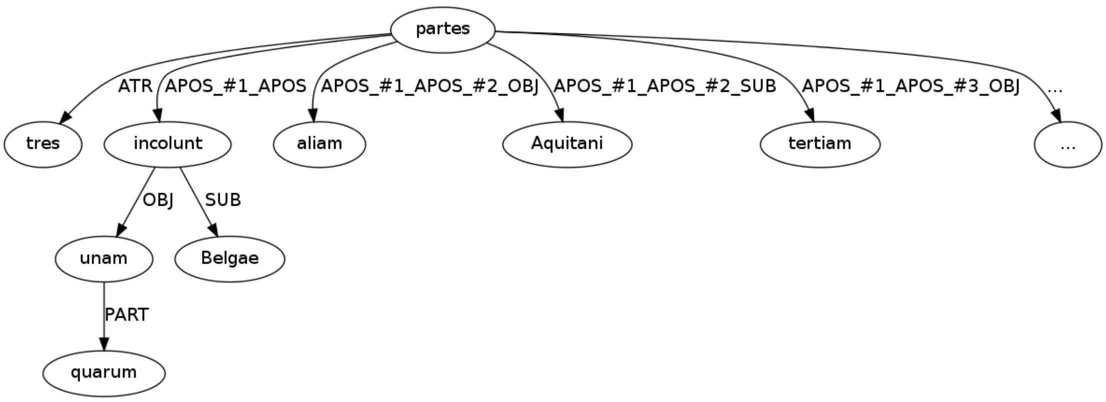


Figure 4: Ellipsis without empty nodes

Function	Label	LFG	Function	Label	LFG
Adverbial	ADV	ADJ	Oblique	OBL	OBJ_{θ}/OBL
Agent	AG	OBL_{AG}	Parenthetical	PARPRED	—
Apposition	APOS	ADJ	Partitive	PART	ADJ
Attribute	ATR	ADJ	Predicate	PRED	—
Auxiliary	AUX	—	Subject	SUB	SUBJ
Complement	COMP	COMP	Vocative	VOC	—
Argument of noun	NARG	OBL	Free predicative	XADV	XADJ
Object	OBJ	OBJ	Open complement	XOBJ	XCOMP

Table 2: Labels in the PROIEL corpus and their LFG equivalents

vocatives (VOC) and parenthetical predications (PARPRED) have no direct counterparts in LFG. In the conversion, these were simply ignored.[3] Predicate (PRED) and auxiliary (AUX) are used in the PROIEL corpus for functions that do not introduce an embedded layer of f-structure, but simply contribute the features of the dependent in the f-structure of the head. We will see how this works in more detail in section 2.

2 Converting to f-structures

F-structures and dependency graphs both encode labelled syntactic dependencies, so conversion is not very difficult. There is also previous work in the LFG tradition that we can lean on, notably Forst (2003).

There are two major differences between f-structures and dependency graphs that conversions typically need to deal with. First, as observed, LFG's structure-sharing runs against DG's unique head principle. This is not a problem for our conversion, since the source corpus already captures structure-sharing through secondary edges. Second, in DG every word introduces depth in the graph, whereas in LFG multiple words can contribute to the same f-structure without nesting. Here, the source corpus uses AUX and PRED for such words, so they are identifiable.

The conversion proceeds by mapping the morphological analysis of each token to a set of features and (unless the token bears the AUX function) its lemma to a semantic form. The subcategorization template for the semantic form is simply retrieved from the argument daughters in the source graph + a subject if none is present, to account for subject pro-drop.[4] This gives us f-structures for each token. In the next step, the f-structure of each token is either made the value of the LFG-equivalent of its function in the f-structure of its head (if the function is not AUX or

tation.

[3]Another simple solution would have been to generate separate f-structures for parenthetical predications.

[4]Before this is done, the lemma is checked against a stop list of impersonal words. Note that no attempt is made to account for pro-dropped objects and obliques, which are much rarer.

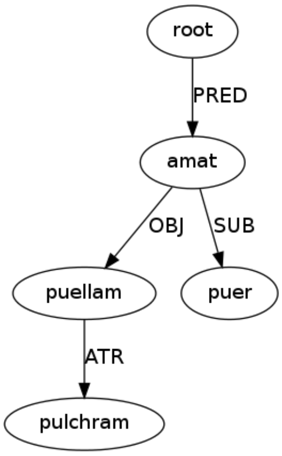


Figure 5: Dependency graph for (3)

PRED), or simply unifies with the f-structure of its head (if its function is AUX or PRED). Consider the simple example in (3).

(3) puer amat puellam pulchram
boy.SG.M.NOM love.3.SG.PRES girl.SG.F.ACC beautiful.SG.F.ACC
'The boy loves the beautiful girl.'

The dependency graph is given in figure 5. The f-structures (omitting NUMBER for the sake of readability) the of the words are as in (4).

(4)

root $\,_{r}\left[\;\right]$

puer $\,_{b}\begin{bmatrix} \text{PRED} & \text{'BOY'} \\ \text{CASE} & \text{NOM} \\ \text{GEND} & \text{MASC} \end{bmatrix}$

amat $\,_{l}\begin{bmatrix} \text{PRED} & \text{'LOVE}\ \langle\text{SUBJ, OBJ}\rangle \\ \text{PERSON} & 3 \\ \text{TENSE} & \text{PRES} \end{bmatrix}$

puellam $\,_{g}\begin{bmatrix} \text{PRED} & \text{'GIRL} \\ \text{CASE} & \text{ACC} \\ \text{GEND} & \text{FEM} \end{bmatrix}$

pulchram $\,_{n}\begin{bmatrix} \text{PRED} & \text{'BEAUTIFUL'} \\ \text{CASE} & \text{ACC} \\ \text{GEND} & \text{FEM} \end{bmatrix}$

In step two, then, n is made the value of the ADJ function in g (embedded inside a set, since we know that (X)ADJ is set-valued), g becomes the value of OBJ in l, b becomes the value of SUBJ in l, and l and r unify, since *amat* has the PRED function. This yields (5).

(5)
$$\begin{bmatrix} \text{PRED} & \text{`LOVE}\ \langle\text{SUBJ, OBJ}\rangle\text{'} \\ \text{TENSE} & \text{PRES} \\ \text{PERS} & 3 \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`BOY'} \\ \text{CASE} & \text{NOM} \\ \text{GEND} & \text{MASC} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`GIRL'} \\ \text{CASE} & \text{ACC} \\ \text{GEND} & \text{FEM} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{`BEAUTIFUL'} \\ \text{CASE} & \text{ACC} \\ \text{GEND} & \text{FEM} \end{bmatrix} \right\} \end{bmatrix} \end{bmatrix}$$

There are a couple of things to notice about the generated f-structures. First, it follows from the method that words can only contribute features to their own f-structure or that of their head – but in the latter case, they contribute *all* their features to the head. This means that it is impossible for the output structures to represent LFG's traditional account of agreement as cospecification of features in one f-structure: instead, *pulchram* bears the agreement features CASE and GEND in its own f-structure. This is not necessarily wrong (and indeed Haug and Nikitina (this volume) argue that such a theory of agreement is needed for Latin). Second, the f-structure is generated separately from the c-structure, which means that the c-structure cannot influence the f-structure. In other words, functions that are typically assigned configurationally, such as TOPIC and FOCUS cannot be accounted for.

3 Converting to c-structures

Unlike f-structures, c-structures contain information that is very different from that found in a dependency graph. This part of the conversion is therefore much more difficult, but also more interesting. As far as I know, there is no LFG work on inferring c-structures from dependencies, but there is more general work on the relationship between phrase structure grammars and dependency grammars going back to at least Gaifman (1965).

Instead of syntactic dependencies, c-structures contain information about word order, category and constituency. None of these need be present in a dependency tree. However, we can reconstruct the linear order by referring to the original string (barring tokenization differences), and in any case nodes will often bear an ordering that lets us reconstruct the original word order. Categorical information is also not too difficult to retrieve, since treebanks typically have morphological annotation.[5] In the following, we will simply assume that the words of the dependency graph are marked for their category. Finally, and this is the most crucial point, constituency and dependency are of course related. The exact relationship is the topic of discussion, which makes the conversion interesting also from a theoretical perspective.

Intuitively, we can look at constituency as combining information about dependencies and word order, i.e. a constituent is a continuous domain of words related by dependency relations. Generally, the phrasal head is also the dependency head of any words inside its phrase, but there are some common mismatches. For example, dependency analyses often (but by no means always) make functional elements such as determiners and auxiliary verbs dependents of their associated lexical element. Such differences are not really due to the difference between the formalisms, but rather to different and sometimes controversial analyses. For example, some phrase structure grammars assume that determiners take their noun as a complement, others that the determiner occurs in the specifier position of the nominal projection. We therefore consider these alterations not to be part of the conversion to c-structure. Instead they are performed in a separate step of preprocessing, where auxiliary verbs are made the head of lexical verbs (but articles remain dependents of their nouns).

In the following we will describe our conversion algorithm. We will not attempt a full-fledged formalization of the linguistic structures and the conversion between them in this context, but we will be explicit enough for it to be possible to see that the algorithm is sound and that it does not lose information, i.e. it is reversible.

3.1 What's in a dependency structure

Words are the cornerstones of our structures: they make up to nodes of our dependency graphs and the terminals of our phrase structure trees. We want the same elements to serve in both structures. More concretely, we will assume that words (and other terminals) are tuples $\langle w, i, c, r, t \rangle$ where w is the form of the word, i is the index (surface string position), c is the category, r is the syntactic function and t is a boolean flag indicating whether the 'word' is a trace. Traces will seem suspicious from an LFG perspective and we will in fact create c-structures that are trace-free, but as we will see, traces are still useful in the intermediate representations between DGs and c-structures. Notice also that we will make no use of the secondary edges in the PROIEL DGs in the conversion procedure, as we want the procedure to be applicable more generally to other corpora. In other words, the input dependency structures will respect the unique head principle. In addition, we

[5] Sometimes, the syntactic function is also necessary for category inference. For example, we assume that adjectives bearing nominal functions such as SUBJ, OBJ etc. have been nominalized and therefore have category N rather than A.

will assume that there is always a single root word, i.e. a unique word that does not have a head. Since we have ignored vocatives and parenthetical predications, this assumption holds good in the source corpus, as it does in most dependency corpora.

Dependency structures, then, will be tuples $\langle \mathcal{W}, r, f\rangle$ where $\mathcal{W}$ is the set of words, $r(\in \mathcal{W})$ is the root and f is a function $\mathcal{W} \setminus \{r\} \mapsto \mathcal{W}$ taking dependents to their heads, such that f forms a tree over $\mathcal{W}$ rooted in r. Notice that while we normally think of the labels in a dependency structure as attaching to the edges of the graph, we have here assumed that the labels attach to words. Because of the unique head principle, this makes no difference.

3.2 Order domain structures

We now introduce the notion of the order domains, a concept we have adapted from Bröker (1998), although it plays a different role in our system. The order domain $\mathcal{D}(w)$ of a node w is the largest subset of $\mathcal{W}$ such that $w \in \mathcal{D}(w)$, all words in $\mathcal{D}(w)$ are dominated (either directly or via nodes that are also in $\mathcal{D}(w)$) by w and $\mathcal{D}(w)$ is continuous, i.e. for any two words in $\mathcal{D}(w)$, all words in between are also contained in $\mathcal{D}(w)$. We will call w the head of the order domain $\mathcal{D}(w)$.

In other words the order domain $\mathcal{D}(w)$ contains w itself as well as all those of its (direct and indirect) dependents that in an intuitive sense are not 'displaced'. This is not far from the concept of a constituent as a continuous domain of words related by dependency relations.

Let us call the set $\mathcal{O}$ of order domains of all the words in a sentence an *order domain structure*. Set inclusion is a partial order on $\mathcal{O}$ and it is easy to see that $\langle \mathcal{O}, \subseteq\rangle$ is a join semi-lattice whose top is $\mathcal{W}$, which is the order domain of r.

Furthermore, for all order domains $\mathcal{D}(w_1), \mathcal{D}(w_2), \mathcal{D}(w_3)$, if $\mathcal{D}(w_1)$ and $\mathcal{D}(w_2)$ are both supersets of the order domain $\mathcal{D}(w_3)$, then by the definition of order domains, w_1 and w_2 must both dominate w_3 in the dependency structure. Hence, since the dependency structure is a tree, either w_1 dominates w_2 or vice versa. Moreover, by the continuity of order domains, all nodes between w_1 and w_3 must be dominated by w_1 and similarly for w_2. It follows that either $\mathcal{D}(w_1) \subseteq \mathcal{D}(w_2)$ or $\mathcal{D}(w_2) \subseteq \mathcal{D}(w_1)$. This means that the order domain structure is always not just a semi-lattice, but in fact a tree in the graph-theoretical sense. Moreover, since the order domains are always continuous, there are no crossing branches in our tree. That is, order domain structures are the same kind of structures as context-free phrase structure trees over nodes that, as we noted above, are very close in conception to constituents. As such, it provides useful intermediate structure between dependency graphs and phrase structures.

Consider the constructed Latin example in (6), which has the dependency graph shown in figure 6.

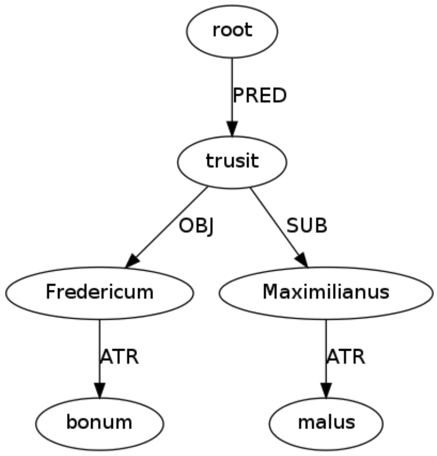


Figure 6: Dependency graph for (6) and (10)

(6) malus Maximilianus trusit bonum Fredericum
bad.SG.NOM.M Max.SG.NOM.M push.SG.NOM.M good.SG.ACC.M Fred.SG.ACC.M
'Bad Max pushed nice Fred.'

malus	Maximilianus	trusit	bonum	Fredericum
bad.SG.NOM.M	Max.SG.NOM.M	push.SG.NOM.M	good.SG.ACC.M	Fred.SG.ACC.M

The order domains of the words in (6) are given in (7).

(7)

malus	{malus}
Maximilianus	{malus,Maximilianus}
trusit	{malus,Maximilianus,bonum,trusit,Fredericum}
bonum	{bonum}
Fredericum	{Fredericum,bonum}

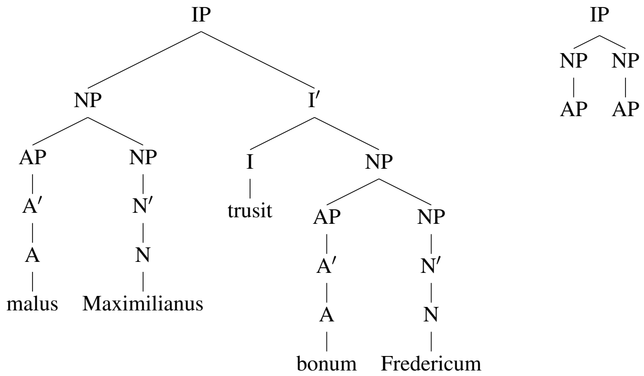
If we order these sets by set inclusion, we get the order domain structure in (8).

(8) {malus,Maximilianus,bonum,trusit,Fredericum}

{malus,Maximilianus} {Fredericum,bonum}

{malus} {bonum}

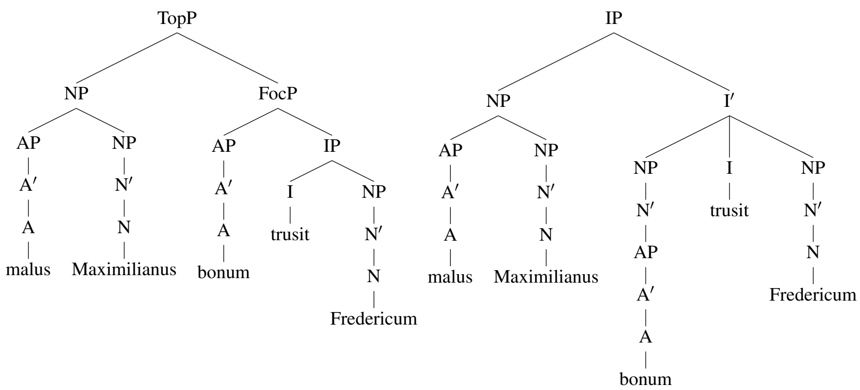
Consider now a possible phrase structure tree for (6), along with a reduced version of this tree, containing only the (uppermost, in the case of adjunction-induced iterations) maximal projections:

(9) 

We observe that the order domain structure is isomorphic to the reduced tree containing only maximal projections. Consider now (10), which gives the same example slightly altered so as to have a discontinuous object NP.

(10) malus Maximilianus bonum trusit
bad.SG.NOM.M Max.SG.NOM.M good.SG.ACC.M push.SG.NOM.M
Fredericum
Fred.SG.ACC.M
'Bad Max pushed nice Fred.'

The dependency tree remains the same, although the indices on the words have changed. It is a theoretically disputed matter how to best represent the phrase structure of an example like (10). Two possible analyses are shown in (11). To the left is a minimalist-style analysis where the discontinuity is accounted for by assuming that the displaced adjective has moved to the specifier of a functional projection (FocP).[6] To the right is an LFG-style style analysis[7] which assumes that the displaced adjective adjoins to a headless object NP which is unified with the other object NP at f-structure.

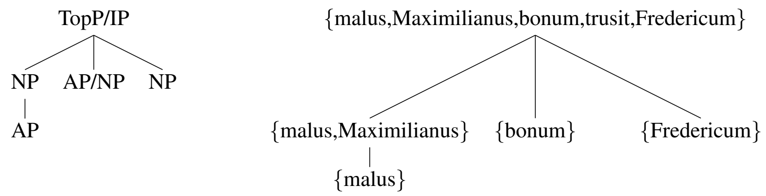
(11) 

However, if we only consider the uppermost maximal projections, both these trees

[6]For an analysis of discontinuities in Latin along these lines, see Devine and Stephens (2006).
[7]This is not to say that LFG *has to* assume the rightmost structure rather than the leftmost.

are isomorphic with each other and with the order domain structure, which are both given in (12).

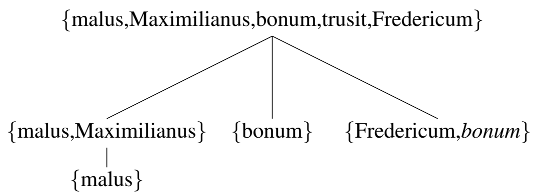
(12)



In a sense, then, the order domain structure represents phrase structure while abstracting away from the internal structure of projections. The internal structure of projections is of course very important in phrase structure grammars, since it determines things such as c-command relations etc. But on the other hand, assumptions about the internal structure of phrases is also where different theories differ most. Therefore, an order domain structure is as close as we can get to a proper phrase structure representation without making theory-internal assumptions. To get to proper phrase structure representations such as those in (11), we need to add those assumptions.

Before we go on to see how that can be done, we need to fix a problem with the order domain structure in (12). Although this structure can be derived from the dependency graph, there is in fact no way to go back from the order domain structure. The problem is that there is no way to retrieve the dependency of *bonum* on *Federicum*. We will solve this by enriching our order domain structures with traces. In addition to the set of nodes defined above, an order domain structure will consist of traces of all nodes that are dominated (directly or indirectly) by a word but are not in its order domain according to the previous definition. The partial order that defines the tree structure of an order domain structure will be the subset relation modulo traces. This yields (13), where traces appear as words in italics.

(13)



Now we can retrieve the dependency tree by ordering the order domains by a subset relation $\subseteq_t$ that consider traces and their overt realizations as identical.

3.3 Adding linguistic knowledge

To take order domain structures to phrase structures there are two steps we must accomplish. First, we must create a projection corresponding to each order domain. Second, we must embed these phrases in each other. To do this, we must know the rules for creating projections and combining them.

Xia and Palmer (2001) identified three questions that any conversion from dependencies to phrase structures must answer:

1. For a category X, what kind of projections can X have?
2. If a category Y depends on a category X in a dependency structure, how far should Y project before it attaches to Xs projection?
3. If a category Y depends on a category X in a dependency structure, to what position on X's projection chain should Y's projection attach?

Our answers to these questions are guided by X′-theory and LFG's approach to discontinuities.

1. All categories X project two levels X′ and XP. However, if the phrase is displaced, it will be embedded inside one or more headless projections corresponding to the path to its functional heads.
2. A dependent Y always projects to Y′ then YP and the YP attaches to the head's projection
3. Dependents are divided into three types using a set of handwritten rules: specifiers, modifiers and arguments. Specifiers are made sisters of X′ and arguments are made sisters of X. Modifiers Chomsky-adjoin to either X′ or XP depending on whether they are restrictive, as indicated by the dependency edge label (ATR or APOS).

The conversion of the order domain structure starts from the root, but recursively converts daughter order domains before mothers. For each order domain we first create a projection according to these rules, i.e. an X – X′ – XP spine, where X is the category of the order domain's head.[8] Next, if there are order domains in the tree that contain a trace of the order domain's head, we order these by $\subseteq_t$ and embed the original projection successively inside headless structures Y′ – YP, where Y is the category of the head of the order domain containing the trace.

Once the full projections of any daughter nodes in the order domain structure have been created, as well as the X – X′ – XP spine of the current node, it is necessary to embed the daughter projections in correct positions in the structure. We rely on hand-written rules to do this. A sample rule for nominal projections is given in table 3.

The rule that governs the relationship between position and grammatical function (e.g. phrase adjuncts should be non-restrictive (APOS) etc.) are for the moment hard-coded in the conversion program, so the handwritten rules only deal with restrictions on categories. The rules are tested from the 'outside in', i.e. the outermost dependents are tested for whether they can be phrasal adjuncts; if yes, the next outermost dependents are also tested and so on; when the test fails, it moves to check

[8]The empty nodes used in ellipsis and asyndetic coordination are represented as phrases without heads.

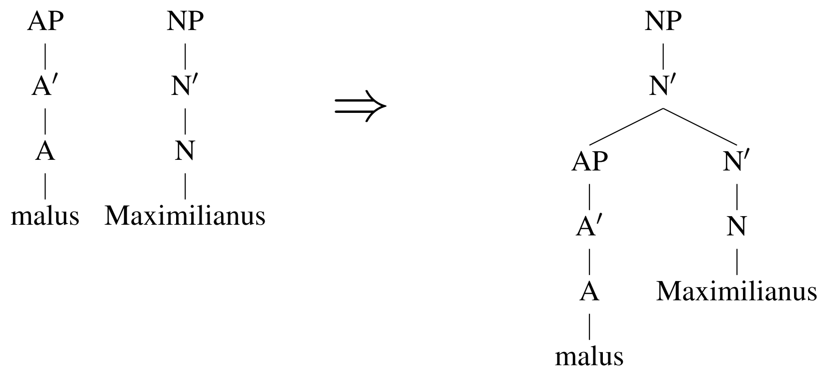
N:		
	:phrase_adjuncts:	NP, AP
	:specifier:	DP
	:bar_adjuncts:	NP, AP
	:complements:	NP, PP, AdvP, AP, CP, IP, VinfP, VptcpP

Table 3: Rules for nominal projections

the next leftmost dependent for specifier-hood; and so on, to bar-level adjunction and complementhood.

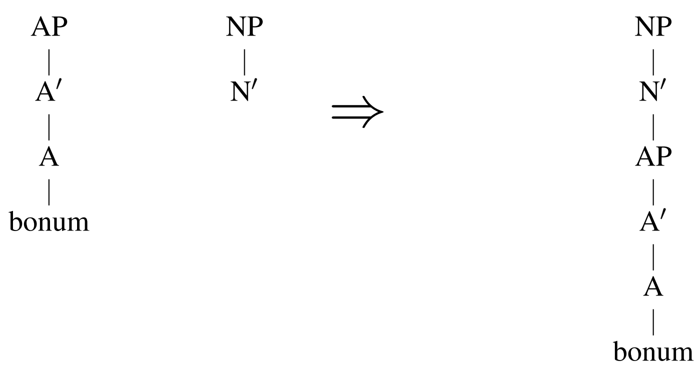
Let us see how this works for (13). In order to create a phrase structure for the top node, we must create projections from the daughter nodes. We start from the left. To create a projection for {malus, Maximilianus} we must create one for {malus} and attach it correctly in the one for {Maximilianus}. This is shown in (14). The two starting projections are determined by our X′-theoretic assumptions, and the correct combination is given by the rules in 3.

(14)



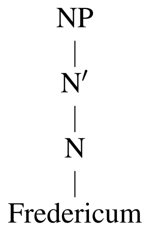
Next, we must construct a projection for *bonum*. Since there is a trace corresponding to *bonum*, we must also create a headless projection for the (possibly multiple, but in this case only one) order domain containing the trace, and embed the first one in the second. This is shown in (15).

(15)



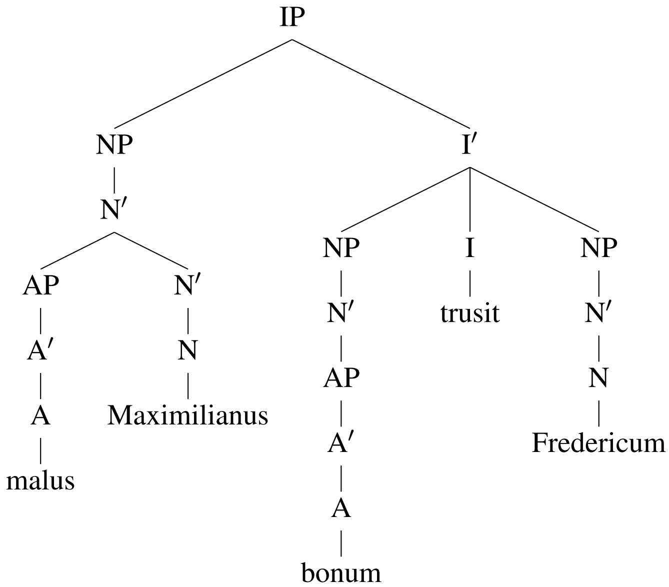
Finally, we create a projection for *Fredericum*. Since this order domain contains only one non-trace element[9] and there is no trace corresponding to *Fredericum*, this is simple:

(16)



Once we have created the projections in (14)–(16), we need to attach them in the projection of the topmost order domain, which is an IP – I′ – I spine. The result is shown in (17).

(17)



So we have arrived at a c-structure which can be motivated within LFG. Observe that in itself, this structure cannot be reverted to the order domain structure, because again there is no way to retrieve the dependency of *bonum* on *Federicum*, which should be represented by a trace in the order domain structure. For our purposes, this is not a problem, since we also generate an f-structure, from which we can see that *bonum* belongs to *Federicum*. For other applications where only c-structures are generated, other options must be considered. One, corresponding to a principles and parameters approach, would be to represent the trace in the order domain as a trace in the phrase structure. Another option is to index maximal projections with the index of the word whose projection it is. This would leave both NPs inside I′

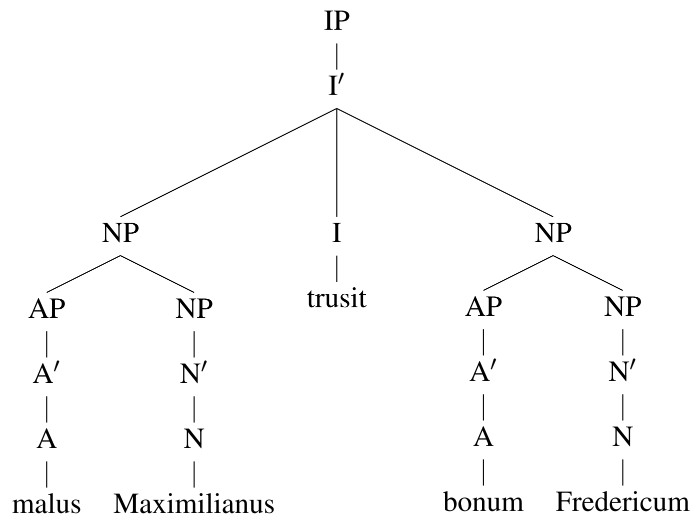
[9]Notice that if we wanted to create constituent structures with traces in them, we could treat the trace of *bonum* as projecting an empty category.

coindexed. For c-structures without such discontinuities, however, the reversion to an order domain is straightforwardly achieved by substituting for each maximal projection the set of terminals it dominates and removing all other nodes.

3.4 Evaluation

The conversion algorithm for c-structures is completely rule-based, and as such it performs no better than the rules it is fed with. But notice that for the most part, the rules do not really have the character of heuristics in the sense that they would guess the most likely alternative. Instead, they are intended as hard constraints embodying linguistic knowledge. The only exception to this is the fact that the rules are always tested from the outside in. This means that if the leftmost dependent is admissible both as a specifier and as a complement, it will always be made a specifier. The reason for this choice is that linguistic theories predict that there are elements such as *wh*-words which have to occur in specifier positions, while with other words it is often hard to see whether they are in a specifier position or not, as the semantic effects of topicalization are vague. Consider (8) again. If we assume that Latin has a non-obligatory specifier position which serves to indicate topicality (and *not* subjecthood), we could equally well assume a phrase structure as in (18).

(18)



Considerations of information structure would tell us that (18) is perhaps be more natural in an ‘all new’ context (answering *What happened?* rather than *What did Maximilian do?*). But such constraints are too vague to be of any use in conversion. This points to a more general problem with evaluating phrase structures, especially when working with ancient languages where the word order is ill understood: there is often no real gold standard to be had.

However, there is another aspect under which our conversion algorithm can be evaluated, namely its preservation of information. The algorithm does not lose linguistic information. There is no room for a formal proof here, but we have already hinted that the order domain structures, once equipped with traces, can be reverted to dependency structures. It is also possible to revert phrase structures to order domain structures.

4 Conclusion

We have seen that it is possible to convert a dependency-annotated corpora to full LFG representations, provided the original annotation is rich enough. Although dependency structures and functional structures encode very similiar kinds of information, which facilitates the conversion, this is also the part which requires the most divergences from a strict phrase structure format. The reason is that dependency grammar (at least in its usual strict form) does not allow structure-sharing. Correct transformation of the data could only be achieved because the dependency format used in the PROIEL source corpus has been extended with structure-sharing.

The conversion from dependency structures to c-structures, on the other hand, is more complicated and challenging both from a technical and a theoretical point of view. On the other hand, it does not require information beyond what is found in normal dependency structures. The conversion algorithm goes beyond previous work in conversion between dependencies and phrase structures in that it deals with non-projectivity in a principled manner, generating structures that are compatible with LFG's treatment of discontinuities.

As we noted, there are difficulties in providing an exact evaluation of the c-structure conversion, since there is room for much disagreement on what a proper c-structure of these old languages should like. Nevertheless, it is an important feature that the algorithm is reversible and does not lose linguistic information.

Finally, we hope that the converted corpora will be of use in future development of LFG grammars for these languages. As mentioned above, their phrase structure is not well understood. It is to be hoped that the converted PROIEL corpus will be a valuable resource and help towards a better understanding of both phrase structure and other aspects of the grammar of these languages.

References

Bröker, Norbert. 1998. A projection architecture for dependency grammar and how it compares to LFG. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG98 Conference*, Stanford: CSLI Publications.

Devine, A. M. and Stephens, L. D. 2006. *Latin Word Order: Structured Meaning and Information*. Oxford: Oxford University Press.

Forst, Martin. 2003. Treebank conversion – creating a German f-structure bank from the TIGER corpus. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG03*, CSLI Publications.

Frank, Anette. 2001. Treebank conversion - converting the NEGRA treebank to an LTAG grammar. In *Proceedings of the Workshop on Multi-layer Corpus-based Analysis, Iasi*, pages 29–43.

Gaifman, Haim. 1965. Dependency systems and phrase-structure systems. *Information and Control* 8(3), 304–337.

Hudson, Richard. 2007. *Language Networks: The New Word Grammar*. Oxford University Press.

Mel'čuk, Igor. 1988. *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press.

Sgall, Petr, Hajičová, Eva and Panevová, Jarmila. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Prague, Czech Republic/Dordrecht, Netherlands: Academia/Reidel Publishing Company.

Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Paris: Klincksieck.

Xia, Fei and Palmer, Martha. 2001. Converting dependency structures to phrase structures. In *Proceedings of the first international conference on human language technology research*, HLT '01, pages 1–5, Stroudsburg, PA, USA.

THE MANY CASES OF NON-FINITE SUBJECTS THE CHALLENGE OF "DOMINANT" PARTICIPLES

Dag Haug and Tanya Nikitina
University of Oslo

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

In this paper we discuss the so-called "dominant" construction found with Latin participles. We argue that this construction instantiates a rare type of subject case assignment where the case of the participle's subject depends on the grammatical function of the participial clause. To capture this in the LFG formalism, we argue for a "copy theory" of agreement, where the information from the agreeing features are present in both the controller and the target of agreement: this theory enables us to offer a uniform account of agreement across all uses of participles. We also discuss the implications for LFG's theory of subject case assignment, in particular the constructive case approach.

1 Introduction

1.1 Morphological case and grammatical function

The relation between grammatical function, thematic role, and morphological case is notoriously complex (see, inter alia, Mohanan 1982; Andrews 1982; Zaenen et al. 1985; Butt and King 1991, 2005). One parameter of variation concerns the case marking "domain": standard single case markers specify the NP's grammatical function in a local clause or phrase whereas "stacked case" specifies the NP's grammatical function in the clause, so that for example a possessive modifier of an ergative noun receives two separate case markers: one for the genitive and one for the ergative (Nordlinger 1998, 2000).

In this paper we argue that the co-called "dominant participle" construction in Latin instantiates a third, rare and non-trivial relation between case and grammatical function: subjects of participial clauses can appear in any case, depending on the grammatical function of the participial clause. In other words, the case markers specify the grammatical function of the entire clause in a larger unit.

1.2 Non-finite forms in Latin

There are five types of non-finite forms in Latin: infinitives, supines, participles, gerunds and gerundives. In this paper we focus on the participles and gerundives, which are the two types that show the full set of nominal features CASE, GENDER and NUMBER.

In addition to these nominal features, participles and gerundives also bear the verbal features VOICE and TENSE. Although the different participles are traditionally named after their finite counterparts perfect, present and future, a tradition that

[†]We thank the audiences of the seminar for theoretical linguistics in Oslo and of the LFG conference 2012 for useful comments. We use standard abbreviations for references to Latin authors. For expository purposes we sometimes use constructed examples. Parallel examples are attested in the corpus and can be found in standard grammars.

name	form	REL-TENSE	VOICE
present participle	*amans*	*simultaneuous*	*active*
perfect participle	*amatus*	*anterior*	*passive*
future participle	*amaturus*	*posterior*	*active*
gerundive	*amandus*	*posterior*	*passive*

Table 1: The inventory of participles/gerundives

we will follow here, they clearly express relative, not absolute tense. The various existing forms of the verb *amare* 'love' are shown in Table 1.[1]

The future participle has a rather restricted distribution, and in classical Latin, it only appears in periphrastic forms. The other forms all have a variety of uses illustrated in (1)-(8): attributive (1), nominalized (2), subject predicative (3), object predicative (4), periphrasis (future in (5), perfect in (6)), free predicative (7) and absolute (8). Notice that the attributive and the free predicative uses do not differ in any way,[2] so the choice of the correct analysis is context dependent.

(1)
rosa	florens	pulchra	est
rose:NOM	bloom:PTCP.PRES.NOM	beautiful:NOM	is

'The blooming rose is beatiful.'

(2)
medici	leviter	aegrotantes	leniter	curant
doctors:NOM	lightly	be.ill:PTCP.PRES.ACC	mildly	cure:PRES.3P

'Doctors cure the lightly ill mildly.' (Cic. de Off 1.83)

(3)
rosa	florens	est
rose:NOM	bloom:PTCP.PRES.NOM	is

'The rose is blooming.'

(4)
vidi	puerum	currentem
see:PERF.1S	boy:ACC	run:PTCP.PRES.ACC

'I saw the boy running.'

(5)
te	sum	visurus
you:ACC	be:PRES.1S	see:PTCP.FUT.NOM

'I will see you.' (Cic. Fam. 9.11.1)

(6)
amatus	est
love:PTCP.PERF.NOM	be:PRES.3S

'He was/has been loved.'

[1] Table 1 simplifies the situation somewhat: the class of verbs known as 'deponents' have an active rather than a passive perfect participle. Some analyses assume that the gerundive is active, rather than passive; here we follow the traditional analysis.

[2] As far as we can tell from the written text, that is. But it is likely that attributive participles, unlike free predicates, formed constituents with their nouns. This constituency could have been marked prosodically, but such evidence is of course no longer available to us.

(7) rosa florens pulchra est
rose:NOM bloom:PTCP.PRES.NOM beautiful.NOM is
'A rose is beautiful when it blooms.'

(8) his pugnantibus illum in equum quidam
them:ABL fight:PTCP.PRES.ABL him:ACC in horse:ACC someone:NOM
ex suis intulit
from his own:ABL mount:PERF.3S
'while they were fighting, one from his [attendants] mounted him on a horse.' (Caes. Gal. 6.30)

Finally, there is the so-called "dominant" use, which is the focus of this paper and is illustrated in (9)–(10).

(9) occisus dictator Caesar aliis pessimum
kill:PTCP.PERF.PASS.NOM dictator:NOM C.:NOM others:DAT worst:NOM
aliis pulcherrimum facinus videretur
other:DAT most.beautiful:NOM deed:NOM perceive:IMPF.SUBJ.PASS.3S
'the slaying of Dictator Caesar seemed to some the worst, and to others, the most glorious deed.' (Tac. Ann. 1.8)

(10) ne eum Lentulus et Cethegus ...
lest him:ACC L.:NOM and C.:NOM
deprehensi terrerent
capture:PTCP.PERF.PASS.NOM.PL frighten.IMPF.SUBJ.3PL
'lest the capture of Lentulus and Cethegus should frighten him.' (Sall., Cat 48.4)

These examples look like attribute uses of the participle; on the surface, *occisus dictator Caesar* and *Lentulus et Cethegus ... deprehensi* look like perfectly normal NPs. But semantically, these examples are clearly different. As the translations show, these constructions have eventive meanings and the participle is typically translated as an event noun.

1.3 Syntactic assumptions

In order to avoid going into irrelevant details of Latin phrase structure we will just assume that finite and non-finite clauses are both S's headed by V. The exact category labels are not important here, but it is crucial to note that the ability to host a subject does not correlate with finiteness. This is well established for the infinitives by the so-called Accusative with Infinitive (AcI) construction and for participles by the absolute construction (8); but it has also been argued to hold for participles in the free predicative construction ('backward control', as argued for Greek in Haug 2011).

For the purposes of this paper, we will assume that the core of the Latin clause is captured by the rule in (11):

finite	$(\uparrow \text{SUBJ CASE}) = \text{NOM}$
infinitives	$((\uparrow \text{SUBJ CASE}) = \text{ACC})$
participles	?
gerundives	?

Table 2: Subject case assignment

(11) $$\text{S} \rightarrow \underset{(\uparrow \text{GF}) = \downarrow}{\text{XP*}} \quad \underset{\uparrow = \downarrow}{\text{(V)}} \quad \underset{(\uparrow \text{GF}) = \downarrow}{\text{XP*}}$$

This is a simplification in various ways: it does not account for the positional licensing of unbounded dependencies nor for auxiliary verb constructions such as (5) and (6). GF stands for a disjunction of all grammatical roles, which means that no particular grammatical function is assigned configurationally at the sentence level. However, XPs are required to have *some* function in their clause. This disallows scrambling across clauses. Inside each clause, the verb selects its grammatical functions and can assign case to the elements filling these functions. For non-subject functions, case assignment typically works on a lemma basis, i.e. if a verb requires a particular argument to be dative, then it does so in all its morphological incarnations. However, subject case assignment is different and depends on the finiteness of the verb form, as shown in Table 2.

Notice that the rule for finite verbs is not optional, i.e. we assume that finite verbs always assign nominative to their subjects.[3] By contrast, the rule for infinitives is a default rule: in functional control constructions, the controlled subject of the infinitive will typically be assigned case outside the infinitive clause, so case assignment from the infinitive itself typically only applies wherever the subject is not structure shared, e.g. in AcI constructions. We will come back to the question of subject case assignment by participles and gerundives, which is the main subject of this article.

At this point we can already see that subject case assignment in Latin raises problems for a constructive case approach along the lines of Nordlinger (1998). Since finite verbs and infinitives specify the subject's case, and the subject itself must specify its case either by a constructive or a constraining equation, case matching is enough to identify grammatical roles. This suggests that constructive case equations are superfluous. We will get back to this problem in more detail in section 5.

1.4 Syntax of participle phrases

The internal syntax of the participle phrase is surprisingly constant across uses shown in (1)-(10). In particular, we note that the subject of the participle is always coreferent with an argument present in the f-structure of the sentence. There are no

[3]There is a debate as to whether the oblique case arguments of some Latin verbs are subjects. If that turns out to be the case, it should be possible to override the FINITE rule on a lexical basis.

‘dangling’ participles controlled by implicit experiencers or agents. Furthermore, there is always agreement in case, number and gender between the participle and its subject. Number and gender are determined by the lexical features of the head of the subject noun phrase, whereas case is assigned externally and in different ways in the different constructions: the crucial point to note is that there is no particular case that the participle assigns to its subject. Notice also that even in cases where there is evidence of lexical conversion (participle $\rightarrow$ adjective), both the agreement facts and the non-subject arguments remain the same, e.g. if the verb is transitive, the adjective can still take an object. To capture the uniform agreement facts, we assume that the subject function is retained in adjective conversion. This allows us to maintain the generalization that participles always agree with their subjects. Since Latin has pro-drop of both referential and generic arguments, we can assume that nominalized participles are really modifiers of such dropped arguments (unless they are lexicalized).

The *configurational* relationship between the participle and the subject can vary: for example, it is natural to assume that in the attributive construction (1), the participle and the noun constitute a phrase, with the participle being adjoined to NP. On the other hand, the participle and its noun would appear to be sisters on the most natural analysis of the c-structure of (3) and (4), since they are co-arguments of the same governing verb. And in (8), it is likely that they make up an S constituent. The best way to capture the constant syntax throughout these various constituent structures is to assume a constant *functional* relationship, i.e. that the noun is always the subject of the participle at the level of f-structure. This of course means that whenever the noun is configurationally not in the subject position of the participle (or there is no configurational subject), then it functionally controls the subject f-structure.

Such an analysis is straighforward for the subject complement use; it is the familiar raising analysis of the copula. If we assume that *be* comes with $(\uparrow \text{SUBJ}) = (\uparrow \text{XCOMP SUBJ})$, we assign the f-structure in (12) to (3).

(12)
$$\left[\begin{array}{ll} \text{PRED} & \text{‘BE}\,\langle\text{SUBJ, XCOMP}\rangle\text{’} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{‘ROSE’} \\ \text{CASE} & \text{NOM} \\ \text{GENDER} & \text{FEM} \\ \text{NUMBER} & \text{SG} \end{array}\right] \\ \text{XCOMP} & \left[\begin{array}{ll} \text{PRED} & \text{‘BLOOM}\,\langle\text{SUBJ}\rangle\text{’} \\ \text{SUBJ} & [\quad] \end{array}\right] \end{array}\right]$$

A similar functional control analysis is also available for the periphrastic constructions,[4] and, with a different control equation, for the object complement case. The

[4]On the assumption that the periphrastic tenses are biclausal. If they are monoclausal, the noun will be the subject of the participle (lexical verb) directly.

functional control analysis can also be extended to the attributive use (1). We assume that the adnominal ADJ function is assigned in the c-structure by the rule in (13).

(13)
$$\begin{array}{llcll} \text{NP} & \rightarrow & \text{(AP)} & , & \text{(NP)} \\ & & \downarrow \in (\uparrow\ \textsc{xadj}) & & \uparrow = \downarrow \\ & & (\downarrow\ \textsc{subj}) = \uparrow & & \end{array}$$

The control equation appears as $(\downarrow\ \textsc{subj}) = \uparrow$ on the adjunct and creates a cyclic f-structure:

(14)
$$\begin{bmatrix} \textsc{pred} & \text{‘}\textsc{rose}\text{’} \\ \textsc{case} & \textsc{nom} \\ \textsc{gender} & \textsc{fem} \\ \textsc{number} & \textsc{sg} \\ \textsc{xadj} & \left\{ \begin{bmatrix} \textsc{pred} & \text{‘}\textsc{bloom}\ \langle \textsc{subj} \rangle\text{’} \\ \textsc{subj} & [\ \ ] \end{bmatrix} \right\} \end{bmatrix}$$

In (13) not only the adjunct but also the head is optional. When there is no head, the PRED of the grammatical function fulfilled by the NP can be contributed by the verb (pro-drop). The nominalized use follows directly from this configuration. Consider (2). The verb *curant* will introduce the equations in (15) (as well as others not directly relevant to the nominalized participle in object position).

(15) **curant**

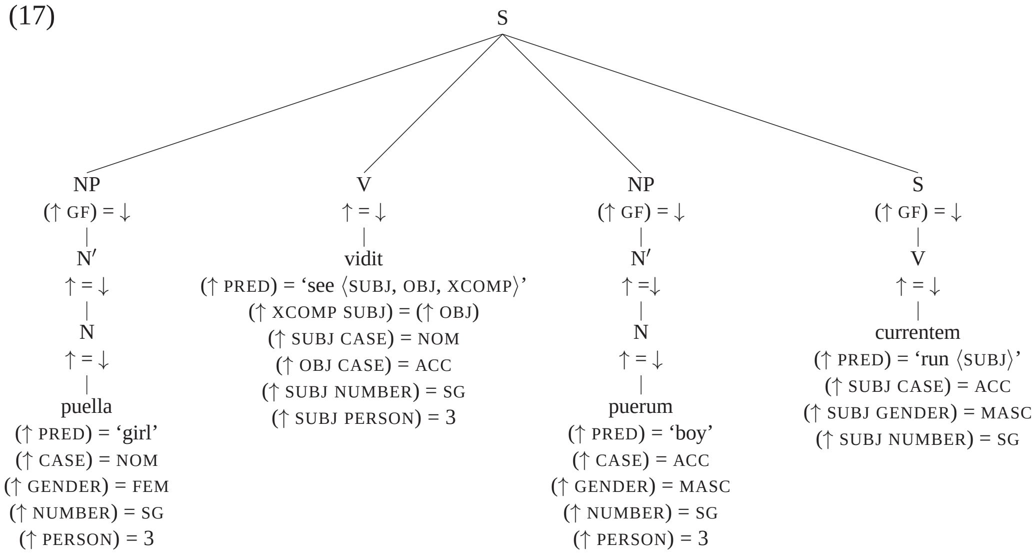
$\uparrow$ PRED = ‘cure $\langle$SUBJ, OBJ$\rangle$’
$\uparrow$ OBJ PRED = ‘PRO’
$\uparrow$ OBJ CASE = ACC

aegrotantes will be introduced by the rule in (13), but there is no head. This yields the partial f-structure in (16) (ignoring non-object functions in the matrix):

(16)
$$\begin{bmatrix} \textsc{pred} & \text{‘}\textsc{cure}\ \langle \textsc{subj}, \textsc{obj} \rangle\text{’} \\ \textsc{obj} & \begin{bmatrix} \textsc{pred} & \text{‘}\textsc{pro}\text{’} \\ \textsc{case} & \textsc{acc} \\ \textsc{gender} & \textsc{masc} \\ \textsc{number} & \textsc{pl} \\ \textsc{xadj} & \left\{ \begin{bmatrix} \textsc{pred} & \text{‘}\textsc{be ill}\ \langle \textsc{subj} \rangle\text{’} \\ \textsc{subj} & [\ \ ] \end{bmatrix} \right\} \end{bmatrix} \end{bmatrix}$$

In this way, it is possible to always analyze the noun as the participle’s f-structural subject. This allows us to capture subject-participle agreement with a single rule. One way of stating this rule would use the standard approach to agreement in LFG, which involves multiple specification of feature values by a controller and target. This is shown in (17) and (18), which give the c- and f-structures for (4).

(17)



(18)

$$\begin{bmatrix} \text{PRED} & \text{'SEE}\ \langle\text{SUBJ, OBJ, XCOMP}\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'GIRL'} \\ \text{CASE} & \text{NOM} \\ \text{GENDER} & \text{FEM} \\ \text{NUMBER} & \text{SG} \\ \text{PERSON} & 3 \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'BOY'} \\ \text{CASE} & \text{ACC} \\ \text{GENDER} & \text{MASC} \\ \text{NUMBER} & \text{SG} \end{bmatrix} \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{'RUN}\ \langle\text{SUBJ}\rangle\text{'} \\ \text{SUBJ} & [\ \ ] \end{bmatrix} \end{bmatrix}$$

Here, the agreement between the complement *currentem* 'running' and its subject *puerum* 'boy' is captured by having both items co-specify the CASE, GENDER and NUMBER features of the f-structure of *puer* 'boy'. This is identical to the way in which the matrix verb *vidit* 'see' and its subject *puella* 'girl' co-specify the PERSON and NUMBER features in the subject f-structure. Notice also that subject case assignment in the finite clause, although not usually thought of as agreement, is captured by exactly the same mechanism of feature cospecification.

2 Dominant participles

Let us now return to the dominant construction. Examples (9) and (10) show the dominant construction in subject function, but it can also appear in a variety of other, typically nominal syntactic functions beside subject such as adnominal genitive (19), object of a preposition (20), or the ablative of comparison (21).

(19) mihi conservatae rei publicae
me:DAT preserve:PTCP.PERF.PASS.GEN affair:GEN public:GEN
dat testimonium
give:PRES.3S testimony:ACC
'He gives testimony, for my sake, of the state having been preserved.' (Cic. Att. 2.1.6)

(20) ante exactam hiemem
before expel:PTCP.PERF.PASS.ACC winter:ACC
'before the winter expired' (Caes. Gal. 6.1)

(21) nullum enim officium referenda gratia magis
none:NOM for duty:NOM render:GRDV.ABL gratitude:ABL more
necessarium est
necessary:NOM is
'For no duty is more imperative than that of rendering one's gratitude.' (Cic. Off. 1.47)

Observe that the case of both the participle and its subject vary with the function of the entire construction. This means that the noun's case is clearly not sensitive to its semantic role or grammatical function, which remain constant as the function of the entire construction changes:

(22) exacta hiems me delectat
expel:PTCP.PERF.PASS.NOM winter:NOM me.ACC pleases.PRES.3SG
'The expiration of the winter pleases me.'

(23) memoria exactae hiemis
remembrance.NOM expel:PTCP.PERF.PASS.GEN winter:GEN
'remembrance of the winter's expiration'

In (22), *hiems* 'winter' is nominative because the entire construction is the subject of *delectat* 'please' and in (23), *hiemis* is genitive because the entire construction is the object of the noun *memoria* 'remembrance'. However, the function of *hiems/hiemis* in both sentences (and of *hiemem* in (20) as well) is the same, namely subject of the passive verb *exagi* 'be expelled, expire'.

We will now examine the properties of this construction more closely.

2.1 Headedness

The fact that the noun phrase agrees with the participle is often taken as an indication of an attributive relation in which the noun is the head (Heick 1936; Bolkestein 1980; Ramat 1994 for Latin; Jones 1939 for Ancient Greek, inter alia). But as we have seen, agreement is characteristic of *all* uses of the participle in Latin, not just the attributive. Moreover, the dominant construction is commonly attested with a pronoun in the nominal slot, as in (24).

(24)

Quibus	latis	gloriabatur
which:ABL	carry:PTCP.PERF.PASS.ABL	glory:IMPF.PASS.3S

'[the laws] in the passing of which he gloried.' (Cic. Phil. 1.10)

Pronouns cannot normally be modified in Latin, so this construction cannot be attributive. Instead, we will pursue an analysis of the dominant construction as a predication where the participle is the semantic predicate and the syntactic head. There are several indications that this is the correct analysis.

First, the meaning of the construction is clause-like, and (9) allows for a number of clausal periphrases, as noted by (Pinkster, 1990, 133):

(25) a.

quod	dictator	occisus	erat
that	dictator:NOM	kill:PTCP.PERF.PASS.NOM	be:IMPF.3S

pulcherrimum	facinus	videbatur
most.beautiful:NOM	deed:NOM	perceive:IMPF.PASS.3S

b.

dictatorem	occisum	esse
dictator:ACC	kill:PTCP.PERF.PASS.ACC	be:PRES.INF

pulcherrimum	facinus	videbatur
most.beautiful:NOM	deed:NOM	perceive:IMPF.PASS.3S

'That the dictator had been killed seemed the most glorious deed.'

As the matrix predicate *pulcherrimum facinus videbatur* 'seemed a glorious deed' indicates, the semantics of *dictator occisus* is specifically eventive, i.e. it entails the existence of an event in which Caesar was killed. This makes it different from constructions such as 'the young Isaac Newton' or 'a more resolute Roosevelt', which are often taken as referring to a stage or a manifestation of the head noun (von Heusinger and Wespel (2006)). In a sentence like *The dead Caesar frightened everyone*, *the dead Caesar* could be argued to refer to Caesar's manifestation as dead. On an analysis where stages and manifestations are inherent in the semantics of nouns it would then be possible to preserve the noun's status as the semantic (and syntactic) head. But in (9) (and its periphrases in (25)), the reference is clearly to an event, which cannot plausibly be inherent in the nominal semantics.

Related to this, it is clear that the participle cannot be omitted without radically changing the semantics.

(26) *dictator pulcherrimum facinus videbatur
dictator:NOM most.beautiful:NOM deed:NOM perceive:IMPF.PASS.3S
*'The dictator seemed a beautiful deed.'

Finally, while the participle is not omissible, the noun *can* be left out if the verb is impersonal, as in (27).

(27) in libris Sibyllinis propter crebrius eo
in books:ABL Sibylline:ABL on.account.of more.frequently that:ABL
anno de caelo lapidatum
year:ABL from sky:ABL stone:PTCP.PERF.PASS.ACC
inspectis
examine:PTCP.PERF.PASS.ABL
'...in the Sibylline books, which were consulted on account of the fact that it rained stones more frequently from the sky that year.' (Liv. 29.10)

The participle *lapidatum* is from the impersonal verb *lapidare* 'to rain stones' and consequently, no noun occurs and the dominant construction consists of the participle alone.

2.2 Category

While the semantics of dominant constructions is clause-like, they typically occur in nominal positions such as subject, object and object of preposition. This suggests that externally, the construction is an NP. There is also evidence from coordination that the construction is an NP, as in (28).

(28) publicum imperium servitium=que
public:NOM dominion:NOM servitude:NOM=and
obversatur animo futura=que
show.oneself:PRES.PASS.3S mind:DAT be:PTCP.FUT.NOM=and
ea deinde patriae fortuna, quam
that:NOM thereafter homecountry:GEN fortune:NOM which:ACC
ipsi fecissent
selves:NOM make:PPF.SUBJ.3PL
'The national sovereignty or servitude were on [their] minds, as well as the fact that the country's fortune would henceforth be such that they themselves had made it.'

Here the dominant construction *futura=que ...fortuna, quam ipsi fecissent* 'the fact that the country's future ... ' is coordinated with the NP *publicum imperium servitiumque'* 'national sovereignty or servitude'. Although in LFG coordination can be based on identity of function rather than of category, we take this as another indication that the construction is externally nominal.

3 Analysis

The fact that the participle is the head of the construction suggests the construction is an S headed by the participle, as in (11).

However, the facts from distribution and coordination suggest that the construction is an NP. We capture this by a syntactic nominalization rule:

(29) $\text{NP} \rightarrow \underset{\uparrow = \downarrow}{\text{S}}$

This rule is also responsible for adding appropriate semantic type-shifting, as we will see in (38). For now we focus on what goes on inside the S.

In dominant participle constructions, the participle and its subject agree in case just as in other participle constructions. But the phenomenon cannot be entirely the same if the participle is the head. In all other participle constructions except absolutes, the *noun* receives case outside the construction, and the participle agrees in case. This is impossible here given the headedness fact: it must be the participle that receives case, and this must somehow be transmitted to the noun.

There is a very simple way to achieve this effect in LFG. As we noted in the introduction, agreement in LFG is usually treated as cospecification of a single set of features by both the controller and the target. In (17)–(18), we implemented this idea in the traditional way by representing the set of agreement features in the f-structure of the controller only and having the target contribute features to this set. As long as the target is also the head, as in typical nominal agreement, this means that the whole construction has the features of the target, which is crucial for CASE to work properly. But in the dominant construction, the target is not the head.

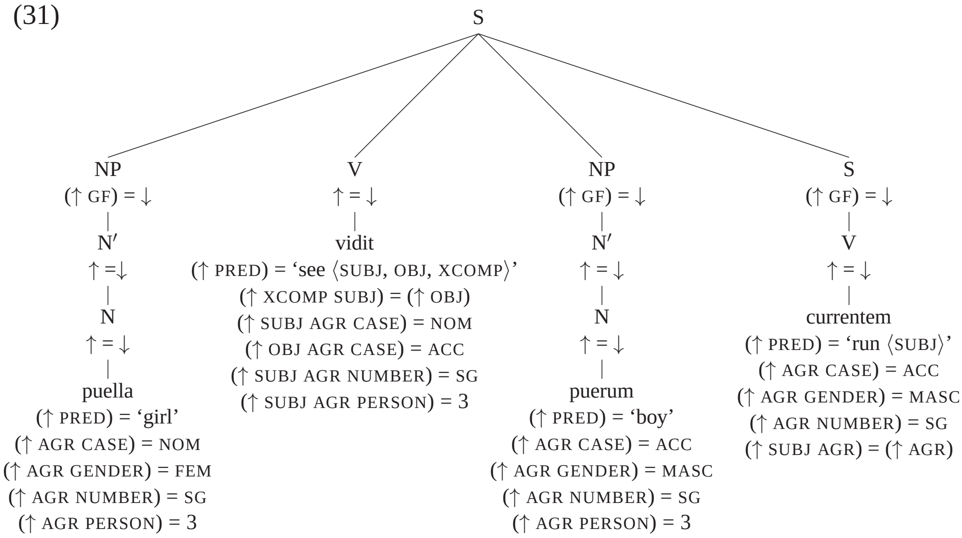
However, cospecification of the set of agreement features can also be implemented through functional control. In this way the set of agreement features can be present in the f-structures of both controller and target. To achieve this we will assume that agreement features are bundled in an f-structure which is the value of AGR in both controller and target.[5] The identification of these is secured by a lexical rule:

(30) $(\uparrow \text{FORM}) =_c \{\text{PTCP} \mid \text{GNDIVE}\} \quad \Rightarrow \quad (\uparrow \text{SUBJ AGR}) = (\uparrow \text{AGR})$

The f-structure of AGR will contain the agreement features CASE, GENDER and NUMBER. (17)–(18) can now be recast as (31)–(32).

[5] The use of a complex AGR feature whose value is an f-structure is in fact not crucial. The important point is that the information provided by the agreeing feature is available in both the controller's and the target's f-structure. This could be captured by equations like (↑ SUBJ CASE) = (↑ CASE), equating the atomic values of CASE, NUMBER and GENDER rather than the complex value of AGR. As far as we can tell, there is no empirical difference between the two approaches. Our AGR approach requires only one identity equation, but with atomic features a similar effect can be achieved by bundling the identity equations in a template.

(31)

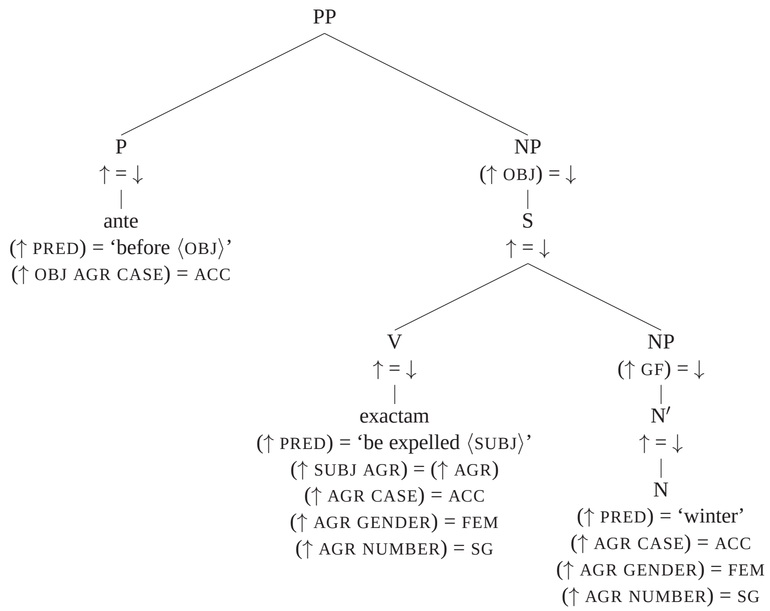


(32)

$$\begin{bmatrix} \text{PRED} & \text{`SEE} \langle\text{SUBJ,OBJ, XCOMP}\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`GIRL'} \\ \text{AGR} & \begin{bmatrix} \text{CASE} & \text{NOM} \\ \text{GENDER} & \text{FEM} \\ \text{PERSON} & 3 \\ \text{NUMBER} & \text{SG} \end{bmatrix} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`BOY'} \\ \text{AGR} & \begin{bmatrix} \text{CASE} & \text{ACC} \\ \text{GENDER} & \text{MASC} \\ \text{NUMBER} & \text{SG} \end{bmatrix} \end{bmatrix} \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{`RUN} \langle\text{SUBJ}\rangle\text{'} \\ \text{AGR} & [\;] \\ \text{SUBJ} & \begin{bmatrix} \vdots \\ \text{AGR} \; [\;] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The fact that the participle bears its own AGR feature lets us exploit the non-directionality of functional control. Case is assigned in the normal way to the NP containing the dominant participle. This is passed on to the S that is a (co-)head of the NP by (29) and then to the participle V that is the head of S. The participle and its subject agree in case, but the external case assignment is to the participle rather than to the NP. This is shown in (33)–(34).

(33)



(34)
$$\begin{bmatrix} \textsc{pred} & \text{‘}\textsc{before}\ \langle\textsc{obj}\rangle\text{’} \\ \textsc{obj} & \begin{bmatrix} \textsc{pred} & \text{‘}\textsc{be expelled}\ \langle\textsc{subj}\rangle\text{’} \\ \textsc{agr} & \begin{bmatrix} \textsc{case} & \textsc{acc} \\ \textsc{gender} & \textsc{fem} \\ \textsc{number} & \textsc{sg} \end{bmatrix} \\ \textsc{subj} & \begin{bmatrix} \textsc{pred} & \text{‘}\textsc{winter}\text{’} \\ \textsc{agr} & [\ \ ] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

We observe that on this ‘copy theory’ of agreement, no special account of the dominant construction is needed beyond the nominalization rule in (29). Or to put it the other way around: the copy theory of agreement, which is needed for the dominant construction, generalizes directly to all agreement in Latin.

4 Semantics

On the traditional analysis of dominant constructions as NPs headed by the subject noun, there is a syntax-semantics mismatch: the noun is the syntactic head, but since the semantics is clausal, the participle must be a semantic predicate taking the noun as its subject. On our analysis, this mismatch disappears: the participle is both the syntactic and the semantic head of the construction. This allows us to give a rather straightforward semantics which crucially relies on a constructional meaning introduced by the nominalization rule in (29). We use Glue semantics (Dalrymple 1999) to combine our syntactic representations with semantic ones, which are cast in Compositional Discourse Representation Theory (CDRT, Muskens 1996); the combination of LFG, Glue semantics and CDRT is also used in van Genabith and Crouch (1999) and Bary and Haug (2011).

The predicate *a glorious deed* in (9) suggests that the denotation of the dominant participle construction is an event-type referent.[6] We can assign the following lexical meaning in (35) to *occisus* and (36) to *Caesar* (c is a constant, following the treatment in Muskens (1996) of proper names as constant discourse referents):

(35) $$\lambda P \lambda e(P(\lambda x. \begin{array}{|l|}\hline \\ \hline kill(e) \\ theme(e,x) \\ \hline \end{array})) : ((E_{(\uparrow\, SUBJ)} \multimap T_{\uparrow}) \multimap T_{\uparrow}) \multimap (EV_{\uparrow} \multimap T_{\uparrow})$$

(36) $$\lambda P \begin{array}{|l|}\hline c \\ \hline \\ \\ \hline \end{array} \oplus P(c) : ((E_{(\uparrow\, SUBJ)} \multimap T_{\uparrow}) \multimap T_{\uparrow})$$

When these two meanings combine, we get (37).

(37) $$\lambda P \lambda e(P(\lambda x \begin{array}{|l|}\hline \\ \hline kill(e) \\ theme(e,x) \\ \hline \end{array}))(\lambda P.P(c)) \equiv \lambda e \begin{array}{|l|}\hline c \\ \hline kill(e) \\ theme(e,c) \\ \hline \end{array}$$

The result is, as we would expect, a set of events, i.e. the same denotation as an event nominal. The nominalization rule should have the same effect as an article, vz. it should pick a discourse referent from this set of events. This is shown in (38), which should be an annotation on (29).

(38) $$\lambda P \lambda Q \begin{array}{|l|}\hline e \\ \hline \\ \\ \hline \end{array} \oplus P(e) \oplus Q(e) : (EV_{\uparrow} \multimap T_{\uparrow}) \multimap (EV_{\uparrow} \multimap T_{GF\uparrow}) \multimap T_{GF\uparrow}$$

Applying (38) to (37) yields (39).

(39) $$\lambda P \lambda Q \begin{array}{|l|}\hline e \\ \hline \\ \\ \hline \end{array} \oplus P(e) \oplus Q(e)(\lambda e \begin{array}{|l|}\hline c \\ \hline kill(e) \\ theme(e,c) \\ \hline \end{array}) \equiv \lambda Q \begin{array}{|l|}\hline e\ c \\ \hline kill(e) \wedge theme(e,c) \\ \hline \end{array} \oplus Q(e)$$

(39) is looking for a property of an event of Caesar being killed. This is supplied

[6]Another obvious option would be to treat the dominant participle construction as denoting a proposition, which would make *a glorious deed* a second-order predicate. We do not pursue this possibility here.

by the denotation of the matrix predicate *be a glorious deed*, which we simplify as in (40).

(40) $$\lambda P \lambda s(P(\lambda e' \begin{array}{|l|}\hline \\ \hline be\_a\_glorious\_deed(s,e') \\ \hline\end{array}))$$

(39) and (40) combine as in (41).

(41) $$\lambda P \lambda s(P(\lambda e'. \begin{array}{|l|}\hline \\ \hline be\_a\_glorious\_deed(s,e') \\ \hline\end{array}))(\lambda Q \begin{array}{|l|}\hline e\ c \\ \hline kill(e) \\ theme(e,c) \\ \hline\end{array} \oplus Q(e)) \equiv$$

$$\lambda s \begin{array}{|l|}\hline e\ c \\ \hline kill(e) \wedge theme(e,c) \\ be\_a\_glorious\_deed(s,e) \\ \hline\end{array}$$

(41) denotes a set of states of an event of Caesar being killed being a glorious deed. From this meaning, matrix tense and aspect will yield the final semantics. Notice that our semantics takes no account of the participle's relative tense. To do this we would probably need a function from sets of events to sets of times, but this would make the event variable inaccessible to the semantics of the nominalization in (38). This is a more general problem in the semantics of participles and cannot be dealt with here.

5 Consequences for the theory of case

There are many ways in which case and its relationship to grammatical function can be treated within LFG and they are suitable for different types of languages. A basic distinction is whether CASE is a syntactic feature at all or whether it is just a morphological phenomenon that serves to construct grammatical (i.e. syntactic) functions, but is otherwise not a syntactic feature and is not represented at f-structure. Some useful criteria are found in Spencer and Otoguro (2005), who point out that a syntactic feature CASE is needed to deal with agreement and some forms of government. The Latin agreement phenomena we have seen clearly demand a syntactic representation of the CASE feature.

Given that CASE is an f-structure feature,[7] we need to ask where it comes from. One option is the c-structure: c-structure rules could introduce CASE features, cf.

[7] In the following we will assume that if CASE is a syntactic feature, it is represented at f-structure. Some LFG theorists (e.g. Falk (2006)) deal with phenomena such as agreement at other (syntactic) levels, and our discussion should carry over to these as well, but for simplicity we maintain a uniform representation at f-structure here.

the notion of 'positional case' in Butt and King (2005). This is particularly likely to happen in configurational languages.

In languages like Latin, however, CASE features are more likely to come from the lexicon. In particular, they could come from the noun that bears the case morphology, or from its governing head, or from both. The dependent-only option is shown in (42).

(42)

nominative	accusative	finite head	infinitive head
(↑ CASE) = NOM	(↑ CASE) = ACC	—	—
((SUBJ ↑) TENSE)	((SUBJ ↑) FORM) = INF ∨ (OBJ ↑)		

Notice that because of the different subject case requirements of finite forms and infinitives, we cannot use simple constructive case features like (SUBJ ↑) – we also need to specify some formal features of the governing f-structures, such as their being finite (having tense) or being an infinitive. In (43) we see how case could be specified by the head only, while the constructive case equations are retained on the dependents.

(43)

nominative	accusative	finite head	infinitive head
((SUBJ ↑) TENSE)	((SUBJ ↑) FORM) = INF	(↑ SUBJ CASE) = NOM	(↑ SUBJ CASE) = ACC
	∨ (OBJ ↑)	(↑ OBJ CASE) = ACC	(↑ OBJ CASE) = ACC

By exploiting properties of the governing f-structure, both these accounts will work. However, it is perhaps less natural to have heads assign case to adjuncts.

The picture is complicated by pro-dropped arguments, which do not have their own lexical entry, but which do have case, as evidenced by e.g. agreement with secondary predicates. This case feature has to be contributed by the verbal head, together with the optional PRED feature. This can be done by equations like (44).

(44)
$$\begin{pmatrix} \uparrow \text{SUBJ PRED} = \text{`PRO'} \\ \uparrow \text{SUBJ CASE} = \text{NOM} \end{pmatrix}$$
$$\begin{pmatrix} \uparrow \text{OBJ PRED} = \text{`PRO'} \\ \uparrow \text{OBJ CASE} = \text{ACC} \end{pmatrix}$$

This means that in a 'pro-drop language' where there is evidence that PRO bears case, a pure dependent-based account of case is not possible. Still, it is possible to limit case specification by the head to pro-drop arguments only and let the CASE feature be contributed by the dependent in all other cases.

Finally, it is possible to have head and dependent cospecify case, as we have assumed in this paper. The equations are shown in Table (45).

(45)

nominative	accusative	finite head	infinitive head
$(\uparrow \text{CASE}) = \text{NOM}$	$(\uparrow \text{CASE}) = \text{ACC}$	$(\uparrow \text{SUBJ CASE}) = \text{NOM}$	$(\uparrow \text{SUBJ CASE}) = \text{ACC}$
		$(\uparrow \text{OBJ CASE}) = \text{ACC}$	$(\uparrow \text{OBJ CASE}) = \text{ACC}$

Notice that when case is specified on both the head and the dependent, the constructive case equations are superfluous – the matching CASE features are enough to establish the grammatical relations.

The dominant participle construction can in fact only be accounted for by this latter approach. We already noted that case assignment in Latin cannot use simple constructive case equations, but must access the features of the head that identify the kind of case it requires (TENSE and FORM in Tables (42) and (43)). For dominant participles, there is no such feature. For example, the equation in (46) for an accusative dependent would overgenerate.

(46) $((\text{SUBJ} \uparrow)\ \text{FORM}) = \text{INF} \vee (\text{OBJ} \uparrow) \vee ((\text{SUBJ} \uparrow)\ \text{FORM}) = \text{PTCP}$

Accusatives can only be subjects of (morphologically) accusative participles, so the constraint that FORM should be PTCP is not restrictive enough. But equations like $((\text{SUBJ} \uparrow)\ \text{CASE}) = \text{ACC}$ would move the locus of the CASE feature to the head, which is unnatural except in a copy theory of agreement. But in such a theory, constructive case equations are redundant. We conclude that constructive case is not able to account for the dominant participle construction.

Moreover, this is just a symptom of a wider problem with the constructive case approach, namely that to deal with case variability, the constructive case marker needs to be able to see some feature in the governing f-structure that controls the choice of case. But this is not always available. ‘Quirky case’ (in the restricted sense of Butt and King (2005), i.e. case that is truly unpredictable and thus must be stated on a lexical basis rather than being derivable from some other feature) is typically captured by equations like $(\uparrow \text{SUBJ CASE}) = \text{DAT}$ on the governing predicate, which leaves no feature in the governing f-structure that reveals the required type of case marking. And because of the global nature of constructive case, it is not possible to do only quirky case via case cospecification and leave the rest to constructive case: if there is a single predicate that requires a dative subject, every dative in the language must be marked with a disjunct $(\text{SUBJ} \uparrow)$, which yields the wrong predictions.

6 Conclusions

We have presented a unified account of the various uses of Latin participles which relies in a crucial way on the formal tools offered by LFG. In particular, the notion of functional control allows us to capture the constant agreement between the participle and its subject, which is the major unifying property of participles.

Our analysis captures the ‘dominant’ construction through a syntactic nominalization rule. Again functional control is crucial, since it allows us to have a non-directional treatment of feature agreement. This means that the unusual case agreement can be treated on a par with gender and number agreement by bundling the features in an AGR structure. The variable case of the participle's subject instantiates a rare type of case assignment which cannot be accounted for in terms of constructive case. It is also non-local, in that it is sensitive to the grammatical function of the entire clause and not just the grammatical function of an NP within that clause. But our copy theory of agreement lets us preserve locality in the analysis, since CASE is a feature of the verbal head itself, which is assigned in the normal way and passed on to the subject by the agreement rules. In this way, we do away with the syntax-semantics mismatch which previous analyses have assumed.

References

Andrews, Avery. 1982. The representation of case in modern Icelandic. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 427–503, Cambridge, MA: The MIT Press.

Bary, Corien and Haug, Dag. 2011. Temporal anaphora across and inside sentences: The function of participles. *Semantics and Pragmatics* 4(8), 1–56.

Bolkestein, A. Machtelt. 1980. The 'ab urbe condita'-construction in Latin: A strange type of raising? *Linguistics in the Netherlands* pages 80–92.

Butt, Miriam and King, Tracy Holloway. 1991. Semantic case in Urdu. In Lise M. Dobrin, Lynn Nichols and Rosa M. Rodriguez (eds.), *Proceedings of CLS*, volume 27, pages 31–46.

Butt, Miriam and King, Tracy Holloway. 2005. The status of case. In Veneeta Dayal and Anoop Mahajan (eds.), *Clause Structure in South Asian Languages*, volume 61 of *Studies in Natural Language and Linguistic Theory*, pages 153–198, Dordrecht: Kluwer.

Dalrymple, Mary (ed.). 1999. *Semantics and Syntax in Lexical Functional Grammar*. Cambridge, MA: MIT Press.

Falk, Yehuda. 2006. On the representation of case and agreement. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG06*, CSLI Publications.

Haug, Dag T. T. 2011. Backward control in Ancient Greek – subsumption or linearization? In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG11*, CSLI Publications.

Heick, Otto William. 1936. *The Ab Urbe Condita-construction in Latin*. Ph. D.thesis, University of Nebraska, Lincoln, Nebraska.

Jones, Frank Pierce. 1939. The ab urbe condita construction in Greek. *Language* 15, 5–96.

Mohanan, K. P. 1982. Grammatical relations and clause structure in Malayalam. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 427–503, Cambridge, MA: The MIT Press.

Muskens, Reinhard. 1996. Combining Montague semantics and discourse representations. *Linguistics and Philosophy* 19, 143–186.

Nordlinger, Rachel. 1998. *Constructive Case: Dependent Marking Nonconfigurationality in Australia*. Stanford, CA: CSLI Publications.

Nordlinger, Rachel. 2000. Australian case Systems: towards a constructive solution. In Miriam Butt and Tracy Holloway King (eds.), *Argument Realization*, Stanford, CA: CSLI Publications.

Pinkster, Harm. 1990. *Latin Syntax and Semantics*. London: Routledge.

Ramat, Paolo. 1994. On Latin absolute constructions. In J. Herman (ed.), *Selected Papers from the 6th International Colloquium on Latin Linguistics*, pages 259–268, Amsterdam: John Benjamins.

Spencer, Andrew and Otoguro, Ryo. 2005. Limits to case – A critical survey of the notion. In Mengistu Amberber and Helen de Hoop (eds.), *Competition and Variation in Natural Languages: The Case for Case*, pages 119–145, Amsterdam: Elsevier.

van Genabith, Josef and Crouch, Richard. 1999. How to glue a donkey to an f-structure. In Harry Bunt and Reinhard Muskens (eds.), *Computing Meaning*, volume 1, pages 129–148, Dordrecht: Kluwer.

von Heusinger, Klaus and Wespel, Johannes. 2006. Indefinite proper names and quantification over manifestations. In E. Puig-Waldmüller (ed.), *Proceedings of Sinn und Bedeutung 11*, pages 332–345, Barcelona: Universitat Pompeu Fabra.

Zaenen, Annie, Maling, Joan and Thrainsson, Höskuldur. 1985. Case and grammatical functions: The Icelandic passive. *Natural Language and Linguistic Theory* 3(4), 441–483, reprinted in Joan Maling and Annie Zaenen, eds., *Syntax and Semantics 24: Modern Icelandic Syntax*, 95–164. New York: Academic Press. 1990.

LEXICAL MAPPING IN YAMI VERBS

One-Soon Her[+]
Dun-Hong Deng[‡]

[+]National Chengchi University
[‡]National Taipei University of Nursing and Health Science

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper demonstrates that the ergative hypothesis works out for Yami, an indigenous language of Taiwan's Orchid Island, where actor voice verbs are either intransitive or antipassive while non-actor voice verbs may be transitive. Crucially, we contend that the so-called genitive voice, marked by *no* or *ni*, is in fact accusative. The account is rendered in the simplified LMT with a unified mapping principle proposed by Her (2012). The Intrinsic Classification *pt/th* → [-*r*] applies to accusative languages like English and Chinese but not to ergative languages like Yami. This minimal parameterization, plus the morphosyntactic classification of θ[voice] →θ[-*r* -*o*], for voice-marking languages like Yami, is sufficient to account for the basic patterns of lexical mapping in Yami verbs.

1 Introduction

A distinctive feature of the Austronesian languages of Taiwan and the Philippines is the voice system, where a verb is marked for an argument role for voice, which receives the nominative case (e.g., Kroeger 1993, Starosta 1997, Li 1997, Arka 2003, Rau and Tung 2006). Such voice markings on the verb can be divided into actor voice (AV) and non-actor voice (NAV). While there is no dispute that NAV verbs are syntactically transitive, whether AV verbs can also be transitive is controversial. In the ergative hypothesis (e.g., Starosta 1997), AV verbs are intransitive, while under the symmetrical voice hypothesis (e.g., Kroeger 1993), AV verbs can be transitive. Chang (2004), however, claims that Formosan languages are not uniform in this regard and that AV verbs are intransitive in Paiwan, Tsou, Atayal, and Kavalan but transitive in Seediq and Saisiyat.

Yami, *a.k.a.* Tao, is a language indigenous to Orchid Island, southeast of Taiwan; yet, Yami belongs to the Batanic, *a.k.a.* Bashiic, group of the northern Philippine islands and thus not the Formosan group (e.g., Li 1997, Rau and Tung 2006). In this paper we will first demonstrate that the ergative hypothesis works out for Yami and then account for the lexical mapping in Yami verbs. The paper is organized as follows. Section 2 establishes the grammatical functions encoded by the various case markers in Yami. Section 3 then describes a revised and streamlined version of LMT, proposed by Her (2009, 2012), and applies it to the argument-function mapping of Yami verbs. Some concluding remarks are given in section 4.

2 Yami Case Markers and Grammatical Functions

The most authoritative reference in the literature on Yami is Rau and Tung's (2006) dictionary and reference grammar, where four different cases are identified, as shown in Table 1.

Table 1. Rau & Tung (2006): Case Markers

	Nom	**Gen**	Loc	Obl
Common noun	*o*	***no***	*do*	*so*
Proper noun	*si*	***ni***	*ji*	ϕ

Under the LFG framework (e.g., Bresnan 2001, Falk 2001), it should be without controversy that the nominative case encodes the grammatical function (GF) of subject (SUBJ), the locative case, the GF $OBL_{locative}$, and the Obl case, the GF OBL_{Θ}, a cover term of the oblique functions of various argument roles, e.g., $OBL_{beneficiary}$, OBL_{goal}, $OBL_{instrument}$, etc. However, the so-called 'genitive' case is rather problematic. First of all, the name 'genitive' is used simply because the two case markers *no* and *ni* are the same two forms marking possessives in the nominal phrase. However, calling it 'genitive' does not tell us anything as to what GF this case encodes. Now that we have identified SUBJ, OBL_{loc}, and OBL_{Θ} in Yami, the most likely GF the genitive case may encode is OBJ. OBJ_{Θ} is out as it is the most marked GF and is found only in languages with the OBJ function. SUBJ and OBL_{Θ} are both unlikely candidates because they have already been represented. Thus, Deng (2004) argues that what is called GEN in fact encodes the accusative case, or the term function OBJ.

Table 2. Deng (2005): Case Markers and Grammatical Functions

	SUBJ	**OBJ**	OBL_{loc}	OBL_{Θ}
Common noun	*o*	***no***	*do*	*so*
Proper noun	*si*	***ni***	*ji*	ϕ

The use of the possessive forms to encode a term GF is in fact not uncommon in Austronesian. In Balinese, for example, the genitive elements are terms (Wechsler and Arka 1998, Arka 2003). This classification is straightforward, as non-terms in Balinese are PP's but genitive agents are not. Deng (2005) demonstrates that genitive agents in Yami are likewise terms unlike the *do*- or *so*-marked non-terms. First of all, genitive agents in some NAV sentences may not be omitted, as in (1). (Yami has three NAVs: patient, locative, and instrucment; the NAV in (1) is locative voice). Yet, *so*-marked patients in AV sentences are optional, as in (2). This also indicates that the NAV clause in (1a) is transitive, while the AV clause in (2) is intransitive.

(1) a. *na-bakbak-an yaken ni yama*
he-beat(LV) I(NOM) GEN father (LV = locative voice)
'I am beat by father.'

b.**bakbak-an yaken*
beat(LV) I(NOM)

(2) *ko man-bakbak (so ino).*
I(NOM) AV-beat OBL dog
'I beat (the dog).'

The second piece of evidence comes from the fact that with agent as the subject, the verbs must always be morphologically marked, as in (2), but with patient subjects, the verb may be unmarked, as in (3) (Ho 1990). This suggests that patient as subject is the basic transitive clause, where the genitive (GEN) agent is thus the object.

(3) *to ko a cita si apen Kalalanet ito*
then I(GEN) LINK see NOM grandfather Kalalanet that
'Then grandpa Kalalanet is seen by me.'

Thirdly, the AV patient marker can only be used on common nouns but not proper nouns or pronouns; in contrast, the PV (patient voice) agent maker, i.e., the genitive marker, can be used on all three types of nouns (Ho 1990). According to Hopper and Thompson (1980), the individuation of the patient could determine the transitivity of a sentence. Proper nouns and pronouns being more specific than common nouns, PV agents are thus more specific, or individuated, than AV patients, which again suggests that PV clauses are transitive. The genitive case thus encodes OBJ in a PA transitive clause.

Finally, there is a special agreement device in Yami in the form of a third-person pronoun, which agrees with a full NP counterpart later in the sentence. This agreement device, crucially, is found only when the pronoun is in the nominative or genitive case (Deng 2005, Rau and Tung 2006: 94).

(4) a. *ni-t-om-anek sira o kanakan*
PAST-stand(AV) they(NOM) NOM children
'The children stood up.'

b. *na-kan-en o soli ni yama.*
he(GEN)-eat-PV NOM taro GEN father
'The taro is eaten by father.'

In (4a) *sira* 'they' agrees with the nominative *o kanakan* 'the children', and in (4b) *na-* 'he' agrees with the genitive *ni yama* 'by father'. According to the accessibility hierarchy in Keenan and Comrie (1977), the subject and the object are the two highest functions in the hierarchy. The evidence that this agreement in Yami applies only to nominative and genitive cases, but not others, thus also suggests that the genitive case is a term GF like the nominative.

In short, the ergative hypothesis is applicable to Yami, where AV verbs are either intransitive (5a) or antipassive (5b) and NAV verbs may be transitive, as in (6)-(8). The argument-function mapping in the four voices are shown in (5)-(8), with the four respective forms of *k-om-an*, *kan-en*, *akan-an*, and *i-akan*, derived from the root *–kan* 'eat', as an example (Deng 2005).

(5) Verbs with Agent Voice (AV, e.g., *k-om-an*):

a. <*ag*> ↓ SUBJ

b. <*ag* *pt*> ↓ ↓ SUBJ OBL

(6) Verbs with Patient Voice (PV, e.g., *kan-en*):

a. <*pt*> ↓ SUBJ

b. <*ag* *pt*> SUBJ OBJ

c. <*pt* *loc*> ↓ ↓ SUBJ OBL

(7) Verbs with Locative Voice (LV, e.g., *akan-an*):

a. <*ag* *pt* *loc*> SUBJ OBJ OBL

b. <*ag* *loc*> SUBJ OBJ

c. <*pt* *loc*> SUBJ OBL

(8) Verbs with Instrument Voice (IV, e.g., *i-akan*):

a. <*ag* *pt* *inst*> SUBJ OBJ OBL

b. <*ag* *inst*> SUBJ OBJ

The voiced role is thus to be assigned [-*r* -*o*], similar to how the Icelandic quirky case is lexically marked (cf., Zaenen and Maling 1990), to ensure mapping to SUBJ. Given the Monotonicity Condition (e.g., Brenan 2001: Chp 5), lexically marked features are preserved in syntax. Mapping principles then determine the grammatical functions of non-voiced roles. However, in the conventional formulations of the Lexical Mapping Theory (LMT), e.g., Bresnan and Zaenen (1990), Bresnan (2001), and Falk (2001), agent is [-*o*] intrinsically, and patient [-*r*] or [-*o*]. Such classifications, where agent is never OBJ and patient is never OBL, are too restrictive for Yami.

Also, there is an inconsistency between the mapping principle of subject roles, which in essence supplies unmarked features, and that of non-subject roles, which essentially assigns marked features. In addition, the Subject Condition and the Argument-Function Biuniqueness Condition should ideally be consequences of a unified mapping principle and not stipulated output constraints. We will therefore adopt the revised and streamlined LMT proposed in Her (2009, 2012).

3 Yami Mapping in a Revised and Simplified LMT

While OBJ[-*r* +*o*] and OBL_θ[+*r* -*o*] were previously on a par in terms of markedness, Her (2012) proposes that [-*r*] be seen as less marked than [-*o*], as only [-*r*] functions can be athematic (e.g, Bresnan 2001). The addition of (9) enables a comprehensive markedness hierarchy of argument functions, as in (10).

(9) Markedness Hierarchy of Grammatical Features (revised):
a. [-*f*] > [+*f*] (*f* = *r*/*o*)
b. [-*r*] > [-*o*]

(10) Markedness Hierarchy of Argument Functions:
SUBJ[-*r* -*o*] > OBJ[-*r* +*o*] > OBL_θ[+*r* -*o*] > OBJ_θ[+*r* +*o*]

For feature assignment in Yami we also propose three generalizations: 1) a parameterized option of no intrinsic assignment of features to any role (11), 2) a morphosyntactic operation assigning [-*r* +*o*] to the role selected for voice (12a), and 3) a universal default morphosyntactic operation assigning [+*r*] to all non-$\hat{\theta}$ roles (12b). We then adopt the spirit of the Unified Mapping Principle (UMP) proposed by Her (2009, 2012) to complete our account. The rephrased single declarative mapping principle we propose is given in (13), which consistently favors the least marked function and can do without the previous output constraints; thus, intuitively, arguments are mapped as high in thehierarchy as possible while maintaining uniqueness.

(11) Intrinsic Classification of Roles:
(*pt/th* → [-*r*]) (✔: English, Chinese; ✘: Yami)

(12) Morphosyntactic Classification of Roles (DC):
a. (θ[voice] → [-*r* -*o*]) (✔: Yami; ✘: English, Chinese); otherwise,
b. $\theta \rightarrow$ [+*r*], if $\theta \neq \hat{\theta}$

(13) Unified Mapping Principle (UMP):

Map a fully specified role θ_1 onto a compatible function F_1, and map a unspecified or underspecified role θ_2 onto the highest compatible function F_2, if $F_2 \neq F_1$ and also F_2 is not the highest compatible function for a role higher than θ_2.

We now present the basic argument-function mapping in the four voices that a Yami verb may be marked for, with illustrative examples.

(14) Verbs with Agent Voice (AV):

a. *ko* *man-bakbak.*
I(NOM) AV-beat
'I beat.'

	(I)
	<***ag***>
AV	[-*r* –*o*]
DC	

	SUBJ
UMP	**SUBJ**

b. *ko* *man-bakbak* *so* *ino.*
I(NOM) AV-beat OBL dog
'I beat the dog.'

	(I)	**(dog)**
	<***ag***	***pt***>
AV	[-*r* –*o*]	
DC		[+*r*]

	SUBJ	OBL_θ/OBJ_θ
UMP	**SUBJ**	$\mathbf{OBL}_{pt}$

(15) Verbs with Patient Voice (PV):

a. *ya* *ma-saray* *si* *ina*.
YA PV-happy NOM mother
'Mother is happy.'

	(mother)
	<***pt***>
PV	[-*r* –*o*]
DC	

	SUBJ
UMP	**SUBJ**

b. *na-kan-en* *o* *soli* *ni* *yama.*
he(GEN)-eat-PV NOM taro GEN father
‘The taro is eaten by father.’

(father) **(taro)**
<***ag*** ***pt***>
PV [-*r* –*o*]
DC

SUBJ/OBJ/ OBL$_{\theta}$/OBJ$_{\theta}$ SUBJ
UMP **OBJ** **SUBJ**

c. ya ma-pno do yala o wakay
YA PV-full LOC basket NOM sweet-potato
’The basket is full of the sweet potatos.’

(sweet potatos) **(basket)**
<***pt*** ***loc***>
PV [-*r* –*o*]
DC [+*r*]

SUBJ OBL$_{\theta}$/OBJ$_{\theta}$
UMP **SUBJ** **OBL$_{loc}$**

(16) Verbs with Locative Voice (LV):

a. ya ko pi-akan-an so among o pasalan ya
Aux I(GEN) eat(LV) OBL fish NOM shore Aux
’This seashore is where I eat fish.’

(I) **(fish)** **(seashore)**
<***ag*** ***pt*** ***loc***>
LV [-*r* –*o*]
DC [+*r*]

SUBJ/OBJ/ OBL$_{\theta}$/OBJ$_{\theta}$ OBL$_{\theta}$/OBJ$_{\theta}$ SUBJ
UMP **OBJ** **OBL$_{pt}$** **SUBJ**

b. ya ko pi-akan-an o pasalan ya
Aux I(GEN) eat(LV) NOM shore Aux
’This seashore is where I eat.’

(I) **(seashore)**
<***ag*** ***loc***>
LV [-*r* –*o*]
DC

SUBJ/OBJ/ OBL$_{\theta}$/OBJ$_{\theta}$ SUBJ
UMP **OBJ** **SUBJ**

c. ya pi-akan-an so among o pasalan ya
Aux eat(LV) OBL fish NOM shore Aux
'The seashore is where fish are eaten.'

	(fish)	**(seashore)**
	<***pt***	***loc***>
LV		[-*r* –*o*]
DC	[+*r*]	
	$OBL_{\theta}/OBJ_{\theta}$	SUBJ
UMP	$\mathbf{OBL_{pt}}$	**SUBJ**

(17) Verbs with Instrument Voice (IV):

a. ya ko ya-kan so among o ipangan ya
Aux I(GEN) IV-eat OBL fish NOM knife Aux
'I eat fish with the knife.'

	(I)	**(fish)**	**(knife)**
	<***ag***	***pt***	***inst***>
IV			[-*r* –*o*]
DC		[+*r*]	
	SUBJ/OBJ/ $OBL_{\theta}/OBJ_{\theta}$	$OBL_{\theta}/OBJ_{\theta}$	SUBJ
UMP	**OBJ**	$\mathbf{OBL_{pt}}$	**SUBJ**

b. ya ko ya-kan o ipangan ya
Aux I(GEN) IV-eat NOM knife Aux
'I eat with the knife.'

	(I)	**(knife)**
	<***ag***	***inst***>
IV		[-*r* –*o*]
DC		
	SUBJ/OBJ/ $OBL_{\theta}/OBJ_{\theta}$	SUBJ
UMP	**OBJ**	**SUBJ**

4 Concluding Remarks

In this paper, we first offered four kinds of evidence for the OBJ function that the so-called genitive case encodes in Yami: 1) genitive agents in an NAV sentence may not be omitted, 2) PA subjects occur with unmarked verbs, suggesting the co-occurring genitive agents are objects, 3) an agreement is found between a nominative (SUBJ) or genitive pronoun and a full NP, indicating the genitive GF is a term, and 4) unlike the AV oblique patient marker, the PV genitive agent maker can be used on all types of nouns, suggesting its objecthood.

We then adopted Her's (2009, 2012) revised, simplified lexical mapping theory to account for the argument-function mapping in Yami verbs. With a minimal parameterization of the intrinsic classification of argument roles, i.e., *pt/th* → [-*r*] does not apply to ergative languages like Yami, and a morphosyntactic classification, i.e., $\theta \rightarrow [-r\ -o]$, if θ is marked for voice, we are able to derive all lexical mappings in the four different voices. The simplified LMT proposed in Her (2009, 2012) is thus shown to be applicable to accusative languages like English and Chinese as well as ergative languages like Yami.

Acknowledgements

We thank the four anonymous LFG12 reviewers and the participants at LFG 12 for their comments and suggestions. We also thank the anonymous LFG12 proceedings reviewer for the detailed comments. The research on Yami verbs reported in the paper is largely based on Deng (2004), an MA thesis by the second author, for which the first author served as the adviser. We thank the other two members of the advising committee, Professor Claire Chang and Professor Jen Ting for their input. However, we are solely responsible for the content of the paper. Research reported in the paper has been partly funded by an NSC grant, 95-2411-H-004-027, awarded to the first author.

References

Arka, Wayan. 2003. *Balinese Morphosyntax: A Lexical-Functional Approach*. Department of Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University.

Bresnan, Joan. 2001. Lexical Functional Syntax. Oxford: Blackwell.

Bresnan, Joan and Annie Zaenen. 1990. Deep unaccusativity in LFG. In K. Dziwirek, P. Farrell, and E. Mejias-Bikandi (Eds.), *Grammatical Relations: A Cross-Theoretical Perspective*, 45-57. Stanford, CA: CSLI Publications.

Chang, Henry Yungli. 2004. AF verbs: Transitive, intransitive, or both? In Ying-chin Lin *et al* (eds.) *Papers in honor of Professor Hwang-Cherng Gong on his seventieth birthday.* Taipei: Institute of Linguistics, Academia Sinica. 95-120.

Deng, Dun-hong. 2004. *Argument-Function Linking in Yami: An Optimality-Theoretic Account.* MA thesis, National Chengchi University.

Falk, Yehuda N. 2001. *Lexical-Functional Grammar: An introduction to parallel constraint-based syntax*. Stanford, California: CSLI Publications.

Her, One-Soon. 2009. Apparent subject-object inversion in Chinese. *Linguistics* 47.5: 1143-1181.

Her, One-Soon. 2012. Lexical Mapping Theory revisited. To appear in *From Quirky Case to Representing Space: Papers in Honor of Annie Zaenen*, edited by Tracy Holloway King and Valeria de Paiva, Stanford, CA: CSLI Publications.

Ho, Arlene Yue-ling. 1990. *Yami Structure: A Descriptive Study of the Yami Language*. M.A. Thesis, National Tsing Hua University.

Kroeger, Paul R. 1993. *Phrase Structure and Grammatical Relations in Tagalog*. CSLI Publications.

Li, Paul Jen-kuei. 1997. A syntactic typology of Formosan Languages – Case markers on nouns and pronouns. In *Typological studies of languages in China*. Symposium series of the Institute of History and Philology, No. 4, 443-477. Taipei: Academia Sinica.

Rau, Der-Hwa V. and Ma-nyu Tung. 2006. *Tao Dictionary: Reference Grammar, Texts, and Dictionary*. Language and Linguistics Monograph. Taipei: Academia Sinica.

Starosta, Stanley. 1997. Formosan clause structure: transitivity, ergativity, and case marking. In Chiu-yu Tseng (ed.) *Chinese Languages and Linguistics IV: Typological Studies of Languages in China*. Taipei: Academia Sinica. 125-154.

Zaenen, Annie and Joan Maling. 1990. Unaccusative, passive and quirky case. In Joan Maling & Annie Zaenen (eds.) *Modern Icelandic Syntax*. San Diego, CA: Academic Press. 137–152.

One-Soon Her
onesoon@gmail.com
Graduate Institute of Linguistics & Research Center of Mind, Brain, and Learning
National Chengchi University

Dun-Hong Deng
dyadeng@yahoo.com.tw
Department of Speech and Hearing Disorders
National Taipei University of Nursing and Health Science

PARTICIPLES, ADJECTIVES, AND THE ROLE OF ARGUMENT STRUCTURE

Anna Kibort

University of Cambridge
& Institute of Computer Science,
Polish Academy of Sciences

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

The focus of this paper is a construction which is surprisingly rarely scrutinised: the impersonal passive of the intransitive. Although it sometimes receives a brief mention in discussions of the passive, and although the impersonal passive label is often wrongly given to morpholexical impersonal constructions, there are as yet no thorough analyses of the impersonal passive available for any language. In this paper, I offer an analysis of this construction in Polish, where it is made up of a tensed auxiliary or copula, and a participle commonly referred to as the passive participle.

1 The impersonal passive of the intransitive – an introduction

A example of an impersonal passive of the intransitive in Polish is given in (1), with the assumption that no neuter singular referent can be found in the context of this sentence which could be interpreted as the antecedent of its 'dropped' subject. The agent in impersonal passives, downgraded to an oblique, is optional and frequently left unexpressed. The *-n-/-t-* participle has the SG.N ending *-e* used in situations when agreement breaks down, such as here, where there is no subject for the participle to agree with:

(1) *Było* *codziennie* *sprzątane* *(przez firmę).*
was.3SG.N daily clean.PART.SG.N (by company)
'There has been cleaning every day (by a company).'

Sentence (1) is an example of a predicative use of the *-n-/-t-* participle in an impersonal construction, i.e. a construction without a subject. Its personal counterpart is the common personal passive, where the participle has to agree with its subject in gender and number; depending on the grammarical gender of the subject, the participle will have one of the following endings, in the singular: masculine *-y*, feminine *-a*, or neuter *-e*; and in the plural: masculine human *-i*, or other than masculine human *-e*. In Table 1 below, (1) is repeated in (3) and shown next to its personal counterpart in (2). Although the subject of the personal variant can bear any number or gender, in the table below it is illustrated with a noun of neuter gender and singular number (this minimises the number of variables for an easier comparison of the examples).

In the typological literature, a participial personal passive is sometimes referred to as an 'objective resultative'. This is in contrast with another type of construction, the so-called 'possessive resultative', in which the participle is a member of a secondary predicate which is part of the clausal object of a personal active verb 'have'. The bottom row of Table 1 shows a relatively familiar personal 'possessive resultative' in (4), and a virtually unstudied impersonal 'possessive resultative' in (5). It is worth noting that Polish possessive resultatives have a more neutral possessive interpretation than their English translations.[1] In the absence of an oblique agent, Polish possessive resultatives do not exclude the interpretation that the possessor may have been the agent.

	personal	impersonal
'objective resultative'	(2) *Mieszkanie było* flat(N).NOM was.3SG.N *sprzątane (przez firmę).* clean.PART.SG.N.NOM (by company) 'The apartment was cleaned (by a [professional] company).'	(3) *Było sprzątane* was.3SG.N clean.PART.SG.N *(przez firmę).* (by company) 'There was cleaning done (by a [professional] company).'
'possessive resultative'	(4) *Ojciec miał mieszkanie* father(M).NOM had.3SG.M flat(N).ACC *sprzątane (przez firmę).* clean.PART.SG.N.ACC (by company) 'Father had the apartment cleaned (by a [professional] company).'	(5) *Ojciec miał* father(M).NOM had.3SG.M *sprzątane (przez firmę).* clean.PART.SG.N (by company) 'Father had the cleaning done (by a [professional] company).'

Table 1. The impersonal passive of the intransitive (3) among the family of 'resultative' constructions; imperfective aspect

The occurrence of the participle in the 'resultative' constructions is independent of the aspect of the verb from which the participle is formed. Table 2 repeats the templatic examples from Table 1, but this time all participles are perfective.

[1] There seems to be a convention in English to use possessive resultatives particularly in the context of services performed by others, where the resultative is closer in meaning to a causative, as in *We had the house painted; I'm having a copy of the report sent to you*; they are also used to express undesirable events, as in *He had his car stolen last night*. These uses are either unavailable in Polish, or only coincidentally available if they are compatible with the basic possessive interpretation of the possessive resultative.

	personal	impersonal
'objective resultative'	(6) *Mieszkanie było* flat(N).NOM was.3SG.N *wysprzątane (przez firmę).* clean-up.PART.SG.N.NOM (by company) 'The apartment was cleaned up (by a [professional] company).'	(7) *Było wysprzątane* was.3SG.N clean-up.PART.SG.N *(przez firmę).* (by company) '[It] was cleaned up (by a [professional] company).'
'possessive resultative'	(8) *Ojciec miał mieszkanie* father(M).NOM had.3SG.M flat(N).ACC *wysprzątane (przez firmę).* clean-up.PART.SG.N.ACC (by company) 'Father had the apartment cleaned up (by a [professional] company).'	(9) *Ojciec miał* father(M).NOM had.3SG.M *wysprzątane (przez firmę).* clean-up.PART.SG.N (by company) 'Father had [it] cleaned up (by a [professional] company).'

Table 2. The impersonal passive of the intransitive (7) among the family of 'resultative' constructions; perfective aspect

Finally, the *-ne/-te* participle is also found in the attributive use, as a modifier, as in (10a,b). This use can only be personal, and it will not be discussed further in this paper:

(10) a. *sprzątane mieszkanie*
clean.PART.SG.N.NOM flat(N).NOM
'[a/the] cleaned apartment'

b. *wysprzątane mieszkanie*
clean-up.PART.SG.N.NOM flat(N).NOM
'[a/the] cleaned-up apartment'

2 Argument structure

In the spirit of 'demotional' approaches, I treat passivisation as an operation on predicate argument structure which results in a different assignment of grammatical functions to the arguments of the predicate than that found in the default active argument structure. Specifically, the highest ranking unergative argument is 'downgraded' to an oblique, while the second argument (corresponding to the active direct object), if there is one, becomes the subject. Detailed argumentation in favour of this analysis of the passive, and extensive references to earlier literature, can be found in Kibort (2001, 2004, 2008). Additionally, following the widely shared understanding that the locus of argument structure is the whole predicate rather than the individual verb, I argue that passivisation is not an operation on a lexical item. Instead, all argument-structure-changing operations are operations on mapping templates within argument structure, producing different mapping templates that fit (semantically and syntactically) certain classes of predicates. Furthermore, I argue in the spirit of LFG that the argument-structure-changing operations do not need to produce (or, derive) one argument structure template from another in a procedural way. Instead, argument-structure-changing operations capture what can be understood as static relations between existing mapping templates – some of which may be 'basic' and others 'derived' – which are available for predicates. If a passive or other argument-structure-changing operation could be alternatively analysed as an operation on a lexical item, this is merely coincidental with the fundamental operation on the mapping template.

Thus, within the argument structure module of the grammar there may be a mapping template that represents the 'active diathesis' and another one that represents the 'passive diathesis', and what we call a passive operation relates the two templates and captures the conditions that have to be fulfilled by a predicate to fit both templates. Therefore, a mapping template can be understood as a constraint on argument structure that specifies a particular voice or diathesis for a class of predicates that fits it.

The following diagrams represent mapping templates for personal and impersonal passive predicates, respectively (the syntactic specifications of argument positions for features are omitted here for the sake of a simpler presentation, as they are not relevant for the points argued in this paper):

(11) personal passive predicate

$$\begin{array}{lccccc} & & \textit{ag} & \textbf{\textit{pat/th}} & \\ & & | & | & \\ \textit{predicate}_{\text{passive}} & \langle & \text{arg} & \text{arg} & \rangle \\ & & | & | & \\ & & (\text{OBL}_\theta) & \textbf{SUBJ} & \end{array}$$

(12) impersonal passive predicate

$$\begin{array}{lccc} & & \textit{ag} & \\ & & | & \\ \textit{predicate}_{\text{passive}} & \langle & \text{arg} & \rangle \\ & & | & \\ & & (\text{OBL}_\theta) & \end{array}$$

The impersonal passive in Polish has no overt subject (either lexical or a 'dummy' expletive one, since Polish does not have expletives); nor does it have a covert subject which could participate in syntactic control or binding. For example, since the reflexive possessive pronoun in (13a) needs to be bound by a subject, it cannot be licensed in a sentence without a subject. By comparison, non-reflexive possessive pronouns in (13b) are locally free, therefore the sentence is well-formed:

(13) a. **Było* *codziennie* *sprzątane* *we* *wszystkich*
was.3SG.N daily clean.PART.SG.N in all.PL.LOC
swoich *pokojach.*
own[REFL].PL.LOC rooms(NONMHUM).LOC
'There was cleaning every day in all of one's own rooms.'

cf. b. *Było* *codziennie* *sprzątane* *we* *wszystkich*
was.3SG.N daily clean.PART.SG.N in all.PL.LOC
naszych/*ich* *pokojach.*
our.LOC/their.LOC rooms(NONMHUM).LOC
'There was cleaning every day in all of our/their rooms.'

While it is understood that passivisation is meaning-preserving, or 'morphosyntactic' (Sadler and Spencer 1998), many other argument-structure-changing operations are meaning-altering. Anticausativisation, for example, is an operation on predicate argument structure that targets the level of argument positions and deletes the first core argument from the valency frame of the base predicate; the anticausative is, then, a lexical detransitiviser (which does not, however, delete the semantic participant of the predicate; see Kibort 2007, 2008 for discussion):

(14) a. *Tomek* *wylał* *zupę.*
Tomek(M).NOM spilt.3SG.M soup(F).ACC
'Tomek spilt the/some soup.'

b. *Zupa* *wylała* *się.*
soup(F).NOM spilt.3SG.F REFL
'The soup spilt.'

The following diagrams represent mapping templates for transitive (causative) and intransitive anticausative (or, inchoative) predicates, respectively (again, without the syntactic specifications of argument positions for features, for the sake of a simpler presentation):

(15) transitive predicate

$$\begin{array}{lccc} & \textit{ag} & \textit{pat/th} & \\ & | & | & \\ \textit{predicate}_{\text{trans}} & \langle\ \text{arg} & \text{arg} & \rangle \\ & | & | & \\ & \text{SUBJ} & \text{OBJ} & \end{array}$$

(16) intransitive anticausative predicate

$$\begin{array}{lccc} & \textit{ag} & \boldsymbol{\textit{pat/th}} & \\ & & | & \\ \textit{predicate}_{\text{anticaus}} & \langle & \text{arg} & \rangle \\ & & | & \\ & & \textbf{SUBJ} & \end{array}$$

Note that both the passive and the anticausative argument structure templates link a SUBJ argument with a *patient/theme* – this is indicated in bold in diagrams (11) and (16).

3 The range and the interpretation of the *-n-/-t-* participle

The impersonal *-ne/-te* construction exemplified in (1) is considered to be an instance of the impersonal passive of the intransitive because it is possible to establish the following active-passive alternation:

(17) a. *Firma* *codziennie* *sprzątała.*
company(F).NOM daily cleaned.3SG.F
'The [professional] company cleaned every day.'

b. *Było* *codziennie* *sprzątane* *(przez firmę).*
was.3SG.N daily clean.PART.SG.N (by company)
'There has been cleaning every day (by a company).'

On the other hand, the impersonal *-ne/-te* construction – particularly the perfective variant, as in (7) – can be argued to be an instance of a non-passive predicative adverbial construction, such as (18), where the *-o* ending on 'clean' is unambiguously adverbial:

(18) *W pokoju było* *czysto.*
in room was.3SG.N cleanly
'[It] was clean in the room.'

This analysis seems particularly appropriate for sentences such as (7), or (19a), since we observe the following analogy:

(19) a. *W pokoju było* *ładnie* ***wysprzątane*** (*przez firmę*).
in room was.3SG.N nicely clean/tidy-up.PART.SG.N (by company)
'[It] was nicely cleaned/tidied up (by a company) in the room.'

b. *W pokoju było* *czysto* *i* ***wysprzątane***.
in room was.3SG.N cleanly and clean/tidy-up.PART.SG.N
'[It] was clean and tidied up in the room.'

The same problem of interpretation occurs in personal sentences with the *-n-/-t-* participle. While it may be reasonable to argue that some personal sentences with an auxiliary/copula and a *-n-/-t-* participle, such as (20a), are quite clearly passive, and others, such as (20b), are quite clearly non-passive, in a vast number of cases the passive versus non-passive interpretation is impossible to establish with any certainty. Examples (20c-d) illustrate that some very common occurrences of the *-n-/-t-* participle are genuinely ambiguous between being passive versus non-passive, and that it is possible to establish an alternation between the *-n-/-t-* construction and either an active transitive variant or an intransitive anticausative (inchoative) variant:[2]

(20) a. *List* *jest/był* ***ukradziony***.
letter(M).NOM is/was.3SG.M steal.PART.SG.M.NOM
'The letter is/was stolen.'
~ Someone stole the letter. <u>passive</u>?

b. *On* *jest/był* ***wyspany***.
he.NOM is/was.3SG.M sleep.PART.SG.M.NOM
'He is well-slept.' (= He has slept well)
~ *He has been slept by someone./*Someone has slept him (e.g. his mother). <u>active</u>?

c. *On* *jest/był* ***zmartwiony***.
he.NOM is/was.3SG.M worry.PART.SG.M.NOM
'He is/was worried.'
~ *On zmartwił się.* 'He has (become) worried.'
~ *Ten problem go zmartwił*. 'This problem has worried him/ got him worried.' <u>both</u>?

d. *Silnik* *jest/był* ***zepsuty***.
engine(M).NOM is/was.3SG.M break.PART.SG.M.NOM
'The engine is/was broken.'
~ *Silnik zepsuł się.* 'The engine has broken.'
~ *Tomek zepsuł silnik*. 'Tomek has broken the engine.' <u>both</u>?

This means that neither personal nor impersonal sentences with the predicatively used *-n-/-t-* participle can be unambiguously assigned the passive argument structure in (11)-(12). Therefore, their syntactic model should not specify whether they instantiate a passive or non-passive construction, but they should remain underspecified with regard to this distinction.

[2] The anticausative construction in Polish uses the derived reflexive form of the verb, hence the presence of the reflexive marker *się* in the anticausative alternants in examples (20c,d).

4 Morphology and lexical entries

The Polish *-n-/-t-* participle underlies a set of forms which make up an inflectional paradigm identical to that of the adjective:[3]

(21)

	'red.NOM'	'clean/tidy.PART.NOM'
[SG].M	*czerwon-y*	*sprzątan-y*
[SG].F	*czerwon-a*	*sprzątan-a*
[SG].N	*czerwon-e*	*sprzątan-e*
[PL].MHUM[4]	*czerwon-i*	*sprzątan-i*
[PL].NONMHUM	*czerwon-e*	*sprzątan-e*

Like adjectives, *-n-/-t-* participles denoting gradable concepts have comparative and superlative forms (analytical), e.g. *bardziej/najbardziej zepsuty/zmartwiony* 'more/the most broken/worried'.

The *-n-/-t-* participle is often labelled a 'passive participle' but at the same time it is widely acknowledged that the 'passive' label does not fit all participles in this class. Authors of academic grammars always clarify that the label 'passive participle' merely groups all participles sharing the same morphological form, including 'morphologically passive participles with active meaning'.

I argue that the *-n-/-t-* participle is not passive, but it should be correctly recognised as the so-called 'resultative participle'. In the typological literature, resultative participles are 'those verb forms that express a state implying a previous event' (Nedjalkov and Jaxontov 1988). A resultative participle characterises its head by expressing a state that (typically) results from a previous event.

Resultative participles have a semantic orientation (Lehmann 1984): they are oriented towards the affected participant. (This is similar to deverbal adjectives which also have a semantic orientation, e.g. *dreadful* vs *fearful*). The affected participant is typically ***patient/theme*** or ***experiencer/undergoer***. However, if a transitive event can be construed as affecting the agent, agent-oriented resultative participles can be formed from transitive unergative verbs – this is widely attested typologically particularly with verbs of obtaining, wearing, ingestion, and 'mental ingestion' (Nedjalkov and Jaxontov 1988: 9, cf. Haspelmath 1994: 174, fn. 10), and is also found in English: *he is drunk*, *a confessed killer, a practised liar, a recanted Chomskyan, a well-read person,* etc. (examples from Bresnan 2001: 34-36).

---

[3] This section summarises the discussion presented in more detail in Kibort (2005 and 2011).

[4] MHUM = masculine human gender; NONMHUM = non-masculine-human gender, i.e. all other than masculine human.

Syntactically, resultative participles modify their head noun or complement their **subject**.

Since not all predicative constructions with the resultative participle can be unambiguously assigned a passive or non-passive argument structure, I do not consider the formation of the adjectival resultative participle to follow the passivisation of the predicate, as is the standard assumption in the LFG tradition. Instead, as in Kibort (2005) and Thomas (2012), I assume the following participle-adjective conversion rule independent of passivisation:

(22) Morphological change: $V \rightarrow [V_{Part}]_{A/V}$
Operation on lexical form: (non-oriented) P → semantically oriented P

In the formation of the resultative participle, the derived lexical form P is semantically oriented towards the affected participant. The semantic orientation does not involve the syntactic notions of subject or object. The participles can be used attributively or predicatively, some being suitable for the passive construction.

Since sentences with 'be' and resultative participles are underspecified with regard to whether they instantiate a passive construction or a non-passive predicative adjectival/adverbial construction, I propose that they are most straightforwardly analysed as having a copula 'be' combined with a predicative element (here, the participle):

(23) 'be' + resultative participle = copula 'be' + predicative element ($Part_{A/V}$)

Recall that I assume that the operations of passivisation and anticausativisation are performed on the argument structure of a predicate, i.e. neither on a syntactic structure (f- or c-structure) nor on an individual lexical item (such as the base verb – which instead undergoes morphological derivations such as the one in (22)). Passivisation produces an argument structure in which the *agent* argument is assigned the grammatical function of the OBLIQUE and is optional, and the *patient/theme* argument if there is one is assigned the function of the SUBJECT (see (1b) for the intransitive variant).

Recall also that the argument structure templates for both a personal passive predicate, as in (11), and an intransitive anticausative predicate, as in (16), make available to the syntax a subcategorisation frame in which the subject argument (**SUBJ**) is interpreted as a ***patient/theme***. Therefore, a resultative participle oriented towards an affected participant (here, the *patient/theme*), and used syntactically as a modifier of its head noun or a complement to its subject, must be listed in the lexicon as having a subcategorisation frame which is compatible with both the passive interpretation and the intransitive anticausative interpretation of the predicate. That is, if its subcategorisation frame does have a subject, the subject is indeed interpreted as a patient/theme.

To check this, let us consider again the participial derivation rule in (22). The lexical entry for the resultative participle formed from *sprzątać* 'to clean' includes the following subcategorisation frame: *sprzątan-* ⟨ SUBJ, (OBL) ⟩. The interpretation of a resultative participle is such that its SUBJECT argument is interpreted as the participant which is affected by the event. This is compatible with passive argument structure, and hence the participle may indeed receive a passive interpretation. The resultative participle *pęknięt-* ⟨ SUBJ ⟩ 'cracked' does not have an optional OBL in its subcategorisation frame, hence the event it expresses cannot be interpreted as caused by any external participant, and the only interpretation available is that the sole SUBJ participant is affected by itself. With the resultative participle *otwart-* ⟨ SUBJ, (OBL) ⟩ 'open(ed)', the SUBJ participant may be interpreted as either affected by the OBL participant, or by itself. The first interpretation is compatible with passive argument structure, while the second with anticausative argument structure. The absence of an overt expression of an OBL argument in the clause makes the construction underspecified with regard to which argument structure it instantiates. This underspecification is possible because in both the passive and the anticausative the same referent is mapped onto the SUBJ. (Note that the same underspecification applies correctly to *sprzątan-* ⟨ SUBJ, (OBL) ⟩ 'clean(ed)'; that is, the grammar does not prevent the construction of an anticausative clause with *sprzątać* 'to clean', however, this usage would constitute a creative, perhaps jocular, extension of its meaning).

5 Distribution of the resultative participle as a predicative element

Tables 1 and 2 in Section 1 showed the impersonal passive of the intransitive among the family of the so-called 'resultative' constructions in Polish. However, it is useful to consider these constructions in the context of the full distribution of the *-n-/-t-* (resultative) participle in Polish.

The resultative participle as a predicative element in personal constructions – that is, with the subject of predication – is found in:

[A] Finite analytic predicates (subject + copula + predicative element):

Mieszkanie [$_V$ *było*] [$_{PartP}$ *sprzątane* *(przez firmę)*].
flat(N).NOM was.3SG.N clean.PART.SG.N (by company)
'Apartment was cleaned (by a company).'

[B] <u>Independent non-finite predicates</u> (no copula, only subject + predicative element):

Mieszkanie [$_{\text{PartP}}$ *sprzątane* *(przez firmę)*].
flat(N).NOM clean.PART.SG.N.NOM (by company)
'Apartment cleaned (by a company).'

[C] <u>Dependent non-finite predicates</u> (no copula, only subject + predicative element), i.e. 'small clauses' fulfilling the functions of subjects, objects, indirect objects, obliques, in main clauses:

(i) <u>in apposition to subject</u>
Mieszkanie [$_{\text{PartP}}$ *sprzątane* *(przez firmę)*] *błyszczało*.
flat(N).NOM clean.PART.SG.N.NOM (by company) shone.3SG.N
'Apartment cleaned (by a company) was shining.'

(ii) <u>in apposition to object</u> = **the personal 'possessive resultative'**[5]
Ojciec *miał* *mieszkanie* [$_{\text{PartP}}$ *sprzątane* (*przez firmę*)].
father(M).NOM had.3SG.M flat(N).ACC clean.PART.SG.N.ACC (by company)
'Father had the apartment cleaned (by a company).'

(iii) <u>in apposition to indirect objects and obliques</u>
Ojciec *zafundował* *swojemu* *mieszkaniu*
father(M).NOM sponsored.3SG.M own[REFL].SG.DAT flat(N).DAT
[$_{\text{PartP}}$ *sprzątanemu* (*przez firmę*)] *wiosenną metamorfozę*.
clean.PART.SG.N.DAT (by company) spring-like.F.ACC transformation(F).ACC
'Father gave his apartment cleaned (by a company) a spring-like transformation.'

Ojciec *zamieszkał* *w mieszkaniu* [$_{\text{PartP}}$ *sprzątanym*
father(M).NOM started-living.3SG.M in flat(N).LOC clean.PART.SG.N.LOC
od *kilku lat* (*przez firmę*)].
from few years (by company)
'Father took accommodation in an apartment cleaned for a few years (by a company).'

Fig. 1. The resultative participle as a predicative element in personal constructions in Polish

Having constructed a catalogue of the occurrences of the resultative participle as a predicative element in personal constructions, the following question can be asked: which of the above types of predicates can be used without the subject of predication, i.e. without a referent of which to predicate (not just without an overt subject)?

[5] This construction might be in the process of being grammaticalised as a new 'perfect' tense in Polish.

Without the subject of predication, the predication has to denote an 'ambient' characteristic or quality and can be general (in an unspecified location) or refer to a particular location specified in the clause or understood from the context. This interpretation restricts the range of possible constructions. Specifically, it does not make sense to create such clauses with independent non-finite predicates, as this would amount to creating independent small clauses without subjects (i.e. subjectless variants of [B]).

However, it is possible to create subjectless clauses with finite analytic predicates made up of a copula + predicative element, i.e. subjectless [A]. The predicative elements which are felicitous in this construction are those which can denote an ambient characteristic – some adverbs, and some resultative participles, e.g.:

(24) a. *Jest/Było czysto/cicho/dobrze/wygodnie.*
is/was.3SG.N clean-ly/quiet-ly/well/comfortabl-y
'[It] was clean/quiet/good/comfortable [there].'

b. *Jest/Było sprzątane/wysprzątane.*
is/was.3SG.N clean/clean-up.PART.SG.N
'[It] was cleaned/tidied up [there].'

Also, it is possible to create subjectless small clauses when the small clause fulfils the direct object function of a main clause, i.e. subjectless [C]. This is possible exclusively with the main verb 'have'. Again, the predicative elements which are felicitous in this construction are adverbs and resultative participles, but their range seems to be wider than in the 'subjectless [A]' type. This is the impersonal variety of the so-called 'possessive resultative', on which research does not yet exist:

(25) a. *Miał dobrze/wygodnie.*
had.3SG.M well/comfortabl-y
'He was fine/comfortable.'

(26) a. *Miał codziennie sprzątane.*
had.3SG.M every-day tidy.PART.SG.N
'He had the cleaning done every day.'

b. *Miał codziennie sprzątane w pokoju.*
had.3SG.M every-day tidy.PART.SG.N in room(M).LOC
'He had the cleaning done in his room every day.'

(27) a. *Miał ładnie posprzątane.*
had.3SG.M nicely tidy-up.PART.SG.N
'He had [it] nicely cleaned/tidied up.'

b. *Miał ładnie posprzątane w pokoju.*
had.3SG.M nicely tidy-up.PART.SG.N in room(M).LOC
'He had it nicely cleaned/tidied up in his room.'

Examples of both variants of the subjectless *-ne/-te* construction are numerous both in Polish corpora and on the web, and many natural, non-constructed examples can be found in Kibort (2011).

To sum up, the resultative participle as a predicative element is also found in impersonal constructions:

SUBJECTLESS VARIANT OF [A]:
in finite analytic predicates made up of a copula + predicative element; denoting an ambient characteristic:
Było sprzątane (przez firmę).
was.3SG.N clean.PART.SG.N (by company)
'[It] was cleaned (by a company).'

SUBJECTLESS VARIANT OF [C]:
in dependent non-finite predicates which fulfil the direct object function of a main clause; exclusively with the main verb 'have':
Ojciec miał sprzątane (przez firmę).
father(M).NOM had.3SG.M clean.PART.SG.N (by company)
'Father had [it] cleaned (by a company).'
= the impersonal 'possessive resultative'

Fig. 2. The resultative participle as a predicative element in impersonal constructions in Polish

5 Functional structure – proposal

The present proposal for the f-structure of construction [B], with the resultative participle as a predicative element in independent non-finite predicates, follows the suggestion made in Dalrymple, Dyvik and King (2004: 191), as there is no evidence in Polish for an empty copula.

Constructions [Ci-iii], with the resultative participle as a predicative element in dependent non-finite predicates, have XADJs adjoined to the phrase like non-restrictive clauses, with their subjects provided from outside the predicate.

[B] PRED ‘sprzątane <SUBJ, (OBL)>’
SUBJ [PRED ‘mieszkanie’]
OBL [PRED ‘przez firmę’]

[Cii] PRED ‘miał <SUBJ, OBJ>’
SUBJ [PRED ‘ojciec’]
OBJ [PRED ‘mieszkanie’
XADJ [PRED ‘sprzątane <SUBJ, (OBL)>’
SUBJ
OBL [PRED ‘przez firmę’]]]

[Ci] PRED ‘pięknialo <SUBJ>’
SUBJ [PRED ‘mieszkanie’
XADJ [PRED ‘sprzątane <SUBJ, (OBL)>’
SUBJ
OBL [PRED ‘przez firmę’]]]

[Ciii] PRED ‘zafundował <SUBJ, OBJ, OBJθ>’
SUBJ [PRED ‘ojciec’]
OBJ [PRED ‘wiosenną metamorfozę’]
OBJθ [PRED ‘mieszkaniu’
XADJ [PRED ‘sprzątanemu <SUBJ, (OBL)>’
SUBJ
OBL [PRED ‘przez firmę’]]]

Fig. 3. Proposed f-structures for constructions [B] and [Ci-iii]

In construction [A], with the resultative participle as a predicative element in finite analytic predicates, the finite verb ‘be’ is analysed as a raising verb (following Przepiórkowski 2001; Dalrymple, Dyvik and King 2004; Ørsnes 2006; and the current LFG grammar for Polish outlined e.g. in Patejuk and Przepiórkowski 2012). Specifically, the SUBJ of ‘be’ is identical to the SUBJ of the subordinate predicate: its SUBJ functionally controls the SUBJ of the subordinate predicative element; however, the SUBJ of ‘be’ is its semantic argument (note that Polish has no expletives).

The form ‘be.3SG.N’ can occur with two different subcategorisation frames, a personal one, used in construction [A], and an impersonal one, used in SUBJECTLESS [A], where it combines with predicative elements that do not subcategorise for a subject, i.e. adverbs and the impersonal resultative participle in SG.N.

The *-n/-t* resultative participle in SG.N likewise has two subcategorisation frames, which combine with the two variants of ‘be.3SG.N’ accordingly:

$$\textbf{[A]}\quad \left[\begin{array}{ll} \text{PRED} & \text{‘było <SUBJ, XCOMP>’} \\ \text{SUBJ} & [\text{PRED} \quad \text{‘mieszkanie’}] \\ \text{XCOMP} & \left[\begin{array}{ll} \text{PRED} & \text{‘sprzątane <SUBJ, (OBL)>’} \\ \text{SUBJ} & \\ \text{OBL} & [\text{PRED} \quad \text{‘przez firmę’}] \end{array}\right] \end{array}\right]$$

Fig. 4. Proposed f-structure for construction [A]

$$\textbf{[sA]}\quad \left[\begin{array}{ll} \text{PRED} & \text{‘było}_{\text{impers}}\text{ <XCOMP>’} \\ \text{XCOMP} & \left[\begin{array}{ll} \text{PRED} & \text{‘sprzątane}_{\text{impers}}\text{ < (OBL)>’} \\ \text{OBL} & [\text{PRED} \quad \text{‘przez firmę’}] \end{array}\right] \end{array}\right]$$

Fig. 5. Proposed f-structure for construction [SUBJECTLESS A]

Finally, all inflectional forms of ‘have’ also need two subcategorisation frames – let us call them object-personal and object-impersonal. The object-impersonal ‘have’ is used in the SUBJECTLESS [C] type of clause, also with predicative elements that do not subcategorise for a subject (i.e. adverbs and the impersonal resultative participle in SG.N):

(28) a. $\textit{miał}_{\text{obj-personal}}$ <SUBJ, OBJ>

b. $\textit{miał}_{\text{obj-impersonal}}$ <SUBJ, XCOMP> **[sC]**

References

Bresnan, Joan. 2001. *Lexical-Functional Syntax.* Oxford: Blackwell.

Dalrymple, Mary, Helge Dyvik and Tracy Holloway King. 2004. 'Copular complements: closed or open?' In: Butt, Miriam and Tracy H. King (eds) *Proceedings of the LFG04 Conference, University of Canterbury.* Stanford, CA: CSLI Publications. 188-198.

Haspelmath, Martin. 1994. Passive participles across languages. In: Fox, Barbara and Paul J. Hopper (eds) *Voice: Form and Function*. Amsterdam: Benjamins. 151-177.

Kibort, Anna. 2001. The Polish passive and impersonal in Lexical Mapping Theory. In: Butt, Miriam and Tracy H. King (eds) *Proceedings of the LFG01 Conference, University of Hong Kong.* Stanford, CA: CSLI Publications. 163-183.

Kibort, Anna. 2004. *Passive and passive-like constructions in English and Polish.* PhD thesis, University of Cambridge.

Kibort, Anna. 2005. 'The INs and OUTs of the Participle-Adjective Conversion Rule'. In: Butt, Miriam and Tracy H. King (eds) *Proceedings of the LFG05 Conference, University of Bergen.* Stanford, CA: CSLI Publications. 205-225.

Kibort, Anna. 2007. 'Extending the applicability of Lexical Mapping Theory'. In: Butt, Miriam and Tracy H. King (eds) *Proceedings of the LFG07 Conference, Stanford University.* Stanford, CA: CSLI Publications. 250-270.

Kibort, Anna. 2008. 'On the syntax of ditransitive constructions'. In: Butt, Miriam and Tracy H. King (eds) *Proceedings of the LFG08 Conference, University of Sydney.* Stanford, CA: CSLI Publications. 312-332.

Kibort, Anna. 2011. The elephant in the room: the impersonal *-ne/-te* construction in Polish. In: Malchukov, Andrej and Anna Siewierska (eds) *Impersonal Constructions: A Cross-linguistic Perspective.* Amsterdam: Benjamins. 357-394.

Lehmann, Christian. 1984. *Der Relativsatz*. Tübingen: Gunter Narr.

Nedjalkov, Vladimir P. and Sergej J. Jaxontov. 1988. 'The typology of resultative constructions'. In: Nedjalkov, Vladimir P. (ed.) *Typology of Resultative Constructions*. Amsterdam: Benjamins. 4-62.

Ørsnes, Bjarne. 2006. 'Creating raising verbs. An LFG-analysis of the complex passive in Danish'. In: Butt, Miriam and Tracy H. King (eds) *Proceedings of the LFG06 Conference, Universität Konstanz*. Stanford, CA: CSLI Publications. 386-405.

Patejuk, Agnieszka and Adam Przepiórkowski. 2012. 'Towards an LFG Parser for Polish. An Exercise in Parasitic Grammar Development'. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.* 3849-3852.

Przepiórkowski, Adam. 2001. 'Case and agreement in Polish predicates'. In: Franks, Steven, Tracy H. King and Michael Yadroff (eds) *Annual Workshop on Formal Approaches to Slavic Linguistics: The Bloomington Meeting 2000*. Ann Arbor, MI: Michigan Slavic Publications. 257-273.

Sadler, Louisa and Andrew Spencer. 1998. Morphology and argument structure. In: Spencer, Andrew and Arnold M. Zwicky (eds) *The Handbook of Morphology*. Oxford: Blackwell. 206-236.

Thomas, Victoria. 2012. 'Stative passives in German'. Talk given at the 7th South of England LFG Meeting, 3 March 2012, SOAS, University of London.

ON THE (UN)BEARABLE LIGHTNESS OF BEING AN LFG STYLE COPULA IN HUNGARIAN

Tibor Laczkó
University of Debrecen

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper develops the first comprehensive LFG analysis of the five most important types of copula constructions in Hungarian. I basically adopt Dalrymple et al.'s (2004) programmatic view, which admits diversity both in c-structure and in f-structure across and within languages, and which postulates that the "postcopular" constituent can be the functional cohead of the copula, in which case the copula itself is only a formative, it can have the open XCOMP function and it can also have the closed PREDLINK function. (This contrasts with Butt et al.'s (1999) and Attia's (2008) uniform PREDLINK approach at the f-structure level.) On the basis of the behavior of the construction types in question, I employ the functional cohead device, the PREDLINK tool (but not as the only uniform tool); however, (contrary to Dalrymple et al. (2004)) I claim that there is no need for the XCOMP treatment. At the same time, I also argue that in the case of some construction types, it is most appropriate to assume that the postcopular constituent has an OBL function.

1 Introduction

Copula constructions (CCs) in Hungarian have received relatively little attention in the Chomskyan generative literature and practically no attention in the LFG literature.[1] In this paper, I propose an outline of the first comprehensive LFG analysis of the five most salient Hungarian CCs, partially reflecting on and capitalizing on empirical and theoretical generalizations and analyses in the relevant LFG literature (e.g., Butt et al. 1999, Dalrymple et al. 2004, Nordlinger & Sadler 2007, Attia 2008, Sulger 2009). This may also result in a meaningful typological and theoretical contribution to LFG's understanding and handling CCs across languages.

The following Hungarian CCs will be analyzed here.

(1)	*Az*	*igazgató*	*okos/tanár*	*volt.*	[attribution or
	the	director.NOM	clever/teacher.NOM	was	classification]

'The director was clever / a teacher.'

(2)	*Az*	*igazgató*	*a*	*szóvivő*	*volt.*	[identity]
	the	director.NOM	the	spokesman.NOM	was	

'The director was the spokesman.'

(3)	*Az*	*igazgató*	*a*	*szobá-ban*	*volt.*	[location]
	the	director.NOM	the	room-in	was	

'The director was in the room.'

[1] For a recent overview of the GB/MP analyses of certain types of Hungarian CCs, see Dalmi (2010).

(4) *Voltak* *boszorkány*-*ok* *(a* *Föld*-*ön).* [existence]
were witch-PL.NOM the Earth-on
'There were witches (on the Earth).'

(5) *Az* *igazgató*-*nak* *volt* *szóvivő*-*je.* [possession]
the director-DAT was spokesman-his.NOM
'The director had a spokesman.'

The structure of the paper is as follows. In section 2, I offer a brief overview of the main LFG approaches to CCs. In section 3, first I present my view of how CCs are best treated in an LFG framework, and then I develop my analysis of the five Hungarian CC types exemplified in (1)-(5). This is followed by a summary and some concluding remarks in section 4.

2 Fundamental LFG approaches

The two main general LFG strategies for the treatment of CCs across languages are best illustrated by Butt et al. (1999) and Dalrymple et al. (2004). In the former approach, CCs are handled in a uniform manner functionally. The copula is always taken to be a two-place predicate, and the two arguments it subcategorizes for have the following two grammatical functions: there is a subject (SUBJ) (which is uncontroversial in any analysis of these constructions), and the other constituent is uniformly assigned a special, designated function designed for the second, "postcopular" argument of the predicate: PREDLINK. By contrast, in Dalrymple et al.'s (2004) approach, the two-place predicate, SUBJ and PREDLINK version is just one of the theoretically available options. In addition, they postulate that the copula can be devoid of meaning (and, hence, argument structure) and it can serve as a pure carrier of formal verbal features: tense and agreement. Finally, it can also be a one-place predicate of the "raising" type: assigning the XCOMP function to its propositional argument and also assigning a non-thematic SUBJ function. When the postcopular constituent has the PREDLINK function, it is closed in the sense that if it has a subject argument, this argument is never realized outside this constituent. For obvious reasons, the XCOMP and the PREDLINK types involve two semantic (and functional) levels (tiers): the copula selects the relevant constituent as an argument. By contrast, when the copula is a mere formative, the two elements are at the same level (tier): the postcopular constituent is the real predicate and the copula only contributes morphosyntactic features. In LFG terms, they are functional coheads. All this is summarized in Figure 1.

postcopular constituent		
open		closed
main **PRED**, the copula is a formative: functional coheads (single-tier)	**XCOMP** of the copula main PRED: ‘be <XCOMP>SUBJ’ (double-tier)	**PREDLINK** of the copula main PRED: ‘be <SUBJ,PREDLINK>’ (double-tier)

Figure 1. Three types of copular constructions

In (7), (8) and (9) I show schematically how the English sentence in (6) can be analyzed along these three different lines.

(6) *She is small.*

(7)
$$\left[\begin{array}{ll} \text{PRED} & \text{‘small < (}\uparrow\text{SUBJ) >’} \\ \text{TENSE} & \text{present} \\ \text{SUBJ} & \text{“she”} \end{array}\right]$$

(8)
$$\left[\begin{array}{ll} \text{PRED} & \text{‘be < (}\uparrow\text{XCOMP) > (}\uparrow\text{SUBJ)’} \\ \text{TENSE} & \text{present} \\ \text{SUBJ} & \text{“she”} \\ \text{XCOMP} & \left[\begin{array}{ll} \text{PRED} & \text{‘small < (}\uparrow\text{SUBJ) >’} \\ \text{SUBJ} & \end{array}\right] \end{array}\right]$$

(9)
$$\left[\begin{array}{ll} \text{PRED} & \text{‘be < (}\uparrow\text{SUBJ) (}\uparrow\text{PREDLINK) >’} \\ \text{TENSE} & \text{present} \\ \text{SUBJ} & \text{“she”} \\ \text{PREDLINK} & \text{“small”} \end{array}\right]$$

One of the most important properties of this approach is that it allows for diversity both in c-structure and in f-structure. Dalrymple et al. (2004) is programmatic: it proposes these three analytical possibilities and assumes that there can be variation across languages and also across constructions within the same language. Only a careful analysis of any single CC in any language can reveal which type it belongs to. Falk (2004) and Nordlinger & Sadler (2007) subscribe to this view and develop their respective analyses in this spirit. By contrast, Attia (2008), inspired by Butt et al. (1999), argues for a generalized PREDLINK approach to CCs within and across languages. Naturally, this means diversity in c-structure and robust uniformity in f-structure, and, for obvious reasons, it radically simplifies the analysis of CCs in the realm of grammatical relations and f-structure. It is in this sense that I consider this PREDLINK approach "light". In addition, the single-tier (formative) use of the copula is also "light" in an obviously different sense.[2] As I will point out when I present my analysis, the PREDLINK lightness in this domain inevitably puts the burden of capturing significant differences of various kinds between CCs on other components of grammar.

3 Analysis of the five Hungarian CC types

Before presenting the details, I discuss the most important general aspects of my analysis.

My approach is along the lines (i.e. analytical philosophy) pursued by Dalrymple et al. (2004), Falk (2004), and Nordlinger & Sadler (2007), as opposed to the path argued for and followed by Butt et al. (1999), Attia (2008) and Sulger (2011). This means that I find it more appropriate to allow for variation in terms of categories, functions, and construction types within and across languages in the CC domain rather than to develop a generalized and unified analysis for the overwhelming majority of CCs within and across languages. In my view, it is more in the spirit of LFG, I consider it is more appealing intuitively, and, furthermore, it is my conviction that the variation and the variety Hungarian CCs exhibit call for a varied and multidimensional treatment.

Naturally, this is not to deny the tenability and potential advantages of the unified approach ("PREDLINK light"); however, I will show that in the case of the investigation of CC phenomena we gain much more by accommodating rich parametric variation in several dimensions. My claim is that although it is elegant to have a uniform treatment at f-structure, it is also the job of f-structure to efficiently feed semantics, and my approach is more useful in this respect. At this point I would also like to emphasize that I do not reject the PREDLINK analysis as such: in the case of two Hungarian CCs (out of the five discussed in this paper) I myself develop a PREDLINK account.

In addition to the PREDLINK strategy, I also employ the single-tier (functional cohead) version. It is important in this connection that in certain Hungarian CCs the copula must be absent in certain cases. Such a fact by

[2] The title of this paper has been inspired by the title of the following book: Milan Kundera, *The Unbearable Lightness of Being*, 1985, Faber & Faber (translated from Czech by Michael Henry Heim).

itself is taken to justify the single-tier analysis in a number of approaches. However, my claim is that the possibility/necessity of having the zero copula (at least in certain paradigmatic slots) is neither a sufficient nor a necessary condition for a single tier analysis. Consider the following two sides of this ±zero-copula-coin.

(A) Compare my accounts is sections 3.1 and 3.2 in this respect: both CC types exibit exactly the same copula-absence behavior; however, I analyze the former in the single-tier, functional cohead manner, while I develop an analysis of the latter along the double-tier, PREDLINK lines.

(B) The obligatory presence of the copula does not necessarily rule out the single-tier analysis: see the more recent LFG analysis of English passive constructions (the copula is merely a formative element without a PRED feature).[3]

Contrary to Dalrymple et al. (2004) (and the views of the overwhelming majority of LFG practitioners), I claim that there is no real need for the double-tier XCOMP analysis of CCs in general. I make this claim on the basis of Dalrymple et al.'s (2004) argumentation (by pointing out that it is not very convincing) and on the basis of the relevant Hungarian facts. I hasten to add that I do not exclude the possibility that certain CC phenomena may call for an XCOMP analysis as the most plausible (or maybe the only feasible) analysis.

Let us take a look at Dalrymple et al.'s (2004) two arguments in favour of XCOMP in certain CCs.

(A) When the English copula is combined with an adjectival "raising" predicate, the well-known control relationships can be captured by dint of the standard LFG control apparatus if the AP is assumed to have the XCOMP function, rather than the PREDLINK function. The crucial aspects of these two different analyses of (10) are shown in (11) and (12).

(10) *It is likely to rain.* (cf. *It seems to rain.*)

(11) a. *is*, V 'be < (↑XCOMP) > (↑SUBJ)'
(↑SUBJ) = (↑XCOMP SUBJ)

b. *likely,* A 'likely < (↑XCOMP) > (↑SUBJ)'
(↑SUBJ) = (↑XCOMP SUBJ)

[3] For instance, Bresnan (2001) adopts this analysis, as opposed to the classical XCOMP analysis in Bresnan (1982).

(12) a. *is*, V ‘be < (↑PREDLINK) > (↑SUBJ)’

b. *likely*, A ‘likely < (↑COMP) > (↑SUBJ)’
(↑COMP SUBJ) = ((PREDLINK↑) SUBJ)

As (12b) shows, only a rather unusual control equation could handle this relation on the PREDLINK account of the copula, while nothing special is required on the XCOMP account, see (11). I fully agree with Dalrymple et al. (2004): the PREDLINK analysis is too costly, and I find this an important argument against a uniform PREDLINK approach to CCs (contra Attia’s (2008) claim to the contrary). However, notice that this is only an argument against the PREDLINK account: a simple single-tier analysis allows for exactly the same standard LFG way of capturing the relevant control relationships. Compare (13) and (14).

(13) a. *is*, V
(↑TENSE) = present
(↑SUBJ PERS) = 3
(↑SUBJ NUM) = sg

b. *likely*, A ‘likely < (↑XCOMP) > (↑SUBJ)’
(↑SUBJ) = (↑XCOMP SUBJ)

(14) *seems*, V ‘seem < (↑XCOMP) > (↑SUBJ)’
(↑SUBJ) = (↑XCOMP SUBJ)
(↑TENSE) = present
(↑SUBJ PERS) = 3
(↑SUBJ NUM) = sg

As these representations demonstrate, on this single-tier account, *is likely* gets exactly the same analysis as *seems* (as is to be expected): the PRED feature is contributed by *likely* and *seem*, respectively, and the general morphosyntactic verbal features are provided by *is* and -*s*, respectively.

(B) Dalrymple et al.’s (2004) second argument is based on subject-adjective agreement in languages like French and Norwegian. Consider their French examples in (15) and their two alternative representations capturing the relevant agreement facts. Needless to say, the PREDLINK approach creates unnecessary complications, as shown in (17).

(15) a. *Elle* *est* *petite.*
she.F.SG is small.F.SG
‘She is small.’

b. *Il* *est* *petit.*
he.M.SG is small.M.SG
‘He is small.’

(16) *petite* (↑PRED) = 'small < SUBJ >'
(↑SUBJ NUM) =c sg
(↑SUBJ GEND) =c fem

(17) *petite* (↑PRED) = 'small'
((PREDLINK↑) SUBJ NUM) =c sg
((PREDLINK↑) SUBJ GEND) =c fem

My comment is the same as in the case of the previous point: this is an absolutely valid argument against the PREDLINK analysis in such cases, but the single-tier analysis is at least as unmarked and straightforward in LFG terms as the XCOMP analysis. Moreover, it may even be taken to be more compelling inasmuch as the adjective imposes its agreement constraints on the subject of the sentence directly (and not through the mediation of an XCOMP style control relationship).

Let me also add that according to several LFG practitioners the XCOMP analysis of the copula in passive sentences in English type languages is no longer tenable, see Footnote 3.[4]

So far, I have pointed out that in my approach I employ both the single-tier analysis and the (double-tier) PREDLINK analysis. In the double-tier domain, however, I reject the use of the XCOMP analysis. At the same time, I will also argue that in this latter domain it is reasonable to assume that in the case of certain CCs the second argument has the OBL (and not the PREDLINK) function. Notice that even with this additional grammatical function in my system the number of the fundamental types of CCs is smaller than that in Dalrymple et al.'s (2004) system. Consider:

(18) Dalrymple et al. (2004):
a. single-tier, functional cohead (open)
b. double-tier, PREDLINK (closed)
c. double-tier, XCOMP (open)

(19) here:
a. single-tier, functional cohead (open)
b. double-tier, PREDLINK or OBL (closed)

Before I present my analysis, I show the most essential features of the account of each type in (20).

[4] The main motivation for dropping the XCOMP analysis and replacing it with the single-tier, functional cohead analysis has been to represent the f-structures of passive sentences in copular passive languages like English and non-copular passive languages like Malayalam in a uniform fashion.

(20)	a. attribution/classification: single-tier, cohead	(section 3.1)
	b. identity: double-tier, PREDLINK	(section 3.2)
	c. location: double-tier, OBL	(section 3.3)
	d. existence: double-tier, OBL	(section 3.4)
	e. possession: double-tier, PREDLINK	(section 3.5)

Although it would be logical to discuss (20b) and (20e) next to each other, because I propose a PREDLINK analysis for both, I find it more important to discuss (20a,b) and (20c,d) next to each other, because these types exhibit some basic Hungarian copula use differences more transparently.

3.1 Attribution or classification

Consider the following examples ((1) is repeated here for convenience).

(1) *Az igazgató* *okos/tanár* *volt.*
the director.NOM clever/teacher.NOM was
'The director was clever / a teacher.'

(21) a. *Az* *igazgató* *tanár.* ⇔ *Én* *tanár* *vagyok.*
the director.NOM teacher.NOM I.NOM teacher.NOM am
'The director is a teacher.' 'I am a teacher.'

b. *Az* *igazgató* *nem* *okos.* ⇔ *Én* *nem* *vagyok* *okos.*
the director.NOM not clever I.NOM not am clever
'The director isn't clever.' 'I am not clever.'

As (21a) shows, in this type the copula must be absent if the sentence is in the present tense and the subject is 3rd person, singular; and the same holds for 3rd person plural subjects (which is not exemplified here). In these paradigmatic slots, negation is done by simply inserting the negative particle *nem*, see (21b). It is a further property of this construction that in neutral sentences, the AP/NP has to occupy the immediately preverbal (precopular) position.[5]

[5] This is the famous VM (verbal modifier) position in Hungarian, normally occupied by separable verbal particles, typically reduced (non-referential) arguments or secondary predicates. This preverbal position is only available to VMs in neutral sentences, because in non-neutral sentences the focussed element must precede the verb immediately, and the VM (if there is one in the sentence) must follow the verb. In other words: VMs and focussed constituents fight for the same immediately preverbal position. Following Laczkó & Rákosi (2011) and others (cited in that paper), I assume, without any justification here, that this special, Janus-faced position is [Spec,VP].

Let us consider predicative APs first. Given the fact that under certain circumstances the copula must be systematically absent, in the spirit of Dalrymple et al. (2004) and Nordlinger & Sadler (2007), we could immediately opt for a single tier analysis. However, as I pointed out above, in my view this fact by itself is not a sufficient condition for a single-tier analysis (for further details, see section 3.2). Thus, in my approach, I need additional (and independent) support for this analysis. This evidence is provided by the fact that Hungarian predicatively used adjectives clearly satisfy Dalrymple et al.'s (2004) criterion for a predicate capable of taking a subject argument. Consider the sentence in (22).

(22) *János* *okos-nak* *tart-ja* *Péter-t.*
John.NOM clever-DAT hold-PRES.3SG Peter-ACC
'John considers Peter clever.'

This is unquestionably a functional control construction: the verb has a SUBJ and an XCOMP argument (realized by the predicative AP bearing dative case in this construction type) and it has a non-thematic OBJ, which can only obey the coherence condition if it functionally controls the AP's thematic SUBJ. It is further evidence for this single-tier analysis that in this construction type (the infinitival form of) the copula cannot even be inserted, as opposed to the English counterpart. Compare the Hungarian example and its English translation in (23).

(23) **János* *okos-nak* *tart-ja* *le-nni* *Péter-t.*
John.NOM clever-DAT hold-PRES.3SG be-INF Peter-ACC
'John considers Peter to be clever.'

The analysis of the NP in this type as the main argument-taking predicate seems to be less intuitive and less unproblematic. In this connection, Attia (2008), agreeing with Dalrymple et al. (2004), for instance, claims that common nouns should not be taken to have an argument structure containing a subject argument.[6] However, in Hungarian such predicative noun phrases can be involved in exactly the same functional control constructions as predicative APs,[7] cf. (22) and (24), which lends considerable support to an analysis along these argument-taking lines.

[6] Dalrymple et al. (2004) point out that in Japanese, adjectives can be used without the copula, but nouns cannot, and this provides partial motivation for them only to analyze adjectives as argument-taking predicates as opposed to nouns in Japanese CCs. By contrast, the corresponding Hungarian facts are partially different, which can justify a partially different approach.
[7] For instance, both categories have the same dative marking.

(24) *János* *géniusz-nak* *tart-ja* *(*le-nni)* *Péter-t.*
John.NOM genius-DAT hold-PRES.3SG be-INF Peter-ACC
'John considers Peter (to be) a genius.'

Also note that the nominal predicate must be non-specific. This fact enables us to define the required (categorial) environment for a predicative (argument-taking) noun: it must occur within an NP and never within a (referring) DP. In (25) and (26), I show the most important lexical aspects of my analysis, using the XLE style formalism. Both lexical forms contain representations capturing the non-zero-copular use of these predicates, and I abstract away from the encoding (and constraining) of tense and agreement.

(25)	*okos*	A,	{ (↑ PRED) = 'clever < (↑ SUBJ) >'	predicative use
			(↑NUM)	must have number
			@FOCUSorVM	focus ⇔VM macro
			| (↑ PRED) = 'clever'	attributive use
			~(↑ NUM)}.	no number feature

The disjunction encodes the predicative vs. the attributive uses of the adjective.[8] It is a fundamental contrast between the two uses that the adjective always has a number feature in the former and never in the latter. The @FOCUSorVM macro captures the FOCUS vs. VM complementarity outlined in Footnote 5: in neutral (non-focussed) sentences the predicative AP must precede the verb. (Technically, this is encoded by dint of XLE's CHECK feature device in the macro.)[9]

(26)	*tanár*	N,	{ (↑PRED) = 'teacher < (↑ SUBJ) >'	predicative use
			(↑ SPECIFIC) = –	non-specific
			@(CAT ↑ NP)	c-structure category: NP
			@FOCUSorVM	focus ⇔VM macro
			| (↑PRED) = 'teacher'}.	non-predicative use

In (26), the disjunction encodes the contrast between the predicative, argument-taking and the ordinary use of a noun. As I pointed out above, non-specificity is intimately related to the predicative use, as is indicated in the first member of the disjunction, and there is also a constraining equation associated with the NP node in the preverbal position: (↓ SPECIFIC) =c –.

[8] In the vein of the (I think) majority LFG opinion, in the attributive representation the adjective does not subcategorize for a SUBJ argument. From the perspective of the present paper this issue is not relevant anyhow.

[9] The behavior of these CCs is even more complex, because the predicative adjective itself can be the focussed element. Space limitations prevent me from discussing this issue here. Suffice it to say that this particular phenomenon can be captured along the lines proposed in King (1997).

The @(CAT ↑ NP) template restricts the category of the nominal predicate to NP (that is, the predicative noun cannot occur in a DP). The function of the @FOCUSorVM macro in (26) is the same as in (25).

3.2 Identity

Consider the following examples ((2) is repeated here for convenience).

(2) *Az igazgató a szóvivő volt.* [identity]
the director.NOM the spokesman.NOM was
'The director was the spokesman.'

(27) a. *Az igazgató a szóvivő.*
the director.NOM the spokesman.NOM
'The director is the spokesman.'

b. *A szóvivő az igazgató.*
the spokesman.NOM the director.NOM
'The spokesman is the director.'

c. *Az igazgató nem a szóvivő (volt).*
the director.NOM not the spokesman.NOM was
'The director is/was not the spokesman.'

d. *A szóvivő nem az igazgató (volt).*
the spokesman.NOM not the director.NOM was
'The spokesman is/was not the director.'

(28) a. *Én a szóvivő vagyok.*
I.NOM the spokesman.NOM am
'I am the spokesman.'

b. *Én a szóvivő voltam.*
I.NOM the spokesman.NOM was.1SG
'I was the spokesman.'

c. **A szóvivő én volt.*
the spokesman.NOM I.NOM was.3SG
cca. 'The spokesman was me.'

In this type, two entities, typically expressed by definite 3rd person DPs, are equated, and as the examples in (27) show, often either of the two DPs can be taken to be the subject and agree with the copula. However, when one of the DPs is not 3rd person (that is, when it is a 1st or 2nd person pronoun) only that DP can function as the subject, see (28). This type and the attribution/classification type share all of the following properties. The copula must be absent if the sentence is in the present tense, and the subject is 3rd person singular, see (27a,b), and the same holds for 3rd person plural subjects (which is not exemplified here). In these paradigmatic slots, negation is done by simply inserting the negative particle *nem*, see (27c,d). In this type, in neutral sentences, the non-subject constituent has to occupy the immediately preverbal (precopular) position.

I propose that this type is most appropriately analyzed in a two-tier approach, despite the fact that the copula must be absent in the present tense, 3SG/PL paradigmatic slots. Thus, here I adopt Butt et al.'s (1999) and Attia's (2008) analysis. The copula is a two-place predicate subcategorizing for a SUBJ and a PREDLINK. Given the nature (semantics) of this construction type, the function (semantics) of this predicate is to equate (or, literally, link) two entities. And, as I pointed out above, there are cases in which the two 3rd person definite DPs can take these two grammatical functions interchangeably. It also has to be encoded in the lexical form of this copula that if one of the DPs is not 3rd person, then it must be the SUBJ and never the PREDLINK.[10]

Even when the copula is not present in the sentence in this type, I postulate that this unexpressed copula is the main predicate. I follow Dalrymple et al.'s (2004) analysis of a Russian construction in this vein, and I assume that the properties of the missing copula are introduced by LFG style (phrase-)structural means:

(29) S →	DP	VCop	∨	ε	DP
	(↑SUBJ)=↓	↑=↓		(↑PRED)='be<SUBJ,PREDLINK>' (↑TENSE)=present (↑SUBJ PERS)=3 (↑SUBJ NUM)={sg | pl} (↑SUBJ PERS)=(↑PREDLINK PERS) (↑SUBJ NUM)=(↑PREDLINK NUM) (↑SUBJ SPECIFIC)=c+ (↑PREDLINK SPECIFIC)=c+	(↑PREDLINK)=↓

In this rule the overt copula (VCop) is in complementary distribution with the special ε (epsilon) symbol, which does not appear in the c-structure representation as an empty category; instead, it contributes its annotations solely to the relevant f-structure. In all the other paradigmatic slots, the appropriate form of the copula encodes all the relevant functional information in its lexical entry.

[10] The simplest and most straightforward way of carrying this out is to use the following constraint: ~(↑PREDLINK PERS) = {1 | 2}.

3.3 Location

Consider the following examples ((3) is repeated here for convenience).

(30) *Az* *igazgató* *a* *szobá-ban* *van.*
the director.NOM the room-in is
'The director is in the room.'

(3) *Az* *igazgató* *a* *szobá-ban* *volt.*
the director.NOM the room-in was
'The director was in the room.'

(31) *Az igazgató* *nincs* *a* *szobá-ban.*
the director.NOM isn't the room-in
'The director isn't in the room.'

(32) *(Én)* *nem* *vagyok* *a* *szobá-ban.*
I.NOM not am the room-in
'I am not in the room.'

(33) *Az* *igazgató* *nem* *volt* *a* *szobá-ban.*
the director.NOM not was the room-in
'The director wasn't in the room.'

The most important properties of this CC are as follows. The copula is normally overt even in the present.3SG/3PL cases, see (30), which exemplifies the present.3SG instance. As is usual in other CCs as well, ordinarily negation takes the form of combining the negative particle and the copula, see (32) and (33). However, in the present.3SG/3PL cases negation is expressed by a special suppletive form (*nincs* 'isn't' and *nincsenek* 'aren't'), see (31), which exemplifies the present.3SG instance. The subject constituent has to be specific, and, in neutral sentences, the locative constituent has to occupy the immediately preverbal (precopular) position, the VM position, see (3) and (30). It is also noteworthy that the locative constituent is not predicative in Hungarian, as opposed to predicative APs and NPs in the attribution/classification type, see section 3.1. For instance, it cannot be the PRED of an XCOMP in a raising construction. Compare (34) with (22), (23) and (24).

(34) **János* *a* *szobá-ban* *tart-ja*
John.NOM the room-in hold-PRES.3SG
(le-nni) *az* *igazgató-t.*
be-INF the director-ACC
'John considers the director (to be) in the room.'

From this fact it follows that the locative constituent in this CC type cannot be analyzed as open: it does not allow the only open version my system applies, the single-tier, functional cohead analysis, but its behavior shown in (34) would not justify the two-tier, XCOMP analysis, either. In theory, it would be possible to assign the PREDLINK function to this locative constituent. However, my alternative solution here is the OBL_{loc} function on the basis of the following considerations. This CC expresses a genuine locative relationship; therefore, it is reasonable to feed semantics directly in terms of grammatical function choice and f-structure representation.[11] Furthermore, as I argue in the next section, the parallel between locative and existential CCs can be captured in a straightforward manner along these lines. In addition, although I myself do accept and use the PREDLINK function in the analysis of certain CC types, in my view this is really motivated and justifiable if it can be assumed that the copula has a genuine "linking" function (semantics). Thus, I take this function (name) at face value.[12]

I represent the lexical form of the locative copula in the following way.

(35)	*van*,	V	$(\uparrow$PRED) = 'BE$_{loc}$ < ($\uparrow$SUBJ) , ($\uparrow$OBL$_{loc}$) >'
			($\uparrow$SUBJ SPECIFIC) =c +
			@FOCUSorVM_OBL.

This copula is a two-place predicate, its SUBJ argument must be specific, and its second argument receives the OBL_{loc} function. The @FOCUSorVM_OBL macro captures the fact that in non-focussed sentences the predicate's OBL argument must occupy the preverbal VM position.

3.4 Existence

Consider the following examples ((4) is repeated here for convenience).

(36)	*Vannak*	*boszorkány-ok*	*(a*	*Föld-ön).*
	are.3PL	witch-PL.NOM	the	Earth-on
	'There are witches (on the Earth).'			
(4)	*Voltak*	*boszorkány-ok*	*(a*	*Föld-ön).*
	were	witch-PL.NOM	the	Earth-on
	'There were witches (on the Earth).'			

[11] It is worth pointing out that Bresnan (2001) and Falk (2004) analyze corresponding locative CCs (in English and in Hebrew, respectively) in exactly the same fashion, assuming that the constituent in question has the OBL function.

[12] My account of identity CCs uses this function (see section 3.2), and I also use it in my analysis of possession CCs (see section 3.5).

(37) *Nincs-enek boszorkány-ok (a Föld-ön).*
isn't-PL witch-PL.NOM the Earth-on
'There aren't witches (on the Earth).'

(38) *Nem voltak boszorkány-ok (a Föld-ön).*
not were.3PL witch-PL.NOM the Earth-on
'There weren't witches (on the Earth).'

In this CC, the copula, as a strict rule, must always be overt, even in the present.3SG/3PL cases, see (36), which exemplifies the present.3PL instance. As is usual in other CCs as well, ordinarily negation takes the form of combining the negative particle and the copula, see (38). However, in present.3SG/3PL negation is expressed by a special suppletive form (*nincs* 'isn't' and *nincsenek* 'aren't'), see (37), which exemplifies the present.3PL instance. The subject constituent must be non-specific. In reality, this CC does not occur in ordinary neutral sentences for the following reason. Even when there is no focussed constituent, the copula itself is the first element and it receives focal stress, see (4) and (36). Very often, this CC does not contain an overt locative constituent, but even in that case the interpretation is that the (non-specific) subject exists in a particular world.

There are, thus, significant similarities and dissimilarities between location and existence CCs. Below I list them.

- In both types, the copula is best treated as a two-place predicate.
- In both types, the second argument is best assigned the closed OBL_{loc} function.
- In the location CC the argument is strictly obligatory, while in the existence CC it is absolutely optional.
- In the location CC the subject must be specific, while in the existence CC it must be non-specific.
- In neutral location CC sentences the OBL_{loc} argument must occupy the preverbal (= precopular) VM position, while in "neutral" existence CC sentences there is no VM option, to begin with, and the copula must receive focal stress.

In my analysis, the existential copula has the following lexical form.

(39) *van*, V (↑PRED) = 'BE_{exist} < (↑SUBJ) , ((↑OBL)) >'
(↑SUBJ SPECIFIC) =c –
{ (↑ FOCUS)
| (↑ PRED FN) = ($↑_i$ FOCUS)}.

The first two lines should be straightforward on the basis of the discussion above. As regards the FOCUS disjunction, it reads as follows: (i) there is a focussed constituent in the sentence (first disjunct); (ii) the copula itself is in focus (second disjunct). The latter case is very special, because the copula is the (functional) head of the entire sentence, so if it received the FOCUS discourse function in the regular LFG way then this would mean that the entire sentence was in focus. However, it is just the predicate that is focussed. This interpretation is encoded, in an XLE way, by the equation in the second conjunct. It is only the copula, its function name (FN), that is in focus (without its arguments), and this focus is represented in information structure ($\uparrow_i$), rather than in f-structure. I have adopted this treatment of focussing predicates from King (1997).[13]

3.5 Possession

Consider the following examples ((5) is repeated here for convenience).

(40)	*Az*	*igazgató-nak*	*van*	*szóvivő-je.*
	the	director-DAT	is	spokesman-his.NOM
	'The director has a spokesman.'			

(5)	*Az*	*igazgató-nak*	*volt*	*szóvivő-je.*
	the	director-DAT	was	spokesman-his.NOM
	'The director had a spokesman.'			

(41)	*Az*	*igazgató-nak*	*nincs*	*szóvivő-je.*
	the	director-DAT	isn't	spokesman-his.NOM
	'The director doesn't have a spokesman.'			

(42)	*Az*	*igazgató-nak*	*nem*	*volt*	*szóvivő-je.*
	the	director-DAT	not	was	spokesman-his.NOM
	'The director didn't have a spokesman.'				

(43)	a.	*az*	*igazgató*	*okos*	*szóvivő-je*
		the	director.NOM	clever	spokesman-his
		'the director's clever spokesman'			
	b.	*az igazgató-nak*	*az*	*okos*	*szóvivő-je*
		the director-DAT	the	clever	spokesman-his
		'the director's clever spokesman'			

In Hungarian, possession is expressed at the sentence level by this peculiar possession CC. First of all, it has a very special agreement pattern. The possessed noun phrase is the subject and its head is inflected in exactly the same way as the noun head of possessive DPs (that is, DPs containing possessor constituents): compare all the sentence level examples in (5), (41), (42) with (43). The possessor in the CC is obligatorily expressed by a DP in the dative case, see (5), (41), (42).[14] The possessed noun phrase is always 3SG

[13] Also see Footnote 9.

[14] Within a DP expressing possession, the dative marking of the possessor is only an option, cf. (43a) and (43b).

or 3PL, and it agrees with the copula in this respect (this is ordinary subject-verb agreement). However, this subject also agrees with the dative possessor for person and number in the same way as the possessed noun head agrees with the (nominative or dative) possessor within possessive DPs: compare, again, (5), (41), (42) with (43).[15]

Some additional properties of this CC are as follows.

- The possessed noun (the subject) is, as a rule, indefinite.
- The copula is strictly obligatory, just like the copula in existence CCs, see section 3.4.
- In "neutral" possession CC sentences the dative possessor is typically a topic, and, more importantly, the copula always gets focal stress, just like the copula in existence CCs, see section 3.4.
- The negation pattern of the copula in this CC type follows that of the copula in location and existence CCs.

I believe that this special CC type is, again, best analyzed along the PREDLINK lines. My intuitive assumption is that the function of the copula here is to link the possessor and the possessed entity at the clause level. In other words, the copula "raises" the possessive relationship which can also be expressed within DPs to a sentential, predicational level.

I propose the following lexical form for the possession copula.

(44) *van*, V (↑PRED) = 'BE_{poss} < (↑SUBJ) (↑PREDLINK) >'
(↑SUBJ DEF) =c – possessee possessor
(↑PREDLINK CASE) =c dat
{(↑ FOCUS)
|(↑ PRED FN) = ($↑_i$ FOCUS)}.

The first two equations about the indefiniteness of the SUBJ (possessee) and about the case constraint of the PREDLINK (possessor) should be straightforward. The FOCUS disjunction here is the same as I postulated in the case of the existence copula in the previous section.

A remark is in order here about the (very special) agreement pattern between the subject and the dative argument in this CC. So far it has been typically assumed in the literature that the dative possessor argument is an OBL. However, this assumption has been criticized by pointing out that it is highly unusual across languages for an OBL to agree with the SUBJ. Now, if we assume that the possessor has the PREDLINK function, this agreement

[15] It is noteworthy that in her GB framework Szabolcsi (1994) treats these possessive sentences as existential sentences. The possessive noun phrase is the sole argument of the existential copula, and the dative marked possessor is obligatorily extracted from the DP. Although such an analysis could also be easily captured in LFG, I claim that my alternative account is more plausible. For lack of space I cannot argue for this in the present paper.

relationship can be argued to be much more justified. It simply follows from the very nature of PREDLINK: it can (or must) enter into an agreement relationship with SUBJ.[16]

4 Conclusion

In this paper I have developed the first comprehensive LFG analysis of the five most important types of copula constructions in Hungarian. The most significant general aspects of my approach are as follows.

- I subscribe to the view, advocated by Dalrymple et al. (2004) and Nordlinger & Sadler (2007), that the best LFG strategy is to examine all CCs individually and to allow for diversity and systematic variation both in c-structure and in f-structure representations across and even within languages. This means that I reject Butt et al.'s (1999) and Attia's (2008) uniform PREDLINK approach at the f-structure level.
- I argue against the two-tier, open, XCOMP analysis of CCs.
- I employ the following analysis types:
  - (i) single-tier, functional cohead (open);
  - (ii) double-tier, PREDLINK or OBL (closed).

Figure 2 (next page) summarizes the most important properties of the five Hungarian CCs and the crucial aspects of my analysis.

Let me conclude this paper with an additional short comment. Interestingly, my claim that the location CC has to be treated differently is (further) independently supported by the fact that out of the five versions of the Hungarian copula analyzed in this paper, it is only the locative version that has a productively used participial counterpart. Compare the location use in (45a) with the attribution use and the possession use in (45b) and (45c), respectively.

(45) a. *a* *szobá-ban* *lévő* *igazgató*
the room-in being director
literally: 'the director being in the room'

b. **az* *okos* *lévő* *igazgató*
the clever being director
literally: 'the director being clever'

[16] In this connection, it is also important that in the XLE implementation of LFG such (special) agreement facts can be rather easily and straightforwardly accommodated. In possessive DPs the tags associated with the noun stem (encoded by the relevant inflectional elements) contribute the following types of equations: (↑POSS PERS) = … and (↑POSS NUM) = … In this particular instance of PREDLINK-SUBJ agreement, we only have to introduce the following alternative annotations: ((SUBJ ↑) PREDLINK PERS) = … ((SUBJ ↑) PREDLINK NUM) = …

c. *a *szóvivő-je* *lévő* *igazgató*
the spokesman-his being director
intended meaning: 'the director having a spokesman'

CC TYPE	PR3: COP	PR3: NEG	COPULA'S FUNCTION	ARGUMENT STRUCTURE	VM	OTHER TRAITS
ATTR/CLASS	–	*nem*	formative	–	AP/NP	NP: –spec
IDENTITY	–	*nem*	predicate	<S, PL>	SUBJ	S: +spec, interch.
LOCATION	+	*nincs*	predicate	<S, OBL>	OBL	S: +spec
EXISTENCE	+	*nincs*	predicate	<S, (OBL)>	–	S: –spec cop: FOC
POSSESSION	+	*nincs*	predicate	<S, PL>	–	S: –def S&PL agr. cop: FOC

Figure 2. Properties and analyses of Hungarian CCs[17]

Acknowledgements

I thank the LFG12 reviewers and the participants of the Bali conference for useful comments and discussions. I also thank the editors of this volume for their assistance. Any errors that remain are solely mine.

I acknowledge that the research reported here has been supported, in part, by the TÁMOP 4.2.1/B-09/1/KONV-2010-0007 project, which is implemented through the New Hungary Development Plan co-financed by the European Social Fund, and the European Regional Development Fund; by OTKA (*Hungarian Scientific Research Fund*), grant number K 72983; and the Research Group for Theoretical Linguistics of the Hungarian Academy of Sciences at the University of Debrecen.

References

Attia, Mohammed. 2008. A unified analysis of copula constructions. In: Butt, Miriam & King, Tracy H. (eds.) *Proceedings of the LFG '08 Conference*. University of Sydney. On-line publication: 2008 CSLI Publications, ISSN 1098-6782, http://cslipublications.stanford.edu/LFG/13/, 89-108.

Bresnan, Joan. 1982. The passive in lexical theory. In: Bresnan, Joan (ed.) *The Mental Representation of Grammatical Relations*. Cambridge: The MIT Press. 3-84.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

[17] I use the following abbreviations in this figure: cop = copula, attr/class = attribution/classification, pr3:cop = is the copula present in the present tense and 3rd person paradigmatic slots? pr3:neg = how is negation expressed in pr3? VM = what element occupies the VM position (if any) in neutral sentences? S = SUBJ, PL = PREDLINK, interch = the two arguments' grammatical functions are interchangeable in the 3rd person, spec = specific, def = definite, FOC = FOCUS, agr = agreement.

Butt, Miriam, Tracy H. King, María-Eugenia Niño & Frédérique Segond. 1999. *A Grammar Writer's Cookbook*. Stanford, CSLI.

Dalmi, Gréte. 2010. *Copular Sentences, Predication and Cyclic Agree*. Habilitation dissertation. Loránd Eötvös University, Budapest.

Dalrymple, Mary, Helge Dyvik & Tracy H. King. 2004. Copular complements: Closed or open? In: Butt, Miriam – King, Tracy H. (eds.) *Proceedings of the LFG '04 Conference*. University of Canterbury, New Zealand. On-line publication: 2004 CSLI Publications, ISSN 1098-6782, http://cslipublications.stanford.edu/LFG/9/lfg04.html, 188-198.

Falk, Yehuda N. 2004. The Hebrew present-tense copula as a mixed category. In: Butt, Miriam & King, Tracy H. (eds.) *Proceedings of the LFG '04 Conference*. University of Canterbury, New Zealand. On-line publication: 2004 CSLI Publications, ISSN 1098-6782, http://cslipublications.stanford.edu/LFG/9/lfg04.html, 226-246.

King, Tracy H. 1997. Focus domains and information structure. In: Butt, Miriam & King, Tracy H. (eds.) *Proceedings of the LFG'97 Conference*. University of California, San Diego. On-line publication: 1997 CSLI Publications, ISSN 1098-6782, http://cslipublications.stanford.edu/LFG/2/lfg97.html.

Laczkó, Tibor & Rákosi, György. 2011. On particularly predicative particles in Hungarian. In: Butt, Miriam & King, Tracy H. (eds.) *Proceedings of the LFG '11 Conference*. Hong Kong: Hong Kong University. On-line publication: 2011 CSLI Publications, ISSN 1098-6782, http://cslipublications.stanford.edu/LFG/16/lfg11.html, 299-319.

Nordlinger, Rachel & Sadler, Louisa. 2007. Verbless clauses: Revealing the structure within. In: Grimshaw et al. (eds.) *Architectures, Rules and Preferences: A Festschrift for Joan Bresnan*. Stanford: CSLI Publications, 139-160.

Rosén, Victoria. 1996. The LFG architecture and "verbless" syntactic constructions. In: Butt, Miriam & King, Tracy H. (eds.) *Proceedings of the LFG '96 Conference*. Rank Xerox, Grenoble. On-line publication: 1996 CSLI Publications, ISSN 1098-6782, http://cslipublications.stanford.edu/LFG/1/.

Sulger, Sebastian. 2011. A parallel analysis of have-type copular constructions in have-less Indo-European languages. In: Butt, Miriam & King, Tracy H. (eds.) *Proceedings of the LFG '11 Conference*. Hong Kong: Hong Kong University. On-line publication: 2011 CSLI Publications, ISSN 1098-6782, http://cslipublications.stanford.edu/LFG/16/lfg11.html, 299-319.

Szabolcsi, Anna. 1994. The noun phrase. In: Kiefer, Ferenc & É. Kiss, Katalin (eds.) *The Syntactic Structure of Hungarian. Syntax and Semantics 27*. New York: Academic Press, 179-274.

IMPLEMENTING LEXICAL FUNCTIONS IN XLE

François Lareau
Macquarie University

Mark Dras
Macquarie University

Benjamin Börschinger
Macquarie University
Universität Heidelberg

Myfany Turpin
University of Queensland

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

Linguistic collocations such as *pay attention* or *heavy rain* are semi-compositional expressions that require a special treatment in symbolic grammars for NLP. Within the Meaning-Text Theory framework, recurrent patterns of collocations have been identified and described via so-called "lexical functions". Building on our previous attempt at importing such lexical functions into LFG using glue semantics, in this paper, we show how a workable approximation to this can be implemented in XLE. We explore and compare three different approaches: using lexical functions in the f-structure, modelling collocations in the σ-projection, and modelling them through transfer rules.

1 Introduction

Linguistic collocations such as *pay attention* or *heavy rain* are semi-compositional expressions that require a special treatment in symbolic grammars for natural language processing (NLP). Such expressions are very common in both oral and written speech, yet, their treatment in NLP remains often *ad hoc* and superficial. Much of the work within the Meaning-Text Theory (MTT) (Mel'čuk, 1973; Kahane, 2003) has focused on the lexicon, and the concept of collocations has been at the heart of this theory since the mid-60s. Consequently, this framework has a well-developed theory of the relations that exist between lexemes in the lexicon. In particular, it provides a convenient tool for the description of collocations, called lexical functions (LFs).

In terms of resources developed and applications within computational linguistics, MTT has not been very prominent outside of natural language generation (NLG), but LFs have proven useful especially for multilingual natural language generation (MNLG) (see, for instance, Heid and Raab, 1989; Bourbeau et al., 1990; Iordanskaja et al., 1992; Lareau and Wanner, 2007; Wanner et al., 2010). As it turns out, the context of our research is an MNLG project where we aim to produce Australian Football League game summaries in both English and Arrernte, an Australian language of the Pama-Nyungan family. The fact that sports news is very rich in collocations, and the bilingual nature of our system, made us look for a way to use LFs in our LFG grammars. More generally, having LFs defined in XLE would allow the LFG community to tap into existing lexical resources that focus on collocations and are built around the concept of LFs: DEC (Mel'čuk et al., 1984–1999), DiCo (Polguère, 2000), DicoInfo (L'Homme, 2005), and RLF (Lux-Pogodalla and Polguère, 2011) for French, DiCE (Alonso Ramos, 2003) for Spanish, the lexical component of ETAP-3 (Apresjan et al., 2003; Boguslavsky et al., 2004) for Russian, English and Arabic, and the multilingual lexical database architecture ILexiCon (LeFrançois and Gandon, 2011).

[†]We wish to thank Melanie Seiss for her help with transfer rules, and John Maxwell and Sina Zarrieß for their insight on generating from s-structures. We also acknowledge the support of ARC Discovery grant DP1095443.

Our first attempt at importing LFs into LFG (Lareau et al., 2011) was a theoretical approach to the problem. We showed that the f-structure alone was not sufficient to handle collocations properly, since the phenomenon pertains to the semantics-syntax interface, and we proposed a way to encode LFs in glue semantics. However, the current version of XLE, the platform that we use for our LFG grammar, does not implement glue semantics, so we searched for ways to put our ideas into practice. In this paper, we will explore and compare three different strategies: using lexical functions in the f-structure (§5), modelling collocations in the σ-projection (§6), and modelling them through transfer rules (§7). But before diving in, we will present briefly what exactly we mean by collocations (§2), introduce the concept of LF (§3), and expose our theoretical, glue-based solution (§4). We take for granted that the reader is familiar with LFG (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001; Falk, 2001), glue semantics (Dalrymple et al., 1993; Dalrymple, 1999, 2001; Andrews, 2010), and XLE (Maxwell and Kaplan, 1993; Crouch et al., 2011).

This paper expands on Lareau et al. (2011), and consequently there is a certain overlap between the two. This one does not entirely subsume the previous one however, since we have left aside all considerations linked too closely to our specific MNLG project, to gain in generality. Hence, although we have in mind the generation of texts rather than their interpretation, the ideas discussed here apply to both tasks (indeed, our implementation presently works better for parsing than generation). Also, although our examples are mostly related to football, our solution is not tied to any particular domain.

2 Collocations

A core phenomenon in the semantics-syntax interface is the mapping between meanings and lexemes—let us take a speech production perspective for a moment and call this *lexicalisation*. The lexicalisation of one meaning is usually independent of that of other meanings in the same sentence, but in the case of collocations, one lexicalisation interferes with another. A collocation is a semi-idiomatic expression where the choice of one lexeme, called the *base*, is free, but the choice of another lexeme, called the *collocate*, is context-sensitive and is constrained by the choice made for the base. Consider for example the phrases *strong preference*, *intense flavour*, *heavy rain* and *great risk*. While the lexemes PREFERENCE,[1] FLAVOUR, RAIN and RISK are chosen freely, the lexemes STRONG, INTENSE, HEAVY and GREAT are not. They carry roughly the same meaning of intensification, but their choice is tied to the lexeme they modify.

The concept of collocation is only fully understood when it is considered in the perspective of speech production rather than interpretation because there are collocations that are semantically transparent, yet the lexicalisation of their collocate is to a certain extent arbitrary. Compare for instance *strong taste* and *intense flavour*. They have very close meaning, and the semantics of STRONG and INTENSE is

[1] We use italics for word forms (e.g., *rain*) and small capitals for lexemes (e.g., RAIN).

transparent. Yet, although *strong flavour* sounds correct, the expression *intense taste* sounds odd. It is not really ungrammatical, nor is it semantically ill-formed, but it does not sound idiomatic; it is just not how English speakers would say it. The meaning of intensification is not lexicalised freely when it is used in the context of the lexeme TASTE: it must be lexicalised as STRONG. As further evidence that the relation between TASTE and STRONG is arbitrary, consider *mild taste*, which sounds idiomatic, as opposed to *weak taste*, which sounds odd, although WEAK is the antonym of STRONG. This is because there is a special relation in the lexicon of English between TASTE and STRONG or MILD that does not exist between TASTE and INTENSE or WEAK. It is these lexical relations that we aim to model; the question is how to describe these relations between lexemes in an LFG dictionary?

There are several types of collocations. We have mentioned examples where the collocate is a modifier of the base and expresses intensification or its opposite. There are adjunctive collocates that express a range of meanings: *black coffee* 'coffee without anything added to it', *green energy* 'energy that does not pollute', *decent meal* 'meal with enough food to satisfy', *wrong decision* 'bad decision', etc.

Another common type of collocation is where the collocate is a verb that takes the base as one of its arguments. These are usually referred to as *light verbs* (Jespersen, 1946). Now, before we go any further, let us get terminological confusion out of the way. The LFG-aware reader may be familiar with the work on light verbs by Kim (1991, 1993), Matsumoto (1996), Butt (2003, 2010), Yokota (2005), Ivana and Sakai (2007) or Seiss (2009). These authors did not confine themselves to collocations, while we are not concerned with light verbs that combine freely, such as MAKE in *Mary made him read the book*. Here, there is no special lexical relation between MAKE and READ; this causative combines with any verb that is semantically compatible. This is different from collocations, where the choice of the causative depends on the choice of the lexeme it combines with, e.g., GIVE in *Mary gave him the flu*. That being said, there is an overlap with the mentioned works, and although LFs are intended to describe collocations, they could also describe non-idiomatic light verbs; see the paper by Dras et al. in this volume.

To illustrate what we have said, much of our discussion will be based on the example below:

(1) Mark kicked a beautiful goal.

The lexemes KICK and BEAUTIFUL, in this context, are both collocates of GOAL. *Beautiful goal* is typical of the "colourful" style that characterises sports news, and does not mean much more than a mere positive appreciation of the goal from the author. It could be replaced with, say, *spectacular goal* or *brilliant goal*, without significantly changing the meaning. BEAUTIFUL, BRILLIANT and SPECTACULAR are all instances of the same collocation pattern (and as we will see in the following section, LFs are all about identifying such patterns). The semantics of these collocations is $(\lambda x.\text{good}(x))(\lambda y.\text{goal}(y))$, where $\lambda x.\text{good}(x)$ gets lexicalised as BEAUTIFUL, BRILLIANT or SPECTACULAR when it is a modifier of GOAL.

Kick a goal may seem more compositional. Indeed, the player does have to kick the ball, but since it is the only way to score a goal in Australian football, the semantic contribution of KICK is marginal. This becomes more apparent when we consider the translation of this phrase into other languages, where the idea of kicking totally disappears: in Spanish, *hacer un gol*, lit. 'do a goal', in Arrernte, *goal arrerneme*, lit. 'put a goal', etc. In English, the phrase *score a goal* means essentially the same as *kick a goal*. Finally, the meaning conveyed by sentence (1) can also be expressed as a noun phrase, as in (2):

(2) a beautiful goal to/by Mark

All these facts suggest that KICK, in this context, is a collocational light verb and that its semantic contribution is weak enough to be considered null. So, the semantics of *kick a goal* in (1) should be the same as that of *goal* in (2), i.e., $\lambda x.\text{goal}(x)$. Note that we are only concerned with lexical situational meaning, leaving aside grammatical meaning and phenomena pertaining to the information structure, such as the difference between (1) and (2). Both expressions denote the same situation, but differ in their communicative perspective—a difference that we want to capture in a different structure.

Now, let us present the concept of LF.

3 Lexical functions

LFs were proposed by Žolkovskij and Mel'čuk (1967) as a way to describe collocations in the context of machine translation. The concept is based on the observation that collocations tend to be instances of a number of recurrent patterns that occur across languages. For example, the expressions *strong preference*, *gravely ill*, *intense flavour* and *win hands down* are all instances of a pattern of semantics-syntax mapping where the meaning of the base is intensified and, syntactically, the collocate is a modifier of the base. Žolkovskij and Mel'čuk's idea, then, was to give names to such patterns. This one was given the name `Magn` (from Latin MAGNUS 'big'). Then, the pattern is modeled as a relation[2] between the base and the collocate. Hence, `Magn`(PREFERENCE)=STRONG, `Magn`(FLAVOUR)=INTENSE, etc. Over the years, more than fifty basic recurrent patterns of this type, and a few hundreds of complex ones, have been identified across languages and given names. Detailed descriptions of LFs can be found elsewhere (Mel'čuk, 1995, 1996, 1998; Wanner, 1996b; Apresjan, 2000; Kahane and Polguère, 2001; Apresjan et al., 2002). We will see other patterns of collocations in the coming sections.

In §2, we identified in sentence (1) two collocations: *beautiful goal* and *kick a goal*. The first one is an instance of a pattern where the semantics of the collocate is $\lambda x.\text{good}(x)$, and where the collocate is realised as a syntactic modifier of the base. The LF for this pattern is called `Bon`, and we would like to have in our dictionary the

[2] In a sense, the term *lexical function* is unfortunate since in fact these are not true mathematical functions because there can be several values for the same relation applied to a given base.

ATTENTION [of X to Y]

Magn	close/whole/complete/undivided $\sim$
$\mathtt{Func}_2$	X's $\sim$ is on Y
$\mathtt{nonFunc}_0$	X's $\sim$ wanders
$\mathtt{Oper}_{12}$	X gives his/pays $\sim$ to Y
$\mathtt{Oper}_2$	Y attracts/receives/enjoys X's $\sim$
$\mathtt{Oper}_2\mathtt{+Magn}_{\mathtt{quant-}X}$	Y is the center of $\sim$ (of many Xs)
$\mathtt{IncepOper}_{12}$	X turns his $\sim$ to Y
$\mathtt{IncepOper}_2$	Y gets X's $\sim$
$\mathtt{ContOper}_2$	Y holds/keeps X's $\sim$
$\mathtt{CausFunc}_2$	Z draws/calls/brings X's $\sim$ to Y
$\mathtt{LiquFunc}_2$	Z diverts/distracts/draws X's $\sim$ from Y

Figure 1: Collocations controlled by ATTENTION described via LFs.

instruction Bon(GOAL)=BEAUTIFUL/BRILLIANT/SPECTACULAR. The second one is an instance of another pattern where the collocate has no meaning but serves only as a support verb to turn a noun into a verbal expression. Syntactically, the collocate takes as its object the base of the collocation, and as its subject the first semantic argument of the base. This pattern corresponds to the LF $\mathtt{Oper}_1$. We would then like to have in our dictionary the instruction $\mathtt{Oper}_1$(GOAL)=KICK/SCORE/BOOT.

In short, the whole idea is to seek patterns in collocations. There will always be subtle nuances between two synonymous collocations because collocates are never entirely stripped of their literal meaning in usage, so in order to recognise patterns we have to somehow reduce the meaning of collocations to rid them of such nuances. For practical purposes however, it is safe to do so, and the benefits we get in terms of grammar engineering and resource maintainability are greater than the loss in granularity (especially given the current state of computational semantics—we are not exactly at the stage of subtle nuances yet). Armed with such LFs, the lexicographer can quickly and conveniently describe collocations in a dictionary. For example, the entry for ATTENTION could look like the one in Fig. 1. We will not discuss each of these LFs here; the point is to illustrate the wide range of collocations that can be captured efficiently with LFs.

The patterns corresponding to each LF must be defined in the grammar. We discuss in the following sections how this can be conceptualised in LFG and implemented in XLE. Once this is done, the patterns can be reused across languages and domains with little or no modifications.

4 Collocations in glue semantics

In the context of LFG, there have been several approaches to developing a compositional notion of semantics derived from the f-structure; we have chosen to base our work on Dalrymple et al. (1999) and Dalrymple (2001)'s view of glue semantics, which we will assume the reader is familiar with. It should be noted that the exact form of the semantic representations does not have to be as in this paper; our analysis is not bound to it. Instead of the simple expressions we show here, one could use event semantics or frames, for instance. What does not vary is the linear logic expressions that control semantic composition. Linear logic differs from classical logic in that premises are treated as resources that are consumed in the process of the proof. This resource-sensitivity is appropriate when dealing with the linguistic expression of semantic content: the contribution of each lexeme and phrase to the meaning of a sentence is usually unique, and there should be no missing or redundant lexemes in terms of the meaning to be expressed. In the expressions that we are concerned with, however, the principle of compositionality is violated, and the mapping between meanings and lexemes can be rather complex.

Let us get back to our example (1). For a literal reading of this sentence, the lexical entries for *goal*, *kicked* and *beautiful* would be as follows:

goal N ($\uparrow$PRED)='goal'
goal : $\uparrow_\sigma$

kicked V ($\uparrow$PRED)='kick$\langle$($\uparrow$SUBJ),($\uparrow$OBJ)$\rangle$'
($\uparrow$TENSE)=past
$\lambda x.\lambda y.\text{kick}(x, y)$: $(\uparrow\text{SUBJ})_\sigma \multimap [(\uparrow\text{OBJ})_\sigma \multimap \uparrow_\sigma]$

beautiful A ($\uparrow$PRED)='beautiful'
$\lambda x.\text{beautiful}(x)$: $(\text{ADJ} \in \uparrow)_\sigma \multimap (\text{ADJ} \in \uparrow)_\sigma$

This would yield the meaning 'kick(Mark, beautiful(goal))' (or, if you prefer, 'kick$(e_1, m, g) \wedge$ Mark$(m) \wedge$ goal$(g) \wedge$ beautiful(e_2, g)', but we will stick to the former, simple notation). What we would like instead is 'good(goal(Mark))'. For this, we need to change the lexical entries for these three word forms.

First, as we said in §2 and §3, *beautiful*, in (1), denotes a vague meaning of positive appreciation, which we could represent as '$\lambda x.\text{good}(x)$'. This is only true when *beautiful* modifies the noun GOAL (and perhaps other nouns with which it forms a collocation), so there must be a constraint in the entry that checks the context in which this modifier is used; this is what the second line does:

beautiful A ($\uparrow$PRED)='beautiful'
((ADJ $\in \uparrow$) PRED)$=_c$'goal'
'$\lambda x.\text{good}(x)$' : $(\text{ADJ} \in \uparrow)_\sigma \multimap (\text{ADJ} \in \uparrow)_\sigma$

Second, *kicked a goal* does not mean more than '$\lambda x.\text{goal}(x)$', i.e., the verb KICK simply recopies its object's meaning, with the constraint that its object is the lexeme GOAL:

kicked V ($\uparrow$PRED)='kick$\langle$($\uparrow$SUBJ),($\uparrow$OBJ)$\rangle$'
($\uparrow$TENSE)=past
($\uparrow$OBJ PRED)$=_c$'goal'
'$\lambda x.x$' : $(\uparrow\text{OBJ})_\sigma \multimap \uparrow_\sigma$

Finally, the meaning of *goal* should be a unary predicate: 'λx.goal(x)', i.e., 'x *goals*', so to speak. In the construction under consideration here, its semantic predicativity is not echoed in syntax, but this should not affect the representation of its meaning. In fact, this is precisely why a support verb is needed in the first place: KICK ties the noun GOAL to its semantic argument MARK and turns the noun into a verbal expression. This is rendered with a meaning constructor that checks that there is a meaning available for the subject of the verb of which GOAL is the object. This entry is only correct when used in the context of a support verb of which it is the object, so we also need a constraining equation here. This is inelegant, but we will see in §6 how we can get rid of it in the implementation.

goal N (↑PRED)='goal'
((OBJ↑) PRED)=$_c$'kick'
'λx.goal(x)' : $((\text{OBJ}\uparrow)\ \text{SUBJ})_\sigma \multimap \uparrow_\sigma$

As we have said above, *kick a goal* could be paraphrased as *score/boot a goal*, and *beautiful goal* could be replaced with *brilliant/spectacular goal*. The entries for the alternative collocates would be nearly identical to the ones we have just shown. There are generalisations to be made here, and we can capture them with lexical rules that we would use in the entries. And this is exactly where LFs come into play. The idea is to define lexical rules for $\texttt{Oper}_1$ and `Bon`, and then use them as follows:

kick V { @(Oper1 *goal*) | … }
boot V { @(Oper1 *goal*) | … }
score V { @(Oper1 *goal*) | … }
beautiful A { @(Bon *goal*) | … }
brilliant A { @(Bon *goal*) | … }
spectacular A { @(Bon *goal*) | … }

In the following sections we will explore three different ways or defining such lexical rules via templates in XLE.

5 An f-structure-based implementation

The simplest way to deal with collocations in XLE is to flatten their semantics by representing it in the f-structure. We achieve this by replacing collocates with the names of LFs in the PRED attribute, thus using in our representations "generalised lexemes", in the sense of Wanner (1996a), in a way similar to MTT's deep-syntactic representations, where LFs appear as nodes like other lexemes (Mel'čuk, 1988). For example, for the sentence (1), we would have in the f-structure $\texttt{Oper}_1$ and `Bon` instead of KICK and BEAUTIFUL, as in Fig. 2. By using LFs instead of lexical items, we abstract away from the collocates used in the sentence, which yields an f-structure that represents a range of paraphrases that have the same syntax and (roughly) the same meaning, but differ in the lexical choice of collocates. In this example, we have two types of collocations: a modifier (`Bon`) and a support verb ($\texttt{Oper}_1$). These correspond to two types of lexical templates in XLE. Let us first look at some modifiers.

```
 [PRED   'Oper1<[1:Mark], [12:goal]>'          ]
 [SUBJ 1 [PRED 'Mark']                         ]
 [        [PRED     'goal'                 ]   ]
 [OBJ     [ADJUNCT  {14[PRED 'Bon']}       ]   ]
 [     12 [DEF      -                      ]   ]
6[TENSE PAST                                   ]
```

Figure 2: An f-structure for (1) with LF names in it.

Below are the templates for three different patterns of collocations where the base is modified by an adjectival or adverbial collocate. The difference between them is semantic, which is captured by the names of the LFs in the PRED attribute: `Bon` for modifiers that denote a positive appreciation from the speaker (***beautiful** goal*, *superior quality*); `Epit` for pleonastic modifiers that contribute little or nothing to the meaning of the phrase, merely repeating something that is included in the meaning of the base (*happy victory*, *safe haven*); and `Magn` for modifiers that denote intensification in a broad sense (*considerable amount*, *intense flavour*). This is what the first instruction of these patterns encodes. It also gives the syntactic valence of the lexeme, which is trivial in the case of modifiers like here (they do not have syntactic arguments). The second instruction ensures that we are really dealing with a collocation, and not a normal modifier. It checks that the adjective/adverb is an adjunct of a specific lexeme, the base of the collocation, passed as an argument to the template. This effectively models restricted lexical co-occurrence.

```
Bon(Base) =
    (^PRED)='Bon'
    ((ADJUNCT ^) PRED FN) =c Base.

Magn(Base) =
    (^PRED)='Magn'
    ((ADJUNCT ^) PRED FN) =c Base.

Epit(Base) =
    (^PRED)='Epit'
    ((ADJUNCT ^) PRED FN) =c Base.
```

Now, describing common adjunct-type collocations in the dictionary is a matter of calling such templates. Adjectives and adverbs have their normal description, corresponding to their literal readings (given below by the trivial templates `@ADJ` and `@ADV`—their exact definition is irrelevant to us), to which are added a number of LF templates for collocations they are involved in. Both the literal and the idiomatic readings are thus available; we leave it to an external process to choose the right interpretation.

```
beautiful A { @ADJ | @(Bon goal) }.
brilliant A { @ADJ | @(Bon goal) }.
happy A { @ADJ | @(Epit victory) }.
spectacular A { @ADJ | @(Bon goal) }.
easily Adv { @ADV | @(Magn win) }.
hands_down Adv @(Magn win).
```

This is equivalent to saying that `Bon`(GOAL)=BEAUTIFUL/BRILLIANT/SPECTACULAR, `Epit`(VICTORY)=HAPPY, and `Magn`(WIN)=EASILY/⌜HANDS DOWN⌝. Note that ⌜HANDS DOWN⌝ has no literal interpretation; it can only be used as an intensifier of WIN (and perhaps a few other lexemes that we have ignored here).

Below are the templates for three different patterns of support verbs: $\texttt{Func}_0$ (*the wind blows*, *the rain pours*) and $\texttt{Oper}_1$ (*perform an operation*, *take a nap*), which are semantically empty verbs, and $\texttt{LiquFunc}_0$ (*snap a streak*, *eradicate a disease*), which means 'to cause the end of'. The templates for support verbs are similar to the ones above, but their PRED must encode their syntactic valence, since such collocates always have at least one syntactic argument; this is what the first line of these templates is for. The second line gives the position of the base in relation to the support verb; unlike the templates above, where the base of the collocation was always the lexeme being modified by the collocate, the base of a support verb can be any of its syntactic arguments. Hence, the difference between a $\texttt{Func}_0$ and an $\texttt{Oper}_1$ is purely syntactic: the former is an intransitive verb that takes the base as its subject, whereas the latter is a transitive verb that takes the base as its object. The difference between an $\texttt{Oper}_1$ and a $\texttt{LiquFunc}_0$, on the other hand, is semantic, and it is captured by the PRED attribute. This is not very transparent however, because patterns that have the same semantics (such as $\texttt{Func}_0$ and $\texttt{Oper}_1$, both semantically empty) also have a different PRED functor.

```
Func0(Base) =
   (^PRED)='Func0<(^SUBJ)>'
   (^SUBJ PRED FN) =c Base.

Oper1(Base) =
   (^PRED)='Oper1<(^SUBJ)(^OBJ)>'
   (^OBJ PRED FN) =c Base.

LiquFunc0(Base) =
   (^PRED)='LiquFunc0<(^SUBJ)(^OBJ)>'
   (^OBJ PRED FN) =c Base.
```

Then, the collocations *boot/kick/score a goal*, *get a mark*, *get the victory* and *snap a streak* are described in the dictionary as below (`@TRANS` is the usual template for transitive verbs with a literal reading):

```
boot V { @TRANS | @(Oper1 goal) }.
get V { @TRANS | @(Oper1 mark) | @(Oper1 victory) }.
kick V { @TRANS | @(Oper1 goal) }.
score V { @TRANS | @(Oper1 goal) }.
snap V { @TRANS | @(LiquFunc0 streak) }.
```

With the lexical entries discussed above, and a few trivial ones not shown here, we can parse (1) to obtain the f-structure in Fig. 2. Because `Oper1` and `Bon` are abstractions on the lexemes KICK and BEAUTIFUL that appeared in the parsed sentence, we can regenerate all paraphrases that only differ in the lexical choice for these collocates:

(3) Mark booted/kicked/scored a beautiful/brilliant/spectacular goal.

This is useful for shallow paraphrasing where lexical items are changed. However, it is not possible to produce paraphrases that differ in their syntactic structure. For this, we need to have a proper semantic structure. The obvious way to get one in XLE is to use the σ-projection mechanism; let us now discuss this approach.

6 An s-structure-based implementation

XLE allows us to define new projections in addition to the built-in ϕ-projection. We use this mechanism to derive from the f-structure an s-structure where we encode the semantics of expressions. Since we now have a separate structure for meaning, the attribute PRED does not have to capture semantic information anymore. It is not obsolete however; it stores lexico-syntactic information about lexemes, that is, the name of the lexeme and its syntactic valence in the expression under consideration. Our f-structure now looks like the usual ones, sticking to a more superficial description of the actual words used in the sentence, regardless of whether they have an idiomatic or a literal reading. Thus, the f-structure for sentence (1) is the one in Fig. 3.

```
 [PRED   'kick<[1:Mark], [12:goal]>'                  ]
 [SUBJ 1 [PRED 'Mark']                                ]
 [         [PRED     'goal'                        ]  ]
 [OBJ      [ADJUNCT  {14 [PRED 'beautiful']}       ]  ]
 [      12 [DEF      -                             ]  ]
6[TENSE PAST                                          ]
```

Figure 3: A normal f-structure for (1).

Our s-structure is a connected direct acyclic graph that encodes predicate-argument relations only. We do not use semantic roles; we only encode the salience of a predicate's arguments by numbering them from 1 up in decreasing order of salience, as is done in MTT's semantic representations (Mel'čuk, 2004). This has the advantage of having more generic relations between meanings, which makes it easier to find recurrent patterns. The graph is encoded in XLE as an attribute-value matrix (AVM) where nodes are rendered as structures, and the relations between the nodes as attributes ARG1, ARG2, etc., that have as their value an embedded structure. The labels of the nodes, given by the attributes SEM in the AVMs, are either the name of a lexeme when it has a literal reading, or the name of a LF when it is a meaningful collocate (e.g., ‘`Bon`’ instead of ‘good’, ‘`Magn`’ instead of ‘big’ or ‘intense’). We use LF names here because we want to underline the fact that we are pointing to an idealised meaning, stripped of the nuances that may exist between instances of a collocation pattern. Hence, for sentence (1), we want to have the s-structure in Fig. 4, equivalent to the expression $(\lambda x.\text{Bon}(x))((\lambda y.\text{goal}(y))(\text{Mark}))$. To the left is a graphical representation of the meaning of (1); to the right is its encoding as an AVM in XLE (note that structure 2 is actually embedded in structure 4 even if it does not appear so visually—this is how XLE displays the result for technical reasons that are irrelevant here). We leave aside grammatical meanings.

$$\equiv \qquad 2\begin{bmatrix}\text{ARG1} & 8\begin{bmatrix}\text{SEM} & \text{Mark}\end{bmatrix}\\ \text{SEM} & \text{goal}\end{bmatrix} \qquad 4\begin{bmatrix}\text{ARG1} & [2] \\ \text{SEM} & \text{Bon}\end{bmatrix}$$

Figure 4: An s-structure for (1).

It is in the projection between f-structure and s-structure that the complex mapping between collocations and their semantics takes place. Again, we use templates to define LFs; we will look below at the templates for `Bon`, `Epit` and $\texttt{Oper}_1$. As in §5 above, modifier-type templates such as `Bon` or `Epit` have a trivial PRED and these collocates are always adjuncts of their base, while support verbs like $\texttt{Oper}_1$ have a PRED that reflects their syntactic sub-categorisation, and their base may be any of their syntactic arguments. This time however, the PRED functor is the lexical stem, since we do not need to represent the semantics of collocates in the f-structure anymore. This is ensured by the `s::` instructions, which construct the σ-projection. In the `Bon` pattern, the first semantic instruction projects the idealised meaning ‘Bon’ as the SEM attribute in the s-structure. The second instruction gives the mapping between the semantic and syntactic relations: the meaning ‘Bon’ is a semantic predicate that becomes in f-structure an adjunct of its first semantic argument—i.e., the σ-projection of the collocate has a first semantic argument in s-structure which is the σ-projection of the lexeme it is an adjunct of in f-structure. In the case of `Epit`, because it is a pleonastic adjunct that does not contribute significant meaning to the phrase, there is simply no semantic information. This is not possible for support verbs like $\texttt{Oper}_1$, although they are also semantically empty, because they stand at the root of the clause, so they must provide a σ-projection for the outermost f-structure. Also, even if they do not contribute meaning, they do perform a rather complex remapping of semantic/syntactic arguments. The meaning of an $\texttt{Oper}_1$ is that of its base, which is always its direct object (cf. the second line of the $\texttt{Oper}_1$ template); this is what the first semantic instruction models. The last instruction handles the remapping of arguments: the first semantic argument in s-structure becomes the subject of the support verb in f-structure.

```
Bon(Base) =
   (^PRED) = '%stem'
```

```
    ((ADJUNCT ^) PRED FN) =c Base
    (s::^ SEM) = Bon
    (s::^ ARG1) = s::(ADJUNCT^).

Epit(Base) =
    (^PRED) = '%stem'
    ((ADJUNCT ^) PRED FN) =c Base.

Oper1(Base) =
    (^PRED)='%stem<(^SUBJ)(^OBJ)>'
    (^OBJ PRED FN) =c Base
    s::^ = s::(^OBJ)
    (s::^ ARG1) = s::(^SUBJ).
```

These templates are used in the same way as in §5 above. Fig. 5 illustrates graphically what is happening in the semantics-syntax interface for the $\texttt{Oper}_1$ and `Bon` patterns. The elements in bold are the ones actively built by the rules above. Note how the $\texttt{Oper}_1$ does not really realise any meaning from the s-structure, but merely links syntactically a predicate to its first semantic argument. These graphs bear a striking resemblance to Polarized Unification Grammars (PUGs) (Kahane and Lareau, 2005; Lareau, 2008).

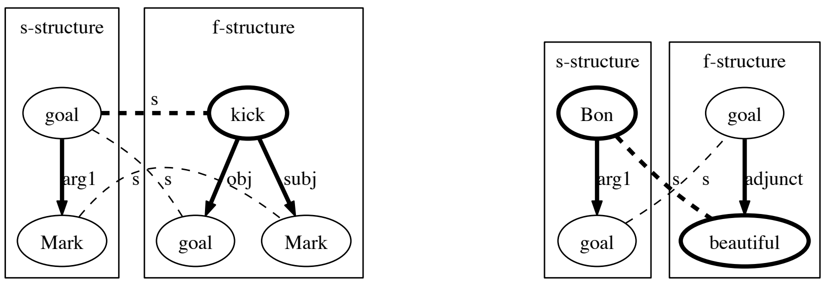


Figure 5: σ-projection and argument mapping for $\texttt{Oper}_1$ and `Bon`.

Let us look at a more complex example involving an idiomatic causative verb:

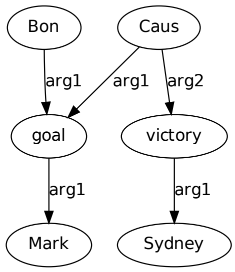
(4) A beautiful goal to Mark gave the victory to Sydney.

Its f-structure, in Fig. 6, has nothing particularly interesting. Its semantics, however, is not completely compositional. Here, *gave* means roughly 'cause', and *beautiful* expresses positive appreciation by the speaker. As we did for the previous example, we replace the collocational meanings by the names of the LFs that correspond to them, to highlight the fact that we are dealing with idealised meanings. Then, the s-structure for (4) should be something like in Fig. 7.

The light verb GIVE, besides having a non-literal meaning in this context, "steals" the first semantic argument of 'victory', which becomes its oblique object (the choice of this syntactic function is flexible). This corresponds to Mel'čuk's `CausFunc1`, which we define below. The first line gives its syntactic sub-categorisation and the

```
 [PRED   'give<[2:goal], [17:victory], [21:Sydney]>'                ]
 [      [PRED     'goal<[8:Mark]>'             ]                    ]
 [      [OBL      8[PRED  'Mark']              ]                    ]
 [SUBJ  [          [PFORM to    ]              ]                    ]
 [      [ADJUNCT {4[PRED 'beautiful']}         ]                    ]
 [     2[DEF      -                            ]                    ]
 [OBJ  17[PRED 'victory']                                           ]
 [       [DEF  +        ]                                           ]
 [OBL  21[PRED  'Sydney']                                           ]
 [       [PFORM to      ]                                           ]
2[TENSE PAST                                                        ]
```

Figure 6: An f-structure for (4).




```
≡   [ARG1   2[ARG1 8[SEM Mark]]          ]
    [         [SEM   goal     ]          ]
    [ARG2  17[ARG1 21[SEM Sydney]]       ]
    [         [SEM    victory    ]       ]
   2[SEM    Caus                         ]

   4[ARG1  [2]]
    [SEM   Bon]
```

Figure 7: An s-structure for (4).

second gives the position of the base that controls this collocation. The last four lines model the tricky semantics of this light verb. First, it projects its meaning, the idealised meaning 'Caus'. Then, it gives the mapping of the first and second semantic arguments of 'Caus', which become respectively the subject and object of GIVE. Finally, it maps the first semantic argument of the σ-projection of its object to its oblique (that is, it "steals" an actant from 'victory'). This template is then used in the entry for GIVE as below. Fig. 8 gives a visual representation of the complex mapping between semantic and syntactic elements performed by this template, where the bold elements are the ones actively involved in the rule.

```
CausFunc1(Base) =
   (^PRED)='%stem<(^SUBJ)(^OBJ)(^OBL)>'
   (^OBJ PRED FN) =c Base
   (s::^ SEM)=Caus
   (s::^ ARG1)=s::(^SUBJ)
   (s::^ ARG2)=s::(^OBJ)
   (s::(^OBJ) ARG1)=s::(^OBL).

give V { @DITRANS | @(CausFunc1 victory) }.
```

The s-structure solution yields a higher level of abstraction than the previous one, so that even expressions with a completely different syntactic structure get the

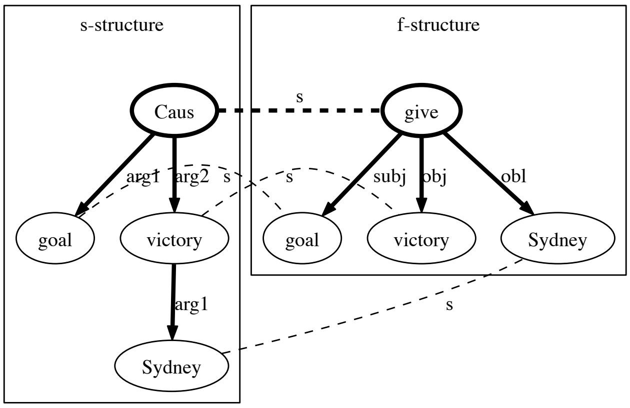


Figure 8: σ-projection and argument mapping for $\texttt{CausFunc}_1$.

same analysis. Consider for example sentence (1) and the noun phrase (2). When we parse them, we get different f-structures, but the same s-structure, as in Fig. 9. However, it is not possible to (re)generate from this s-structure, because additional projections in XLE do not have their own structure but are instead encoded as special attributes in the f-structure. This means that generating from an s-structure amounts to generating from an underspecified f-structure where all the attributes except the S:: ones are missing. XLE can accept underspecified f-structures as input for generation, but there must be a PRED attribute in each structure, because of choices made during the implementation of the platform. So this approach works for parsing, but will not work for generation or paraphrasing in the current implementation of XLE. In order to generate from s-structures like the ones we have just looked at, we have tried a different approach, using transfer rules.

```
 [PRED   'kick<[1:Mark], [12:goal]>'              ]
 [SUBJ 1[PRED 'Mark']                             ]     [PRED     'goal<[8:Mark]>'          ]
 [       [PRED     'goal'                       ] ]     [OBL      [PRED  'Mark']            ]
 [OBJ    [ADJUNCT  {14[PRED 'beautiful']}       ] ]     [        8[PFORM to    ]            ]
 [     12[DEF      -                            ] ]     [ADJUNCT {4[PRED 'beautiful']}      ]
6[TENSE PAST                                      ]    2[DEF      -                         ]

                                        ↘  ↙
                                  [ARG1 8[SEM Mark]]
                                 2[SEM   goal      ]

                                  [ARG1   [2]]
                                 4[SEM    Bon]
```

Figure 9: Two synonymous expressions yielding the same s-structure.

7 A transfer-based implementation

The Packed Rewriting System (PRS) was first intended as the transfer module of an XLE-based machine translation system. It applies rewrite rules that produce new f-structures from existing ones. Since s-structures are in fact encoded as f-structures, it is possible to hijack this mechanism to model the semantics-syntax interface. Indeed, Crouch and King (2006) and Zarrieß and Kuhn (2010) have done it for, respectively, parsing and generation. We have tried to use it in a similar way to generate collocations, with mixed results.

Transfer rules rewrite attributes of an f-structure. Rules can be optional, which allows for backtracking and provides a set of output f-structures. Our input is an s-structure like the one in Fig. 4, where the only attributes are SEM and ARG1, ARG2, etc. These will be the elements on the left-hand side of the transfer rules. On the right-hand side, we have attributes of a normal f-structure: PRED, SUBJ, OBJ and ADJUNCT. There is also a predicate ARG that encodes the position of the syntactic arguments in PRED's value, as well as IN_SET, which builds sets of adjuncts.

Below are the transfer-based templates for `Bon` and $\texttt{Oper}_1$. Whereas in the s-structure approach we looked for the base of the collocation in the f-structure with the instruction `((ADJUNCT ^) PRED FN) =c Base`, in the transfer-based approach we must encode lexical restrictions in semantics. Indeed, the left-hand side of a rule can have contextual elements, denoted by the `+` symbol; these elements are not consumed by the rule. There is no way to specify such contextual elements on the right-hand side of the rules to avoid building new items in the f-structure. This means that in this version of our grammar, collocations are controlled by meanings rather than lexemes, which is a dangerous setting since most of the lexical relations that we are trying to describe actually exist between lexemes, not meanings. For the rest, these templates do essentially the same things as the ones we have discussed in §6 (cf. Fig. 5). In the case of `Bon`, the meaning 'Bon' and its first semantic argument are consumed by the rule to produce an adjunct in f-structure. For $\texttt{Oper}_1$, only the relation ARG1 is consumed, and it is realised in syntax by the support verb and its subject and object.

```
bon(%Base, %Collocate) ::
  SEM(%X, Bon),
  ARG1(%X, %Y),
  +SEM(%Y, %Base)
  ?=>
  PRED(%X, %Collocate),
  ADJUNCT(%Y, %Adjs), in_set(%X, %Adjs).

oper1(%Base, %Collocate) ::
  +SEM(%X, %Base),
  ARG1(%X, %Y)
  ?=>
  PRED(%Z, %Collocate),
  arg(%Z, 1, %Y), SUBJ(%Z, %Y),
  arg(%Z, 2, %X), OBJ(%Z, %X).
```

The lexical items in this kind of grammar contain only the information necessary for the semantics-syntax interface. Below are some of the entries relevant for the meaning 'goal' (`@n` and `@n_obl` are trivial templates that simply realise a meaning as a noun or a noun with an oblique complement).

```
@bon(goal, beautiful).
@bon(goal, brilliant).
@bon(goal, spectacular).
@oper1(goal, boot).
@oper1(goal, kick).
@oper1(goal, score).
@n_obl(goal).
@n(goal).
```

The most interesting characteristic of this approach is that it allows us to group the collocations by their base instead of having them described in the lexical entries for the collocates as in §5 and §6. This is a lot more convenient for the lexicographer. There are three main drawbacks, however. First, as we have mentioned above, the collocations in this grammar have to be controlled by meanings instead of lexemes. Second, we have to set all our rules as optional. When there is no obligatory rule, it is always a valid solution to leave some or all of the elements of the input structure untouched in the output. We end up with chimeras that contain stock from the semantic and functional levels of representation, and we cannot do anything with these. While it would be possible to filter out the invalid output structures, this is obviously not an elegant solution. Finally, the transfer rules are applied in the order in which they appear in the file. This is unfortunately incompatible with what we are trying to do. So we have not yet overcome the limitations of the current implementation of XLE so as to allow for a direct application of the analysis presented here to NLG. We are currently exploring other strategies to solve this problem by means of external processing that are of no particular interest from a linguistic perspective.

8 Conclusion

MTT's lexical functions offer an elegant and convenient solution to the treatment of collocations in NLP. We showed that this device could be used also in the LFG framework, both from a conceptual and implementational point of view. Glue semantics offers the kind of expressive power we need to handle the complex mappings between meanings and lexemes, as well as their arguments, in semi-idiomatic expressions. However, our glue-based solution involves adding information in the lexical entry of the base of collocations in a way that is not entirely satisfactory, since it spreads out the information across several entries. In XLE, we tried and compared three different strategies for the implementation of LFs.

The f-structure based implementation lacks the depth needed to fully model a phenomenon that indeed belongs to the semantics-syntax interface. By using abstract LF names instead of actual lexemes in the f-structure, we managed to represent the functional structure of a range of paraphrases that differ in the choice of the collocates. Despite its inherent limitations, this implementation works for both parsing and generation, and allows shallow paraphrasing.

The s-structure based implementation is the one that most closely matches the power of glue semantics. It is the most elegant of our three solutions to the problem of describing collocations, as it allows us to parse synonymous expressions with radically different lexico-syntactic structures and get the semantic representation that we expect. However, because of how XLE is implemented, we cannot generate from s-structures using this approach.

As a complementary strategy for generation, we have used transfer rules to handle the semantics-syntax interface. This strategy is different from the other two in that it isolates the semantics-syntax interface from the rest of the model. It is the most compelling solution for lexicographers because it allows them to group collocations by their base, rather than forcing them to describe the collocations in the entries of the collocates. However, technical considerations force us to seek other strategies to handle generation from s-structures. One strategy that we plan to explore is to use Bohnet et al. (2000)'s generic graph-transducer, MATE, to handle the mapping between s-structure and f-structure. This approach could be combined with our f-structure implementation.

We view this work as developing a set of tools for the LFG community, rather than an actual grammar. For the f-structure and s-structure strategies, we have defined 222 templates corresponding to a range of LFs (modifiers, dummy support verbs, causative light verbs, inchoative light verbs, etc.), which we will make available at `http://web.science.mq.edu.au/~ayeye`.

For future work, we plan to extend the coverage of our patterns to include more LFs. We also want to explore how we can make these patterns even more generic, and we hope to make them compatible with Butt et al. (2002)'s ParGram architecture.

References

Alonso Ramos, Margarita. 2003. Hacia un diccionario de colocaciones del español y su codificación. In M.A. Martí (ed.), *Lexicografía computacional y semántica*, pages 11–34, Barcelona: Edicions de l'Universitat de Barcelona.

Andrews, Avery. 2010. Propositional glue and the correspondence architecture of LFG. *Linguistics and Philosophy* 33, 141–170.

Apresjan, Juri. 2000. *Systematic lexicography*. Oxford: Oxford University Press.

Apresjan, Jury, Boguslavsky, Igor, Iomdin, Leonid, Lazursky, Alexander, Sannikov, Vladimir, Sizov, Victor and Tsinman, Leonid. 2003. ETAP-3 linguistic processor: a full-fledged NLP implementation of the Meaning-Text Theory. In *Proceedings of MTT 2003*, pages 279–288, Paris.

Apresjan, Jury, Boguslavsky, Igor, Iomdin, Leonid and Tsinman, Leonid. 2002.

Lexical functions in actual NLP applications. In *Computational linguistics for the new millennium: divergence or synergy?*, pages 55–72, Frankfurt: Peter Lang.

Boguslavsky, Igor, Iomdin, Leonid and Sizov, Victor. 2004. Multilinguality in ETAP-3: reuse of lexical resources. In *Proceedings of the Workshop on Multilingual Linguistic Ressources (MLT'04)*, pages 7–14, Stroudsburg, PA.

Bohnet, Bernd, Langjahr, Andreas and Wanner, Leo. 2000. A development environment for an MTT-based sentence generator. In *Proceedings of the first international conference on Natural Language Generation*, volume 14, pages 260–263, Mitzpe Ramon.

Bourbeau, Laurent, Carcagno, Denis, Goldberg, Eli, Kittredge, Richard and Polguère, Alain. 1990. Bilingual generation of weather forecasts in an operations environment. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, pages 90–92.

Bresnan, Joan. 2001. *Lexical-functional syntax*. Oxford, UK: Blackwell.

Butt, Miriam. 2003. The light verb jungle. In G. Aygen, C. Bowern and C. Quinn (eds.), *Harvard Working Papers in Linguistics. Papers from the GSAS/Dudley House Workshop on Light Verbs*, pages 1–49.

Butt, Miriam. 2010. The light verb jungle: still hacking away. In M. Amberber, B. Baker and M. Harvey (eds.), *Complex predicates: cross-linguistic perspectives on event structure*, pages 48–78, Cambridge: Cambridge University Press.

Butt, Miriam, Dyvik, Helge, King, Tracy Holloway, Masuichi, Hiroshi and Rohrer, Christian. 2002. The parallel grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Crouch, Dick, Dalrymple, Mary, Kaplan, Ron, King, Tracy, Maxwell, John and Newman, Paula. 2011. *XLE documentation*. Palo Alto Research Center, Palo Alto.

Crouch, Dick and King, Tracy Holloway. 2006. Semantics via f-structure rewriting. In *Proceedings of the LFG06 Conference*, Konstanz.

Dalrymple, Mary (ed.). 1999. *Semantics and syntax in lexical functional grammar: the resource logic approach*. Cambridge, MA: The MIT Press.

Dalrymple, Mary. 2001. *Lexical functional grammar*, volume 42 of *Syntax and Semantics Series*. New York: Academic Press.

Dalrymple, Mary, Gupta, Vineet, Lamping, John and Saraswat, Vijay. 1999. Relating resource-based semantics to categorial semantics. In Dalrymple (1999).

Dalrymple, Mary, Lamping, John and Saraswat, Vijay. 1993. LFG semantics via constraints. In *Proceedings of the 6th Meeting of the European Association for Computational Linguistics*, Utrecht.

Dras, Mark, Lareau, François, Börschinger, Benjamin, Dale, Robert, Motazedi, Yasaman, Rambow, Owen, Turpin, Myfany and Ulinski, Morgan. 2012. Complex predicates in Arrernte. In M. Butt and T. H. King (eds.), *Proceedings of the LFG12 Conference*, Stanford: CSLI Publications.

Falk, Yehuda. 2001. *Lexical-functional grammars*. Stanford: CSLI Publications.

Heid, Ulrich and Raab, Sybille. 1989. Collocations in multilingual generation. In *Proceedings of the fourth conference of the European chapter of the Association*

for Computational Linguistics (EACL'89), pages 130–136.
Iordanskaja, Lidja, Kim, Myunghee, Kittredge, Richard, Lavoie, Benoît and Polguère, Alain. 1992. Generation of extended bilingual statistical reports. In *Proceedings of COLING'92*, pages 1019–1023, Nantes.
Ivana, Adrian and Sakai, Hiromu. 2007. Honorification and light verbs in Japanese. *Journal of East Asian Linguistics* 15(3), 171–191.
Jespersen, Otto. 1946. *A modern English grammar on historical principles. Part VI: morphology.*. London: Allen & Unwin.
Kahane, Sylvain. 2003. The Meaning-Text theory. In V. Agel, L. M. Eichinger, H.-W. Eroms, P. Hellwig, H. J. Heringer and H. Lobin (eds.), *Dependenz und Valenz: Ein internationales Handbuch der zeitgenössischen Forshung/ Dependency and Valency: An International Handbook of Contemporary Research*, volume 1, pages 546–570, Berlin / New York: Walter de Gruyter.
Kahane, Sylvain and Lareau, François. 2005. Meaning-Text Unification Grammar: modularity and polarization. In *Proceedings of MTT 2005*, pages 163–173, Moscow.
Kahane, Sylvain and Polguère, Alain. 2001. Formal foundation of lexical functions. In *Proceedings of ACL 2001*, Toulouse.
Kaplan, Ronald M. and Bresnan, Joan. 1982. Lexical-functional grammar: a formal system for grammatical representation. In J. Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 173–281, Cambridge, MA: The MIT Press.
Kim, Jay. 1991. *A lexical-functional grammar account of light verbs*. Ph.D. thesis, University of Hawaii.
Kim, Jeong-Ryeol. 1993. Parsing light verb constructions in lexical-functional grammar. *Language Research* 29(4), 535–566.
Lareau, François, Dras, Mark, Börschinger, Benjamin and Dale, Robert. 2011. Collocations in multilingual natural language generation: lexical functions meet lexical functional grammar. In *Proceedings of ALTA'11*, pages 95–104, Canberra, Australia.
Lareau, François and Wanner, Leo. 2007. Towards a generic multilingual dependency grammar for text generation. In *Proceedings of Grammar Engineering Across Frameworks (GEAF'07)*, pages 203–223, Stanford.
Lareau, François. 2008. *Vers une grammaire d'unification sens-texte du français: le temps verbal dans l'interface sémantique-syntaxe*. Ph.D. thesis, Université de Montréal/Université Paris 7.
LeFrançois, Maxime and Gandon, Fabien. 2011. ILexicOn: toward and ECD-compliant interlingual lexical ontology described with semantic web formalisms. In *Proceedings of MTT 2011*, pages 155–164, Barcelona.
L'Homme, Marie-Claude. 2005. Conception d'un dictionnaire fondamental de l'informatique et de l'Internet : sélection des entrées. *Le langage et l'homme* 40(1), 137–154.
Lux-Pogodalla, Veronika and Polguère, Alain. 2011. Construction of a French lexical network: Methodological issues. In *Proceedings of the International*

Workshop on Lexical Resources (WoLeR 2011), ESSLLI, pages 55–62, Ljubljana.

Matsumoto, Yo. 1996. A syntactic account of light verb phenomena in Japanese. *Journal of East Asian Linguistics* 5(2), 107–149.

Maxwell, John T. and Kaplan, Ronald M. 1993. The interface between phrasal and functional constraints. *Computational Linguistics* 19(4), 571–590.

Mel'čuk, Igor. 1973. Towards a linguistic "Meaning-Text" model. In F. Kiefer (ed.), *Trends in Soviet Theoretical Linguistics*, pages 33–57, Dordrecht: Reidel.

Mel'čuk, Igor. 1988. *Dependency syntax: theory and practice*. SUNY series in linguistics, Albany: State University of New York Press.

Mel'čuk, Igor. 1995. The future of the lexicon in linguistic description and the Explanatory Combinatorial Dictionary. In I.-H. Lee (ed.), *Linguistics in the morning calm*, volume 3, Seoul: Hanshin.

Mel'čuk, Igor. 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. In Wanner (1996b), pages 37–102.

Mel'čuk, Igor. 1998. Collocations and lexical functions. In A. P. Cowie (ed.), *Phraseology. Theory, analysis, and applications*, pages 23–53, Oxford: Clarendon.

Mel'čuk, Igor. 2004. Actants in semantics and syntax. I: Actants in semantics. *Linguistics* 42(1), 1–66.

Mel'čuk, Igor et al. 1984–1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques*, volume I–IV. Montréal: Presses de l'Université de Montréal.

Polguère, Alain. 2000. Une base de données lexicales du français et ses applications possibles en didactique. *LIDIL* 21, 75–97.

Seiss, Melanie. 2009. On the difference between auxiliaries, serial verbs and light verbs. In M. Butt and T. H. King (eds.), *Proceedings of LFG'09*, pages 501–519, Trinity College, Cambridge, UK.

Wanner, Leo. 1996a. Introduction. In Wanner (1996b), pages 1–36.

Wanner, Leo (ed.). 1996b. *Lexical functions in lexicography and natural language processing*. Amsterdam/Philadelphia: John Benjamins.

Wanner, Leo, Bohnet, Bernd, Bouayad-Agha, Nadjet, Lareau, François and Nicklaß, Daniel. 2010. MARQUIS: generation of user-tailored multilingual air quality bulletins. *Applied Artificial Intelligence* 24(10), 914–952.

Yokota, Kenji. 2005. The structure and meaning of Japanese light verbs. *Language Sciences* 27(3), 247–280.

Zarrieß, Sina and Kuhn, Jonas. 2010. Reversing f-structure rewriting for generation from meaning representations. In *Proceedings of the LFG10 Conference*, Ottawa.

Žolkovskij, Aleksandr and Mel'čuk, Igor. 1967. O semantičeskom sinteze. *Problemy kibernetiki* 19, 177–238.

IN SEARCH OF A NOMINAL COMP

Helge Lødrup
University of Oslo

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King
(Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

Dalrymple and Lødrup (2000) proposed that complement clauses can have two different syntactic functions. Those with the external syntactic properties of noun phrase objects are OBJs, while other complement clauses are COMPs. The idea of a COMP function has been criticized. One argument against it is that COMP differs from other syntactic functions in that it can only be filled by a clause. This paper attempts to show, on the basis of Norwegian, that there might also be nominal COMPs.

1. Introduction[1]

It is well known that complement clauses differ concerning their external syntactic behavior (Stowell 1981, Webelhuth 1992, Bošković 1995, Dalrymple and Lødrup 2000, Lødrup 2004). Some complement clauses have the external syntactic properties of nominal objects. They can topicalize, and correspond to a subject in the passive. One example is the complement of *believe*, as in 1-2.

(1) That the earth is round, everybody believed.
(2) That the earth is round was not believed.

Other complement clauses lack these properties. An example is the complement of *hope*, as in 3-4.

(3) *That it would rain, everybody hoped.
(4) *That it would rain was hoped.

Traditional LFG assumed that all clausal complements have the syntactic function COMP. Dalrymple and Lødrup (2000) proposed that a clausal complement is an object if it behaves syntactically like a nominal object, and a COMP if it does not. This idea has been discussed and criticized (Berman 2003, Alsina et al. 2005, Forst 2006; see also Börjars and Vincent 2008). The critics do not deny the need to distinguish between complement clauses with different syntactic behavior. However, they do not accept the need for a separate syntactic function COMP, preferring an analysis in which a COMP

[1] For input and discussion, I would like to thank the anonymous reviewers for LFG12, members of the conference, and colleagues at the Oslo theoretical linguistics seminar. Thanks are also due to the Proceedings editors.

is really an OBLΘ (see also Zaenen and Engdahl 1994, Bresnan 2001:309, 317).[2]

This paper approaches the question of COMP from a new angle. Both in traditional LFG and in Dalrymple and Lødrup (2000), COMP differs from other syntactic functions in that it can only be filled by a clause. It is not clear, however, why this should be the case (see Alsina et al. 2005). If only clauses can be considered COMPs, this constitutes an important argument against the COMP function. This paper suggests that noun phrases can also be COMPs. COMP is then a syntactic function that can be realized by a noun phrase or a clause, in the same way as the OBJ function.

It will be shown, on the basis of Norwegian, that some verbs that take a clausal COMP alternatively take a noun phrase complement that could be argued to be a nominal COMP. Possible cases of nominal COMPs that do not alternate with clausal COMPs are also discussed.

The structure of the paper is as follows: Part 2 discusses properties that distinguish objects and COMPs. Part 3 gives examples of verbs that seem to take clausal and nominal COMPs, while part 4 gives examples of verbs that seem to take nominal COMPs without taking clausal COMPs. Some general properties of nominal COMPs are discussed in part 5, while part 6 takes up the important question of how COMP behaves in unbounded dependency constructions.

2. Object properties

A nominal COMP could be compared to the 'new' complement function proposed in Postal (2010). Postal suggests, mainly on the basis of English, a function that he calls an 'array 1 object' (and also a '4 object'); see Postal (2010:56-64, 2004:264-75). Postal assumes that both nominal and clausal arguments can have this function.

Postal (2010) describes a group of properties for array 1 objects. A central property is that they "are not passivizable" (Postal 2010:56). This wording reveals a problem in his reasoning (which maybe follows from exceptions to the passive being his point of departure; see Postal (2004:264-75)). If an object can be a subject in a passive, the precondition is that the verb can passivize. Most verbs passivize, but there are exceptions that vary

[2] Other frameworks have also assumed analyses that could be seen as equivalent to considering COMP an OBLΘ. In the German descriptive grammar tradition, a COMP has the same syntactic function as OBLΘ, namely *Präpositionalobjekt* 'prepositional object', (Breindl 1989, Zifonun et al. 1997:1097). A different implementation of this intuition is to assume that COMP is really a PP with a deleted preposition (see Rosenbaum 1967:83 on English, Ralph 1975 on Swedish).

between languages. Exceptions involve partly idiosyncrasies, and partly semantic properties of the verbs (Jackendoff 1972:43-46, Lødrup 2000). The question of passivizability arises for all verbs, independently of what complements they take. It cannot in general be connected to the nature of complements; this is especially clear in a language with impersonal passives such as Norwegian. In Norwegian, there is no requirement that an object become the subject of a passive verb, as shown by 5. (To be more exact, this is only true of an indefinite object, because of the definiteness restriction in impersonal sentences.)

(5) De leste en bok. - Det ble lest en bok.
they read a book - there was read a book
They read a book. - A book was read.

For these reasons, the passivizability of a verb cannot be a criterion for differentiating nominal arguments. On the other hand, corresponding to a subject in the passive is a traditional criterion for an object. The inability of a complement to correspond to a passive subject is then an argument against its being an object – but only if the verb can passivize.

The intuition behind the COMP function could be verbalized this way:[3] COMP differs from the other complement functions by not having their properties; it is a complement that just 'is there', and does not take part in grammatical processes.

To distinguish a nominal COMP from an object, the object properties below will be used (selected and modified from the list in Postal 2010:58-59[4]). The premise is that an object should have these properties, and that their absence is indicative of non-object status. A clausal COMP does not have the properties in question (to the extent that they are applicable to clauses), as shown below. Note that the list of properties below does not include syntactic behavior in unbounded dependency constructions; this is discussed in part 6.

---

[3] This way of thinking about (clausal) COMP comes from a discussion with Joan Bresnan in 2006. There is no implication here that this is (or was) her actual view.

[4] Not all of the properties on the list in Postal (2010:58-59) are relevant to Norwegian. Sentences with array 1 objects have no corresponding middles, and do not allow object deletion with *too* or *enough*. Norwegian does not have these constructions. Postal also includes not allowing *tough* movement in his list. Postal (2010:59) writes that not allowing parasitic gaps is a relevant property for array 1 objects, but he puts this property aside "for simplicity".

• Object property a): The argument corresponds to the subject of a passive (of a passivizable verb). An object does, a clausal COMP does not, as shown in 6-7.

(6) Boka ble lest.
book.DEF was read
The book was read.
(7) *At han var skyldig ble svart.
that he was guilty was answered
It was answered that he was guilty. [intended]

It should be noted that the impersonal passives in 8-9 are not relevant with respect to this object property, because no subjectivization has taken place. Both 8 and 9 have an expletive subject in functional structure. The analysis assumed is that the noun phrase in 8 is an object, while the clause in 9 is syntactically ambiguous between an object and an 'extraposed' COMP (Lødrup 1999).

(8) Det ble lest en bok.
there was read a book
A book was read.
(9) Det ble sagt at han var skyldig.
it was said that he was guilty
It was said that he was guilty.

• Object property b): The argument corresponds to the 'subject' of an adjectival passive (of a passivizable verb), as in 10.

(10) en lest bok
a read book

• Object property c): The argument can be a parasitic gap. An object can, a clausal COMP can not, cf. 11-12.[5] (Even if 12 is possible with *svare* 'answer' as a one-place verb, it is not acceptable when interpreted with a parasitic gap following the verb.)

[5] For expository purposes, I have put in an *e* and a *t* in example sentences with parasitic gaps. No theoretical claims are implied.
Using parasitic gaps as a criterion should ideally be supported by a theory that predicts that they cannot correspond to a COMP. Unfortunately, our understanding of parasitic gaps does not seem to have reached a stage where this is possible.

(11) Denne boka vil jeg kaste *t* uten å lese *e*.
this book.DEF will I throw without to read
This book, I want to throw away without reading.
(12) *At man er skyldig kan man akseptere *t* uten å svare *e*.
that one is guilty can one accept without to answer
One can accept that one is guilty without answering that. [intended]

The properties below are also included (modified from Postal 2010:58-59), even if they are not criterial, only typical for objects.

• Object property d): The argument corresponds to a PP with the preposition *av* 'of' in a nominalization, as in 13.

(13) lesing av bøker
reading of books

• Object property e): The argument corresponds to the first part of a synthetic compound with the nominalized verb as a head, as in 14.

(14) boklesing
book.reading

• Object property f): The argument corresponds to the subject of an adjective that is derived from the verb with the suffix *-bar* '-able', as in 15.

(15) Teksten er ikke lesbar.
text.DEF is not readable
The text is not readable.

The last properties d)-f) are not decisive for object status, because there are clear cases of objects that do not have them. (An example is the object of *se* 'see'.) Even so, they could be seen as typical properties of typical objects.

Some cases of verbs that might be argued to take a nominal COMP will now be discussed. Verbs that do not passivize are left out, for the reasons mentioned above.

3. Examples

3.1 The verb *stønne* 'moan'

A good place to start could be one of the groups of verbs considered by Postal (2010:61) to take array 1 objects, namely verbs of manner of speaking (Zwicky 1971). Some of these verbs take a clausal COMP (Lødrup 2004). An example is the verb *stønne* 'moan', as in 16; example 17 shows that the clausal complement does not correspond to a passive subject. As an alternative to the clausal COMP, the verb can take a nominal argument that could be a nominal COMP, cf. 18. The verb can passivize, as shown by the impersonal passive in 19, but the nominal argument cannot be a passive subject, cf. 20. There is also no adjectival passive, cf. 21.

(16) Han stønnet at alt var slutt.
he moaned that everything was over
He moaned that everything was over.
(17) *At alt var slutt ble stønnet.
that everything was over was moaned
(18) Han stønnet noen uforstålige ord.
he moaned some incomprehensible words
He moaned some incomprehensible words
(19) Det ble stønnet noen uforstålige ord.
there were moaned some incomprehensible words
Some incomprehensible words were moaned.
(20) *Noen uforstålige ord ble stønnet.
some incomprehensible words were moaned
(21) *stønnede ord
moaned words

A parasitic gap is unacceptable, cf. 22.

(22) *Bannord kan man like *t* uten å måtte stønne *e*.
swearwords can one like without to have.to moan
One may like swearwords without having to moan them. [intended]

It might be objected that the status of 22 follows from general requirements on the referentiality or specificity of parasitic gaps. However, Engdahl (2001) shows that parasitic gaps do not have this kind of requirements in Scandinavian.

The nominalization and the synthetic compound are marginal, cf. 23-24, while the derived adjective is unacceptable, cf. 25.

(23) ??stønning av bannord
moaning of swearwords
(24) ??bannordstønning
swearword.moaning
(25) *Ordene er ikke stønnbare.
words.DEF are not moanable
The words cannot be moaned. [intended]

Manner of speaking verbs sharing properties with *stønne* 'moan' include those in 26.[6]

(26) *hvese* 'hiss', *hyle* 'howl', *brøle* 'roar', *bjeffe* 'bark', *grynte* 'grunt', *kvitre* 'tweet'

Manner of speaking verbs often take a resultative particle, such as *frem* 'forward' or *ut* 'out'. A nominal complement then behaves like an ordinary object. It corresponds to a passive subject, cf. 27, there is an adjectival passive, cf. 28, and a parasitic gap is possible, cf. 29.

(27) Noen uforstålige ord ble stønnet frem.
some incomprehensible words were moaned forward
Somebody moaned some incomprehensible words.
(28) ?fremstønnede ord
forward.moaned words
(29) Bannord kan man like *t* uten å måtte stønne *e* frem.
swearwords can one like without to have.to moan forward
One may like swearword without having to moan them.

This behavior is expected in LFG. The resultative particle is an XCOMP whose subject is controlled by an object. It has been established that a resultative can only be controlled by an object, or to be more exact, by an 'underlying' object (see Simpson 1983, Bresnan and Zaenen 1990).

3.2 The verb *leke* 'play'

The verb *leke* 'play' is used of children's play (not of for example chess, music or theatre). It can take a clausal COMP, cf. 30 (which does not correspond to a passive subject, cf. 31), or a nominal argument that denotes what the subject pretends to be, cf. 32.

(30) De lekte at de var lingvister.
they played that they were linguists
They played that they were linguists.

[6] There are also manner of speaking verbs that do not behave this way, such as *hviske* 'whisper', cf. (i)-(ii).
(i) Tre ord ble hvisket i øret mitt.
three words were whispered in ear.DEF my
Three words were whispered in my ear.
(ii) hviskede ord
whispered words

(31) *At de var lingvister, ble lekt.
that they were linguists was played
(32) De lekte lingvist / lingvister.
they played linguist / linguists
They played linguists.

This nominal complement might look like a nominal predicate, i.e. an XCOMP. First, it seems to denote a property. Second, the optional plural in 32 might be seen as agreement. There is, however, a decisive argument against predicate status: These sentences have impersonal passives, cf. 33.

(33) Det ble lekt lingvist hele dagen.
there was played linguist all day
They played linguists all day.

The nominal argument seems to be a nominal COMP. It cannot be a passive subject, and there is no adjectival passive, cf. 34-35.

(34) *Lingvist ble lekt.
linguist was played
(35) *en lekt lingvist
a played linguist

Parasitic gaps are unacceptable, cf. 36.

(36) *Lingvist(er) kan vi like *t* uten å måtte leke *e*.
linguist(s) can we like without to have.to play
We may like linguists without having to play linguists. [intended]

The nominalization and the synthetic compound are unacceptable, cf. 37-38. This is also true of the derived adjective, cf. 39.

(37) *leking av lingvist
playing of linguist
(38) *lingvistleking
linguist.playing
(39) *Lingvist er ikke lekbart.
linguist is not playable
Linguist cannot be played. [intended]

The verb *leke* 'play' can alternatively take an argument that denotes an established game. This argument behaves like a regular object; for example, it can be a subject in the passive, as in 40.

(40) Sisten lekes med stor entusiasme. (www)
tag play.PASS with great enthusiasm
Tag is played with great enthusiasm.

The verb *agere* 'act' behaves like *leke* 'play'. The verb *spille* 'act' also behaves this way when what is played is not 'established' in advance, as in *spille idiot* 'play idiot' (differing from *spille Hamlet* 'play Hamlet').

3.3 The verb *svare* 'answer'

The verb *svare* 'answer' can take an object that denotes the person who is answered. In addition, or instead, it can take an argument denoting the answer. This argument can be an OBLΘ or a clausal COMP, cf. 41 (which does not correspond to a passive subject, cf. 42), or an argument that could be a nominal COMP, cf. 43. The verb can passivize, as shown by the impersonal passive in 44, but the nominal argument cannot be a passive subject, and there is no adjectival passive, cf. 45-46.

(41) Jeg svarte (ham) at de var hjemme.
I answered (him) that they were home
I answered (him) that they were at home.
(42) *At de var hjemme ble svart (ham).
that they were home was answered (him)
It was answered him that they were home. [intended]
(43) De svarte noe tull.
they answered some nonsense
They answered some nonsense.
(44) Det ble svart noe tull.
there was answered some nonsense
Some nonsense was answered.
(45) *Noe tull ble svart.
some nonsense was answered
Some nonsense was answered. [intended]
(46) *noe svart tull
some answered nonsense

Parasitic gaps are unacceptable, cf. 47.

(47) *Slikt tull kan man tenke på *t* uten å måtte svare *e*.
such nonsense can one think of without to have.to answer
One may think of such nonsense without having to answer that. [intended]

The nominalization and the synthetic compound are unacceptable, cf. 48-49. This is also true of the derived adjective, cf. 50.

(48) *svaring av noe tull
answering of some nonsense
(49) *tullesvaring
nonsense.answering
(50) *Det tullet er ikke svarbart.
that nonsense is not answerable
One cannot answer that nonsense. [intended]

We see, then, that some verbs that take a clausal COMP can be argued to take a nominal COMP as an alternative. Other possible examples include the verbs *håpe* 'hope', *fantasere* 'fantasize', and *spørre* 'ask' (for some speakers; see note 6 below). It must be admitted, however, that it is not easy to find many clear cases of verbs that take either a clausal or a nominal COMP.

4. Mismatches between clausal and nominal COMP
4.1 Verbs with concealed questions

Some verbs seem to take a nominal, but not a clausal COMP.

Concealed questions are nominal complements that are interpreted as questions when they are headed by a verb that takes an embedded question (Grimshaw 1979). Typical examples are 51-52.

(51) Ola ville ikke fortelle tidspunktet.
Ola would not tell time.DEF
Ola would not tell the time.
(52) Jeg husker ikke hovedstaden i Sverige.
I remember not capital.DEF in Sweden
I do not remember the capital of Sweden.

Example 52 is ambiguous. If the complement is not a concealed question, it means that I cannot remember the city of Stockholm. If it is a concealed question, it means that I cannot remember which city is the capital of Sweden. In the latter case, it is possible to use the neuter pronoun *det* 'it, that' to refer to the definite masculine *hovedstaden* 'capital.DEF'. (This pronoun is used to refer to propositions and certain non-individuated nominals; see Lødrup 2012.)

The concealed questions in 51-52 behave like nominal COMPs. They show the expected properties (even if an impersonal passive is not possible because concealed questions are definite nominals): The concealed question cannot be a passive subject, and there is no adjectival passive, cf. 53-54.

Parasitic gaps are unacceptable, cf. 55, and so are the nominalization and the synthetic compound, and the derived adjective, cf. 56-58.

(53) *Tidspunktet ble fortalt.
time.DEF was told
(54) *det fortalte tidspunktet
the told time.DEF
(55) *Tidspunktet skal jeg bestemme *t* uten å fortelle *e*.
time.DEF shall I decide without to tell
I will decide the time without telling it. [intended]
(56) *fortelling av tidspunktet
telling of time.DEF
(57) *tidspunktfortelling
time.telling
(58) *Tidspunktet er ikke fortellbart.
time.DEF is not tellable
One cannot tell the time. [intended]

The verbs in 51-52, *fortelle* 'tell' and *huske* 'remember', do not take other COMPs than the concealed questions. Other nominal complements behave like objects, and this is also the case with complement clauses.

4.2 *prate* 'talk'

The verbs *prate* 'talk' and *snakke* 'talk' can take an OBLΘ that denotes what is being talked about, as in 59. As an alternative to this OBLΘ, they can take an indefinite bare noun that might be a nominal COMP, cf. 60. They cannot take a clausal COMP, however.

(59) prate om dop
talk about drugs
(60) prate dop
talk drugs

Again, the indefinite argument shows the expected properties. The verb can passivize, as shown by the impersonal passive in 61, but the nominal argument cannot be a passive subject, cf. 62, and there is no adjectival passive, cf. 63. Parasitic gaps are unacceptable, cf. 64. So are the nominalization, and the synthetic compound, and the derived adjective, cf. 65-67.

(61) Det ble pratet dop.
there was talked drugs
People talked about drugs.
(62) *Dop ble pratet.
drugs were talked
People talked about drugs. [intended]
(63) *pratet dop
talked drugs
(64) *Dop kan man like *t* uten å måtte prate *e*.
drugs can one like without to have.to talk
One may like drugs without having to talk about them. [intended]
(65) *prating av dop
talking of drugs
(66) *dopprating
drugs.talking
(67) *Dop er ikke pratbart.
drugs are not talkable
One cannot talk about drugs. [intended]

Another group of verbs that seem to take a nominal, but not a clausal COMP is verbs for emitting a substance from the body, such as *hoste* 'cough' or *spy* 'vomit'.

4.3 Clausal COMP, but no nominal COMP

There are also verbs that take a clausal COMP that do not take a nominal COMP. Some of these verbs do not take a (thematic) nominal complement at all, for example *henstille* 'request', *regne* 'assume', *akte* 'intend'.

Other verbs that take a clausal COMP take a nominal argument with the properties of a regular object, such as *anslå* 'estimate', or *erklære* 'declare'.[7] The verb *anslå* 'estimate' takes a clausal COMP, cf. 68 (which does not correspond to a passive subject, cf. 69). As an alternative to the clausal

---

[7] Another example is *spørre* 'ask', which takes a clausal COMP. Its nominal argument has the properties of regular object — for many language users. Examples such as (i)-(ii), which show verbal and adjectival passives, are easy to find on the www. However, I and other native speakers I have consulted do not accept them; we seem to have a nominal COMP with this verb.
(i) Alle spørsmål ble spurt på en veldig høflig og grei måte. (www)
all questions were asked in a very polite and nice way
All questions were asked in a very polite and nice way.
(ii) ofte spurte spørsmål (www)
frequently asked questions

COMP, the verb takes a nominal argument, which seems to be an object, cf. 70. This argument can be a passive subject, cf. 71, and there is an adjectival passive, cf. 72. A parasitic gap seems to be possible, cf. 73. The nominal argument must be an object, then, even if it does not satisfy the object criteria d), e) and f).

(68) Jeg anslår at han har ti katter.
I estimate that he has ten cats
I estimate that he has ten cats.
(69) *At han har ti katter ble anslått.
that he has ten cats was estimated
It was estimated that he has ten cats. [intended]
(70) Han anslo antallet katter.
he estimated number.DEF cats
He estimated the number of cats.
(71) Antallet katter ble anslått.
number cats was estimated
The number of cats was estimated.
(72) det anslåtte antallet
the estimated number.DEF
(73) (?)Utgiftene måtte han betale *t* uten å kunne anslå *e* på forhånd.
expenses.DEF must he pay without to be.able.to estimate in advance
He had to pay the expenses without being able to estimate them in advance.

The picture given of the selection of complements is complicated. This complexity seems to be difficult to avoid, however. A related area in which the complexity of syntactic selection is generally acknowledged concerns the selection of the formal categories of XCOMPs. It has been pointed out several times that it depends upon the individual verb, as illustrated in 74-75 (from Pollard and Sag 1987:122-23).

(74) Kim grew poetical / *a success.
(75) Kim ended up poetical / a success.

5. Properties of nominal COMPs

Some possible properties of nominal COMPs will be considered, based on the background of the cases discussed above.

5.1 Referentiality

The cases of nominal COMPs discussed were low in referentiality. With the verbs discussed, the nominal COMP could hardly be definite. (The concealed questions are different, however.) With some of the verbs discussed, the referent of the nominal COMP does not exist in advance of the verbal action. For example, what is moaned or answered only exists through the action denoted by the verb. In these cases, the nominal COMP is what has been called an object of result (see e.g. Jespersen 1963:159-60). It is not the case, however, that objects of result are always COMPs; they often behave as regular objects. (For example, *grave en grøft* 'dig a ditch' has a regular passive.)

Some of the nominal COMPs are bare nominals, i.e. indefinite nominals without determiners. Bare nominals are reluctant to be subjects in Norwegian, and this might explain that they do not correspond to passive subjects. There is no absolute restriction against bare nominals as (passive) subjects, however. (Cf. the following sentence from the www: *Plass kan bestilles på forhånd* 'seat can order-PASS in advance'.) Furthermore, there are also nominal COMPs with quantifiers, as in examples 18 and 43 above.

5.2 Thematic roles

Most nominal COMPs discussed are abstract participants in the verbal event, and have a neutral, theme- or patient-like role. This role is realized as an object with many verbs that are close to COMP-taking verbs in meaning. For example, *hviske* 'whisper' takes an ordinary object, even if *stønne* 'moan' and other manner of speaking verbs take a COMP (see footnote 5).

Traditional Lexical Mapping Theory (LMT, Bresnan and Kanerva 1989) cannot account for nominal COMPs. This is not necessarily an argument against the idea, however, because traditional LMT cannot account for the traditional clausal COMP or XCOMP either. Furthermore, traditional LMT can be extended to include COMP and XCOMP; see Falk (2001:136-41).

A possible alternative to the idea of a nominal COMP is that the arguments in question could be OBJΘs.[8] It is sometimes assumed that an object that does not show the core object properties is an OBJΘ, even if it does not co-occur with an OBJ (Lødrup 1995, Dalrymple and Nikolaeva 2011). In Dalrymple and Nikolaeva (2011:ch 8), OBJΘ is the unmarked, non-topical object without the properties of core grammatical functions. This

[8] Postal (2010:106-11) considers the second object in ditransitive sentences to be an 'array 1 object', with English-specific arguments.

might seem to fit the nominal argument that is seen here as a nominal COMP. However, such an analysis would not capture the relation between a nominal COMP and a clausal COMP.

5.3 Is nominal COMP an OBLΘ?

There is an affinity between nominal COMPs and OBLΘs (Lødrup 2004). With some verbs, a nominal COMP alternates with an OBLΘ. In those cases, however, the OBLΘ does not have restrictions on definiteness or referentiality corresponding to those of a nominal COMP. An example is 76.

(76) Vi pratet *(om) dopen.
we talked about drugs.DEF
We talked about the drugs.

It does not seem motivated to identify a nominal COMP with an OBLΘ. One argument is that there are almost no clear cases of a nominal OBLΘ in Norwegian. The cases that exist are very different from nominal COMPs. Example 77 has a temporal OBLΘ, while 78 has a locative OBLΘ. They allow the insertion of a preposition, while this is not necessarily the case with nominal COMPs.

(77) Konserten varer (i) tre timer.
concert.DEF lasts (for) three hours
The concert lasts for three hours.
(78) Ola har bodd (på) mange steder.
Ola has lived (in) many places
Ola has lived in many places.

Another argument is that a nominal COMP can be pronominalized, and it is asked for with an interrogative pronoun, cf. 79-80. This is not the case with OBLΘ.

(79) Det svarte jeg.
that answered I
I answered that.
(80) Hva svarte du?
what answered you
What did you answer?

Finally, coordination could give an argument against identifying a nominal COMP with an OBLΘ. A clausal COMP can be coordinated with a nominal

COMP, as in 81. Because the ability to coordinate cannot be due to phrasal structure here, it must be due to syntactic function.

(81) Han svarte noe tull og at han måtte rekke trikken.
he answered some nonsense and that he must catch tram.DEF
He answered some nonsense and that he had to catch the tram.

6. Unbounded dependencies

An important property of COMP that was decisive for the analysis in Dalrymple and Lødrup (2000) is that a clausal COMP cannot take part in an unbounded dependency. This generalization has been known as 'Higgins's Generalization' (Higgins 1973). Dalrymple and Lødrup (2000) paraphrased it in the following way:

"A clausal argument can enter into an unbounded dependency only if it is in an NP position, i.e. a position in which an NP is possible as an alternative to the clausal argument."

The predicted situation is illustrated in 82-85.

(82) That the earth is round, everybody believed.
(83) Everybody believed it.
(84) *That it would rain, everybody hoped.
(85) *Everybody hoped it.

Higgins's Generalization has been discussed within different frameworks (Stowell 1981, Kaplan and Bresnan 1982, Kaplan and Zaenen 1989, Webelhuth 1992, Bošković 1995, Berman 1996). Dalrymple and Lødrup (2000) proposed that COMP is an exception to the general option for a syntactic function to enter into an unbounded dependency (or, more technically, that COMP cannot be the 'bottom' of a functional uncertainty equation $\uparrow$DF = $\uparrow$GF* GF).

The question is then if a nominal COMP can take part in an unbounded dependency. The simple answer is that it can, like all other nominal arguments in Norwegian (even if some cases might sound a bit strange — the reason is probably that a nominal COMP is not a good topic). Examples are 86-87.

(86) Lingvister tror jeg vi leker hver dag.
linguists think I we play every day
I think that we play linguists every day.

(87) Tull tror jeg ikke det er noen som svarer.
nonsense think I not there is anybody that answers
I don't think that there is anybody who answers nonsense.

This means that one important point of Dalrymple and Lødrup (2000) is no longer valid. Higgins's Generalization can no longer be stated as referring to syntactic function only if COMPs can be clausal or nominal. This might be considered an important argument against the idea of a nominal COMP. This is not the case, however. Postal (2004:279-282) argues that Higgins' Generalization is not empirically correct for English. It cannot be correct for Norwegian either. Part 3 discussed verbs that take either a clausal or a nominal COMP. These clausal COMPs cannot enter into an unbounded dependency, as shown by 88-89.

(88) *At alt var slutt stønnet han.
that everything was over moaned he
That everything was over, he moaned. [intended]
(89) *At de var hjemme svarte jeg ham.
that they were home answered I him
That they were at home, I answered him. [intended]

These clauses alternate with noun phrases, however. Given Higgins' Generalization, it is impossible to see any reason that these clauses are exceptions to the general option of taking part in an unbounded dependency. (A parallel case is verbs such as *anslå* 'estimate' in section 4.4, which take clausal COMPs that do not topicalize, even if they alternate with a nominal object.)

Given these premises, the generalization that a clausal COMP cannot take part in an unbounded dependency must be stated referring both to form and function. This generalization would seem to be unnecessarily complicated. It is difficult to see an alternative, however, and it is striking that there is another syntactic function in LFG whose ability to take part in an unbounded dependency depends upon its form (at least in some languages): An XCOMP can take part in an unbounded dependency only when it is non-verbal, as shown in 90-91.

(90) Redd vil jeg ikke si han virker.
afraid will I not say he seems
I would not say that he seems afraid.
(91) *Å være redd vil jeg ikke si han forekommer meg.
to be afraid will I not say he seems to.me
I would not say that he seems to me to be afraid.

7. Conclusion

This paper has presented some evidence from one language that COMP can be nominal. However, it must be admitted that there are also problems involved. First, even if possible examples of a nominal COMP can be found, it is not easy to find many clear cases of verbs that take either a clausal or a nominal COMP. Second, the differentiation of the various complement functions raises problems in general, as is well known form discussion inside and outside LFG.

It is possible that Norwegian does not give the best point of departure for investigating the grammar of COMP. Many Norwegian verbs that take clausal complements take clausal complements that show object properties (Lødrup 2004). The existence of clausal complements without object properties was not acknowledged at all by traditional Norwegian grammar.

More work on different languages is needed before it can be established if COMP should be a part of the inventory of syntactic functions in grammatical theory.

REFERENCES

Alsina, Alex, KP Mohanan and Tara Mohanan 2005. How to get rid of the COMP. In Miriam Butt and Tracy Holloway King (eds.) *Proceedings of the LFG05 Conference*, 21-41. CSLI Publications.

Berman, Judith, 1996. Topicalization vs. left dislocation of sentential arguments in German. In Miriam Butt and Tracy Holloway King (eds), *Proceedings of the First LFG Conference, 1996*. CSLI Publications.

Berman, Judith 2003. *Clausal Syntax of German*. CSLI Publications.

Börjars, Kersti and Nigel Vincent 2008. Objects and OBJ. In Miriam Butt and Tracy Holloway King (eds.) *Proceedings of the LFG08 Conference*, 150-168. CSLI Publications.

Bošković, Željko, 1995. Case properties of clauses and the Greed Principle. *Studia Linguistica* 49, 1, 32–53.

Breindl, Eva, 1989. *Präpositionalobjekte und Präpositionalobjektsätze im Deutschen*. Niemeyer.

Bresnan, Joan, 2001. *Lexical-Functional Syntax*. Blackwell Publishers.

Bresnan, Joan and Jonni M. Kanerva 1989. Locative inversion in Chichewa: A case study of factorization in grammar. *Linguistic Inquiry* 20, 1-50.

Bresnan, Joan and Annie Zaenen 1990. Deep unaccusativity in LFG. In Katarzyna Dziwirek et al. (eds.) *Grammatical Relations: A Cross-Theoretical Perspective*, 45-57. CSLI Publications.

Dalrymple, Mary and Helge Lødrup 2000. The grammatical functions of complement clauses. In Miriam Butt and Tracy Holloway King (eds.) *Proceedings of the LFG00 Conference*. CSLI Publications.
Dalrymple, Mary and Irina Nikolaeva 2011. *Objects and Information Structure*. Cambridge University Press.
Engdahl, Elisabet 2001. Versatile parasitic gaps. In Peter W. Culicover and Paul M. Postal (eds.) *Parasitic Gaps*, 127-45. MIT Press.
Falk, Yehuda 2001. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-based Syntax*. CSLI Publications.
Forst, Martin 2006. COMP in (parallel) Grammar Writing. In Miriam Butt and Tracy Holloway King (eds.) *Proceedings of the LFG06 Conference*. CSLI Publications.
Grimshaw, Jane 1979. Complement selection and the lexicon. *Linguistic Inquiry*, 10, 2, 279-326.
Higgins, F. R., 1973. On J. Emonds's analysis of extraposition. In John P. Kimball (ed.), *Syntax and Semantics 2*, 149–195. Academic Press.
Jackendoff, Ray S. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press.
Jespersen, Otto 1963 [1924]. *The Philosophy of Grammar*. Allen & Unwin.
Kaplan, Ronald M. and Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan (ed.) *The Mental Representation of Grammatical Relations*, 173-281. MIT Press.
Kaplan, Ronald M. and Annie Zaenen 1989. Long-distance dependencies, constituent structure, and functional uncertainty. In Mark Baltin and Anthony Kroch (eds.) *Alternative Conceptions of Phrase Structure*, 17–42. Chicago University Press.
Lødrup, Helge 1995. The realization of benefactives in Norwegian. In Audra Dainora et al. (eds.) *Papers from the 31. Regional Meeting*, 317-328. Chicago Linguistic Society.
Lødrup, Helge 1999. Linking and optimality in the Norwegian presentational focus construction. *Nordic Journal of Linguistics* 22, 2, 205-229.
Lødrup, Helge 2000. Exceptions to the Norwegian passive: Unaccusativity, aspect and thematic roles. *Norsk lingvistisk tidsskrift* 18, 1, 37-54.
Lødrup, Helge 2004. Clausal complementation in Norwegian. *Nordic Journal of Linguistics* 27, 1 61-95.
Lødrup, Helge 2012. Some Norwegian 'type anaphora' are surface anaphora. *Journal of Germanic Linguistics* 24, 1 23–52.
Pollard, Carl and Ivan A. Sag 1987. *Information-Based Syntax and Semantics*. CSLI Publications.
Postal, Paul M. 2004. *Skeptical Linguistic Essays*. Oxford University Press.
Postal, Paul M. 2010. *Edge-based Clausal Syntax: A Study of (Mostly) English Object Structure*. MIT Press.

Ralph, Bo, 1975. On the nature of preposition deletion in Swedish. In Karl-Hampus Dahlstedt (ed.) *The Nordic Languages and Modern Linguistics 2*, 666–684. Almqvist and Wiksell.
Rosenbaum, Peter, 1967. *The Grammar of English Predicate Complement Constructions*. MIT Press.
Simpson, Jane 1983. Resultatives. In Lori Levin, Malka Rappaport, and Annie Zaenen (eds.) *Papers in Lexical-Functional Grammar*, 143-57. Bloomington, IN: Indiana University Linguistics Club. Reprinted in Miriam Butt and Tracy Holloway King (eds.) *Lexical Semantics in LFG*, 149-61. CSLI Publications, 2006
Stowell, Timothy, 1981. *The Origins of Phrase Structure*. Ph.D. thesis, MIT.
Webelhuth, Gert, 1992. *Principles and Parameters of Syntactic Saturation*. Oxford University Press.
Zaenen, Annie and Elisabet Engdahl, 1994. Descriptive and theoretical syntax in the lexicon. In B. T. S. Atkins and Antonio Zampolli (eds.), *Computational Approaches to the Lexicon*, 181–212. Oxford University Press.
Zifonun, Gisela, Ludger Hoffmann, and Bruno Strecker, 1997. *Grammatik der deutschen Sprache*. de Gruyter.
Zwicky, Arnold M. 1971. In a manner of speaking. *Linguistic Inquiry*, 2, 2, 223-233.

Helge Lødrup
University of Oslo
Department of Linguistics and Scandinavian Studies
Pb 1102, Blindern, NO-0317 Oslo, Norway
helge.lodrup@ilf.uio.no
http://folk.uio.no/helgelo

INESS-SEARCH:
A SEARCH SYSTEM FOR LFG
(AND OTHER) TREEBANKS

Paul Meurer
Uni Computing, Bergen

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper describes INESS-Search, a new search tool for constituency, dependency and LFG treebanks. The tool is derived from TIGERSearch and has been extended to encompass full first-order predicate logic over node variables. In addition, several operators have been implemented that are specific for querying c- and f-structures. The original TIGERSearch syntax has been extended and considerably simplified, thus making a graphical query input device less necessary. The search index is dynamically updated when the treebank is modified. The INESS-Search tool is usable via a Web interface as an integrated part of INESS, the Norwegian Infrastructure for the Exploration of Syntax and Semantics.

1 Introduction

In the last two decades, many tools for querying traditional dependency and constituency treebanks have been developed. They all differ in expressiveness, query language and formalism, ease of use, and applicability to specific kinds of treebanks. But no tool has been developed previously that can handle LFG treebanks, since LFG treebanks of a size that calls for a dedicated and powerful search tool have only recently been emerging.

The structural representation of syntactical analyses in Lexical Functional Grammar is quite different from and more complex than the tree-like structures that we encounter in traditional treebanks. Whereas c-structures in fact are proper (ordered) trees, f-structures can be described as unordered directed graphs, possibly with cycles. In addition, c- and f-structures are interconnected by virtue of the projection relation holding between c-structure nodes and sub-f-structures, and thus must be seen in combination (again, formally describable as a directed graph). In contrast, the structures that are prominent in traditional treebanks are the following:

- proper ordered trees, with or without labeled edges (e.g. the Penn Treebank)
- proper unordered trees as used in dependency treebanks derived from Constraint Grammar (e.g. the Sami treebanks in INESS)
- trees equipped with some additional structure peculiar to a specific framework or treebank (e.g. secondary edges and crossing edges in the Tiger treebank)
- unordered trees with or without secondary edges in some dependency treebanks (e.g. the PROIEL treebank (Haug, 2008) in INESS)

None of those tree varieties are equivalent to general directed graphs.

As a consequence, existing treebank search tools, which are designed to operate on traditional treebanks, are unable to cope with fully general directed graphs. Some of those tools are designed or implemented in a way that makes them in principle unsuited for general directed graphs, as there is no way to extend them beyond proper trees. Examples are LPath^{+} (Lai & Bird, 2005) and MonaSearch (Maryns, 2009).

Other tools are in principle extensible to directed graphs, like TIGERSearch, Emdros (Petersen, 2005) and fsq (Kepser, 2003).

Among those tools, TIGERSearch (Koenig, Lezius, 2003) was identified as a suitable basis for the implementation of INESS-Search for the following reasons:

- TIGERSearch is equipped with an elegant and concise query language that can easily be extended to meet the needs of a more general search tool.

- The implementation of dominance relations in TIGERSearch via Gorn addressing naturally extends to directed graphs and lends itself to an elegant implementation of circularity detection.

- The Java implementation of TIGERSearch is reasonably fast, so one could expect that a reimplementation would have acceptable query execution speed.

INESS-Search contains extensions necessary to query fully general directed graphs like LFG f-structures, but also implements the full functionality of TIGERSearch and thus can be used to query constituency and dependency treebanks.

Whereas the expressive power of the query language of TIGERSearch can be characterized as roughly equivalent to the existential fragment of first-order predicate logic over node variables, the query language of INESS-Search is equivalent to full first-order predicate logic.

The INESS-Search tool is useable via a Web interface as an integrated part of INESS, the Norwegian Infrastructure for the Exploration of Syntax and Semantics (Rosén et al., 2012).

2 Abbreviated syntax and specialized operators

In order to make the syntax of the query language concise and easy to use, the original TIGERSearch syntax has been extended with convenient abbreviations and path-like concatenation of operators. Variables in operator expressions can be omitted when they are not needed for coreferencing in other relations. Examples for full and abbreviated syntax are given below.

- *Terminal nodes*

full:	[word="Sophie"]
abbreviated:	"Sophie"

- *Node labels*

full:	#c:[cat="NP"]
abbreviated:	#c:NP

- *Operator concatenation*

full:	[cat="IP"] $>$ #x:[cat="NP"] & #x $>$ [cat="N"]
abbreviated:	IP $>$ NP $>$ N

- *Omission of variables in relations*

full:	#f $>$OBJth #g
abbreviated:	$>$OBJth

In addition, several operators have been implemented that are specific for querying complex tree node and f-structure constellations:

- A *rule operator*, which has the shape of a derivation rule and makes it possible to specify relations between mothers and daughters

  #c $\rightarrow$ AP .$*$ PP

- A *c-command operator*

  #n $>$c$>$ #c

Some operators are specific to LFG c- and f-structures:

- A *path operator* specifying regular expressions over f-structure attributes:

  g is either the value of OBJth of f or contained in the ADJUNCT set of f

  #f $>$(OBJth | ADJUNCT $) #g

  g is both OBJ and TOPIC of f

  #f $>$(OBJ & TOPIC) #g

- A *projection operator*: tree node c projects to the f-structure f

  #c $>>$ #f

- A *projective equivalence operator*: nodes c_1 and c_2 are in the same projective domain

  #c1 $>><<$ #c2

- An *extended-head operator*: n is the extended head of c according to the definition given in Bresnan (2001)

  #n $>$h$>$ #c

Many of these operators could in principle be expressed and implemented using more primitive relations like dominance and labeled dominance. Defining them as dedicated operators however has two advantages: queries can be expressed more concisely, and the operators can be hard-coded, resulting in dramatically improved performance.

The syntax of INESS-Search is sufficiently compact and intuitive to make elaborate graphical query devices unnecessary, especially in the case of relatively simple searches. Moreover, in the case of more complex searches involving advanced operators and quantification, a GUI would face expressiveness challenges. Instead, we will in further work explore the possibilities offered by predefined examples and cached previous queries.

3 Querying parallel treebanks

INESS-Search is being extended with a parallel search mode (Dyvik, Meurer, Rosén & De Smedt, 2009). This mode is still in an experimental stage. The main idea is that for aligned sentence pairs, certain nodes (tree nodes or c-structure nodes and sub-f-structures) will be aligned. To make alignment searchable, an alignment relation has been introduced as shown in (1).

(1) #s >>> #t

This relation holds if s is instantiated by a node in the source c- or f-structure, t is instantiated by a node in the target c- or f-structure, and those nodes are aligned. Thus, query (2) will match all aligned pairs of analyses in a Norwegian–English parallel treebank where a source c-structure lexical node “jente” is aligned with a target c-structure lexical node “girl”.

(2) #s:“jente” >>> #t:“girl”

An alignment relation can of course be part of a more complex query expression, as (3) illustrates. This query will match a source c-structure node dominating a lexical node “jente”, aligned with a target c-structure node dominating a lexical node “girl”.

(3) #s > “jente” & #t > “girl” & #s >>> #t

Our approach is influenced by Volk, Lundborg & Mettler (2007), who were the first to devise a syntax for querying node alignment based on TIGERSearch, which they implemented in the Stockholm Tree Aligner tool.

4 Expressivity

The expressive power of the original TIGERSearch query language is equivalent to the existential fragment of first-order predicate logic over node and value variables.

In TIGERSearch, all variables are implicitly existentially quantified and universal quantification is not available. Unfortunately, with existential quantification alone, many seemingly basic queries cannot be expressed, as we will see below.

Therefore, the query language of INESS-Search has been equipped with unrestricted universal quantification over node variables and a couple of new predicates and operators including the equality operator. Its expressivity is equivalent to full first-order predicate logic over node variables (with the less important addition of value variables, which are always existentially quantified).

The introduction of universal quantification increases the complexity of the query language; new notational devices have to be introduced, and they have to be provided with an interpretation in terms of predicate calculus. Since in TIGER-Search all variables are existentially quantified, quantification does not have to be specified explicitly, that is, no quantifier expressions (i.e., $\exists x \exists y$: ...) are needed. When both existential ($\exists x$: ...) and universal quantification ($\forall y$: ...) are possible, quantification has to be specified explicitly. This, however, can clutter a query expression considerably. Therefore, notational conventions are introduced that make the use of explicit quantifiers unnecessary in most cases.

First, the variable marker $\#$ is interpreted as an existential quantifier marker; each variable occurring with a $\#$ (and being in a positive context; see below) introduces an existential quantifier in prenex form (i.e., standing to the left and scoping over all terms of the expression). Also implicit variables, variables that are tacitly introduced via an abbreviated syntax construction, are existentially quantified. Thus, a query expression like (4) is translated into the logical form (5). Since both quantifiers are of equal type, the quantifier order is insignificant.

(4) #x $>$ #y

(5) $\exists x \exists y : x > y$

In order to express universal quantification, a new variable marker $\%$ is introduced.[1] A variable marked with $\%$ is universally quantified and introduces a universal quantifier in prenex form. The expression (6) translates to the logical form (7).

(6) #x $>$ $\%$y

(7) $\exists x \forall y : x > y$

When existential and universal variables cooccur in one query expression as in (6), quantifier order is no longer arbitrary. If the quantifier order is not specified explicitly, a default scoping rule determines that all universal variables are in the scope of all existential variables.

If the default scoping order is not the intended one, scoping can be specified explicitly by stating the intended quantifier order in parentheses at the beginning of the query expression:

(8) (%y #x): #x $>$ $\%$y

[1]See Marek, Lundborg & Volk (2008), who first introduced the use of $\%$ as a notational device for universal quantification, but gave it a different interpretation.

Query (8) translates to the logical form (9).

(9) $\forall y \exists x : x > y$

It is also important to note how constraints on variables are interpreted in the case of universal quantification. A constraint like #x:[cat='NP'] (stating that x should be an NP node) can either be realized as a predicate clause in the logical form: $\exists x$: cat(x, 'NP'), or it could be interpreted as a *restricted quantifier*[2]: $\exists x$.cat(x, 'NP'). In the case of existential quantification, the two interpretations are equivalent.

However if we consider an example like (10) that involves universal quantification, the two interpretations given in (11) and (12) are no longer equivalent.

(10) *Find all sentences where each NP directly dominates an N*

(%x #y): %x:NP $>$ #y:N

(11) $\forall x \exists y$: cat(x, 'NP') $\wedge$ cat(y, 'N') $\wedge$ $x > y$

(12) $\forall x$.cat(x, 'NP') $\exists y$.cat(y, 'N'): $x > y$

In interpretation (11), variable x ranges unrestrictedly over all nodes, and the predicate cat(x, 'NP') requires that every node be an NP node, which is clearly not the intended interpretation of (10). In interpretation (12) however, x ranges over the restricted domain of NP nodes, and only for each of those, a dominated N node has to exist.

Thus, the restricted quantifier interpretation of constraints is the intended one, and the one that is implemented. To make this interpretation more explicit, the constraints can also be placed together with the quantifiers, as in (13).

(13) (%x:NP #y:N): %x $>$ #y

Further complications arise when we introduce negation. Consider example (14), where the node variable z is only mentioned in a negative context.

(14) *A* PP *node dominating an* N *node with no intervening* PP *node*

#x:PP $>*$ #y:N & !(#x $>*$ #z:PP $>*$ #y)

The intended meaning of the query, phrased in prose, is: "Find nodes x (PP) and y (N) such that there is no node z (PP) lying between x and y." Thus, z is interpreted as existentially quantified in the scope of the negation. (Note that x and y are already existentially quantified outside the scope of the negation.) This leads to the logical form (15).

(15) $\exists x$.cat(x, 'PP') $\exists y$.cat(y, 'N'): $x >* y \wedge \neg(\exists z$.cat($z$, 'PP'): $x >* z >* y)$

[2] A restricted quantifier expresses a restriction on the domain over which the variable in question ranges.

This logical form can be transformed into prenex form (16), which is the canonical form underlying the implementation of the query expressions.

(16) $\exists x$.cat(x, 'PP') $\exists y$.cat(y, 'N') $\forall z$.cat(z, 'PP'): $x >\!\!* \; y \wedge \neg(x >\!\!* \; z >\!\!* \; y)$

Observe that by moving it out of the scope of the negation, the existential quantifier is transformed into a universal quantifier. In the same way, a negated universal quantifier resurfaces as an existential quantifier in prenex form.

We should keep in mind that the TIGERSearch query language does allow constraint variables and value variables, in addition to node variables. In query (17), c is a value variable that is used to express that x and y should have equal *cat* values. The corresponding logical form is given in (18).

(17) #x:[cat=#c] $>*$ #y:[cat=#c]

(18) $\exists x \exists y \exists c$: $x >* y \wedge$ cat(x, c) $\wedge$ cat(y, c)

INESS-Search allows constraint and value variables to occur only with existentially quantified node variables that are not in the scope of a universal quantifier since it is otherwise difficult to give a sensible interpretation.

The rules that determine the interpretation of quantification and constraints in the extended query language of INESS-Search can be summarized as follows:

- **Prenex form**: all quantifiers precede the body of the logical form
- **Existentially quantified** are: $\#$-variables and implicit variables in a positive context; $\%$-variables in a negated context
- **Universally quantified** are: $\%$-variables in a positive context; implicit variables and $\#$-variables in a negated context that are not mentioned in a positive context
- **Default scoping**: universal variables are in the scope of all existential variables by default
- **Explicit scoping**: quantifier scoping can be explicitly specified in prenex form
- **Constraints on variables** are interpreted as restricted quantifiers

One could ask what the practical value of the increased expressiveness of INESS-Search might be. In their survey of treebank query systems, Lai & Bird (2004) list typical queries that a query system should be able to express. Among those queries that are relevant in our setting (Q1–Q5), TIGERSearch is not able to handle Q2 and Q5:

(19) Q2: Find sentences that do not include the word "saw".

Q5: Find the first common ancestor of sequences of a noun phrase followed by a verb phrase.

These queries can easily be expressed in INESS-Search as:

(20) Q2: !(#x:"saw" = #x)

Q5: #c $>*$ #n:NP !$>*$ #v &
#c $>*$ #v:VP !$>*$ #n &
!(#c $>*$ #x $>*$ #n & #x $>*$ #v)

The formulation of Q2 might seem slightly odd at first glance, but its meaning becomes clearer when we look at the corresponding logical form (21), where the constraint is transformed into a restricted quantifier.

(21) Q2: $\forall x$.word(x,"saw"): $\neg(x = x)$

A tree matches the query Q2 if every node whose word attribute has the value "saw" is not equal to itself. Since $x = x$ is tautologically true for every node instantiation of x, this means that the restricted domain defined by word(x,"saw") must be empty, that is, the tree must not contain any such node.

One might consider introducing a more intuitive abbreviated syntax for Q2, e.g., !"saw".

Full first-order predicate logic is not the most powerful logical system conceivable. Most importantly, transitive closure of binary relations cannot be expressed in first-order predicate logic. Since the transitive closure of some basic relations, notably direct dominance and direct precedence, are of crucial importance in a linguistic querying system, they are normally implemented as basic operators (dominance and precedence).

Other useful complex relations like the c-command relation and the extended-head relation that could hardly be defined efficiently using more basic relations have been implemented in INESS-Search as hard-coded relations.

It is however not possible to define transitive closures of arbitrary ad-hoc relations. Maryns (2009) mentions as an example the transitive closure of the dominance relation PP $>$ NP, which could be used to find arbitrarily long chains of embedded PPs dominating NPs. This query cannot be expressed in first-order predicate logic, but it can be expressed in MonaSearch, which is based on an implementation of Monadic second-order logic. It is not clear to me whether such queries are of great practical importance. MonaSearch, however, cannot be extended to general directed graphs; the tree automata that MonaSearch query expressions are compiled into can only handle proper trees.

INESS-Search is not the only attempt to extend TIGERSearch with universal quantification. In their paper entitled "Extending the TIGER query language with universal quantification", Marek, Lundborg & Volk (2008) point out the lack of expressive power in TIGERSearch and try to outline a design of a universal quantification extension to TIGERSearch. They introduce the notion of a “node set”; variables instantiated by node sets are marked with a %. Marek et al. do not explicitly equate node set variables with universally quantified variables, although their definition makes it clear that the concepts are the same. Unfortunately, by not seeing this equivalence, they also do not see how %-variables interact with negation and implication, and instead try to extend their “node set” notion in a rather complicated way by introducing “subqueries” in order to cope with queries of type Q5.

Marek et al. seem to have partially implemented the “node set” extension in their adaptation of TIGERSearch, whereas “subqueries” are only proposed as an extension. While they state that their approach is easy to implement, they also mention that it is very slow, and they cite the arguments of the developers of TIGERSearch for not having implemented universal quantification:

> The use of the universal quantifier causes computational overhead since universal quantification usually means that a possibly large number of copies of logical expressions have to be produced. For the sake of computational simplicity and tractability, the universal quantifier is (currently) not part of the TIGER language. (TIGERSearch Help, section 10.3)

This, however, is a misconception; as I show in the outline of the implementation, the computational complexity introduced by a universally quantified variable is not significantly higher than the complexity originating from existential variables.

5 Implementation

INESS-Search is written in Common Lisp. The implementation is heavily inspired by the TIGERSearch implementation, and parts of the query parser are a reimplementation of the code of the Stockholm Tree Aligner (Marek, Lundborg & Volk, 2008).

5.1 Static and dynamic indices

In INESS-Search, the various search indices are static and are stored in files on disk. Using the Unix system call *mmap*, those index files are mapped onto virtual memory addresses. Since *mmap* implements demand paging, only those parts (pages) of the index files that are actually needed are loaded into main memory in a lazy manner. This obviates the need for loading the files entirely into main memory, as is done in TIGERSearch.

The treebank index consists of inverted indices for the various features that are represented in the treebank (including *word*, *cat*, *parent-edges* and *child-edges*), and a graph file encoding the graphs of the entire treebank. Whereas the graph file can only be traversed sequentially, the inverted indices allow a quick lookup of all graphs containing a node with a given feature value, and of all nodes with a given feature value. In addition, since the lexicon part of the inverted index is organized as a *suffix array* (Manber & Myers, 1991), sentences and nodes whose feature values satisfy a given regular expression can be looked up equally quickly.[3] This ability to look up all and only those graphs and nodes that satisfy given constraints is crucial in the implementation of an efficient query evaluation strategy.

An alternative to storing the treebank index in static files which is persued in some query tools (e.g., ANNIS2[4]) is to use a relational database. The advantages of a relational database approach are immediate: index lookup and joins are built-in functionality and do not have to be implemented in the tool, and, most importantly, relational databases are dynamic; it is easy to add trees to the treebank index, or to delete trees from it. This flexibility, however, comes at a price. When querying a relational database, there is some overhead connected to keeping track of transactions and concurrency, and to client-server communication. This means in practice that querying a database is potentially much slower than reading from an *mmap*-ed file with a dedicated index structure.[5] On the other hand, as most treebanks that have been constructed so far are quite static in nature, there is little need to change them dynamically.

The LFG treebanks stored in the INESS system are in fact an exception in that respect. Since it is possible to disambiguate the parses of a given sentence in the treebank, an INESS LFG treebank is quite dynamic while it is being constructed. In order to keep the treebank index synchronized with the evolving treebank and make it seem dynamic, the index has been divided into two layers. The main index layer is a static index reflecting the treebank state at the time when the index was generated. In addition, there is an incremental layer which indexes only those sentences that have been added or edited since the main index layer was compiled. It also keeps track of deleted sentences. Since the incremental index is quite small, it can be compiled very fast, and thus can be regenerated every time the treebank changes. To keep the incremental index small, the main index is regenerated offline when the incremental index exceeds a certain size.

5.2 Query evaluation strategy

Every INESS-Search query is equivalent to a logical form Q such that all quantifiers are in prenex form, all node constraints are expressed as quantifier restrictions, and the body of the form is a boolean combination of binary relations and predicates. We can assume that the body is normalized, in the sense that it is equal

[3]See Meurer (2012) for a detailed account on the indexing techniques used here.

[4]See http://www.sfb632.uni-potsdam.de/d1/annis/.

[5]Experience from the ANNIS2 project (Rosenfeld, 2010) suggests that this can be compensated for by using a sophisticated indexing strategy, which, however, results in long indexing times and a large on-disk index.

to a disjunction of unions of relations, predicates and negated terms, where each negated term is the negation of a union of relations and predicates.

A query is parsed into an internal representation that is close to the logical form, but where auxiliary node, constraint and value variables are introduced that make it possible to represent the query in a flat form.

A *match* of a query Q$(x_1, ..., x_n)$ with variables $x_1, ..., x_n$ is a graph Γ together with an instantiation of all the existential variables up to the first universal variable with nodes $X_1, ..., X_k$ from Γ such that Q$(X_1, ..., X_k,\ x_{k+1}, ..., x_n)$ evaluates to true.

Let us look at the example query (22), which corresponds to the logical form (23) and has the internal representation (24). The slashes /... / denote a regular expression; plus and minus signs mark whether a variable or value occurs in an existential context.

(22) (#x:IP %s:S$*$ #y:PROP): #x $>*$ %s $>*$ #y

(23) $\exists x$.cat(x, 'IP') $\forall z$.cat(z, /S.$*$/) $\exists y$.cat(y, 'PROP'): $x >* s\ \&\ s >* y$

(24)
```
node-order: #x, %s, #y
node-var:   #x, node: [#fc_1] (+)
node-var:   %s, node: [#fc_2] (-)
node-var:   #y, node: [#fc_3] (+)
fc-var:     #fc_1, constraint: cat=#fv_1/+
fc-var:     #fc_2, constraint: cat=#fv_2/+
fc-var:     #fc_3, constraint: cat=#fv_3/+
fv-var:     #fv_1, value: 'IP' (+)
fv-var:     #fv_2, value: /S.*/ (+)
fv-var:     #fv_3, value: 'PROP' (+)
relations:  %s >* #y, #x >* %s
```

A simple-minded algorithm for evaluating a query on a set of graphs (a treebank) would be to go through the graphs one by one, and check for every possible instantiation of the variables (by doing a depth-first traversal of the search space) whether the body of the logical form evaluates to true. This algorithm is actually correct, although not necessarily very efficient, when all variables are existentially quantified.

Some improvements are immediate: We only have to consider graphs that for every quantifier contain nodes that match the node constraints (i.e., that are lying in the domain of the restricted quantifier), and each node variable again only needs to be instantiated with those nodes that match the respective restrictions. As has been shown, finding those candidate graphs and nodes can be done very efficiently by a reverse-index lookup.

The set of candidate nodes can be restricted further by using relation and predicate signatures. For a given relation or predicate, certain types of nodes can be excluded a priori from the set of node candidates. For instance, in the dominance relation $x >\!* \, y$, x can only be instantiated by non-terminal nodes, and in the projection relation $c >\!> f$, c must be a c-structure node and f an f-structure node. The restrictions on the node types of a relation or a predicate is called the *signature* of the relation or predicate. Since the type of a node is coded in the inverted index, the signature information can effectively be used in reverse-index lookup.

When there are universally quantified node variables involved, a correct algorithm is substantially more complex, since it is not sufficient to evaluate the body of the logical form for each instantiation of the variables in isolation. The outline given below is quite close to the actual implementation, although it does not spell out details of the technically rather intricate treatment of dependent disjunctions, negation of unions of relations, and of variable binding and backtracking for value variables.

- Let Q be a query with node variables x_1, ..., x_n, constraints, predicates and relations.
- Begin by calculating candidate graphs using reverse index lookup for existential constraints up to the first universal variable (in (24): sentences having an IP).
- For each candidate graph Γ, calculate candidate node sets for each variable that match the constraints (in (24): all IP, S$*$, PROP nodes for x, s, y), or a dummy node for a universal variable if it is not instantiable.

The matches of Q for a given graph Γ can be calculated by recursion over the candidate node sets. We first need some definitions:

- A *partial matching tuple* $(X_1, ..., X_i)$ of nodes in Γ for some $i \leq n$ is an instantiation of $x_1, ..., x_i$ such that all constraints and relations involving $x_1, ..., x_i$ are satisfied.
- If x_{i+1} is *existential*, then $(\Gamma, X_1, ..., X_i)$ is a *partial match* if $(X_1, ..., X_i)$ is a partial matching tuple and *there is* an instantiation X_{i+1} of x_{i+1} such that $(X_1, ..., X_{i+1})$ is a partial matching tuple.
- If x_{i+1} is *universal*, then $(\Gamma, X_1, ..., X_i)$ is a *partial match* if $(X_1, ..., X_i)$ is a partial matching tuple and *for all* instantiations X_{i+1} of x_{i+1} matching the constraints on x_{i+1}, the tuple $(X_1, ..., X_{i+1})$ is a partial matching tuple.
- $(\Gamma, X_1, ..., X_n)$ is a *partial match* if $(X_1, ..., X_n)$ is a partial matching tuple.

Then, $(\Gamma, X_1, ..., X_k)$ is a *match* of Q if $(\Gamma, X_1, ..., X_k)$ is a partial match and k is maximal such that all $x_1, ..., x_k$ are existentially quantified. This includes the case $k = 0$.

It is left as an exercise to the reader to verify that the outlined algorithm is correct.

When the first variable in a query Q is existential and a match does not consist of a graph alone ($k > 0$), there might exist more than one match of Q for the same graph Γ. The given algorithm will enumerate all such matches. In the search interface however, a lazy evaluation strategy is used: For every graph, only the first match is calculated, which can speed up the calculation of the set of matching graphs considerably. Only when the user inspects a particular graph, the remaining matches for that graph are calculated.

An informal evaluation of INESS-Search against some treebank search systems (i.e., TIGERSearch, MonaSearch and Emdros) based on the TIGER treebank indicates that our system is as fast or significantly faster on most types of queries.

5.3 Gorn adressing of directed graphs

In TIGERSearch, node dominance and precedence are coded using Gorn addresses (Gorn, 1967). Each node has a Gorn address, which is an encoding of the path starting from the tree root and leading to the node. In concrete terms, a Gorn address is a sequence of integers, each one telling which child node to chose when traversing the path through the tree.

Using Gorn addressing, dominance and precedence relations are straightforward to check: node X dominates node Y if g(X) (the Gorn address of X) is a proper prefix of g(Y), and X precedes Y if g(X) is alphabetically smaller than g(Y).

This addressing scheme extends easily to directed acyclic graphs. As opposed to trees, there may be more than one path from the root to a given node in a graph. So we simply associate to each graph node the set of Gorn addresses that describe the possible paths from the root to the node. (Note that this addressing scheme assumes that the children of every node are ordered.) With these extended Gorn addresses in place, a graph node X dominates node Y if there is an address in g(X) that is a proper prefix of some address in g(Y).

Determining the Gorn addresses of the nodes in a graph is done by traversing the graph in a depth-first traversal; each step corresponds to one path to the node in focus and contributes to the extended Gorn address of that node.

When we try to extend this algorithm to arbitrary directed graphs, the problem arises that circularity would give rise to infinitely many Gorn addresses, each being a prefix of infinitely many others, since a path can wind arbitrarily often around a cycle. For all practical purposes however, given any two nodes in a cycle, we only need to be able to detect that they dominate each other along that cycle.

A query like (25) that explicitly specifies a double cycle in an f-structure would in fact fail to match that f-structure (e.g., 27), but such queries are quite unintuitive and artificial.

(25) #x >(ADJUNCT $ SUBJ ADJUNCT $ SUBJ) #x

Here is an outline of the algorithm that assigns Gorn addresses in directed graphs with cycles.

- Do a depth-first traversal of the cyclic structure;
- Assign Gorn addresses to nodes as you proceed;
- Stop and backtrack when you detect that **two** assigned Gorn addresses would be prefixes of the new Gorn address. (It is **not** sufficient to stop when one assigned Gorn address is already a prefix of some other assigned Gorn address.)

Consider example (26) and its f-structure in (27). Figure 1 illustrates the Gorn addressing for such a circular f-structure. The boxed numbers are node IDs, and the number sequences below are the calculated Gorn addresses.

(26) *Jagede hunder bjeffer.* “Chased dogs bark.”

(27)
$$\begin{bmatrix} \textsc{pred} & \text{‘bjeffe<[8:hund]>’} \\ \textsc{topic} & {}_{8}\begin{bmatrix} \textsc{pred} & \text{‘hund’} \\ \textsc{adjunct} & \left\{ \begin{bmatrix} \textsc{pred} & \text{‘jage<NULL, [8:hund]>’} \\ \textsc{subj} & [\mathit{8}] \\ \textsc{vform} & \text{pastpart} \end{bmatrix} \right\} \end{bmatrix} \\ \textsc{subj} & [\mathit{8}] \end{bmatrix}$$

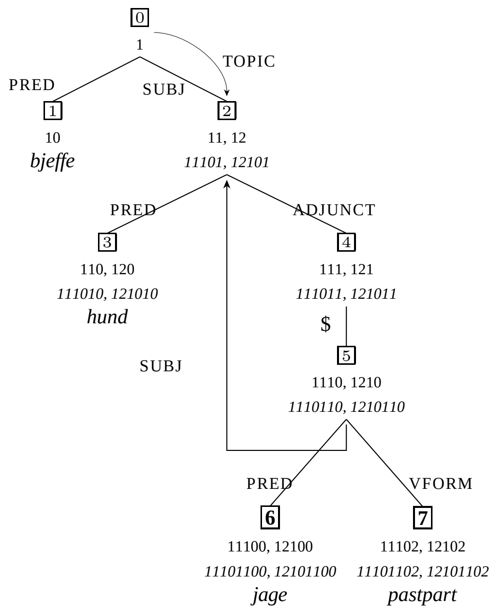


Figure 1: Gorn addressing of a circular f-structure

6 Interface and visualization

The INESS-Search tool is an integrated part of INESS, the Norwegian Infrastructure for the Exploration of Syntax and Semantics[6] and can be used to query all treebanks hosted in that infrastructure via a Web interface. In the display of the search results, matching tree/c-structure and sub-f-structures are hightlighted, and the user can choose to see one sub-match at a time, or all possible matches at once. Figure 2 illustrates the display of a match to the query (28) in the German Tiger LFG treebank.

(28) V >>(TNS-ASP TENSE) "pres"

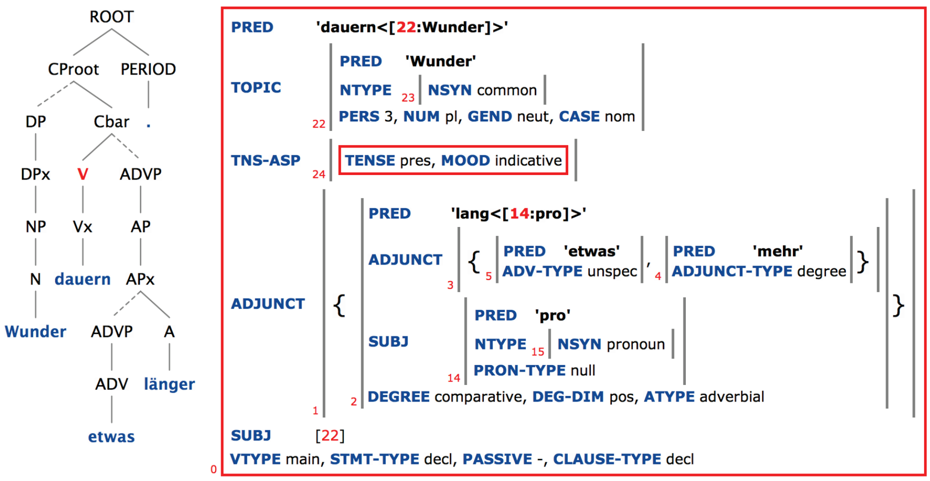


Figure 2: Visualization of a query match

7 Future plans

INESS-Search is still work in progress. Even though the basic functionality as described in this article is implemented and stable, there are many conceivable extensions that would make the tool even more useful. Below is a list of those features that will be implemented in the course of the ongoing INESS project.

Query refinement. Instead of writing a complex search expression, it is often easier to start with a simple expression and refine it by searching in the set of graphs matching the first expression. Since queries operate on single graphs in isolation, query refinement is well-defined and easy to implement. This stands in contrast to query refinement in a traditional corpus, where the scope of a query expression can span over arbitrarily many corpus positions.

[6]See http://iness.uib.no.

Search in cross-sentential annotation. Some linguistic phenomena, such as discourse structure and anaphora resolution, are not restricted to isolated sentences, since they may cross sentence boundaries. In INESS, the PROIEL treebank is an example of a treebank featuring such cross-sentential annotation. With a slight adaptation of the search algorithms and the index layout, INESS-Search will be able to handle cross-sentential search.

Search in metadata. Large treebanks often consist of several different analyzed documents, where each document comes with its own set of metadata such as title, author, publishing year, and so on. These metadata have to be searchable in combination with syntactic queries, thus enabling the user to restrict the scope of a query to a subset of the documents.

Aggregation and export of query results. For many purposes, it is not sufficient to be able to browse through the matches of a query. One should be able to aggregate the query results in tabular form in order to feed them into a statistics package or the like. The anchor points for aggregation would be the matching graphs and the matching nodes in each graph, and the table entries could be a user-selectable function of the graph and the nodes, such as for instance the node label or the value of any other node feature, or some more complicated expression that can be calculated on a match.

HPSG support. Starting with the Redwoods treebank in 2001, quite large treebanks have been compiled in the HPSG framework.[7] To our knowledge, there exists no dedicated query tool for searching in HPSG treebanks. We are planning to adapt the INESS infrastructure and the INESS-Search tool to accommodate HPSG treebanks.

8 Acknowledgements

INESS is a project cofunded by the Norwegian Research Council and the University of Bergen.

I would like to thank Victoria Rosén, Koenraad De Smedt and the reviewers for valuable comments and suggestions.

References

Bresnan, Joan: Lexical Functional Syntax. Blackwell Publishers, 2001.

Dyvik, Helge, Paul Meurer, Victoria Rosén and Koenraad De Smedt: Linguistically Motivated Parallel Parsebanks. In: Passarotti et al. (eds.): Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories, Milano, 2009, pp. 71–82.

[7] See e.g. http://www.delph-in.net.

Gorn, Saul: Explicit Definitions and Linguistic Dominoes. Systems and Computer Science, Eds. J. Hart & S. Takasu. 77-115. University of Toronto Press, Toronto Canada, 1967.

Haug, Dag Trygve Truslew and Marius Jøhndal. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Paris, 2008.

König, Esther, Lezius, Wolfgang, Voormann, Holger: Tigersearch 2.1 User's Manual. Technical report, IMS Stuttgart, 2003.

Kepser, Stephan: Finite Structure Query – A Tool for Querying Syntactically Annotated Corpora. In EACL 2003, Ann Copestake and Jan Hajič (eds.), pp. 179–186.

Lai, Catherine and Steven Bird: Querying and Updating Yreebanks: A Critical Survey and Requirements Analysis. In: Proceedings of the Australasian Language Technology Workshop, Sydney, Australia, 2004.

Lai, Catherine and Steven Bird: LPath$^+$: A First-Order Complete Language for Linguistic Tree Query. In: Proceedings of PACLIC 19, 2005.

Manber, Udi and Gene Myers: Suffix Arrays: A New Method for On-line String Searches. In: SIAM Journal on Computing, Vol. 22/5, 1993, pp. 935–948.

Marek, Torsten, Joakim Lundborg and Martin Volk: Extending the TIGER Query Language with Universal Quantification. In: KONVENS 2008: 9. Konferenz zur Verarbeitung natürlicher Sprache, Berlin, pp. 5–17.

Maryns, Hendrik and Stephan Kepser: MonaSearch – A Tool for Querying Linguistic Treebanks. In: Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT 2009), Groningen, Netherlands, 2009.

Meurer, Paul: Corpuscle – A New Corpus Management Platform for Annotated Corpora. In: Gisle Andersen (ed.): Exploring Newspaper Language, Studies in Corpus Linguistics 49, Benjamin, 2012.

Petersen, Ulrik: Evaluating Corpus Query Systems on Functionality and Speed: TIGERSearch and Emdros. In: Angelova et al. (eds): International Conference on Recent Advances in Natural Language Processing 2005, Proceedings, Borovets, Bulgaria, 21-23 September 2005, pp. 387–391.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer and Helge Dyvik: An Open Infrastructure for Advanced Treebanking. In: META-RESEARCH Workshop on Advanced Treebanking at LREC2012, İstanbul, 2012, pp. 22–29.

Rosenfeld, Victor: An Implementation of the Annis 2 Query Language, Master thesis, Humboldt University Berlin, 2010.

Volk, Martin, Joakim Lundborg and Maël Mettler: A Search Tool for Parallel Treebanks. In: Proceedings of the Linguistic Annotation Workshop, 2007, pp. 85–92.

NUMBER MARKING IN THE DALY LANGUAGES (AUSTRALIA)

Rachel Nordlinger

University of Melbourne

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu

Abstract

In this paper, presented as part of the workshop on the morphosyntax of number marking, I discuss subject and object number marking in languages of the Daly region of the Northern Territory of Australia, especially Ngan'gityemerri and Murrinh-Patha. In these languages number is frequently marked by multiple elements distributed throughout the verbal word. I argue that the interpretation of number marking needs to take into account the full morphological context in which it appears, causing difficulties for the constructed number analysis of Arka (2011) (as well as the disjunctive approach of Nordlinger (2011)).

1. Introduction

In this paper I discuss the common patterns of argument number marking across the Daly languages of northern Australia in the context of recent LFG work on constructed number systems (Sadler 2010, Arka 2011).[1] Although the basic patterns of number marking appear to lend themselves to a constructed number analysis (Arka 2011), I argue that examination of the full range of data makes such an approach problematic. Rather, the interpretation of argument number marking can only be determined within the context of the full verbal word, not by composing features from individual morphs.[2] This finding is somewhat unsurprising given the well-known nature of inflectional paradigms, in which 'words as wholes' are arranged and contrasted according to grammatical categories (Matthews 1991:187). The data from the Daly languages thus suggests that the constructed number approach may be less useful when the complex number marking interactions fall within an inflectional paradigm, as opposed to capturing the interaction of number marking across syntactic categories (as in Hopi (Corbett 2000, Sadler 2010)).

---

[1] I would like to thank Mary Dalrymple and I Wayan Arka for inviting me to present in the workshop on Number and Plurals at the LFG2012 conference, and for interesting and insightful discussion on the issues. I would also like to thank the people of Wadeye for their hospitality and support for my research, especially Carmelita Perdjert and family. My fieldwork on Murrinh-Patha has been funded by the Australian Research Council (DP0343354, DP0984419, DP110100961) and the University of Melbourne.

[2] Here and throughout the paper, I use 'morph' as a neutral term for 'piece of grammatical form', rather than morpheme, which assumes a pairing of form and function. This is in order to put aside for present purposes the theoretical debate concerning the status of morphemes in morphological theory (see, for example, Spencer (2004) for discussion of the central issues).

2. Daly languages

The Daly languages are an areal grouping of Australian languages that are traditionally spoken in the Daly River region, south-west of Darwin in Australia's Northern Territory. These languages include Western Daly languages such as Marrithiyel (Green 1989) and Marri Ngarr (Preston 2012), and the Southern Daly languages Murrinh-Patha (Blythe 2009, Nordlinger 2010, 2011, Seiss & Nordlinger 2010) and Ngan'gityemerri (Reid 1990), among many others. Although these internal subgroups have been well-established (Green 2003), the Daly languages as a whole have not been shown to form a single family, but do share a number of areal similarities, including templatic verbal structures characterized by complex predicates (McGregor 2002), multiple exponence and discontinuous dependencies (Nordlinger 2010). In this paper I will focus primarily on data from Ngan'gityemerri and Murrinh-Patha.

2. Typological overview

All Daly languages have a four-way number marking system which distinguishes singular, dual, plural and trial/paucal (depending on the language). There is no number marking on nouns (in fact, very little nominal inflectional morphology at all), so number is encoded only in (optional) pronouns and on verbal agreement morphology. In this paper, I will be focusing on verbal agreement morphology, but I provide the table of Murrinh-Patha pronominals below simply to show the full range of categories distinguished (verbal agreement morphology distinguishes these same categories).

Table 1 Murrinh-Patha pronouns

		sing	**dual**		**paucal**		**Plural**
			sib[3]	**nsib**	**sib**	**nsib**	
1	**exc**	ngay	nganku	ngankunintha (m) ngankungintha (f)	nganki	ngankuneme (m) ngankungime (f)	nganki
	inc	–	neki		nganki	nekineme (m) nekingime (f)	neki
2		nhinhi	nanku	nankunintha (m) nankungintha (f)	nanki	nankuneme (m) nankungime (f)	nanki
3		nukunu(m) nigunu (f)	piguna	penintha (m) peningintha (f)	pigunu	peneme (m) peningime (f)	pigunu

[3] Murrinh-Patha is unusual in grammatically encoding a distinction between groups of siblings ('sib') and groups who are not siblings ('nsib').

Daly languages are characterized by complex verbs consisting of discontinuous complex predicates, and multiple exponence of tense/aspect/mood and number marking throughout the verbal word. The verbal template for Murrinh-Patha is reflective of this general structure, and is provided in Table 2. The forms given in italics (slots 1 and 5) are the two parts of the complex predicate – referred to here as the classifier stem (CS) and the lexical stem (LEXS). These are (for the most part) bound morphemes that together form the verbal predicate. Of particular relevance to this paper are the number marking elements, which are given in bold. Subject number marking occurs in slot 1 (via the classifier stem), as well as slot 2 (when there is no object marker) and possibly slot 8 (when there is an object marker). Object number marking appears in slot 2 (encoded via the object bound pronoun) and slot 8. For detailed discussion of the Murrinh-Patha verbal template and the details of its morphological structure see Nordlinger (2010).

Table 2 Murrinh-Patha verbal template

1	2	3	4	5	6	7	8	9
CS.SUBJ.TAM	**SUBJ.NUM/ OBJ/**	RR	IBP	*LEXS*	TAM	ADV	**NUM**	ADV

As will be illustrated below, slot 8 can encode either subject or object number and the same elements are involved in each case. The interpretation of the number marker in slot 8 depends on the broader morphological context, namely whether or not it is semantically compatible with other elements in slots 1 and/or 2.

3. Number marking in the verb

3.1 Ngan'gityemerri

Number of both subject and object is marked in the verb, using a combination of multiple markers distributed throughout the verbal word. In Ngan'gityemerri, Reid (1990:114) states that "[i]ncreasingly marked number categories are derived by taking simpler number categories as a base and adding additional morphological marking in new verbal slots". Consider the examples in (1), which illustrate the basic system (Reid 1990: 118):[4]

[4] In the examples in (1) the first three elements in the verbal word (e.g. *nge-Ø-beny* in (1a)) constitute the classifier stem, and therefore correspond with slot 1 in Table 2. In the rest of

(1a) *ngayi* *nge-**Ø-**beny-da*

1sg 1S-**sg-**Bash.Perf-hit

I hit it. **= 1 only**

(1b) *ngagurr* *nge-**rr**-beny-da*

1pl 1S-**pl.exc**-Bash.Perf-hit

We (pl.exc) hit it. **= 4 or more**

(1c) *ngarrgu* *nge-**rr**-beny-**gu**-da*

1du 1S-**pl.exc**-Bash.Perf-**du**-hit

We (du.exc) hit it. **= 2 only**

(1d) *ngarrgu-nime* *nge-**rr**-beny-**gu**-da-**nime***

1du-tr 1S-**pl.exc**-Bash.Perf-**du**-hit-**tr**

We (tr.exc) hit it. **= 3 only**

As these examples demonstrate, the classifier stem in the Ngang'ityemerri verb shows a basic contrast between -Ø singular (as in 1a) and *-rr-* non-singular. In the absence of any further number markers in the verb, the nonsingular marker is interpreted as plural which, in this case, means 4 or more as in (1b). Dual is formed by adding the dual marker to the non-singular verb form, as in (1c), and then the trial category is built on the dual, with the addition of *–nime* 'tr(ial)', as in (1d).

The dual marker *–gu*, however, appears in the same slot in the verbal template as the object agreement markers, which take priority (Reid 1990: 128, 135). To encode a dual subject in the presence of an object pronoun, a special 'dual-subject' form of the indirect or direct object marker is used instead. The examples in (2) and (3) illustrate this with an intransitive and transitive verb respectively. In the (a) examples we see a regular dual subject, containing the dual subject marker *-gu,* and no object marker (since the object in (3a) is third person singular, and therefore unmarked). In the (b) examples we see the presence of an object marker in the second verbal slot, leading to a plural interpretation for the subject (since the dual marker is not present). Finally, in the (c) examples, the special 'dual-subject' object

---

the examples in this paper they are written as a single morph, but are separated here to make clearer their internal structure.

marker is used to express dual number for the subject, as well as the singular object information.

(2a) *werrmen'geny-**gu***
3plArrive.Perf-du
They (du) have arrived.

(2b) *werrmen'geny-**ngiti***
3plArrive.Perf-1sgIO
They (pl) came to me.

(2c) *werrmen'geny-**ngeterr***
3plArrive.Perf-duS/1sgIO
They (du) came to me.

(3a) *warriny-**gu**-pawal*
3plPoke.Perf-du-spear
They (du) speared him.

(3b) *warriny-**nyi**-pawal*
3plPoke.Perf-2sgO-spear
They (pl) speared you.

(3c) *warriny-**nyerr**-pawal*
3plPoke.Perf-duS/2sgO-spear
They (du) speared you.

Note, however, that these special ‘dual-subject’ object bound pronouns are only available for singular objects. When the object is non-singular, dual subjects are not specified (Reid 1990: 129), and thus the distinction between plural and dual subject number is lost. Example (4) is therefore ambiguous between an interpretation with a plural subject and one with a dual subject, as the translation indicates.[5]

[5] A trial interpretation is not possible however, as discussed below.

(4) ***warriny-ngirr-tyerr-pu***, *nyinyi tyagani derrigidi-yerim*
3plPoke.Perf-1pl.excO-mouth-ask 2sg what want-2sgHands.Pres
They (du or pl) (airline hostesses) asked each of us (pl. exc), "What would you like (to drink)?"

As shown in (1d), the trial subject form is usually built on the dual form, including the dual marker *-gu*. In the presence of an object pronoun, in which case the dual marker is absent as shown above, the trial marker combines just with the non-singular form – with no dual marker required (Reid 1990:224):

(5) *alayi warrakma kinyi* ***werrme*** *-ngi-pul-**nime**-tye*
mother three this 3**pl**Hands.PImp-1sgO-clean-**tr**-Past
These three mothers of mine used to wash me.

Interestingly, the special 'dual subject' object pronouns, as in (3d), are *not* possible here, even though the trial is generally built on a dual subject form of the verb in other constructions. So, here the plural form of the classifier can function as trial in the presence of a *singular* object as well, but only if the verb later includes the trial number marker *-nime*.

We can therefore summarize the Ngang'ityemerri number marking facts so far as follows:

SING: (i) singular classifier (+ regular object marker) (1a)

DUAL: (ii) plural classifier + dual marker (1c) OR
(iii) plural classifier + 'dual' singular object marker (2c) OR
(iv) plural classifier + non-singular object marker (4)

TRIAL: (v) plural classifier + dual marker + trial marker (1d) OR
(vi) plural classifier + regular object marker + trial marker (5)

PLURAL (vii) plural classifier (+regular object marker) (1b)

These facts appear to lend themselves nicely to a constructed number analysis (e.g. Corbett 2000, Sadler 2010, Arka 2011), in which the different number marking elements contribute different features to the overall number category. Assuming the feature combinations for the four different number categories shown in (6), the number marking facts described above can be accounted for as shown below, assuming that [AUG –] is applied by default, in the case that the AUG feature is underspecified by the morphology.

(6) **Constructed number analysis:**

SINGULAR : $\begin{bmatrix} \text{SG} & + \\ \text{DU} & - \\ \text{AUG} & - \end{bmatrix}$

DUAL: $\begin{bmatrix} \text{SG} & - \\ \text{DU} & + \\ \text{AUG} & - \end{bmatrix}$

TRIAL: $\begin{bmatrix} \text{SG} & - \\ \text{DU} & + \\ \text{AUG} & + \end{bmatrix}$

PLURAL: $\begin{bmatrix} \text{SG} & - \\ \text{DU} & - \\ \text{AUG} & - \end{bmatrix}$

On this analysis, we can assume that the singular classifier stems (as in 1a) contribute [+ SG, –DU], and the non-singular classifier stems (as in 1b-d) contribute [–SG]. The dual marker contributes [+DU], and the trial marker contributes [+AUG]. The dual example (1c) follows straightforwardly, as shown in (7), and the trial example (1d), as shown in (8).[6]

(7) *ngarrgu* *nge-**rr**-beny-**gu**-da*
1du 1S-**pl.exc**-Bash.Perf-**du**-hit
[–SG] [+DU]
'We (du.exc) hit it.'

DUAL: $\begin{bmatrix} \text{SG} & - \\ \text{DU} & + \\ \textit{AUG} & - \end{bmatrix}$

[6] The [AUG –] feature is given in italics in (7) to show it has been applied by default.

(8) *ngarrgu-nime* *nge-**rr**-beny-**gu**-da-**nime***
1du-tr 1S-**pl.exc**-Bash.Perf-**du**-hit-**tr**
[-SG] **[+DU]** **[+AUG]**
'We (tr.exc) hit it.'

TRIAL: $\begin{bmatrix} \mathrm{SG} & - \\ \mathrm{DU} & + \\ \mathrm{AUG} & + \end{bmatrix}$

The fact that the trial marker cannot co-occur with the special 'dual-subject' object markers shown in (2) and (3) is captured by assuming that these object markers also contribute the [AUG –] feature to the subject's number category, thereby making them incompatible with the trial marker (which carries [AUG +]). The analysis of (3c) is given in (9):

(9) *wa**rr**iny-**nyerr**-pawal*
3**pl**Poke.Perf-**duS**/2sgO-spear
[–SG] **[+DU, –AUG]**
They (du) speared you (sg) (cf. 3b)

DUAL: $\begin{bmatrix} \mathrm{SG} & - \\ \mathrm{DU} & + \\ \mathrm{AUG} & - \end{bmatrix}$

Examples such as (4) and (5), however, present some difficulties. In (4) we saw that the non-singular classifier can alone mark a dual subject (ambiguously with a plural subject) *just in the case* that there is a non-singular object marker in the verb (blocking the appearance of the dual marker). The absence of the dual marker will leave the DU feature underspecified, as shown in (10), which captures the ambiguity nicely.

(10) *wa**rr**iny-ngirr-tyerr-pu,* *nyinyi* *tyagani* *derrigidi-yerim*
3**pl**Poke.Perf-1pl.excO-mouth-ask 2sg what want-2sgHands.Pres
[–SG]
'They (du or pl) (airline hostesses) asked each of us (pl), 'What would you like (to drink)?''

PLURAL/DUAL: $\begin{bmatrix} \mathrm{SG} & - \\ \mathrm{DU} & \\ \mathit{AUG} & - \end{bmatrix}$

However, the difficulty is that we need to ensure that the combination of [SG –, AUG –] (with the DU underspecified) is *only* possible in the presence of a non-singular object marker in the verb. Example (1b), for example, can never have a dual interpretation, despite the fact that it has the same feature array as (10), as shown in (11):

(11) *ngagurr* *nge-**rr**-beny-da*
1pl 1S-**pl.exc**-Bash.Perf-hit
[–SG]
'We (pl.exc) hit it.'
NOT 'We (du.exc) hit it'

*PLURAL/DUAL: $\begin{bmatrix} \text{SG} & - \\ \text{DU} & \\ \textit{AUG} & - \end{bmatrix}$

A similar issue arises with the trial category in the presence of an object marker, as in (5). In this case the trial marker contributes [AUG +], but there is nothing to contribute the dual feature, leaving it underspecified:

(12) *alayi* *warrakma* *kinyi* *we**rr**me -ngi-pul-**nime**-tye*
mother three this 3**pl**Hands.PImp-1sgO-clean-**tr**-Past
[–SG] **[+AUG]**
'These three mothers of mine used to wash me.'

TRIAL: $\begin{bmatrix} \text{SG} & - \\ \text{DU} & \\ \text{AUG} & + \end{bmatrix}$

Once again, we are left with the difficulty of how to ensure that an underspecified dual feature is *only* possible in the trial category when there is an object marker in the verb, but that otherwise the dual marker is required to contribute [DU +], as in (1d). Thus, it appears that the subject number feature array for a Ngan'gityemerri verb cannot be constructed compositionally, but can only be determined once the morphological structure of the whole verb is taken into consideration, such as whether or not there is an object marker present.

This general point is further reinforced once we examine the first person inclusive category, which behaves differently with respect to subject number marking than was shown above. In the first person inclusive subject forms, the addition of *–nime* marks plural number (i.e. 3 or more), and not trial (Reid 1990:114):

(13a) *nayin ngi-**mbi**-bem*
1du.inc 1S-**du.inc**-Lie
'We (du.inc) are lying down.' **(= 2 only)**

(13b) *nayin-nime ngi-**mbi**-bem-**nime***
1du-pl 1S-**du.inc**-Lie-**pl**
'We (pl.inc) are lying down.' (**= 3 or more)**

Thus, when the subject is first person inclusive, the *-nime* number marker encodes plural, whereas with other subjects the *–nime* number marker *only* encodes trial number. Thus, in order to correctly interpret a verb containing the *–nime* number marker, it is necessary to know the full feature specification of the verb – i.e. whether the subject is first person inclusive, or not.

3.2 Murrinh-Patha

Number marking in the Murrinh-Patha verb raises similar issues to those presented for Ngan'gityemerri, but also adds a number of additional complexities which make a constructed number analysis (and indeed, any morpheme-based analysis) difficult to maintain.

The Murrinh-Patha verb is similar in structure to Ngan'gitymerri, but with a few key differences. Firstly, the number category corresponding to the trial in Ngan'gityemerri marks paucal number (approximately 3-10) in Murrinh-Patha. Furthermore, a grammatical distinction between sibling and non-sibling groups in the dual and paucal categories has led to skewing in the paradigm, as we shall see below.

The basic subject number facts are given in (14) (see also Nordlinger 2011):

(14a) *bamkardu*
bam-ngkardu
3**sg**S.SEE.nFut-see
'He/she saw him/her.'

(14b) ***bam-ngintha***-*ngkardu*

3**sg**S.SEE.nFut-**du.f**-see

'They two (female non-siblings) saw him/her.'

(14c) ***pubamka***-*ngkardu*

3**du**S.SEE.nFut-see

'They two (siblings) saw him/her.'

(14d) ***pubamka***-*ngkardu*-***ngime***

3**du**S.SEE.nFut-see-**pauc.f**

'They paucal (female non-siblings) saw him/her.'

(14e) *pubamkardu*

pubam-ngkardu

3**pl**S.SEE.nFut-see

'They (paucal siblings / plural) saw him/her.'

In Nordlinger (2011) (see also Dalrymple, this volume), I provided an analysis of these number marking facts that made use of disjunctive features and constraining equations to capture the patterns. The singular classifier stems, as in (14a), for example, were analysed as carrying the disjunctive features in (15):

(15) $\{(\uparrow \text{ SUBJ NUM}) = \text{SG} \vee (\uparrow \text{ SUBJ NUM}) =_c \text{DU}\}$

Thus, the singular classifier either provides the value SG for the number of the subject (as in (14a), or requires the value DU to be provided by some other element in the construction (as in (14b). Similarly the dual classifier stem was analysed as either providing a dual number value (as in 14c) or requiring the paucal number feature (as in 14d).

Arka (2011:22) showed that the Murrinh-Patha facts could also be viewed in terms of a constructed number approach, and provides the following analysis for the subject marking facts:[7]

[7] This table has been taken directly from Arka (2011:22).

Figure 8: the number system in Murrinh-Patha

	CONSTRUCTED NUMBER CATEGORIES	NUMBER EXPONENTS	
		CLASSIFIER STEMS	EXTRA NUMBER FORMATIVES
(i)	*singular*	*singular* [+SG]	∅ [–AUG]
(ii)	*augmented dual non-sibling*	*singular* [+SG]	*ngitha* (F)/ *nitha* (M) [+AUG]
(iii)	*dual sibling*	*dual* [+DU]	∅ [–AUG]
(iv)	*augmentedpaucal non-sibling*	*dual* [+DU]	*ngime* (F)/ *neme* (M) [+AUG]
(v)	*plural*or*paucal sibling*	*plural* [+PL]	∅ [–AUG]

This analysis has the advantage of neatly accounting for the skewing we see in the dual and paucal categories – each of these categories can be constructed in two ways. The dual category can be expressed with a dual classifier and no augment, in which case is it interpreted as dual sibling (14c); or it can be expressed as an augmented singular (combining a singular classifier and the dual number marker *-ngintha*/*-nintha*) in which case it is interpreted as dual non-sibling (as in 14b). Likewise for the paucal category, which is constructed either as a plural (paucal sibling, 14e) or as an augmented dual (paucal non-sibling, 14d).

There are a number of additional wrinkles that need to be addressed in order for this analysis to fully capture the facts in (14): for example, we need to account for the fact that the dual classifier encodes 'sibling' when it constructs dual number, but 'non-sibling' when it constructs paucal number. We also need to constrain *-ngintha*/*-nintha* to only occur with singular classifiers, and *-ngime*/*-neme* to only occur with dual classifiers (Arka 2011:16). We could do this by associating the augments with the following feature arrays, for example (represented informally):

-ngintha: [+AUG, +NSIB, $+SG_c$]
-ngime: : [+AUG, +NSIB, $+DU_c$]

However, the real difficulty comes when we consider object marking. The dual non-sibling marker *-ngintha* is used to express dual object number also. Whereas it co-occurs with a singular classifier to mark a dual non-sibling subject as we saw in (14b), when encoding *object* number it must combine with a *dual* object marker to mark a dual object as shown in (16). Example (17) shows that if a singular object marker is used instead, then the dual marker cannot be interpreted as referring to the object; in this case it is interpreted as referring to the subject.

(16) *ma-**nanku**-rdarri-purl-nu-**ngintha***
1sgS.HANDS.Fut-**2duO**-back-wash-Fut-**du.f**
'I will wash your (dual non-sibling) backs.'

(17) ***ma**-nhi-rdarri-purl-nu-**ngintha***
1**sg**S.HANDS.Fut-2sgO-back-wash-Fut-**du.f**
'We (du.excl.nsib) will wash your (sg) back.'
NOT 'I will wash your (dual non-sibling) backs.'

Thus, whatever constraint we use to restrict the occurrence of *-ngintha* to *singular* classifier forms when expressing subject number will not adequately capture its behaviour with objects, where it must co-occur with *dual* forms only.

Furthermore, even if we resolve this particular issue, it remains problematic to treat *-ngintha* as contributing an [AUG +] feature (as in the constructed number analysis presented above) when we consider its behaviour with object marking, since in this case it co-occurs with a *dual* form, so we would expect the addition of the [AUG +] feature to result in an augmented dual interpretation, i.e. paucal. But in fact, this is not what we find – paucal objects, like subjects, require the paucal marker *–ngime/-neme* to combine with the dual object marker, as in (18):

(18) *ma-**nanku**-rdarri-purl-nu-**ngime***
1sgS.HANDS.Fut-**2duO**-back-wash-Fut-**pauc.f**
'I will wash your (paucal non-sibling) backs.'

In fact, *-ngintha* consistently marks dual number irrespective of which other verbal elements it is combining with, and *–ngime/-neme* consistently marks paucal; thus weakening the motivation for a constructed number analysis in the first place.

Furthermore, the dual marker in an example like (16) could refer to either the subject *or* the object – so this example is actually ambiguous between the following interpretations:

(19) ***ma-nanku***-*rdarri-purl-nu*-***ngintha***

1**sg**S.HANDS.Fut-**2duO**-back-wash-Fut-**du.f**

(i) 'I will wash your (dual non-sibling) backs.' [-*ngintha* refers to object]

(ii) 'We (du.exc.nsib) will wash your (du.sib) backs' [-*ngintha* refers to subject]

(iii) 'We (du.exc.nsib) will wash your (du.nsib) backs' [-*ngintha* refers to object and dual number for subject is unspecified]

In the latter case, we have a dual interpretation for subject, without any dual subject number marker (since –*ngintha* in this case is marking dual object number). This is problematic both for the constructed number analysis, and for the Nordlinger (2011) analysis, both of which rely on an overt morpheme to contribute the dual number feature to the subject's f-structure. Nordlinger's (2011) analysis would predict that (19) could not have the reading in (iii), since there is nothing in the structure to provide the dual number feature, so the singular classifier stem would have to be interpreted as contributing singular subject number. The constructed number analysis, would likewise fail since there would be nothing to provide the [+AUG] feature to the subject, leading to the application of the [AUG –] default:[8]

(19') ***ma***-*nanku-rdarri-purl-nu-ngintha*

1**sg**S.HANDS(8).Fut-2duO-back-wash-Fut-du.f

[+SG]

SUBJECT: [+SG, –*AUG*] (should be **singular)**

Once again, we are confronted with the fact that the number features for subject and object in the Daly verb can only be properly interpreted in the context of the full morphological word. A singular classifier stem as in (19) can encode dual subject number, but *only* when the possible number marking slots are filled with object information. If there were no object marker in (19), for example, or if the -*ngintha* number marker were incompatible with the object marker (as in (17)), the singular classifier stem could not be interpreted as expressing *dual* subject number.

[8] Another possibility is that we assume the [AUG] feature remains underspecified here, which captures the ambiguity, but then we are left with the problem of how to ensure that this is only possible in the presence of an object number. This is the same issue that was discussed with regards to the examples (10)-(12) above.

3.3 Summary

We have seen that a constructed number approach to number marking in the Ngan'gityemerri and Murrinh-Patha verbs is initially appealing for capturing the basic facts, but runs into difficulties when we examine the full range of number marking facts. Issues identified in the above discussion include:

(i) The distinction between dual and plural subjects is lost in the presence of an object marker (e.g. (4));

(ii) Trial/paucal subjects *must* co-occur with dual marking, but *only* when there is no object marker present (e.g. (1d) vs (5));

(iii) Trial marking (in Ngan'gityemerri) is interpreted differently depending on whether the subject is first person inclusive, or not (e.g. 13);

(iv) The Murrinh-Patha dual number marker is constrained to occur with a singular classifier form to mark dual non-sibling *subjects* but a dual form to mark dual non-sibling *objects* (e.g. (14b) vs (16)).

While none of these issues appears particularly devastating for an analysis in and of itself, together they amount to accumulated evidence that constraints on number marking patterns in the Daly verbs are context-dependent, in that the contribution of the various number marking elements varies according to the morphological context in which the number marker appears. Thus, in order to interpret the subject or object number values for any given verb, one needs to consider the complete morphological structure of the verb in order to interpret the various number markers within it. This is extremely problematic for the morpheme-based accounts of Nordlinger (2011) and Arka (2011), which rely on individual morphemic elements to contribute number features in a consistent and independent manner.

4. Conclusion

The verbal number marking systems in Daly languages such as Ngan'gityemerri and Murrinh-Patha at first appear to lend themselves to a constructed number analysis (e.g. Sadler 2010, Arka 2011), with different parts of the verbal word contributing different (combinations of) number features and co-constructing the overall number category. However, in the above discussion we have seen that this approach breaks down when we examine the full range of empirical facts, including the interaction of number marking elements with other aspects of the morphological word. In fact, the data shows that in many cases the subject or object number value can only be determined within the context of the whole verbal word; and cannot be deduced from simple composition of the contributions of different morphemic elements, as the constructed number approach assumes.

The difficulties arise from the fact that individual morphs do not contribute fixed, invariable feature values, but can be interpreted in different ways depending on whether or not other morphs are present in the verbal word. The singular classifier in Murrinh-Patha, for example, can alone denote a dual non-sibling subject, but *only* when there is an object marker and associated object number marker present in the verbal word, as in (20), since in this case there is no available position in the verbal word for the subject dual number marker to appear:

(20) *ma-**nanku**-rdarri-purl-nu-**ngime***
1sgS.HANDS.Fut-**2duO**-back-wash-Fut-**pauc.f**
'I will wash your (pauc.nsib) backs.' **OR**
'We (du.exc.nsib) will wash your (pauc.nsib) backs.'

In other cases, the singular classifier can only express a singular subject (unless the dual number marker is present):

(21) *ma-**nanku**-rdarri-purl-nu*
1sgS.HANDS.Fut-**2duO**-back-wash-Fut
'I will wash your (du.sib) backs.'
NOT 'We (du.exc.nsib) will wash your (du.sib) backs.'

These empirical facts are not particularly surprising, given that they form part of a (very complex) inflectional verbal paradigm. It is well-known that inflectional paradigms often challenge incremental, morpheme-based analyses, as discussed in the extensive body of work on word-and-paradigm approaches (e.g. Matthews 1972, Anderson 1992, Aronoff 1994, Stump 2001, Spencer 2004, among many others). It may be that the constructed number approach is more insightful in cases, such as Hopi, where the number categories are constructed across syntactic categories, i.e. where the interaction arises through agreement in the syntax (e.g. Sadler 2010). In the Daly languages, on the other hand, the interaction of number morphology is internal to the verb's inflectional paradigm, and is therefore subject to the properties characteristic of paradigmatic systems, in which whole words are interpreted through their place in the paradigm and their opposition to other related word-forms, rather than as incrementally-composed bundles of features.

References

Anderson, Stephen. 1992. *A-morphous morphology*. Cambridge: CUP.

Aronoff, Mark. 1994. *Morphology by itself*. Cambridge: MIT Press.

Arka, I Wayan. 2011. Constructive number systems in Marori and beyond. In Miriam Butt & Tracy H King (eds) *Proceedings of the LFG11 conference*. Stanford: CSLI Publications, 5-25.

Blythe, Joseph. 2009. *Doing referring in Murriny Patha conversation*. PhD thesis, University of Sydney.

Corbett, Greville G. 2000. *Number*. Cambridge: CUP.

Green, Ian. 1989. *Marrithiyel: a language of the Daly River region of Australia's Northern Territory*. PhD thesis, Australian National University, Canberra.

Green, Ian. 2003. The genetic status of Murrinh-patha. In Nicholas Evans (ed.) *The non-Pama-Nyungan languages of northern Australia*. Canberra: PL, 280-310.

Matthews, P. H. 1972. *Inflectional morphology*. Cambridge: CUP.

Matthews, P. H. 1991. *Morphology* [Second Edition]. Cambridge: CUP.

McGregor, William. 2002. *Verb classification in Australian languages*. Berlin: Mouton de Gruyter.

Nordlinger, Rachel & Patrick Caudal. 2012. The tense, aspect and modality system in Murrinh-Patha. *Australian Journal of Linguistics* 34(1): 73-113.

Nordlinger, Rachel. 2010. Verbal morphology in Murrinh-Patha: evidence for templates. *Morphology* 20(2): 321-341.

Nordlinger, Rachel. 2011. LFG and language documentation. *Plenary paper presented at LFG11*, Hong Kong, July 2011.

Preston, Asher. 2012. *Marri Ngarr morphosyntax and the expression of reflexivity and reciprocity*. Unpublished honours thesis, University of Melbourne.

Reid, Nicholas. 1990. *Ngan'gityemerri: a language of the Daly River region, Northern Territory of Australia*. PhD thesis, Australian National University, Canberra.

Sadler, Louisa. 2010. Indeterminacy, complex features and underspecification. *Morphology* 21(2): 379-417.

Seiss Melanie & Rachel Nordlinger. 2010. Applicativizing complex predicates: A case study from Murrinh-Patha. In Miriam Butt & Tracy H King (eds) *Proceedings of the LFG10 conference*. Stanford: CSLI Publications, 416-436.

Spencer, Andrew. 2004. Morphology – an overview of central concepts. In Louisa Sadler & Andrew Spencer (eds) *Projecting Morphology*. Stanford: CSLI Publications, 67-109.

Stump, Gregory T. 2001. *Paradigm function morphology: A theory of paradigm structure*. Cambridge: CUP.

NOMINAL NUMBER IN MESO-MELANESIAN

Bill Palmer

University of Newcastle

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper, presented in the workshop on number marking, presents details of nominal and pronominal number marking in the Meso-Melanesian group of Austronesian languages. Languages of this group display a range of morphosyntactic and morphosemantic phenomena relating to number that require accounting for by any theory of grammar. These include hierarchies of number categories; the interaction of hierarchies of animacy with number; the role of number markers as syntactic heads; inversion in number marking; and patterns of indexing target for number in possessive constructions. This paper does not attempt to account for these phenomena from an LFG perspective, but presents details of the phenomena requiring accounting for.

1. Introduction

Meso-Melanesian (henceforth MM) is a second-order subgroup of the large Oceanic branch of Austronesian. Oceanic is regarded as having six first-order subgroups: Yapese; Admiralties; St Matthias; Temotu; Central Eastern Oceanic; and Western Oceanic. MM is a subgroup within Western Oceanic, the other Western Oceanic subgroups being North New Guinea and Papuan Tip, which are now thought to belong to a single sister group to MM. The 69 languages within MM are spoken in Island Melanesia to the east of mainland New Guinea: in New Britain, New Ireland, Bougainville, and the western Solomon Islands. Languages of the group are highly diverse in a range of lexical, phonological, morphological and syntactic ways, and in many cases are quite divergent from typical Oceanic structures. This diversity is assumed to have resulted from long periods of bilingualism with neighbouring Papuan languages.

This paper presents details of number marking in nominal constructions across MM, including number in pronominal systems, and number marking with nominal heads. In discussing pronominal number categories it discusses number in verb agreement. Verbal number phenomena such as pluractionality are, however, outside its scope and are not discussed.

2. Pronominal number

2.1 Independent pronouns

MM languages typically recognize more number categories than singular and plural in independent pronouns. However, a few do distinguish only those categories. As is typical for MM, Bannoni (Bannoni-Piva, Lincoln 1976; Lynch & Ross 2002) distinguishes four person categories, with the standard first person exclusivity distinction in non-singular. However, unlike most of its near relatives, Bannoni independent pronouns distinguish only singular and plural, as in (1).

(1) Bannoni:

	1EXC	1INC	2	3
SG	*na*		*no*	*nna*
PL	*ɣamam*	*ɣata*	*ɣamu*	*nari*

More specific number categories are expressed by means of periphrasis in a possessive construction in which the enumerated entity is expressed as a possessor adjunct modifying a possessum number, as in (2a). This is not a specifically pronominal construction, but one of the language's strategies for enumerating nominals, as (2b) shows.

(2) a. *ɣata* *ɣe-ra* *toom*
weINC POSS-1INCPL.PSSR two
'we two' (lit. 'our (inc.) two') (Bannoni)

b. *bekeu* *ɣe-ri* *toom*
dog POSS-3PL.PSSR two
'two dogs' (lit. 'the dog's two') (Bannoni)

The syntax of the construction in (2) involves the numeral as head, with the possessor pronoun or noun as an adjunct. As is standard in NWS languages, the adjunct expressing the possessor may be omitted, as in (3).

(3) *ɣe-ri* *ɣinima*
POSS-1INC.PL.PSSR five
'the five of them' (lit. 'their five') (Bannoni)

However, in Bannoni this possessive construction is unusual in that in non-enumerating phrases, it only occurs with pronominal possessors expressed only by agreement on the possessive particle *ɣe-*, although the interrogative pronoun may occur as the possessor, as in (4). A separate possessive construction is required if the possessor is an NP. NP possessors as in (2b) are only permitted in this construction if the head is a numeral and the function of the phrase is to enumerate the possessor.

(4) *hee* *ɣe-na* *moono*
who POSS-3SG.PSSR woman
'whose wife?' (Bannoni)

In MM, more number distinctions than singular and plural are typically expressed in independent pronouns. Several distinguish dual in addition to plural, for example Babatana (Choiseul, Money 2002), as in (5):

(5) Babatana:

	1EXC	1INC	2	3
SG	*ra*		*re*	*gii*
DU	*raru*	*zituru*	*raburu*	*ziru*
PL	*rami*	*zita*	*ramu*	*zira*

Despite the regularity of the final syllable in the dual forms, in Babatana the dual pronouns are not synchronically morphologically complex and are non-transparent – the standard Babatana numeral 'two' is *kere*.

Many MM languages distinguish a fourth number category. This may be a trial as in Kubokota (New Georgia, Chambers 2009):

(6) Kubokota:

	1EXC	1INC	2	3
SG	*ara*		*ao*	*aza*
DU	*ɣami-kori*	*ɣita-kori*	*ɣamu-kori*	*ari-kori*
TR	*ɣami-kue*	*ɣita-kue*	*ɣamu-kue*	*ari-kue*
PL	*ɣami*	*ɣita*	*ɣamu*	*ria*

In Kubokota the dual and trial pronouns are morphologically complex and are semantically transparent – the standard Kubokota numerals are *kori* 'two' and *kue* 'three', and the dual and trial pronouns are transparently constructed on the plural forms as their base. Some other languages with a trial as well as dual have non-transparent forms in both number catgeories. In Vinitiri (Patpatar-Tolai, New Ireland, Van Der Mark 2007) the numerals are *uruə* 'two', *utulu* 'three', with the dual and trial pronouns not synchronically morphologically composed:

(7) Vinitiri:

	1EXC	1INC	2	3
SG	*iau*		*iəβəu*	*iə*
DU	*iamiru*	*iadori*	*iamuru*	*idiru*
TR	*iəmitalu*	*iədətalu*	*iəmutalu*	*iditalu*
PL	*iəməmami*	*iadə*	*iamui*	*idi*

Several MM languages with four number categories in their independent pronouns have a paucal rather than trial category, as Siar (Patpatar-Tolai, New Ireland, Frowein 2011) illustrates in (8). The Siar dual and paucal forms are again not morphologically transparent. As with the Vinitiri dual and trial they represent an irregular diachronic derivation from numerals. In the Siar case the dual is irregularly derived from the numeral *ru* 'two', and, predictably, the paucal is derived from *tol* 'three'. Siar is interesting in that the plural forms are similarly derived diachronically from *at*, the numeral for 'four'.

(8) Siar:

	1EXC	1INC	2	3
SG	*ya(u)*		*u*	*i*
DU	*mara(u)*	*dara(u)*	*amra(u)*	*dira(u)*
PC	*mato(l)*	*dato(l)*	*amto(l)*	*diat*
PL	*mèt*	*dat*	*amat*	*dit*

Sematically the Siar paucal is representative of MM paucals in which there is no specific upper bound to the number of items that may be included. The Siar paucal may refer to three or more items, up to several dozen depending on context (see Corbett 2000:22).

The additional number category quadral is attested in MM. Sursurunga (Patpatar-Tolai, New Ireland, Hutchisson 1975), for example, displays a quadral category:

(9) Sursurunga:

	1EXC	1INC	2	3
SG	*yau*		*u*	*a*
DU	*giur*	*gitar*	*gaur*	*diar*
TR	*gimtul*	*gitul*	*gamtul*	*ditul*
QD	*gimat*	*gitat*	*gamat*	*diat*
PL	*gim*	*git*	*gam*	*di*

These forms are semi-transparent – the standard Sursurunga numerals are *ru* ‘two’, *tul* ‘three’, *hat* ‘four’. A quadral similar to that in Sursurunga is presumably the origin of the Siar plural forms, suggesting that at an earlier stage Sursurunga had a quadral. Logically the loss of the quadral must have preceded a shift from trial to paucal in the Siar forms derived from the numeral for three. The Siar forms are therefore interesting for several reasons: the language originally had more number categories in independent pronouns than the modern language displays – one category has been lost; the semantic category quadral has been lost, but the formal category lost is the plural set - the function of the quadral has expanded to take over the semantic territory of the plural; and the trial shifted to a paucal function.

Corbett (2000:26-29) argues that usage demonstrates that the Sursurunga trial is actually a paucal, while the quadral is actually an extended paucal. However, the historical relationship between the forms given as trial and quadral in (9) and the numerals for three and four in Sursurunga demonstrates that the synchronic paucal had its origins in a construction meaning ‘they three’, and the extended paucal in a construction meaning ‘they four’, etc. The extended paucal function of the quadral in synchronic Sursurunga therefore represents a likely middle point in the diachronic development of the Siar plural with its origin in the numeral for four.

2.2 Argument agreement

Verb agreement for subject and object occurs in many but not all MM languages. In most languages this agreement distinguishes number as well as person.

Many MM languages typically display what is referred to in the Oceanist literature as a "verb complex" – a sequence that includes serializable verbs, negation, a number of adverbial categories, and a preverbal particle or proclitic encoding subject agreement along with modality or tense and sometimes aspect, and a postverbal position occupied by "object agreement", in fact often object clitic pronouns. NWS also display an unusual phenomenon in which possessive or former possessive morphology occurs postverbally agreeing with the subject as well as expressing aspectual categories (Palmer 2011). In addition to verb argument agreement, MM languages display possessor agreement in the NP.

Argument agreement in MM languages displays as many number distinctions as independent pronouns in that language, or fewer categories, but never more.

2.2.1 Preverbal subject agreement

Some MM languages have no preverbal subject agreement. In all cases this is a diachronic development in languages or subgroups that at an earlier stage displayed preverbal subject agreement. In Cheke Holo (Isabel, Palmer 2011:702), for example, preverbal former person and number indexing morphology has developed more fine-grained modal, aspect and tense functions and completely lost its subject agreement role.

In some MM languages preverbal subject agreement has lost its number agreement function, while retaining its person agreement function (in addition to modal functions). This is the case in Kokota (Isabel, Palmer 2009). The realis set is shown in (10).

(10) Kokota:

<table>
<tr><td></td><td>1EXC</td><td>1INC</td><td>2</td><td>3</td></tr>
<tr><td>SG</td><td rowspan="2">na</td><td></td><td rowspan="2">no</td><td rowspan="2">ne</td></tr>
<tr><td>PL</td><td>da</td></tr>
</table>

Some languages with multiple number categories in independent pronouns distinguish the same categories in subject agreement. In Vinitiri, for example, the independent pronouns distinguish four number categories - singular, dual, trial and plural, as discussed above. Preverbal subject agreement in the language makes the same distinctions:

(11) Vinitiri:

	1EXC	1INC	2	3
SG	*iə*		*u*	*i*
DU	*miru*	*təru*	*muru*	*diru*
TR	*mitalu*	*tulu*	*mutulu*	*ditalu*
PL	*mi*	*də / təu*	*mui*	*di*

In some person and number categories the relationship between the Vinitiri subject agreement particles and independent pronouns is regular and transparent. In others it is not.

The Vinitiri situation where preverbal subject agreement displays the same number categories as independent pronouns is not typical. MM languages with multiple number categories in independent pronouns typically distinguish only singular versus plural in subject agreement. This is the case in Kubokota, for example, as in (12).

(12) Kubokota:

	1EXC	1INC	2	3
SG	*ga*		*gu / go*	*za / gi*
PL	*ɣami*	*tage*	*ɣamu*	*gari / ge*

In several languages, primarily of the North Bougainville subgroup, the third singular subject agreement particle, usually *e*, has been generalized to all person and number categories. Typically this involves what might be termed ‘creeping neutralisation’. Rather than the entire function of person and number agreement being neutralised, categorical distinctions are progressively formally neutralised. In Hanahan Halia (North Bougainville, Allen 1987), for example, past tense realis has neutralised number in first exclusive and second persons, with those person distinctions also being neutralised, and person distinctions between first inclusive plural and third plural also neutralised, as (13) shows. Only third singular has retained a dedicated combination of person and number. In nonpast realis all person and number categories have been neutralised, as in (14).

(13) Hanahan:

	1EXC	1INC	2	3
SG	*u*		*u*	*e*
PL	*u*	*i*	*u*	*i*

(14) Hanahan:

	1EXC	1INC	2	3
SG	*e*		*e*	*e*
PL	*e*	*e*	*e*	*e*

Other languages are less far or further along a path of category neutralisation. Torau (Mono-Torau, Palmer 2007), for example, has only neutralized number in second person.

2.2.2 Object, postverbal subject, and possessor agreement

In terms of number the same phenomenon is seen in object, postverbal subject and possessor agreement as in preverbal subject agreement. Some languages with multiple number categories in independent pronouns distinguish the same categories in one or more of these types of agreement, while others display fewer categories, typically only singular versus plural. No languages display more number distinctions in these types of agreement than in independent pronouns.

2.3 Number hierarchies and animacy hierarchies

Number hierarchies and animacy hierarchies (see Corbett 2000:90-94) play interacting roles in the expression of pronominal categories in several MM languages. This phenomenon is probably more widespread in the group than is known as it is not typically reported in grammatical descriptions.

Vinitiri provides a good example. When the referent is human, the distinction between singular and plural is obligatorily expressed. The dual and trial forms are optional – plural forms may be used instead, demonstrating that in this language dual and trial are subcategories of plural, rather than discrete number categories. However, the use of these two subcategories is not equivalent. Dual, while optional, is used more frequently when there are two referents than trial is when there are three. To put this the other way around, plural is more likely to be used instead of trial than instead of dual. This suggests a number hierarchy as follows, in terms of likelihood of expression: PL > DU > TR. This hierarchy is exemplified in (15), where the same group of three actors is expressed in the first clause using a plural form, and in the second clause using a trial form.

(15)

Mi	*mutu*	*βuse*	*burəsi*	*u-ra=ra*	*pisa*
1EXC.PL.SBJ	chop	throw.away	fall	to-DIR=ART	ground

‘We chopped [it] away onto the ground.

na-muru	***mitalu***	*mutu-iau*	*a*	*uruə-na-pəkanə.*
LOC-follow	1EXC.TR.SBJ	chop-1SGOBJ	ART	two-LIG-piece

Then we three chopped me a piece.’ (Vinitiri)

The situation described above holds in relation to human referents. It also appears to hold with non-human animates, although the facts are not entirely clear. However, the situation is different with inanimate referents. In this situation plural marking in pronominal forms is not merely optional, it is impossible. In (16), for example, the fact that multiple tunnels are involved is explicitly established in the first clause by marking the noun with a plural marker. However, the subsequent pronominal reference to these tunnels in the third clause involves an otherwise singular agreement form.

(16) *Supu di gə kəli ra=**umənə** tuŋu.*
PURP 3PL.SBJ PST dig ART=PL tunnel
'They were supposed to dig tunnels.

*βare mi gə kisi, mi gə launu ta-**nə**.*
PURP 1EXC.PL.SBJ PST stay 1EXC.PL.SBJ PST live LOC=3SG.PSSR
So that we stayed, we lived in it [the tunnels].' (Vinitiri)

Data from other languages in the group suggests that a similar interaction of hierarchies of number and animacy may be at work, but descriptions typically do not make this explicit.

3. Pluralizing nouns

3.1 Lexical plurals

In most MM languages a handful of referents are expressed with distinct forms for the singular and plural. These lexical plurals usually occur with important human terms. In some cases the singular and plural terms are formally similar but irregularly related. For example Halia plural *tohaliou* 'women' is related to the singular *tahol* 'woman' by vowel metathesis and the addition of suffixed phonological material. More typically, the plural forms are suppletive, as the examples in (17) from Mono (Mono-Torau, Boch n.d.) illustrate.

(17) Mono:
a. *tioŋ* 'man' *hanua* 'men'
b. *batafa* 'woman' *talaiβa* 'women'
c. *tauii* 'child' *aanana* 'children'

3.2 Pluralization

Many MM languages lack a dedicated plural marker to accompany nouns. However, plural is also not expressed by inflection on nouns in MM, with the exception of reduplication discussed below. Instead, MM languages employ a range of strategies for expressing plurality with nouns, including accompanying pronouns, articles, demonstratives, quantifiers, and numerals. In all cases these are optional. As a result, in the majority of noun phrases in any MM language number is formally ambiguous and must be recovered from context.

3.2.1 Pronominal heads

Languages without an overt plural marker typically pluralize noun phrases periphrastically by making the NP the complement of a third person plural pronoun. This gives the pronoun the superficial appearance of a plural article, and descriptions of some languages analyse forms such as these as both a pronoun and a polysemous (or homophonous) plural article. In some languages this occurs with no article, reinforcing the appearance of the pronoun as an article. In others, such as Kubokota, the pronoun co-occurs with an article in the embedded NP, making it clear that the pronoun is not, itself, an article, as Chambers (2009) recognizes. In (18) the embedded NP is bracketed.

(18) *ria* [*na tinoni paleka=di*]
they ART person wound=3PL
'the wounded people' [lit. 'they the wounded people'] (Kubokota)

3.2.2 Articles

No number marking in articles common in MM languages. In Kubokota, for example, the common article *na* occurs with singular, as in (19a), and plural (19b) referents. Equally common are languages where the articles distinguish singular and plural, as in Kokota in (20).

(19) a. *Ani* ***na*** *toa=na.*
PROX.SG ART live=3SG.PSSR
'This is a live one.'

b. *Za kubo* ***na*** *seru.*
3SG.SBJ.RL be.many ART star
'There are lots of stars.' (lit. 'The stars are many.') (Kubokota)

(20) a. ***Ia*** *puku ba,* **ia** *do ba, n-e kati=nau ara.*
ART.SG fly ALT ART.SG mosquito ALT RL-3SBJ bite=1SG.OBJ I
'A fly or a mosquito bit me.' (Kokota)

b. *koɽ̥o ma=di* ***ira*** *l̥ol̥oguai=na.*
pull come=3PL.OBJ ART.PL coil=3SG.PSSR
'…[he] pulled his coils towards him.' (Kokota)

Several languages of the North Bougainville subgroup employ a noun class system that interacts with number in interesting ways involving the phenomenon of inverse number marking (see Corbett 2000:163-165). In Teop (North Bougainville, Mosel & Thiesen 2007) this system also interacts with the animacy hierarchy in a system somewhat more complex than as discussed by Corbett (2000:164-165). Teop has three classes, referred to as the A-class, O-class and E-class. The E-class involves what is referred to in the Oceanist literature as a personal article. Many Oceanic languages have an article that is used with personal names, and often also with pronouns. The exact extent of the coverage of the personal article varies from language to language. Typically the personal article does not distinguish number, but in a few MM languages, including Teop, it does. The inversion applies to the other two classes – the A-class and O-class. As (21) shows, the form that functions as the singular article in the A-class functions as the plural article in the O-class, and vice versa.

(21) Teop:

	E-class	A-class	O-class
SG	*e*	*a*	*o*
PL	*ere*	*o*	*a*

The basis of membership of the A-class and O-class has not been fully worked out for any MM language in which the phenomenon occurs. However, one of the factors at work appears to be an animacy hierarchy. For example in Teop, the A-class includes terms for humans, vertebrate animates, and invertebrate animates that have legs. The O-class, on the other hand, includes invertebrates without legs and plants. However, the A-class and O-class both include a number of items that do not conform to that system, as (22) shows.

(22) A-class: humans, vertebrates, legged invertebrates, food (inc. fruit), non-plant utensils, landmarks, possessed parts.
O-class: legless invertebrates, plants, plant parts (except fruit), plant material utensils, masses, etc.
E-class: personal names, pronouns, kinship terms, important humans (e.g. ‘chief’, ‘friend’), domestic animals

One factor at play in class membership appears to be cultural importance. Another appears to be a distinction between count and mass nouns. The latter suggests that one overarching factor may be individuation – items that are normally or readily individuated belong to the A-class, while items that are not normally individuated (or are not individuatable) belong to the O-class. From this perspective, the function of the articles may be collapsed from a system where each expresses both singular and plural depending on the class, into a system where the two articles each have a single function: *a* expresses ‘expected number’ (singular with items normally individuated, plural with things not normally individuated), while *o* expresses ‘unexpected number’ (plural with items normally individuated, singular with things not normally individuated). This hypothesis has yet to be tested against the data in North Bougainville languages displaying this inversion.

The animacy hierarchy clearly does come into play with the E-class in Teop. Interestingly, membership of the E-class differs in singular and plural, or to put it another way, the boundary between the singular personal *e* and human/animate *a* is at a different point in the animacy hierarchy compared to that of the plural personal *ere* and human/animate *o*:

(23) Teop:

	SG	PL
personal names	***e***	***ere***
kin terms	***e***	***ere***
important humans	***e***	*o*
domestic animals	***e***	*o*
other humans	*a*	*o*
wild vertebrates	*a*	*o*
legged invertebrates	*a*	*o*
legless invertebrates	*o*	*a*
plants	*o*	*a*

The data in (23) shows the role of the animacy hierarchy in Teop class membership, and the differential boundary between E-class and A-class for singular and plural. As discussed above, other factors such as cultural salience come into play with membership of the A-class and O-class. Examples in (24)-(25) illustrate the articles in use:

(24) a. ***a*** *moon*
ART woman
'the woman'

b. ***o*** *moon*
ART woman
'the women' (Teop)

c. ***o*** *hoi*
ART basket
'the basket'

d. ***a*** *hoi*
ART basket
'the baskets' (Teop)

(25) a. ***e*** *subu-na=e*
ART grandparent-POSS-3SG.PSSR
'his/her grandparent'

b. ***ere*** *subu-na=e*
ART grandparent-POSS-3SG.PSSR
'his/her grandparents' (Teop)

c. ***e*** *magee te=naa*
ART friend LOC=I
'my friend'

d. ***o*** *magee te=naa*
ART friend LOC=I
'my friends' (Teop)

3.2.3 Plural marking lexeme

Many MM languages have one or more independent lexeme that accompanies nouns to express plural. Vinitiri has one plural marker *umənə*. As is the case with most MM languages with a plural marker, *umənə* is optional - most NPs with plural referents have no overt plural marking. However, when *umənə* does occur, it must be accompanied by an article, as in (26). Syntactically *umənə* appears to be a noun, as it functions as head of the NP in which it occurs, as in (27).

(26)	*a*=***umənə***	*ŋətiŋəti*
	ART=PL	mosquito
	'mosquitos' (Vinitiri)	

(27)	*a*=***umənə***
	ART=PL
	'some [of something]' (Vinitiri)

While *umənə* indicates only plurality, the term is a member of a closed lexical class of four quantifiers:

(28)	a.	*umənə*	'plural'	b.	*kəβuanə*	'plenty/lots'
	c.	*sələβuru*	'various'	d.	*paupau*	'few' (Vinitiri)

Several MM languages have more than one plural marking lexeme. Teop has two: *maa*, a general plural, and *ba*, a plural used with kin terms:

(29)	a.	*a*=***maa***	*hoi*	*ohita*	b.	*a*=***ba***	*keara*	*te=naa*
		ART=GEN.PL	basket	galip.nut		ART=KIN.PL	brother	LOC=I
		'galip nut baskets'				'my brothers' (Teop)		

Again the Teop plural markers must occur with an article. However, as the article is the head of a DP and the plural marker occurs within the NP, with conjoined NPs like those in (30) the article has scope over both NPs, meaning the plural marker in the second NP superficially appears to occur without an article.

(30)	***a***=[[***maa***	*bebeahu*]	*bara*	[***maa***	*sun*	*hiaβa*	*sana*]]
	ART=GEN.PL	be.long	and	GEN.PL	stand	up	very
	'long and very high ones' (Teop)						

Bannoni has three distinct plural markers recognising distinctions of animacy: human *na*, animate *ne*, and inanimate *kare*. In (31c-d) the same lexical form has animate and inanimate meanings indicated solely by the articles.

(31)	a.	***na***	*taβana*	b.	***o***	*boroɣo*
		HUM.PL	person		ANIM.PL	pig
		'people'			'pigs' (Bannoni)	
	c.	***kare***	*pipito*	d.	***o***	*pipito*
		INANIM.PL	star		ANIM.PL	firefly
		'stars'			'fireflies' (Bannoni)	

As with the varying optionality of number marking determined by animacy in pronominal forms discussed in §2.3, Bannoni plural markers vary in optionality and likelihood of use depending on animacy. Unlike the plural markers in Vinitiri and Teop, in Bannoni the human plural marker *na*, shown in (31a), is obligatory with plural referents, while the animate plural in (31b) is optional. The inanimate plural *kare*, as in (31c), is rarely used, and is absent from some dialects.

3.2.4 Plural by reduplication

The one morphological strategy MM employs to mark plural on nouns involves reduplication. In Teop, for example, nouns may reduplicate to indicate plurality. However, with the exception of pluralization of some human terms in a handful of MM languages (see §4.1 below), this is optional. When reduplication does occur, the noun must be accompanied by an article, which itself indicates number, as in (32a). Reduplicated nouns may also optionally be accompanied by one of the plural marking lexemes discussed above, adding further marking for plurality, as in (32b). A noun in Teop may therefore carry as little as no overt marking for number, or as much as three forms – reduplication, a plural article, and a plural marking lexeme.

(32) a. *o* ***kari****~kariβana* *te=βe* *o* *beera…*
ART PL~scale LOC=3SG ART be.big
'[This fish,] its scales are big…' (Teop)

b. *a=****maa*** ***nahu****~nahu* *guu,* *a=****maa*** *meha* *nahu* *muu*
ART=PL PL~pot pig ART=PL other pot taro
'pots with pork, other pots with taro' (Teop)

As in Teop, in Vinitiri the reduplicative plural is optional, must occur with an article, and may or not occur with a plural marker. However, the Vinitiri reduplicative plural gives a distributive reading, as in (33b).

(33) a. *a=umənə* *dəβə~dəβəi*
ART=PL PL.DSTR~plant
'the plants' (Vinitiri)

b. *pətai* *a* *pisə* *i* *gə* *βanə* *parukə* *ta=ra* ***kani****~kaniəni*
NEG ART ground 3SG.SBJ PST go all LOC=ART PL.DSTR~home
'No earth [from the volcano] fell on any of the homes.' (Vinitiri)

3.2.5 Demonstratives

While demonstratives are employed to mark number in many MM languages, many others do not distinguish number in the demonstrative system. In Sisiqa (Choiseul, Ross 2002), for example, three spatial categories corresponding to person categories are recognized, but no number distinctions made, as in (34).

Kubokota, on the other hand, distinguishes singular and plural in its demonstratives (that in spatial terms are distance based, not person-based), as in (35). No MM language distinguishes more number categories than singular versus plural in its demonstratives.

(34) Sisiqa:

	Speaker proximal	Hearer proximal	Distal
SG/PL	*gəti*	*ta*	*gei*

(35) Kubokota:

	Proximal	Medial	Distal
SG	*ani*	*zana*	*nari*
PL	*ari*	*zara*	*rari*

In Kubokota distinct demonstrative forms express each number category. In Bannoni, on the other hand, plural demonstratives are constructed using demonstrative base forms, followed by a form identical to the inanimate plural article *kare*. However, in this context *kare* is not an article and follows the demonstrative rather than precedes it in two of the categories.

(36) Bannoni:

	Proximal	Medial	Distal
SG	*ie*	*nana*	*io*
PL	*ie kare*	*nana kare*	*kare io*

In MM languages the demonstratives are not determiners so do not occur in DET and freely co-occur with articles. In some languages, such as Bannoni in (37a) and Kubokota in (37b), demonstratives distinguishing number occur with number-invariant articles. Note that in some languages (e.g. Bannoni), the demonstrative is prenominal, while in others (e.g. Kubokota), it is postnominal.

(37) a. *tama-na*[*=**i*** ***ie*** *megara*
father-3SG.PSSR=ART PROX.SG child
'this child's father' (Bannoni)

b. ***na*** *tina=gu* *ara* ***ani***
ART mother=1SG.PSSR I PROX.SG
'this mother of mine' (Kubokota)

In other MM languages, such as Kokota in (38), number-distinguishing demonstratives occur with number-distinguishing articles.

(38) ***ira*** *naitu* *toke* ***aro***
ART.PL devil arrive PROX.PL
'these arriving devils' (Kokota)

4. Indexing target – number in possession

The typological dimension of indexing target has recently begun to attract attention (Evans & Fenwick 2010). This section examines the indexing of possessor number in relation to both marking locus and indexing target. 'Marking' is the location in a construction where a dependency is expressed – on the head, the dependent, neither or both. This is independent of 'indexing' of grammatical or lexical properties (Nichols 1986:58). Comparable to (but independent of) marking, the morphology may index features of the head, dependent, neither, or both. The locus of the marking may or may not be the same as the target of the indexing.

In MM all dependencies discussed so far in the present paper involve head-marking. This is typical for MM, and for Oceanic in general. For example, the dependency between a verb and its core arguments is expressed by marking the head verb with particles or clitics (i.e. by agreement), and not by marking the dependent argument (i.e. not by case).[1] However, while the head is the locus of the marking, it is features (person and number) of the dependent that are indexed. This is agreement. The notion of agreement may be most parsimoniously defined as morphology that is head-marking and dependent-indexing, while case may be defined as dependent-marking and dependent-indexing.

In MM the possessive dependency within NPs is also typically head-marking and dependent-indexing (i.e. it involves possessor agreement, not genitive case). Here the head possessum noun is marked with morphology indexing the number and person of the dependent possessor. This may be exemplified with Mono (Mono-Torau, Evans & Palmer 2011). Mono's relative Uruava (Evans & Palmer 2011) exemplifies an atypical situation for MM where the possessive dependency is marked on the head, but number of both the head and dependent are indexed.

4.1 Number in possession in Mono – head-marking:dependent-indexing

Typically for MM languages, and Oceanic in general, Mono has two possessive constructions – a 'direct' construction and an 'indirect' construction. In the direct

[1] This is the standard analysis. In fact in some MM languages apparent object agreement actually involves weak accusative pronouns. In some languages subject agreement may actually involve nominative pronouns. In these cases the morphology is the argument, not agreement. Discussion of this is beyond the scope of the present paper. Outside MM some Oceanic languages do have case marking particles, clitics or affixes (e.g. Polynesian). However, head-marking is typical.

construction, typically associated with inalienable possession, an affix attaches directly to the head possessum noun to index the number and person of the dependent possessor, as in (39).

(39) *batafa* *ifa-**na***
woman sis.in.law-3SG.PSSR
'the woman's sister(s)-in-law' (Mono)

Note that the number of the possessum is ambiguous – (39) may refer to one sister-in-law, or multiple sisters-in-law. The morphology therefore indexes the number (and person) only of the dependent possessor, and not of the head possessum. However, the morphology is located on the noun expressing the head possessum, while the dependent possessor noun is unmarked. This morphology is therefore head-marking and dependent-indexing, and thus agreement.

In the indirect construction, a particle precedes the head possessum noun, and it is this particle that carries the dependent possessor-indexing morphology. The particle marks the head noun as being in a possessive dependency, while the feature-indexing morphology again indexes the number (and person) of the dependent possessor, so again the construction involves head-marking and possessor-indexing. As is typical for Oceanic languages, more than one particle participates in the indirect construction, the particles encoding different possessive relationships. The distinctions they encode are not categories of possessum nouns, but categories of possessive relations (see Lichtenberk 1983). In Mono two such particles occur. One, *e-*, typically expresses a possessive relation in which the possessed item has been or is intended to be eaten, drunk, or consumed in some other way, as in (40a). The other, *sa-*, expresses general (i.e. non-consumed) alienable possession, as in (40b) and (41). Like a handful of other related languages but atypically for MM, in Mono certain human terms are obligatorily reduplicated when plural, as in (41b).

(40) a. ***e-gu*** *niunu*
CONS.POSS-1SG.PSSR coconut
'my coconut(s) [to eat]'

b. ***sa-na*** *mauto*
GEN.POSS-3SG.PSSR basket
'his/her basket(s)' (Mono)

(41) a. ***sa-gu*** *kanega*
GEN.POSS-1SG.PSSR husband
'my husband' (Mono)

b. *mani* ***sa-ma*** ***ka**~kanega*
weEXC.PL GEN.POSS-1EXC.PL.PSSR PL~husband
'our husbands' (Mono)

Note that in (40), as with the direct construction in (39), non-human possessums are ambiguous with respect to number, as the morphology indexes the number of the dependent possessor only. In a few MM languages such as Mono, the obligatory reduplication of certain human terms when plural means that examples like those in (41) are not ambiguous as to possessum number. However, this reduplication is entirely independent of the possessive construction and occurs whether a participating human term is possessed or not. The reduplication represents nominal number-marking. The number-indexing on the possessive particle represents head-marking and dependent-indexing for number within the possessive construction.

4.2 Number in possession in Uruava – head-marking:double indexing

Nominal possession outside of possessive constructions in Uruava displays the same kinds of phenomena seen in other MM languages. Terms with non-human referents are not marked for number by dedicated pluralizers or reduplication, and are typically ambiguous as to number. Number relating to such nouns is optionally expressed by forms such as demonstratives (which distinguish singular and plural) or numerals. With terms referring to humans plurality is marked. In the case of kin terms it is marked by reduplication, as in (42f), while with non-kin human terms a preposed pluralizer *buri* occurs, as in (42d).

In possessive constructions number is indexed in ways that in some respects are typical for MM and resemble that described for Mono above. However, in several important respects Uruava differs from other MM languages. As with Mono, Uruava employs a direct and an indirect possessive construction.

4.2.1 Number in Uruava indirect possession

In the indirect construction, the possessive dependency is expressed by a preposed particle that head-marks the possessum noun, as in (42). This particle carries morphology that indexes the number and person of the dependent possessor. As with Mono, two paradigms of preposed particle occur. However, unlike Mono, these two paradigms do not distinguish different categories of possession in the way that Mono *e-* and *sa-* do, and the categories of possessive relations have been neutralized. The two paradigms have been retained, but they have been co-opted to perform the function of indexing number of the head possessum. One paradigm occurs when the possessum is singular, as in (42a), (42c) and (42e), the other when it is plural, as in (42b), (42d) and (42f).

(42) a. ***e-gu*** *soni*
SG.PSSM-1SG.PSSR man
'my man'

b. ***go-gi*** *bere*
1SG.PSSR-PL.PSSM spear
'my spears' (Uruava)

c. ***e-mau*** *soni*
SG.PSSM-1EXC.PL.PSSR man
'our man'

d. ***mo-gi*** *buri* *soni*
1EXC.PL.PSSR-PL.PSSM PL man
'our men' (Uruava)

e. ***e-gu*** *aβutei*
SG.PSSM-1SG.PSSR bro.in.law
'my brother-in-law'

f. ***go-gi*** *aβu~aβutei*
1SG.PSSR-PL.PSSM PL~bro.in.law
'my brothers-in-law' (Uruava)

The Uruava indirect possessive particles as a whole therefore index number of both the dependent possessor and the head possessum. In this way they represent double indexing for the feature number. This represents real double indexing, rather than a co-occurrence of nominal number marking on the one hand and possessive dependent-marking for number on the other as seen with reduplicated human terms in Mono, as the Uruava number-indexing of the possessor and possessum are confined to and inseparable parts of a single construction, rather than separate types of number indexing that may independently turn up in the same phrase, as is the case with Mono reduplicated human terms. This double indexing for number in possession is highly atypical for MM, and for Oceanic in general. It appears to also be found in neighbouring Torau (Mono-Torau, Evans & Palmer 2011), and has been reported for Gabadi (Papuan Tip, Ross pers. comm.), but is not attested elsewhere in Oceanic, although it is reported in Austronesian outside Oceanic in Biak (Dalrymple, this volume). In Uruava and Torau development of double indexing for number in possession is a recent metatypic change under the influence of the neighbouring Papuan Naasioi language (Evans & Palmer 2011)

4.2.2 Number in Uruava direct possession

With a singular possessum, direct possession in Uruava resembles that seen in (39) for Mono. In (43) the head posssessum noun carries a suffix indexing the number and person of the dependent possessor.

(43) *aro* *patu-**mu***
youSG head-2SG.PSSR
'your head' (Uruava)

However, the crucial difference between Uruava and other MM (and Oceanic) languages is that (43) is not ambiguous for number – it can only have a singular reading. As with the indirect construction, Uruava directly possessed phrases must index the number of the head possessum as well as that of the dependent possessor. In (43) no overt morphology is present indexing the head as singular. However, (43) is singular because it lacks morphology indexing it as plural.

With kin terms, the number of the possessor is expressed by reduplication, as in (44). No dedicated possessive morphology is needed to index head number.

(44) a. *tama-**gu***
father-1SG.PSSR
'my father'

b. ***tama**~tama-**gu***
PL~father-1SG.PSSR
'my fathers' (Uruava)

However, when a noun with a non-human referent is directly possessed, no such strategy is available to indicate that the head is plural, as Uruava has no pluralizing morphology for non-humans, as discussed above. Uruava resolves this by employing the particle from the indirect construction, as in (45).

(45) a. *kabe-**gu***
hand-1SG.PSSR
'my hand'

b. ***go-gi*** *kabe-**gu***
1SG.PSSR-PL.PSSM hand-1SG.PSSR
'my hands' (Uruava)

This extraordinary strategy, apparently unique in Austronesian languages that have direct and indirect possessive constructions, results in multiple marking and indexing. Directly possessed non-human nouns in the plural are marked for the possessive dependency twice: once by the direct suffix and once by the indirect particle. The number (and person) of the dependent possessor is indexed twice: once by the suffix on the head noun itself and once by the prefix on the particle. The pressure to index the number of the head in possessive constructions, absent elsewhere in Oceanic, is so great in Uruava it causes the introduction of the particle from the indirect construction, with its resulting multiple marking and indexing of other categories.

5. Future Research

This paper has examined a range of phenomena of wider typological significance that require accounting for by theories of grammar. Some may be straightforward to account for within LFG. Others will pose greater challenges. In many cases, more work is needed to give a detailed enough understanding of the phenomena to allow theoretical modelling. Several issues discussed above in particular invite more detailed investigation. In each case, the phenomenon has been described to only a very limited extent, and in only a very few of the relevant languages. One such issue is the interaction of hierarchies of animacy and number discussed in §2.3. It is likely this plays a role in many or all MM languages, but is under-reported or unreported throughout the group. Similarly, while inversion in MM (§3.2.2) has been described to some extent, its interaction with animacy hierarchies and noun class systems warrant considerably more detailed investigation. A further issue relates to the possible syntactic status of quantifiers and number markers as nouns (touched on in §3.2.3), in turn relating to the pervasive problem of lexical category membership in Oceanic languages. Finally, alignments of marking-locus and indexing-target for the feature number in possession (§4) remains under-investigated. It is hoped that this paper will encourage further investigation of these issues in this theoretically and typologically significant group of languages.

References

Allen, J. 1987. *Halia grammar.* Ukarumpa: SIL.

Boch, A. n.d. *Short Alu grammar.* Typescript.

Chambers, M.R. 2009. *Which way is up? Motion verbs and paths of motion in Kubokota, an Austronesian language of the Solomon Islands.* PhD thesis: SOAS.

Corbett, G. 2000. *Number.* Cambridge: Cambridge University Press.

Evans, B. & B. Palmer. 2011. Contact-induced change in southern Bougainville. *Oceanic Linguistics* 50/2:483-524.

Evans, N. & E. Fenwick. 2010. *Marking vs indexing: revisiting the Nichols marking-locus typology.* Paper presented at ANU.

Frowein, F.M. 2011. *A grammar of Siar, an Oceanic language of New Ireland Province, Papua New Guinea.* PhD thesis: La Trobe University.

Hutchisson, D. 1975. *Sursurunga grammar essentials.* ms <www.sil.org/pacific/png/show_work.asp?id=928474543876>

Lichtenberk, F. 1983. Relational classifiers. *Lingua* 60:147–76

Lincoln, P.C. 1976. *Describing Banoni, an Austronesian language of southwest Bougainville.* PhD thesis: University of Hawai'i.

Lynch, J. & M.D. Ross. 2002. Banoni. In Lynch et al. 440-455.

Lynch, J., M.D. Ross & T. Crowley. 2002. *The Oceanic languages.* London: Routledge.

Money, L. 2002. *Babatana language lessons.* ms

Mosel, U. & Y. Thiesen. 2007. *The Teop sketch grammar.* ms. <www.linguistik.uni-kiel.de/Teop_Sketch_Grammar_May07.pdf>

Nichols, J. 1986. Head-marking and dependent-marking grammar. *Language* 62/1:56-119

Palmer, B. 2007. Imperfective aspect and the interplay of aspect, tense and modality in Torau. *Oceanic Linguistics* 46/2:499-519.

---- 2009. *Kokota grammar.* Honolulu: University of Hawai'i Press.

---- 2011. Subject-indexing and possessive morphology in Northwest Solomonic. *Linguistics* 49/4. 685–747.

Ross, M.D. 2002. Sisiqa. In Lynch et al. 456-466.

Van Der Mark, S.C. 2007. *A Grammar of Vinitiri. An Austronesian Language of Papua New Guinea.* PhD thesis: La Trobe University.

LEXICO-SEMANTIC COORDINATION IN POLISH

Agnieszka Patejuk and Adam Przepiórkowski
Institute of Computer Science, Polish Academy of Sciences

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

On the basis of data extracted from the largest currently available corpus of Polish, this paper discusses a variety of coordination under which, contrary to previous (mostly implicit) assumptions, particular conjuncts may correspond to distinct grammatical functions at the level of f-structure as long as they represent the same restricted semantic class (*wh*-words, *n*-words and items expressing quantifiers). Moreover, it demonstrates that dependents coordinated in this way may additionally belong to entirely different f-structures (depend on distinct heads) and it offers a formal analysis which was successfully implemented in a large XLE grammar of Polish.

1 Introduction

It was assumed for a long time that coordinated items should belong to the same c-structure category. When coordination of unlikes came to the attention of LFG, the new assumption was that it is possible to coordinate different categories but the coordinate structure bears the same grammatical function as a whole (Dalrymple and Lødrup 2000 discusses an example from Sag et al. 1985 where a nominal is coordinated with a clause and together they correspond to the object grammatical function). However, over the years it was noticed in different formalisms that, under certain circumstances, it is possible to coordinate dependents which bear different grammatical functions. This phenomenon was first discussed in Sannikov 1979, 1980 on the basis of Russian data, its existence was mentioned (though largely disregarded) in Mel'čuk 1988 and later a dependency-like analysis was provided for Polish by Kallas (1993); other analyses include Chaves and Paperno 2007 for Russian in the framework of HPSG (where this phenomenon is referred to as 'hybrid coordination') and quite recently Gazdik 2010 and 2012 in LFG for French and Hungarian.

This paper presents attested examples selected from abundant data extracted from the National Corpus of Polish (NKJP, Przepiórkowski et al. 2010, 2012; `http://nkjp.pl/`) and it shows how generalisations stemming from presented data were formalised and implemented in an LFG grammar of Polish (Patejuk and Przepiórkowski, 2012). It demonstrates, providing relevant corpus evidence, that it is possible to coordinate dependents which correspond to grammatical functions belonging to various levels in the f-structure (particular conjuncts depend on different heads), and it employs different formal representations of lexico-semantic coordination, monoclausal vs biclausal, depending on which items are involved in such coordination. Finally, it provides some discussion of controversial issues.

2 Data and generalisations

Though most often examples of lexico-semantic coordination include question words (*wh*-words), this phenomenon is by no means limited to these:

(1) czy komukolwiek, kiedykolwiek i do czegokolwiek przydał się poradnik
PART anybody.DAT anytime and for anything come in handy guide
'Has a(ny) guide ever come in handy to anybody for anything?' (NKJP)

(2) Obiecać można wszystko i wszystkim.
promise may everything.ACC and everyone.DAT
'One may promise everything to everyone.' (NKJP)

In (1) all coordinated items contain a pronoun which expresses an existential quantifier. This is the only similarity: particular conjuncts belong to different categories (noun phrase, adverbial phrase and a prepositional phrase, respectively) and bear distinct grammatical functions: indirect object (OBJ$_{\theta}$), adjunct (ADJ) and oblique object (OBL), respectively. In (2) both conjuncts contain a pronoun expressing a universal quantifier: the first corresponds to the direct object (OBJ) while the other is the indirect object (OBJ$_{\theta}$). Unlike in the previous example, both conjuncts happen to belong to the same category (noun phrase).

It is also possible to coordinate phrases containing pronouns which belong to another semantic class, namely *n*-words:

(3) nic i nikogo nie może tłumaczyć.
nothing.NOM and nobody.GEN NEG can excuse
'Nothing can excuse anybody.' (NKJP)

This example is interesting because particular conjuncts not only correspond to distinct grammatical functions but they also belong to different predicates: the first conjunct (*nic*) is the subject (SUBJ) of the main clause verb *może*,[1] while the other (*nikogo*) is the object (OBJ) of *tłumaczyć* in the embedded infinitival clause (XCOMP). There are further, more sophisticated examples of coordination where conjuncts depend on different heads:

(4) Skąd i jakie otrzymujemy informacje?
whence and what receive information
'What information and from where do we receive?' (NKJP)

In (4) both conjuncts are modifiers though they depend on different heads: the first conjunct is an adjunct (ADJ) of the verb (*otrzymujemy*), the other modifies the verb's object (*informacje*). Furthermore, it is possible that one lexico-semantic conjunct may be the head of the other one:

(5) ile i czego znaleźli.
how much.ACC and what.GEN found
'How much, and what, did they find?' (NKJP)

[1] As a result of structure sharing under raising, it is also the subject of *tłumaczyć* at the same time.

Polish numeral phrases are headed by a numeral while the accompanying nominal is analysed as its dependent.[2] In (5) the first conjunct is a numeral, analysed as the object (OBJ) of the verb (*znaleźli*), while the other is the object of the numeral (*ile*) – together they constitute a complete numeral phrase with the following f-structure representation:

(6) $$\begin{bmatrix} \text{PRED} & \text{`HOW\_MUCH}\langle\boxed{1}\rangle\text{'} \\ \text{OBJ}\ \boxed{1} & \begin{bmatrix} \text{PRED} & \text{`WHAT'} \end{bmatrix} \end{bmatrix}$$

Conjuncts taking part in lexico-semantic coordination in (4) and (5) belong to yet another semantic class, namely *wh*-words. Let us consider one more example featuring such conjuncts:

(7) Nie wiadomo było, czy *(i) kiedy wróci.
NEG know was whether and when returns
'It was not clear whether and when he would return.' (NKJP)

At first glance (7) appears similar to previous examples as all conjuncts represent the same semantic class, *wh*-words in this case: the first conjunct is a question particle (*czy*), the other is an adverb (*kiedy*). The particle is analysed as a marker (marking interrogative clauses), the other conjunct is treated as an adjunct of the verb. There is a crucial difference, though: when the conjunction (*i* 'and') is removed, (7) becomes ungrammatical, while all other examples presented so far remain grammatical even if the conjunction is deleted. It is possible, however, to use a biclausal construction as an alternative to (7), with roughly the same meaning:

(8) Nie wiadomo było, [czy wróci] i [kiedy wróci].
NEG know was whether returns and when returns
'It was not clear whether he would return and when he would return.'

This suggests that the representation of sentences such as (7), where the conjunction cannot be omitted without making the utterance ungrammatical, should be biclausal, i.e., based on the coordination of two clauses headed by the same main predicate. On the other hand, the remaining sentences, where the conjunction may be dropped without affecting the grammaticality of the utterance, will be analysed as essentially monoclausal, with only one occurrence of the main predicate in the representation.

Before proceeding to how lexico-semantic coordination is formalised in LFG and implemented in XLE, let us briefly summarise its properties: particular conjuncts bear distinct grammatical functions (arguments, adjuncts) or bear no grammatical function at all (as in the case of *czy*, the question marker), they may also belong to different levels of f-structure, sometimes even to different clauses (biclausal constructions featuring the question marker) as long as each conjunct represents the same semantic type (pronouns expressing a universal quantifier, existential

[2] This is the standard analysis in Polish linguistics, e.g., in the textbook of Saloni and Świdziński (2001); see also arguments for such a structure of Polish numeral phrases in Przepiórkowski and Patejuk 2012, Section 2 (in these proceedings).

quantifier, *n*-words or *wh*-words). Finally, particular conjuncts may correspond to different categories at the level of c-structure.

3 Formalisation and implementation

Lexical entries of items of a particular semantic type bear the attribute TYPE which may take one of four values: ANY (existential quantifier; cf. (1)), ALL (universal quantifier; cf. (2)), NEG (n-word; cf. (3)) or WH (question word; cf. (5), (4) and (7)). This feature has independent motivation: it is used in the grammar for the purposes of direct and embedded questions, free relatives and for handling negative concord. Simplified lexical entries of selected *n*-words are provided below:

(9) a. nic N $(\uparrow \text{PRED})=\text{`NOTHING'}$
$(\uparrow \text{TYPE})= \text{NEG}$

b. nigdy ADV $(\uparrow \text{PRED})=\text{`NEVER'}$
$(\uparrow \text{TYPE})= \text{NEG}$

Using parameterised c-structure rules, such elements are rewritten to phrases whose name contains, apart from category, a parameter whose value corresponds to its semantic type (represented below as a subscript in italics):

(10) $\text{NP}_{neg} \rightarrow \{ \text{ nic} \mid \text{nikt } \}$

(11) $\text{ADVP}_{neg} \rightarrow \{ \text{ nigdy} \mid \text{nigdzie } \}$

Parameters make it possible to use such semantic information at the level of c-structure without resorting to checking f-structure attributes (which is considerably more costly when it comes to measuring parser performance). It is possible to use parameters to ensure that certain categories in a given rule represent the same type:

(12) $$\text{XPextr}_{type} \quad \rightarrow \quad \underset{(\uparrow \text{ XPATH GF}^{+})=\downarrow}{\text{XP}_{type}}$$

The rule in (12) is also independently motivated as it is used for the purposes of handling extraction. Its left-hand side rewrites to a disjunction of phrases of the same type; the XP category used in (12) is in fact a metacategory; its expansion rule is provided in (13), with the definition of allowed types given in (14):

(13) $\text{XP}_{type} \equiv \{\text{NP}|\text{PP}|\text{ADVP}|\text{AP}\}_{type}$

(14) $type \equiv \{ \ all \mid any \mid wh \mid neg \ \}$

The annotation attached to XP in (12) makes it possible for dependents representing relevant semantic types to appear at the level of c-structure outside the clause containing their f-structure head. There are two important elements of this annotation: XPATH, defined in (15), provides the extraction path, while GF, defined in (16), corresponds to grammatical functions which may be assigned:

(15) $\textsc{xpath} \equiv \textsc{xcomp}^*$

(16) $\textsc{gf} \equiv \{\textsc{subj}|\textsc{obj}|\textsc{obj}_\theta|\textsc{obl}|\textsc{adj} \in\}$

Together, these allow the dependent to be extracted from infinitival clauses:

(17) uśmiecha się nieśmiało, bo nikogo nie chce krępować
smiles REFL shyly because nobody NEG wants intimidate
'She smiles shyly as she does not want to intimidate anybody.' (NKJP)

In (17) it is *nikogo* that undergoes extraction: even though it belongs at the level of c-structure to the main clause (with the verb *chce*), it is an argument of the embedded infinitival clause headed by *krępować*. Sometimes, however, it is possible to extract dependents from sentential complements:

(18) Kogo powiedziała, że nie chce więcej widzieć?
who said that NEG wants anymore see
'Who did she say she does not want to see anymore?'

In (18) the *wh*-word *kogo* is placed in the main clause while in terms of f-structure it is an argument of the infinitival complement (*widzieć*) of the sentential complement (*chce*) of the main clause (*powiedziała*). To account for such data, the extraction path is extended for relevant items, namely for (phrases containing) *wh*-words:[3]

(19) $\textsc{xpath} \equiv \textsc{comp}^*\ \textsc{xcomp}^*$

3.1 Monoclausal coordination

After particular conjuncts have been assigned appropriate functional annotation, they are fed into rules handling lexico-semantic coordination. The rule provided in (20) serves the purposes of handling sentences such as (1)–(4). Since the conjunction may be removed from these examples without any loss in grammaticality or any obvious change in meaning, these are assumed to have a monoclausal structure.

(20) $$\text{XPlxm}_{type} \rightarrow \underset{\uparrow=\downarrow}{\text{XPextr}_{type}} \quad [,\ \underset{\uparrow=\downarrow}{\text{XPextr}_{type}}]^* \quad \underset{\uparrow=\downarrow}{\text{CONJ}} \quad \underset{\uparrow=\downarrow}{\text{XPextr}_{type}}$$

Let us see how the f-structure corresponding to (4), repeated in (21) below for convenience, is constructed in a stepwise manner.

(21) Skąd i jakie otrzymujemy informacje?
whence and what receive information
'What information and where from do we receive?' (NKJP)

[3]The provided extraction path is trivial since closer investigation of Polish extraction phenomena remains outside of the scope of this paper. To account for attested data, it may require certain adjustments, including imposing additional constraints on some of its parts.

Particular conjuncts build their own partial f-structures thanks to the rule provided in (12). It assigns each conjunct its own, independent grammatical function annotation and although this annotation is very general (it may in theory generate a path consisting of the extraction path and any sequence of grammatical functions), one must bear in mind that its output is constrained by the f-structure of the rest of the utterance. As a result, the rule may generate infinitely many structures, but only the following f-structures built by individual conjuncts may be unified with the rest:

(22) a. $\left[\textsc{adj} \quad \left\{\left[\textsc{pred} \quad \text{`}\textsc{whence}\text{'}\right]\right\}\right]$

b. $\left[\textsc{obj} \quad \left[\textsc{adj} \quad \left\{\left[\textsc{pred} \quad \text{`}\textsc{what}\text{'}\right]\right\}\right]\right]$

Since all conjuncts in (20) bear the co-head annotation ('↑=↓'), unlike under the standard account of coordination (using the '↓∈↑' annotation), no set is created. Instead, f-structure fragments built by particular conjuncts, (22), are placed in one top-level f-structure:

(23)
$$\left[\begin{array}{ll} \textsc{adj} & \left\{\left[\textsc{pred} \quad \text{`}\textsc{whence}\text{'}\right]\right\} \\ \textsc{obj} & \left[\textsc{adj} \quad \left\{\left[\textsc{pred} \quad \text{`}\textsc{what}\text{'}\right]\right\}\right] \end{array}\right]$$

Finally, (23) is unified with the f-structure of the rest of the utterance to yield the full f-structure provided in (24):[4]

(24)
$$\left[\begin{array}{ll} \textsc{pred} & \text{`}\textsc{receive}\langle\boxed{1},\boxed{2}\rangle\text{'} \\ \textsc{subj}\ \boxed{1} & \left[\begin{array}{ll} \textsc{pred} & \text{`}\textsc{pro}\text{'} \\ \textsc{num} & \textsc{pl} \\ \textsc{pers} & 1 \end{array}\right] \\ \textsc{obj}\ \boxed{2} & \left[\begin{array}{ll} \textsc{pred} & \text{`}\textsc{information}\text{'} \\ \textsc{adj} & \left\{\left[\textsc{pred} \quad \text{`}\textsc{what}\text{'}\right]\right\} \end{array}\right] \\ \textsc{adj} & \left\{\left[\textsc{pred} \quad \text{`}\textsc{whence}\text{'}\right]\right\} \end{array}\right]$$

[4]Note that the f-structure provided in (24) includes an implicit subject (first person, plural).

3.2 Biclausal coordination[5]

A slightly different coordination rule, provided in (25)[6] below where the PART$_{wh}$ category corresponds to the question particle *czy* (see the lexical entry in (26)), is designed for examples such as (7), repeated in (27) below, which are considered biclausal, as discussed above.

(25) $\text{XPlxb}_{wh} \rightarrow \underset{\downarrow\in\uparrow}{\text{PART}_{wh}} \quad [,\ \underset{\downarrow\in\uparrow}{\text{XPextr}_{wh}}]^{*} \quad \underset{\uparrow=\downarrow}{\text{CONJ}} \quad \underset{\downarrow\in\uparrow}{\text{XPextr}_{wh}}$

(26) czy PART$_{wh}$ ($\uparrow$ CLAUSE-TYPE)= INT

(27) Nie wiadomo było, czy *(i) kiedy wróci.
NEG know was whether and when returns
'It was not clear whether and when he would return.' (NKJP)

To represent the fact that such utterances are not monoclausal, all conjuncts bear the set membership annotation ('↓∈↑'). As a result, partial f-structures constructed by individual conjuncts provided in (28) are placed inside a set, as shown in (29):

(28) a. $\begin{bmatrix}\text{CLAUSE-TYPE} & \text{INT}\end{bmatrix}$

b. $\begin{bmatrix}\text{ADJ} & \left\{\begin{bmatrix}\text{PRED} & \text{`WHEN'}\end{bmatrix}\right\}\end{bmatrix}$

(29) $\left\{\begin{bmatrix}\text{CLAUSE-TYPE} & \text{INT}\end{bmatrix}, \begin{bmatrix}\text{ADJ} & \left\{\begin{bmatrix}\text{PRED} & \text{`WHEN'}\end{bmatrix}\right\}\end{bmatrix}\right\}$

When the structure in (29) is unified with the f-structure of the rest of the utterance provided in (30), a biclausal coordinate structure results, as in (31):[7]

(30) $\begin{bmatrix}\text{PRED} & \text{`RETURN}\langle\boxed{1}\rangle\text{'} \\ \text{SUBJ}\ \boxed{1} & \begin{bmatrix}\text{PRED} & \text{`PRO'} \\ \text{NUM} & \text{SG} \\ \text{PERS} & 3\end{bmatrix}\end{bmatrix}$

[5]The type of lexico-semantic coordination described here is referred to as *biclausal* coordination despite the fact that the rule provided in (25) may generate structures containing more than two clauses. It seems, however, that examples illustrating this phenomenon tend to feature two conjuncts, leading to a biclausal representation, hence the name.

[6]This rule accounts for examples where the question particle is the first conjunct. Such examples seem to be most frequent; there exist, however, examples in which *czy* serves as the last conjunct:

(i) Będą sprawdzać kto i czy miał zezwolenie
will check who and PART had permission
'They will check whether (they had permission) and who had permission.' (NKJP)

Such cases may be handled by applying simple word order modifications to the rule in (25).

[7]Note that the implicit subject (third person, singular) in (31) is structure-shared: it belongs to both clauses at the same time.

(31)
$$\left\{ \begin{bmatrix} \text{PRED} & \text{‘RETURN}\langle\boxed{1}\rangle\text{’} \\ \text{SUBJ}\ \boxed{1} & \begin{bmatrix} \text{PRED} & \text{‘PRO’} \\ \text{NUM} & \text{SG} \\ \text{PERS} & 3 \end{bmatrix} \\ \text{CLAUSE-TYPE} & \text{INT} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{‘RETURN}\langle\boxed{1}\rangle\text{’} \\ \text{SUBJ} & \boxed{1} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{‘WHEN’} \end{bmatrix} \right\} \end{bmatrix} \right\}$$

The structure in (31) is biclasal because of the interaction of properties of its partial f-structures: (29) is a set and the PRED attribute in (30) is a distributive feature. When these structures are combined, (30) distributes to particular elements of the set in (29): ‘copies’ of (30) are merged with respective elements of (29), the question particle *czy* and the adjunct *kiedy* (‘when’), and the resulting structures are enclosed in a set, as in (31).

3.3 Argument saturation under the biclausal analysis

The biclausal analysis of certain instances of lexico-semantic coordination forces the introduction of some changes to relevant rules in order to account for independent argument saturation in coordinated clauses. While modifications are not required by examples such as (27) because there is an intransitive predicate whose only argument is shared (the implicit subject in (31)), argument saturation turns out to be an issue with sentences such as the following:

(32) czy *(i) ile będzie mogła zarobić tego typu placówka?
PART and how much AUX be able earn such institution
‘Will such an institution be able to earn and how much will it be able to earn?’ (NKJP)

In (32) the interrogative particle (*czy*) is coordinated with one of the arguments of the verb *zarobić*, which expresses a two-place predicate, taking a subject and a direct object. The former is overt (*placówka*) and it is shared by both coordinated clauses. According to the analysis provided above, dependents coordinated under biclausal lexico-semantic coordination belong to different clauses. As a result, *ile* may only fill the object grammatical function of one of the clauses. To avoid the violation of the completeness principle, the object of the other clause must also be filled in some way. This can be achieved using the following statement to handle implicit argument saturation:

(33)
$$\begin{array}{lll} \text{PRODROP} & \equiv & ((\uparrow \text{SUBJ PRED})=\text{‘PRO’}) \\ & & ((\uparrow \text{OBJ PRED})=\text{‘PRO’}) \\ & & \ldots \\ & & ((\uparrow \text{GF PRED})=\text{‘PRO’}) \end{array}$$

The statement provided above consists of a set of equations optionally (they are enclosed in brackets) filling a given grammatical function with an implicit argument (represented as the PRO value of the PRED attribute). The last line of (33) is to be treated as an abbreviation for all other appropriate grammatical functions, as defined in (16), with the exception of adjuncts.

It must be noted that the place of attachment of such statements is of importance – attaching (33) to the entire coordinate structure would give rise to a shared implicit dependent, which could cause violations of the uniqueness condition – a given grammatical function could be filled with a lexical dependent, leading to a clash with the implicit argument attempting to fill the same slot. For this reason (33) must not be placed inside the rule adding conjuncts to a set, it should instead be placed so that implicit arguments attach inside individual clauses. To achieve this, (33) should be attached at an intermediate level, so that its partial f-structure is merged with the f-structure fragment built by a given conjunct:

(34) $\text{XPextrbicl}_{type} \quad \rightarrow \quad \begin{array}{c}\text{XPextr}_{type}\\ \uparrow=\downarrow \\ \text{PRODROP}\end{array}$

Furthermore, care must be taken in order to ensure that conjuncts with prodrop statements are only used with biclausal lexico-semantic coordination. One of possible means to this end is to introduce additional categories for biclausal conjuncts exclusively, as in (34) above and (35) below:

(35) $\text{PARTbicl}_{type} \quad \rightarrow \quad \begin{array}{c}\text{PART}_{type}\\ \uparrow=\downarrow \\ \text{PRODROP}\end{array}$

Finally, the rule provided in (25) must be rewritten, replacing XPextr and PART categories with XPextrbicl and PARTbicl, respectively:

(36) $\text{XPlxb}_{wh} \quad \rightarrow \quad \begin{array}{c}\text{PARTbicl}_{wh}\\ \downarrow\in\uparrow\end{array} \quad [, \begin{array}{c}\text{XPextrbicl}_{wh}]^{*}\\ \downarrow\in\uparrow\end{array} \quad \begin{array}{c}\text{CONJ}\\ \uparrow=\downarrow\end{array} \quad \begin{array}{c}\text{XPextrbicl}_{wh}\\ \downarrow\in\uparrow\end{array}$

Let us now construct the f-structure representing (32) stepwise to see the modifications discussed above at work. First, individual conjuncts construct their partial f-structures using (12):[8]

(37) a. $\begin{bmatrix}\text{CLAUSE-TYPE} & \text{INT}\end{bmatrix}$

b. $\begin{bmatrix}\text{OBJ} & \begin{bmatrix}\text{PRED} & \text{‘HOW\_MUCH’}\end{bmatrix}\end{bmatrix}$

Subsequently, optional implicit arguments are added as a result of attaching (33) inside particular conjuncts, (34) and (35). The f-structure fragment provided in (38) shows how an implicit argument fills the object grammatical function in the f-structure which contains the question particle:

(38) $\begin{bmatrix}\text{CLAUSE-TYPE} & \text{INT}\\ \text{OBJ} & \begin{bmatrix}\text{PRED} & \text{‘PRO’}\end{bmatrix}\end{bmatrix}$

[8]Note that (37b) is simplified: the implicit object of the numeral is not represented.

Next, conjuncts are added to a set using the modified rule handling biclausal lexico-semantic coordination provided in (36):

(39)
$$\left\{ \begin{bmatrix} \textsc{clause-type} & \textsc{int} \\ \textsc{obj} & \begin{bmatrix} \textsc{pred} & \text{`}\textsc{pro}\text{'} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \textsc{obj} & \begin{bmatrix} \textsc{pred} & \text{`}\textsc{how\_much}\text{'} \end{bmatrix} \end{bmatrix} \right\}$$

Finally, the partial f-structure built by lexico-semantic coordination is merged with the f-structure of the rest of the utterance. The following full f-structure results:

(40)
$$\left\{ \begin{bmatrix} \textsc{pred} & \text{`}\textsc{earn}\langle \boxed{1},\boxed{2} \rangle\text{'} \\ \textsc{subj}\ \boxed{1} & \begin{bmatrix} \textsc{pred} & \text{`}\textsc{institution}\text{'} \end{bmatrix} \\ \textsc{obj}\ \boxed{2} & \begin{bmatrix} \textsc{pred} & \text{`}\textsc{pro}\text{'} \end{bmatrix} \\ \textsc{clause-type} & \textsc{int} \end{bmatrix}, \begin{bmatrix} \textsc{pred} & \text{`}\textsc{earn}\langle \boxed{1},\boxed{3} \rangle\text{'} \\ \textsc{subj} & \boxed{1} \\ \textsc{obj}\ \boxed{3} & \begin{bmatrix} \textsc{pred} & \text{`}\textsc{how\_much}\text{'} \end{bmatrix} \end{bmatrix} \right\}$$

While the lexical subject is shared by both clauses (as indicated by appropriate structure sharing of relevant f-structure fragments), the object of the first clause (it contains the question particle, the first lexico-semantic conjunct) is filled with an implicit argument, while the object of the other clause is filled with a lexical argument, the second lexico-semantic conjunct.

4 Issues

While previous sections discussed key facts concerning lexico-semantic coordination, the aim of this section is to address potential doubts as to the standing of this phenomenon as a variety of coordination, as well as some less obvious (though important) issues and, finally, possible extensions.

4.1 Is this coordination?

Since lexico-semantic coordination is a potentially very surprising variety of coordination, it seems natural to question whether it is indeed an instance of coordination. While typical tests such as agreement seem inapplicable, there is fortunately some other potentially convincing evidence.

First, it is possible to use such constructions with items which are unambiguous and uncontroversial conjunctions in Polish:

(41) [kto oraz kiedy] miałby płacić za postawiony budynek
who and when should pay for erected building
'Who and when would be supposed to pay for the erected building?' (NKJP)

(41) features *oraz* ('and'), an entirely unambiguous conjunction as there is no other available interpretation of this word.

Furthermore, it is possible to find examples where a preconjunction is used, as in 'both... and...' coordinate structures:

(42) A jest i co, i gdzie eksportować.
and is and what and where export
'There (certainly) is what and where to export to.' (NKJP)

While all examples presented so far featured conjoining, *and*-type conjunctions (mostly *i*), there exist examples with alternative conjunctions:

(43) [kto lub czego] będzie w Wikipedii szukał.
who or what will in Wikipedia seek
'Who will seek what in Wikipedia?' (NKJP)

While the word *lub* ('or') is not perfectly unambiguous, its other interpretation, the imperative form of the verb *lubić* 'like', is not an option in this context, leaving the conjunction interpretation. The LFG account of coordination using an alternative conjunction is exactly the same as for phrases coordinated using a conjoining conjunction and consequently the same convention was used for lexico-semantic coordination with such conjunctions. However, as in the case of more standard coordination, there is a difference in the semantics, perhaps less evident under lexico-semantic coordination. For this reason, it is important to record the shape of the conjunction involved, which is discussed in more detail in Section 4.2.

4.2 Representing the conjunction

The f-structures provided in Section 3 did not include the contribution of the annotation of the conjunction in any way. As mentioned in Section 4.1, the form of the conjunction, namely whether it belongs to the conjoining or the alternative type, is of importance from the perspective of semantics. Such information may be provided using a dedicated attribute, COORD-FORM for instance:

(44) a. i CONJ $(\uparrow \textsc{coord-form}) = \textsc{and}$

b. lub CONJ $(\uparrow \textsc{coord-form}) = \textsc{or}$

When conjunctions annotated in this way are used with rules such as (20), the rule handling monoclausal lexico-semantic coordination, the conjunction is represented in the top-level f-structure. The relevant fragment corresponding to the entire lexico-semantic coordinate phrase (including the conjunction) from example (4), repeated later as (21), is provided below:

(45)
$$\begin{bmatrix} \textsc{adj} & \left\{ \begin{bmatrix} \textsc{pred} & \text{`}\textsc{whence}\text{'} \end{bmatrix} \right\} \\ \textsc{obj} & \begin{bmatrix} \textsc{adj} & \left\{ \begin{bmatrix} \textsc{pred} & \text{`}\textsc{what}\text{'} \end{bmatrix} \right\} \end{bmatrix} \\ \textsc{coord-form} & \textsc{and} \end{bmatrix}$$

When this fragment is unified with the f-structure of the rest of the utterance, the following structure results:[9]

[9]The f-structure of the implicit subject is simplified in (46) and the following examples.

(46)
$$\left[\begin{array}{ll} \text{PRED} & \text{‘RECEIVE}\langle\boxed{1},\boxed{2}\rangle\text{’} \\ \text{SUBJ}\ \boxed{1} & \left[\begin{array}{ll}\text{PRED} & \text{‘PRO’}\end{array}\right] \\ \text{OBJ}\ \boxed{2} & \left[\begin{array}{ll}\text{PRED} & \text{‘INFORMATION’}\\ \text{ADJ} & \left\{\left[\begin{array}{ll}\text{PRED} & \text{‘WHAT’}\end{array}\right]\right\}\end{array}\right] \\ \text{ADJ} & \left\{\left[\begin{array}{ll}\text{PRED} & \text{‘WHENCE’}\end{array}\right]\right\} \\ \text{COORD-FORM} & \text{AND}\end{array}\right]$$

Such a representation is potentially vulnerable to interference caused by dependent sharing whereby a single phrase is shared by more than one head, as in the following modified version of (4):

(47) [Skąd i jakie] [otrzymujemy lub kradniemy] informacje?
whence and what receive or steal information
‘What information and where from do we receive or steal?’

When, as in (47), verbs are coordinated, the conjunction is represented at the same level as the set containing particular verbal heads. The structure provided below represents (47) with the exclusion of the lexico-semantic coordinate phrase (it corresponds to the following fragment: *[otrzymujemy lub kradniemy] informacje*):

(48)
$$\left[\begin{array}{l} \left\{\left[\begin{array}{ll} \text{PRED} & \text{‘RECEIVE}\langle\boxed{1},\boxed{2}\rangle\text{’} \\ \text{SUBJ}\ \boxed{1} & \left[\begin{array}{ll}\text{PRED} & \text{‘PRO’}\end{array}\right] \\ \text{OBJ}\ \boxed{2} & \left[\begin{array}{ll}\text{PRED} & \text{‘INFORMATION’}\end{array}\right]\end{array}\right], \left[\begin{array}{ll} \text{PRED} & \text{‘STEAL}\langle\boxed{1},\boxed{2}\rangle\text{’} \\ \text{SUBJ} & \boxed{1} \\ \text{OBJ} & \boxed{2}\end{array}\right]\right\} \\ \text{COORD-FORM}\quad \text{OR}\end{array}\right]$$

While *i* (‘and’) would be (accidentally) unproblematic as the element conjoining verbs because the same conjunction is used in the lexico-semantic coordinate phrase (see (45) and (46)), using *lub* (‘or’) as the conjunction in the coordinate verb phrase results in the clash ($\neq$) of values of COORD-FORM in the top-level f-structure:

(49)
$$\left[\begin{array}{l} \left\{\left[\begin{array}{ll} \text{PRED} & \text{‘RECEIVE}\langle\boxed{1},\boxed{2}\rangle\text{’} \\ \text{SUBJ}\ \boxed{1} & \left[\begin{array}{ll}\text{PRED} & \text{‘PRO’}\end{array}\right] \\ \text{OBJ}\ \boxed{2} & \left[\begin{array}{ll}\text{PRED} & \text{‘INFORMATION’}\\ \text{ADJ} & \left\{\left[\begin{array}{ll}\text{PRED} & \text{‘WHAT’}\end{array}\right]\right\}\end{array}\right] \\ \text{ADJ}\ \boxed{3} & \left\{\left[\begin{array}{ll}\text{PRED} & \text{‘WHENCE’}\end{array}\right]\right\}\end{array}\right], \left[\begin{array}{ll} \text{PRED} & \text{‘STEAL}\langle\boxed{1},\boxed{2}\rangle\text{’} \\ \text{SUBJ} & \boxed{1} \\ \text{OBJ} & \boxed{2} \\ \text{ADJ} & \boxed{3}\end{array}\right]\right\} \\ \text{COORD-FORM}\quad \text{AND}\neq\text{OR}\end{array}\right]$$

In (49) conjunctions used in two coordinate phrases, lexico-semantic (*i*) and verbal (*lub*), set conflicting COORD-FORM values, AND and OR, respectively, represented as inequality: AND$\neq$OR.

A related problem is caused by the embedding of coordination within lexico-semantic coordination, as in the two examples below:

(50) Nigdy nie wiadomo, [[kto lub co], skąd i kiedy] zaatakuje.
never NEG know who or what whence and when attacks
'You never know who or what, where from and when may attack.' (NKJP)

(51) kombinowaniem [kto, kogo, kiedy i jak], [z kim przeciw komu] albo
plotting who whom when and how with whom against whom or
[od kogo i za co]
from whom and for what
'[...] plotting about who, whom, when and how, with whom against whom or from whom and for what [...]' (NKJP)

Two varieties of coordination are involved in (50): the first conjunct of lexico-semantic coordination is at the same time a regular coordinate NP (both its elements bear the subject grammatical function), while the remaining lexico-semantic conjuncts are adjuncts (ablative and temporal). This is unproblematic representationally, because the conjunction is represented inside the coordinate NP.

Example (51), is considerably more interesting as it presents embedded lexico-semantic coordination: two edge conjuncts are also instances of such coordination. The first conjunct contains a subject, an object and two adjuncts (temporal and manner), the middle conjunct features multiple *wh*-phrases (two obliques), and the last conjunct consists of another oblique coordinated with an adjunct. It is possible to construct a less complicated example, though:

(52) [Kto i kogo] lub [kiedy i gdzie] poznał?
who and whom or when and where met
'Who did meet whom, or when and where?'

In the above example the first conjunct contains a subject and an object while the other conjunct consists of two coordinated adjuncts.

There is a potential solution to problems posed by such examples in terms of discourse functions. The value of a discourse function, a hybrid structure, would represent lexico-semantic coordination: it would gather particular conjuncts inside a set and, as under standard coordination, the conjunction would be represented outside the set. Furthermore, particular conjuncts would be structure-shared with relevant parts of the main f-structure. In this way, the conjunction inside the lexico-semantic coordinate phrase would only be represented in the f-structure corresponding to the discourse function, making it impossible to conflict with the value of the COORD-FORM attribute (if present) of the top-level f-structure. Finally, embedding could be handled using standard coordination rules which give rise to embedded hybrid structures. Substantiating this general idea should be a matter of future work.

4.3 More types of lexico-semantic coordination?

Section 2 presented data focused on four types of items involved in lexico-semantic coordination, namely pronouns expressing an existential quantifier, (1), or a universal quantifier, (2), *n*-words, (3), and *wh*-words, (5)–(7). It seems, however, that the inventory of lexico-semantic types could be extended.

Kallas (1993) discusses the following example:[10]

(53) Jan pamięta tyle i takich oskarżeń.
John remembers that many and such accusations
'John remembers that many (of) such accusations.' (Kallas, 1993, p. 53)

There are similar attested examples:

(54) my nie mogłybyśmy zapewnić naszym podopiecznym tylu i takich
we NEG could provide our charges so many and such
materiałów do pracy
resources for work
'We would not be able to provide our charges with so many (of) such work resources.' (NKJP)

(55) Że będzie i jest tyle i takich afer?
that will and is that many and such scandals
'That there is and will be that many (of) such scandals?' (NKJP)

In (53) and (54) particular conjuncts correspond to the object (*tyle* and *tylu*,[11] respectively) and the modifier of the object's object (*takich*). (55) shows lexico-semantic coordination of the subject and the modifier of the subject's object. It seems that the common feature of lexico-semantic conjuncts in the examples presented above is the fact that they belong to the class of demonstratives.

Another possible class is consistuted by free relatives. Recently, Citko and Gracanin-Yuksek (2012) discussed such coordination on the basis of data from Polish, English and Croatian. They provide the following Polish example:

(56) Jan je cokolwiek i kiedykolwiek Maria gotuje.
John eats whatever and whenever Mary cooks

In (56) an object (*cokolwiek*) and an adjunct (*kiedykolwiek*) are coordinated.

Finally, though Citko and Gracanin-Yuksek (2012) claim that there is 'a more general constraint that rules out two relative pronouns in a relative clause modifying a single head, regardless of whether the relative pronouns are coordinated or not' and provide example (57) in support of this claim, there are examples such as (58) which seem to be grammatical when coordination is used:

[10] The glosses and free translation in (53) are our own.

[11] In (54) the object is marked for genitive case, unlike in (53), as a result of object case assignment in the syntactic scope of negation. See Przepiórkowski 1999 for an extensive discussion of Genitive of Negation (GoN) in Polish.

(57) *student którego (i) któremu Maria przedstawiła
student who and whom Mary introduced

(58) człowiek, z którym i o którym lubię mówić
man with whom and about whom like talk
'the man with whom and about whom I like to talk'

Incorporating examples such as the ones presented above (demonstratives, free relatives, relatives) in the analysis proposed in this paper should not pose any problems. Necessary changes would include extending the inventory of allowed semantic conjunct types in relevant rules and assigning demonstratives a type (DEM for instance). Relative pronouns and pronouns expressing an existential quantifier already bear appropriate types (REL and ANY, respectively) for independent reasons, namely for the purposes of handling relative clauses and free relatives. It must be noted, however, that such examples, especially ones including relative pronouns, are not as numerous, varied and productive as other examples presented in previous sections.

5 Conclusion

There is a growing interest in lexico-semantic coordination, also within LFG. In comparison to previous work, the main contributions of this paper include: showing that coordinated elements may be dependents of different heads, distinguishing between monoclausal and biclausal lexico-semantic coordination, and providing a relatively comprehensive analysis of lexico-semantic coordination in Polish.

However, some loose ends remain. The most pressing is the question of the representation of the conjunction, with some preliminary ideas suggested at the end of Section 4.2, but also the exact repertoire of semantic classes which may participate in lexico-semantic coordination, especially, the possibility of such coordination in free relatives and in ordinary relative clauses, mentioned in Section 4.3. Finally, the fundamental issue of why exactly such semantic classes make it possible to violate the overwhelming constraint that only the same grammatical functions can be coordinated has not been addressed. The present interest in lexico-semantic coordination will certainly continue to grow in the near future.

Acknowledgements

This research is supported by the POIG.01.01.02-14-013/09 project which is co-financed by the European Union under the European Regional Development Fund. The publication of the resulting implemented grammar is supported by the CESAR project (European project CIP ICT-PSP 271022, part of META-NET). The authors are grateful to the editors of this volume for valuable comments on the previous version of this paper.

References

Chaves, Rui Pedro and Paperno, Denis. 2007. On The Russian Hybrid Coordination Construction. In Stefan Müller (ed.), *The Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*, pages 46–64, Stanford: CSLI Publications.

Citko, Barbara and Gracanin-Yuksek, Martina. 2012. Wh-Coordination in Free Relatives. Abstract of a talk presented at the 43rd annual meeting of the North East Linguistic Society, City University of New York, 19–21 October 2012; `http://nels2012.commons.gc.cuny.edu/?page_id=14`.

Dalrymple, Mary and Lødrup, Helge. 2000. The grammatical functions of complement clauses. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG'00 Conference*, University of California, Berkeley.

Gazdik, Anna. 2010. Multiple Questions in French and Hungarian: An LFG Account. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG'10 Conference*, pages 249–269, Ottawa, Canada.

Gazdik, Anna. 2012. *Multiple Questions in French and Hungarian. A Lexical-Functional Analysis with Special Emphasis on the Syntax-Discourse Interface*. Ph. D.thesis, Université Paris 7.

Kallas, Krystyna. 1993. *Składnia współczesnych polskich konstrukcji współrzędnych*. Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika.

Mel'čuk, Igor. 1988. *Dependency Syntax: Theory and Practice*. Albany, NY: The SUNY Press.

Patejuk, Agnieszka and Przepiórkowski, Adam. 2012. Towards an LFG parser for Polish: An exercise in parasitic grammar development. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 3849–3852, ELRA, Istanbul, Turkey.

Przepiórkowski, Adam. 1999. *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Ph.D. dissertation, Universität Tübingen, Germany.

Przepiórkowski, Adam, Bańko, Mirosław, Górski, Rafał L. and Lewandowska-Tomaszczyk, Barbara (eds.). 2012. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Warsaw: Wydawnictwo Naukowe PWN.

Przepiórkowski, Adam, Górski, Rafał L., Łaziński, Marek and Pęzik, Piotr. 2010. Recent Developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, ELRA, Valletta, Malta.

Przepiórkowski, Adam and Patejuk, Agnieszka. 2012. The puzzle of case agreement between numeral phrases and predicative adjectives in Polish. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG'12 Conference*, pages 490–502, Stanford, CA: CSLI Publications.

Sag, Ivan A., Gazdar, Gerald, Wasow, Thomas and Weisler, Steven. 1985. Coordination and How to Distinguish Categories. *Natural Language and Linguistic Theory* 3, 117–171.

Saloni, Zygmunt and Świdziński, Marek. 2001. *Składnia współczesnego języka polskiego*. Warsaw: Wydawnictwo Naukowe PWN, fifth edition.

Sannikov, Vladimir Z. 1979. Sočinitel'nye i sravnitel'nye konstrukcii: ix blizost', ix sintaksičeskoe predstavlenie I. *Wiener Slawistischer Almanach* 4, 413–432.

Sannikov, Vladimir Z. 1980. Sočinitel'nye i sravnitel'nye konstrukcii: ix blizost', ix sintaksičeskoe predstavlenie II. *Wiener Slawistischer Almanach* 5, 211–242.

ON CASE ASSIGNMENT AND THE COORDINATION OF UNLIKES: THE LIMITS OF DISTRIBUTIVE FEATURES

Adam Przepiórkowski and Agnieszka Patejuk
Institute of Computer Science, Polish Academy of Sciences

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper discusses the notion of distributivity of features in coordinate structures and demonstrates its limits in unlike coordination constructions involving conjuncts bearing different grammatical cases. Two solutions are presented, one – termed "liberal" – necessitating certain extensions to the formal machinery of LFG, and another – termed "conservative" – which recycles the mechanism of off-path constraints.

1 Introduction

The aim of this paper is to demonstrate a certain weakness of the standard mechanism of distributive and non-distributive features in LFG analyses of coordination, and to propose a straightforward extension of this mechanism to account for the problematic data. An alternative solution of the same problem, by Mary Dalrymple (p.c.), is also presented, which does not require any modifications to the formal apparatus of LFG and relies instead on the mechanism of off-path constraints.

The problem addressed here may be summarised as follows. For various languages, it makes sense to posit general statements taking care of so-called structural case assignment (as opposed to lexical – or inherent – case assignment). For example, in Polish, such statements may require that case-bearing subjects must be in the nominative, with the exception of a class of numeral phrases, which must occur in the accusative.[1] However, given standard LFG assumptions, such statements fail in cases of unlike coordination. Since CASE is a distributive feature, such statements would assign the same case (if any) to all conjuncts, even if one of them is a nominal phrase (and, hence, should occur in the nominative in the subject position), and another one is a numeral phrase (accusative) or a clause (caseless). The problem is not limited to subject positions.

After outlining standard LFG assumptions regarding coordination, distributive features and case assignment in Section 2, we present Polish case facts in more detail in Section 3. Then, the problem such facts present to the current analyses – and to standard LFG assumptions – is described in Section 4. Two solutions to this problem are then proposed: our solution introducing an extension of the standard approach to distributivity (Section 5) and a more conservative solution suggested to us by Mary Dalrymple (Section 6). Finally, Section 7 discusses the relative pros and cons of these approaches, and concludes the paper.

2 Standard LFG assumptions

In LFG analyses of coordination, following Dalrymple and Kaplan 2000 (who credit John Maxwell with the basic idea), all conjuncts are elements of a set in a hybrid feature structure which, apart from representing this set, may also contain its own features. When f is a hybrid feature structure, the interpretation of a functional

[1] See Przepiórkowski and Patejuk 2012 (in these proceedings), especially fn. 2 there.

description such as "(f FEATURE) $=$ VALUE" depends on the status of the FEATURE: in case of distributive features, like CASE, such a statement applies to all elements in the set represented by the feature structure, but not to the feature structure itself, while in case of non-distributive features, like NUMBER, GENDER and PERSON, the statement pertains to the whole feature structure and it does not affect the elements of the set.

Here, we assume recent LFG approaches to the representation and assignment of case proposed in Dalrymple et al. 2009; for example, the unambiguously accusative German pronoun *ihn* 'him' will be specified for CASE as in (1a), giving rise to the feature structure in (1b), while the German *was* 'what', syncretic between the nominative and the accusative, will be specified as in (2a), satisfied, e.g., by the feature structure (2b) when interpreted as accusative:

(1) a. CASE NOM = −
CASE GEN = −
CASE DAT = −
CASE ACC = +

b. $$\left[\text{CASE}\ \begin{bmatrix}\text{NOM} & - \\ \text{ACC} & + \\ \text{GEN} & - \\ \text{DAT} & -\end{bmatrix}\right]$$

(2) a. CASE GEN = −
CASE DAT = −
CASE {NOM|ACC} = +

b. $$\left[\text{CASE}\ \begin{bmatrix}\text{NOM} & \\ \text{ACC} & + \\ \text{GEN} & - \\ \text{DAT} & -\end{bmatrix}\right]$$

On this approach, the usual case assignment statements look as in (3) below:

(3) a. ($\uparrow$ SUBJ CASE NOM) = +
b. ($\uparrow$ OBJ CASE ACC) = +

An application of these statements to the nominative/accusative syncretic *was* in a German free relative construction (4a), where *was* is simultaneously an object (of the matrix verb) and a subject (of the embedded verb), results in the structure in (4b):

(4) a. Ich habe gegessen was übrig war.
I have eaten what left was
'I ate what was left.'

b. *was* in (4a):

$$\left[\text{CASE}\ \begin{bmatrix}\text{NOM} & + \\ \text{ACC} & + \\ \text{GEN} & - \\ \text{DAT} & -\end{bmatrix}\right]$$

The distributivity of CASE is crucial in LFG analyses such as Dalrymple and Kaplan 2000 and Dalrymple et al. 2009, where case assignment annotations such as "ACC $\in$ ($\uparrow$ OBJ CASE)" (in the former) or "($\uparrow$ OBJ CASE ACC) $=$ +" (in the latter) apply to all conjuncts in coordinate objects. Interestingly, in case of verbs with indeterminate case requirements, e.g., the Russian *proždat'* 'wait', taking accusative or genitive objects, the statement “($\uparrow$ OBJ CASE {ACC|GEN}) = +” applies to all conjuncts, but the “{ACC|GEN}” uncertainty is resolved separately for each conjunct, giving rise to the possibility of coordination of differently cased NPs, as in (5) from Levy and Pollard 2001, p. 221 (cited by Dalrymple et al. 2009, p. 41):

(5) Včera ves’ den’ on proždal svoju podrugu Irinu i zvonka ot
yesterday all day he waited-for self’s girlfriend.ACC Irina and call.GEN from
svoego brata Grigorija.
self’s brother Gregory
‘Yesterday he waited all day for his girlfriend Irina and for a call from his brother Gregory.’

3 Case in Polish

We adopt the basic distinction between structural and lexical case assignment, widely assumed, e.g., in transformational theories (starting with Rouveret and Vergnaud 1980 and Vergnaud 1982, and adopted in Chomsky 1981, *inter alia*) and in Head-driven Phrase Structure Grammar (e.g., Heinz and Matiasek 1994, Pollard 1994, Przepiórkowski 1996), and traceable back at least to Kuryłowicz 1949. In the particular implementation of this idea assumed here (roughly that of Przepiórkowski 1999), predicates mark their case-bearing arguments with a specific morphological case or they leave the case of the argument unspecified, marked only with the diacritic “SC = +” (SC stands for *structural case*).[2]

The basic facts of structural case assignment in Polish are as follows:[3]

(6)
- **subjects** bearing structural case are in the nominative,
- with the exception of numeral phrase subjects, headed by so-called governing numerals (see below), which are in the accusative;
- **objects** bearing structural case are in the accusative,
- unless they are in the syntactic scope of sentential negation, in which case they are in the genitive (so-called Genitive of Negation, GoN).

These facts may be modelled in a straightforward way by case assignment statements like the following (for the first two bullets above), on the assumption that S = $\uparrow$ SUBJ:[4]

(7) a. $(\mathrm{S}\ \textsc{sc}) =_c + \ \wedge\ (\mathrm{S}\ \textsc{acm}) \neq \textsc{rec} \quad \rightarrow \quad (\mathrm{S}\ \textsc{case nom}) = +$

b. $(\mathrm{S}\ \textsc{sc}) =_c + \ \wedge\ (\mathrm{S}\ \textsc{acm}) =_c \textsc{rec} \quad \rightarrow \quad (\mathrm{S}\ \textsc{case acc}) = +$

[2]Note that SC is largely a bookkeeping feature, which may be avoided at the cost of complicating the analysis.

[3]We ignore here case assignment to adjuncts; see Przepiórkowski 1999 for an extensive discussion.

[4]Implication is understood here as in Andrews and Manning 1993, pp. 17–18 (they in turn give credit to Ron Kaplan and John Maxwell), and Bresnan 2000, p. 62, i.e., $A \rightarrow B$ is equivalent to $\neg A \vee (A_c \wedge B)$, where A_c is the constraining (‘nonconstructive’) version of A.

As implied above, "SC $=_c$ +" distinguishes arguments assigned case via syntactic statements from those inherently case marked or not case marked at all. Moreover, ACM represents *accommodability*, a lexical feature introduced for Polish by Bień and Saloni (1982) to distinguish numeral forms governing the genitive noun (the value of ACM in such cases is REC) from numeral forms agreeing with the following noun (CONGR) (see also Przepiórkowski and Patejuk 2012 in these proceedings).

4 Problem

Like many other languages, Polish allows for the coordination of unlikes, e.g.:

(8)

Janka	szokowała	umowa	ACTA	i	że	polski
Janek.ACC	shocked.3.SG.F	agreement.NOM.SG.F	ACTA	and	that	Polish

rząd	ją	w	ogóle	podpisał.
government	it	at all		signed

'Janek was shocked by the ACTA agreement and the fact that Polish government signed it at all.'

(9)

Janek	i	jego	pięć	córek	głosowali
Janek.NOM.SG.M	and	his	five.ACC.PL.F	daughters.GEN.PL.F	voted.3.PL.M

przeciw	ACTA.
against	ACTA

'John and his five daughters voted against ACTA.'

As shown in (8) above, involving verb agreement with the closest conjunct,[5] the subject of the verb *szokować* 'shock' may be nominal or sentential, so it may also be realised by a coordinated structure containing an NP – apparently receiving the nominative case via the first statement above – and a CP. In (9), on the other hand, the subject of the verb may only be nominal (in the broad sense of the word), but just as in case of other verbs taking structurally-cased subjects, it may be realised by a noun phrase or by a numeral phrase, among others, so it may also be realised as a coordination of an NP and a NumP, which should be assigned case, respectively, via the statements (7a–b) above.

Unfortunately, given current LFG assumptions, these statements do not in fact handle such cases of unlike coordination. In order to see the problem more clearly, let us simplify (7) to (10), assuming for a moment that we deal with structurally case-marked elements only (i.e., that "SC $=_c$ +" is true):

(10) a. $(S\ \text{ACM}) \neq \text{REC} \quad \rightarrow \quad (S\ \text{CASE NOM}) = +$

b. $(S\ \text{ACM}) =_c \text{REC} \quad \rightarrow \quad (S\ \text{CASE ACC}) = +$

[5]Polish is a relatively free word order language with dominating SVO order, but (8) happens to exhibit the OVS order.

Let us now try to apply the simplified statements (10) to the assumed structure of (9). Without going into details, "(S ACM) $=_c$ REC" is either true or false, so exactly one of the antecedents in the statements (10a–b) is true.[6] Let us assume that "(S ACM) $=_c$ REC" is false. This means that (10b) has no effect (as the antecedent is false), while (10a) has the effect of requiring all conjuncts to be nominative (because CASE is distributive). This is contrary to fact, as one conjunct in (9) – *Janek* – is nominative, and the other one – *jego pięć córek* – is accusative.

A similar reasoning can be carried out with the assumption that "(S ACM) $=_c$ REC" is true – in such a case both conjuncts in (9) are required to bear the accusative case.

Which of these two possible assumptions is true here? The accommodability feature ACM is never assigned syntactically, it is a lexical feature of a class of numerals, so it does not make much sense on a coordinate structure as a whole, i.e., it should rather be considered a distributive feature. Then, "(S ACM) $=_c$ REC" would require that both conjuncts have the appropriately valued ACM feature, which is not true for (9), as the noun phrase *Janek* has no ACM. Hence, "(S ACM) $=_c$ REC" is false here, and both conjuncts are assigned the nominative via (10a).

Let us now return to the original subject case assignment statements, as given in (7). On the reasonable assumption that SC is a distributive feature valued "+" on both broadly nominal conjuncts in (9), "(S SC) $=_c$ +" is true, and the discussion based on (10) carries over. On the other hand, the statement "SC $=_c$ +" distinguishes between the two conjuncts in (8): the nominal phrase *umowa ACTA* satisfies it, while the clause *że polski rząd ją w ogóle podpisał* has no SC feature at all. The latter implies that "(S SC) $=_c$ +" is false for the whole coordinate construction in (8), so the antecedents of both statements (7a–b) are false, and the statements are vacuously true without the constructive consequent having any effect. This means that *umowa ACTA* is not constrained to be nominative and could bear any case, contrary to fact.

Note that the problem is not limited to subject positions. Consider (11) below (from Kallas 1993, p. 93, translation and glosses ours), involving coordination between an accusative noun, *wyjazd*, and a finite clause, *żeby nie wracał*.

(11) Doradził mu wyjazd i żeby nie wracał.
advised him.DAT departure.ACC and that not return
'(He) advised him to leave and not to come back.'

The last two bullets of (6) may be formalised in a way similar to the statements in (7), giving rise to the same problems in case of unlike coordination in the object position, as in (11).

[6] Apparently, in XLE, the platform for implementing LFG grammars (Maxwell and Kaplan 1996; `http://www2.parc.com/isl/groups/nltt/xle/`), distributivity and negation are encoded in a way that makes both "(S ACM) $=_c$ REC" and "(S ACM) $\neq$ REC" false when applied to a coordinate structure with one conjunct satisfying "(S ACM) $=_c$ REC" and the other having no ACM feature and, hence, satisfying "(S ACM) $\neq$ REC". Under this interpretation, both constraints in (10a–b) – and in (7a–b) – would be vacuously satisfied, without any constructive effect. Note that this still leads to an undesirable interpretation of case assignment constraints.

To summarise, the intuitively clear case assignment statements such as (7) fail in cases of unlike coordination. In particular, instead of meaning "for each conjunct: if the conjunct is structural and numeral, it must be accusative", (7b) currently means "if all conjuncts are structural and all are numeral then all must be accusative" (and analogously for (7a)).

5 Liberal solution

In order to handle structural case assignment in Polish (and, no doubt, many other languages) naturally, we propose to understand (non-)distributivity not as a property of features, but as a property of statements.

More precisely, we rewrite (7a–b) above as (12a–b) below, which should be read as "the f-structure(s) referred to as ($\uparrow$ SUBJ) must satisfy the following properties…".

(12) a. ($\uparrow$ SUBJ): (SC $=_c$ + $\wedge$ ACM $\neq$ REC $\rightarrow$ CASE NOM $=$ +)
b. ($\uparrow$ SUBJ): (SC $=_c$ + $\wedge$ ACM $=_c$ REC $\rightarrow$ CASE ACC $=$ +)

By default, all such statements are understood as distributive with respect to the path specified before ":" (i.e., with respect to "($\uparrow$ SUBJ)" in (12)), while non-distributive statements are explicitly marked as such. This means that, in case of, say, (12b), if the value of "($\uparrow$ SUBJ)" is a vanilla feature structure, it must satisfy the condition "(SC $=_c$ + $\wedge$ ACM $=_c$ REC $\rightarrow$ CASE ACC $=$ +)" as a whole, but if it is a hybrid feature structure, each element of the set represented by this hybrid feature structure must satisfy this condition. This ensures that the implication is applied to each conjunct separately, giving the desired results: if the conjunct is structural and numeral (in the appropriate sense), it will be assigned the accusative case; otherwise (12b) has no effect.

An interesting consequence of this proposal is that a given feature may behave distributively in some ways and non-distributively in others. This seems to be required for the full analysis of examples like (9) above, repeated as (13) below.

(13) Janek i jego pięć córek głosowali przeciw ACTA.
Janek.NOM.SG.M and his five.ACC.PL.F daughters voted.3.PL.M against ACTA
'John and his five daughters voted against ACTA.'

In Polish, as in other Indo-European languages, verbs only agree with nominative subjects, otherwise occurring in the default 3rd person singular neuter form. In particular, sentential subjects and accusative numeral subjects trigger such "default agreement"; compare (13) with (14) below.

(14) Pięć córek głosowało przeciw ACTA.
five.ACC.PL.F daughters voted.3.SG.N against ACTA
'(The) five daughters voted against ACTA.'

Now, if CASE were always distributive, then the subject in (13) should be caseless, so the verb should be in the default 3.SG.N form, as in (14). Note that, unlike in (8), (13) does not involve agreement with the closest conjunct: the form of the verb is plural masculine, unlike either of the two conjuncts. Rather, the verb agrees with the coordinated phrase as a whole, which bears the features of gender and number resolved to masculine and plural, just as in many other languages (cf., e.g., Wechsler and Zlatić 2003 and Dalrymple and Kaplan 2000). But in order to keep the generalisation that verbs only agree with nominative subjects, the coordinated phrase as a whole must also bear a non-distributive feature CASE equal to NOM. In effect, we assume the f-structure of the subject in (13) as in (15) below:

(15) $$\begin{bmatrix} \textsc{case} & \begin{bmatrix}\textsc{nom} & +\end{bmatrix} \\ \textsc{number} & \textsc{pl} \\ \textsc{gender} & \textsc{masc} \\ \multicolumn{2}{l}{\left\{ \begin{bmatrix} \textsc{case} & \begin{bmatrix}\textsc{nom} & +\end{bmatrix} \\ \textsc{number} & \textsc{sg} \\ \textsc{gender} & \textsc{masc} \end{bmatrix}, \begin{bmatrix} \textsc{case} & \begin{bmatrix}\textsc{acc} & +\end{bmatrix} \\ \textsc{number} & \textsc{pl} \\ \textsc{gender} & \textsc{f} \end{bmatrix} \right\}} \end{bmatrix}$$

One way to formalise these observations is to posit a default *non-distributive* statement of nominative case assignment to subjects (where '@' marks non-distributivity of the statement):

(16) $(^{@}(\uparrow \textsc{subj}) \colon (\textsc{case nom} = +))$

Because this is a default statement (as indicated by the outer parentheses), it does not conflict with the explicitly accusative case of numeral subjects or the explicitly caseless clausal subjects, but at the same time it expresses the prevalent intuition that Polish subjects are nominative.[7]

6 Conservative solution (by Mary Dalrymple)

There is, however, a solution which does not require extending the formal apparatus of LFG, although it is based on a relatively rarely used LFG mechanism, namely, the so-called off-path constraints (Dalrymple 2001, p. 148).[8]

Off-path constraints make it possible to restrict the path (or, more importantly, its part) used by other statements. For example, while the minimal feature structure satisfying (17a) is that of (17b), the statement (18a), with an off-path constraint added to the attribute A, specifies (18b).[9]

(17) a. $(\uparrow \textsc{a b c}) =_c +$ b. $\begin{bmatrix} \textsc{a} & \begin{bmatrix} \textsc{b} & \begin{bmatrix} \textsc{c} & + \end{bmatrix} \end{bmatrix} \end{bmatrix}$

[7]See, e.g., an agitated defence of this position in Saloni 2005.

[8]This solution was suggested to us by Mary Dalrymple after our presentation of the analysis of Section 5 at the LFG 2012 conference in Denpasar.

[9]Note that off-path constraints are written *below* the attribute to which they apply.

(18) a. $(\uparrow \underset{(\leftarrow\ \text{D}) =_c \text{E}}{\text{A}}\ \text{B C}) =_c +$ b. $\left[\begin{array}{ll} \text{A} & \left[\text{B} \quad \left[\text{C} \quad +\right]\right] \\ \text{D} & \text{E} \end{array}\right]$

More formally, '←' denotes the f-structure which contains the attribute to which it is attached, while '→' denotes the f-structure which is the value of the attribute to which it is attached. Hence, (19a) (i.e., with '←' above replaced by '→') specifies the structure in (19b).

(19) a. $(\uparrow \underset{(\rightarrow\ \text{D}) =_c \text{E}}{\text{A}}\ \text{B C}) =_c +$ b. $\left[\text{A} \quad \left[\begin{array}{ll} \text{B} & \left[\text{C} \quad +\right] \\ \text{D} & \text{E} \end{array}\right]\right]$

Now, the idea of Dalrymple's solution is to attach the constraint that should be distributed to all conjuncts, e.g., the constraint "SC $=_c$ + ∧ ACM $=_c$ REC → CASE ACC = +", to a distributive feature guaranteed to be present on all conjuncts. A new feature could be created specifically for this purpose, but it is also possible to recycle the standard feature PRED:

(20) $(\uparrow \text{SUBJ} \underset{(\leftarrow\ \text{SC}) =_c +\ \wedge\ (\leftarrow\ \text{ACM}) =_c \text{REC}\ \rightarrow\ (\leftarrow\ \text{CASE ACC}) = +}{\text{PRED}})$

(20) says that – in Polish – there are no semantically vacuous (expletive) subjects, i.e., each subject has a PRED value. This part of the statement is trivial. The main import of the statement is given in the off-path constraint part: for each such PRED, if the value of SC (at the same level as the PRED) is "+" and the value of ACM (again, at the same level) is "REC", then the value of CASE ACC (again, starting from the same level as PRED) must be "+". This way the whole implication is interpreted independently for each conjunct.

7 Comparison and Conclusion

The aim of this article was to demonstrate the limits of the standard approach to distributivity in coordination. The main problem concerns the possibility of distribution of complex statements to all conjuncts in a hybrid feature structure. One solution, discussed in Section 5, is to extend the LFG formalism so that such non-trivial distribution can be stated explicitly. The obvious disadvantage of this solution is the need to tweak the well-established machinery of LFG.

A more conservative solution, due to Mary Dalrymple, is to let such complex statements "piggyback" on distributive features, e.g., on PRED, as discussed in Section 6. This recycles the mechanism of off-path constraints and makes it possible to retain the standard LFG approach to distributivity, where it is some features, e.g., CASE, not statements, that are distributive. While this solution seems rather technical, it solves the basic problem stated in Section 4.

Ultimately, the choice between the two solutions must be made on the basis of empirical facts, and the possibility of a regular subject-verb agreement with coordinated subjects, where one of the conjuncts is an accusative numeral phrase (see the discussion at the end of Section 5), seems to favour the more liberal solution, which allows for the whole coordinate structure to be nominative even if one of the conjuncts is accusative. But this preference is only as strong as the generalisation that subject-verb agreement in Polish (and other Indo-European languages) involves solely nominative subjects; a conservative analysis violating this generalisation is readily available.[10]

Acknowledgements

This research is supported by the POIG.01.01.02-14-013/09 project which is co-financed by the European Union under the European Regional Development Fund. The publication of the resulting implemented grammar is supported by the CESAR project (European project CIP ICT-PSP 271022, part of META-NET).

The authors are grateful to Tracy Holloway King for valuable comments on the previous version of this paper.

References

Andrews, Avery D. and Manning, Christopher D. 1993. Information Spreading and Levels of Representation in LFG. CSLI Technical Report CSLI-93-176, Stanford University, Stanford, CA.

Bień, Janusz S. and Saloni, Zygmunt. 1982. Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna). *Prace Filologiczne* XXXI, 31–45.

Bresnan, Joan. 2000. *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics, Blackwell.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Dalrymple, Mary. 2001. *Lexical-Functional Grammar*. Academic Press.

Dalrymple, Mary and Kaplan, Ronald M. 2000. Feature Indeterminacy and Feature Resolution. *Language* 76(4), 759–798.

Dalrymple, Mary, King, Tracy Holloway and Sadler, Louisa. 2009. Indeterminacy by underspecification. *Journal of Linguistics* 45, 31–68.

Heinz, Wolfgang and Matiasek, Johannes. 1994. Argument Structure and Case Assignment in German. In Nerbonne et al. (1994), pages 199–236.

Kallas, Krystyna. 1993. *Składnia współczesnych polskich konstrukcji współrzędnych*. Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika.

[10] See also Przepiórkowski and Patejuk 2012 (in these proceedings) for an analysis of another phenomenon where the non-distributivity of CASE is crucially assumed.

Kuryłowicz, Jerzy. 1949. Le problème du classement des cas. *Biuletyn Polskiego Towarzystwa Językoznawczego* 9, 20–43.

Levy, Roger and Pollard, Carl. 2001. Coordination and Neutralization in HPSG. In Frank van Eynde, Lars Hellan and Dorothee Beermann (eds.), *Proceedings of the HPSG 2001 Conference*, pages 221–234, Stanford, CA: CSLI Publications.

Maxwell, III, John T. and Kaplan, Ronald M. 1996. An Efficient Parser for LFG. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG'96 Conference*, Grenoble.

Nerbonne, John, Netter, Klaus and Pollard, Carl (eds.). 1994. *German in Head-Driven Phrase Structure Grammar*. CSLI Lecture Notes, No. 46, Stanford, CA: CSLI Publications.

Pollard, Carl. 1994. Toward a Unified Account of Passive in German. In Nerbonne et al. (1994), pages 273–296.

Przepiórkowski, Adam. 1996. Case Assignment in Polish: Towards an HPSG Analysis. In Claire Grover and Enric Vallduví (eds.), *Studies in HPSG*, volume 12 of *Edinburgh Working Papers in Cognitive Science*, pages 191–228, Centre for Cognitive Science, University of Edinburgh.

Przepiórkowski, Adam. 1999. *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Ph.D. dissertation, Universität Tübingen, Tübingen.

Przepiórkowski, Adam and Patejuk, Agnieszka. 2012. The puzzle of case agreement between numeral phrases and predicative adjectives in Polish. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG'12 Conference*, pages 490–502, Stanford, CA: CSLI Publications.

Rouveret, Alain and Vergnaud, Jean-Roger. 1980. Specifying Reference to the Subject: French causatives and conditions on representations. *Linguistic Inquiry* 11(1), 97–202.

Saloni, Zygmunt. 2005. O przypadkach w języku polskim (na marginesie artykułu Adama Przepiórkowskiego). *Biuletyn Polskiego Towarzystwa Językoznawczego* LXI, 27–48.

Vergnaud, Jean-Roger. 1982. *Dépendances et Niveaux de Representation en Syntaxe*. Ph. D. dissertation, Université Paris 7.

Wechsler, Stephen and Zlatić, Larisa. 2003. *The Many Faces of Agreement*. Stanford, CA: CSLI Publications.

THE PUZZLE OF CASE AGREEMENT BETWEEN NUMERAL PHRASES AND PREDICATIVE ADJECTIVES IN POLISH

Adam Przepiórkowski and Agnieszka Patejuk
Institute of Computer Science, Polish Academy of Sciences

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper addresses the optionality of case agreement between a numeral phrase in the subject position and its modifying or predicating adjectives in Polish: such adjectives either agree with the numeral or – apparently – reach into the numeral phrase and agree with the noun phrase within it. While previous analyses of this phenomenon postulated special agreement mechanisms, we account for the troublesome facts by assimilating Polish numeral phrases to coordinate structures.

1 Introduction

The puzzle addressed in this paper concerns the two agreement patterns exhibited in Polish constructions such as (1).

(1) Ostatnie dziesięć lat było fatalne / fatalnych.
last.ACC ten.ACC years.GEN was.3.SG.N terrible.ACC/GEN
'The last ten years were terrible.'

On the assumption, justified below, that *ostatnie dziesięć lat* is an accusative subject in (1), the accusative form of the predicative adjective *fatalne* is expected, but the genitive form, *fatalnych*, is completely surprising – it appears as if the predicative adjective *fatalnych* looked into the subject numeral phrase to agree in case with the genitive noun *lat* within it.

To the best of our knowledge, all previous analyses, when they get the facts right, stipulate special agreement machinery to account for such constructions. In this paper, on the other hand, we propose an LFG account which does not assume any new agreement mechanisms. Rather, what is special about constructions exemplified in (1) is the internal structure of Polish numeral phrases, which is essentially the structure assumed in LFG analyses of coordination.

The outline of this paper is as follows. Section 2 justifies various assumptions concerning Polish numeral phrases in the subject position. Then, Section 3 presents the problem at hand in a more detailed way and Section 4 provides an LFG analysis. Finally, Section 5 sketches previous attempts at dealing with this problem and Section 6 concludes the paper.

2 Assumptions

When talking about Polish numerals, we restrict ourselves to a morphosyntactically defined subclass of cardinal numerals, i.e., those which have number (always plural) and inflect for case and gender. So, for example, ordinal numerals, which *inflect* for number, are excluded from our considerations.

Such cardinal numerals may agree in case with the accompanying noun or they may require the noun to occur in the genitive case. In Polish linguistics, a new category, *accommodability*, was introduced (Bień and Saloni 1982) to distinguish between the two classes of numeral forms: the value of this category is CONGR in case of agreeing numerals and REC in case of numerals governing the genitive case. Roughly speaking, governing numeral forms are:

- all forms of non-paucal (5 and above) numerals and some forms of paucal (2–4) numerals
- which occur in – loosely speaking – accusative and apparently nominative positions.[1]

The pattern exemplified in (1) involves governing numerals in subject positions. It is completely uncontroversial that such numeral phrases are in fact subjects: they bind anaphors, participate in control constructions and may be coordinated with nominative subjects.

It is slightly less clear that they are really headed by the numeral. One argument is that only the noun may be elided in such constructions, as shown below:

(2) a. Pięć kobiet stało.
five.F women.F.GEN stood.3.SG.N
'Five women were standing.'

b. Pięć stało.
five.F stood.3.SG.N
'Five were standing.'

c. *Kobiet stało.
women.F.GEN stood.3.SG.N
'Women were standing.' (putative)

Another argument is that phrases such as *ostatnie dziesięć lat* (in (1)) and *pięć kobiet* (in (2)a) cannot occur in genitive positions, which would be surprising if the genitive noun were the head. On the other hand, they can occur in accusative positions, which is expected if the accusative numeral is the head.

This brings us to the last assumption to be introduced here, namely, that such numeral subjects are in fact accusative, not nominative.[2] This has been noted repeatedly at least since Małecki 1863,[3] and is defended at length in Przepiórkowski 1999 (within HPSG). One argument comes from examples like (3)–(4), where the sudoku-like puzzle posed by the syncretic case forms may be satisfactorily solved only if the numeral bears the accusative case and agrees with the accusative feminine (F.ACC) *te* in (3) and the accusative human-masculine (HM.ACC) *tych* in (4).

[1]It is largely a matter of convention whether numeral forms in genitive positions, co-occurring with genitive nouns, should be analysed as agreeing or governing, but see the HPSG analysis of Przepiórkowski 1999, § 5.3.1.3, which implies that some such occurrences of genitive numerals should be analysed as governing, and others as agreeing.

[2]We leave unresolved the question of whether the numeral has a nominative form at all. In the XLE implementation of the grammar of Polish, Patejuk and Przepiórkowski 2012, we assume that such governing numerals are defective and do not have a nominative form.

[3]Other (numerous) bibliographical references are provided in Przepiórkowski 2004.

(3) {Tych / te} pięć kobiet stało.
these.F.**GEN** these.F.NOM/**ACC** five.F.NOM?/**ACC** women.**GEN** stood.3.SG.N
'These five women were standing.'

(4) {Tych / *ci} pięciu mężczyzn
these.HM.**ACC/GEN** these.HM.NOM five.HM.NOM?/**ACC**/GEN men.**GEN**
stało.
stood.3.SG.N
'These five men were standing.'

Tych in (3) is also possible and reflects the agreement with the genitive noun; this implies that (4) is structurally ambiguous, with *tych* agreeing either with the accusative numeral or with the genitive noun. Crucially, the common assumption that Polish nominal subjects are always in the nominative is contradicted by the ungrammaticality of (4) with the unambiguously nominative *ci*. Hence, in the two examples above, the possible case values are those marked in bold.

The conclusion that *pięć* in (3) and *pięciu* in (4) are accusative is corroborated by the default (non-agreeing) features on the verb, which patterns with non-nominative subjects in other Indo-European languages, e.g., Icelandic, see (5):

(5) Drengina vantar mat. (Andrews 1982)
boys-ACC lack-3.SG.N food-ACC
'The boys lack food.'

Also, in the process of numeralisation, it is the accusative form of the noun that becomes a numeral, as in (6), where the nominative noun *masa* in (6a) enters into the subject-verb agreement, but the accusative *masę* in (6b) shows the default agreement patterns typical of numeral subjects.[4]

(6) a. Masa ludzi przyszła / *przyszło.
mass.NOM people.GEN came.3.SG.F came.3.SG.N
'Lots of people came.'

b. Masę ludzi przyszło / *przyszła.
mass.ACC people.GEN came.3.SG.N came.3.SG.F

3 Problem

The problem of dual agreement between a numeral phrase and its modifier, exemplified in (1), is not limited to predicative adjectives. In fact, (3) above shows similar duality with respect to an attributive adjectival form: accusative *te* and genitive *tych*.

In case of attributive modifiers the problem does not seem acute, as an explanation for the preposed *tych* in terms of simple word order rules within an NP or a

[4]Note that there is no separate part of speech *quantifier* in the repertoire of grammatical classes assumed here. Semantic quantifiers may be expressed as numerals and nouns, among other parts of speech.

Numeral Phrase is in principle possible. However, as already shown in (1), there are analogous facts involving predicative adjectives that do not seem to be amenable to such an NP-internal discontinuity analysis; see the attested examples below:[5]

(7) Następne kilkadziesiąt metrów było czyste.
next.ACC several tens.ACC metres.GEN was.3.SG.N clean.ACC
'The next few tens of metres were clean.'

(8) Pięć osób zostało rannych.
five.ACC persons.GEN became.3.SG.N wounded.GEN
'Five people were wounded.'

(9) Kolejnych jedenaście zarzutów było podobnych.
further.GEN eleven.ACC charges.GEN was.3.SG.N similar.GEN
'Further eleven charges were similar.'

While (7) illustrates the expected agreement between the accusative numeral phrase and the accusative predicative adjective *czyste*, (8)–(9) are totally unexpected: it seems as if the genitive predicative adjectives *rannych* and *podobnych* can reach into the subject numeral phrase and agree with the genitive noun within it.

Note that both in (7) and in (9) the attributive adjective and the predicative adjective bear the same case: accusative in (7) (*następne*, *czyste*) and genitive in (9) (*kolejnych*, *podobnych*). Examples following these ACC/ACC and GEN/GEN patterns may be easily found in the corpus, and the same holds for the ACC/GEN pattern, illustrated in (10). While rarer, the GEN/ACC pattern is also attested, cf. (11).[6]

(10) Kolejne pięćdziesiąt aut zostało uszkodzonych.
further.ACC fifty.ACC cars.GEN became.3.SG.N damaged.GEN
'Further fifty cars became damaged.'

(11) Minionych dwanaście miesięcy było najgorsze w historii.
past.GEN twelve.ACC months.GEN was.3.SG.N worst.ACC in history
'The past twelve months were the worst in history.'

It must be stressed that the possibility of dual agreement concerns only morphosyntactically numeral phrases headed by a governing numeral. The example (1) above, repeated as (12) below, should be contrasted with (13) and (14).

[5] While examples (3)–(4) are constructed (but uncontroversial), the following examples – with the exception of (16) from Kallas 1974 – are based on attested uses found in the National Corpus of Polish (`http://nkjp.pl/`).

[6] The query `[pos=adj] [pos=num] [pos=adj]* [pos=subst] (było | zostało) [pos=adj]` (see `http://nkjp.pl/poliqarp/help/en.html` for the query syntax) on the complete 1.8-billion-segment National Corpus of Polish gives 113 results, including the following numbers for the total of 34 true positives: 23 for the ACC/ACC pattern, 5 for ACC/GEN, 4 for GEN/GEN and 2 for GEN/ACC. Other examples may be found by changing the word order, considering adjectival participles instead of adjectives, etc.

(12) Ostatnie dziesięć lat było fatalne / fatalnych.
last.ACC ten.ACC years.GEN was.3.SG.N terrible.ACC/GEN
'The last ten years were terrible.'

(13) Tuzin kropel był przepisywany / *przepisywanych.
dozen.NOM.M drops.GEN was.3.SG.M prescribed.NOM/*GEN
'A dozen drops were prescribed.'

(14) Coś takiego jest potrzebne / *potrzebnego.
something.NOM such.GEN is.3.SG.N needed.NOM/*GEN
'Something like this is needed.'

Example (13) involves *tuzin*, a form of a lexeme which has a numeral meaning but is a morphosyntactic noun here, as evidenced by the singular masculine agreement with the verb.[7] Unlike in (12), only the adjectival form agreeing with *tuzin* is possible. Similarly, in (14), involving a special construction consisting of the indefinite non-human pronoun *coś* and an NP-internal genitive modifier, the external predicative adjective *potrzebne* must agree with the pronoun and, hence, cannot occur in the genitive.

4 Analysis

4.1 Idea

The main pre-theoretical idea of the analysis is that Polish numeral phrases of the kind considered here are somewhere between being single-headed, like typical NPs, and somewhat multi-headed, like coordinate structures. In other words, they seem to be 1.5-headed – single-headed for the purpose of being assigned case (they may occur only in accusative – not genitive – positions), but bi-headed for the purpose of agreement (they may agree with accusative and genitive adjectives).

In terms of LFG, we postulate that such phrases, like coordinate structures, are represented by a hybrid feature structure (Dalrymple and Kaplan 2000), where the numeral and the noun (with any immediate modifiers) are elements of the set encoded by such a structure. However, unlike in the case of coordination, one of the elements of this set, representing the numeral, is at the same time the whole hybrid structure, i.e., the relevant Polish numeral phrases are represented by cyclic structures of the form given schematically in (15).

[7] In fact, just as many other nouns of this kind, *tuzin* may also be used numeratively, in which case the genitive form of the adjective is possible:

(i) Tuzin kropel było przepisywanych.
dozen.ACC drops.GEN was3.SG.N prescribed.GEN
'A dozen drops were prescribed.'

Note that, due to the nominative/accusative syncretism of many inanimate masculine nouns in Polish, the nominative noun in (13) and the accusative denominal numeral in (i) have the same form: *tuzin*.

(15) $\boxed{1}\left[\left\{\boxed{1}\begin{bmatrix}\text{CAT} & \text{NUM}\\ \text{CASE} & \text{ACC}\end{bmatrix}, \boxed{2}\begin{bmatrix}\text{CAT} & \text{NOUN}\\ \text{CASE} & \text{GEN}\end{bmatrix}\right\}\right]$

Given such a structure, any case assignment or case checking mechanisms will target the whole feature structure, whose case is structure-shared with the numeral element, i.e., accusative (in constructions considered here).[8]

4.2 Empirical consequences

The real advantage of this analysis over alternatives mentioned in Section 5 is that nothing special needs to be said about the accusative / genitive optionality of case agreement with predicative adjectives: the accusative case marking on the predicative adjective, as in (7) or (11), represents agreement with the accusative numeral phrase, while the genitive marking in (8)–(10) is handled by whatever mechanism is responsible for the single conjunct agreement (e.g., Kuhn and Sadler 2007), a phenomenon which independently occurs in Polish (Kallas 1974):

(16) Pachniał wiatr i morze.
smelled.SG.M wind.SG.M and sea.SG.N
'Wind and sea smelled.'

Also, the analysis naturally extends to agreement with *attributive* adjectives, without the need to assume discontinuous structures of numeral phrases in the relevant variants of (3)–(4) or in (9) and (11).

One potential problem for this analysis is that, to the best of our knowledge, Polish linguistic literature only reports cases of *closest* conjunct agreement, as in (16), while some of the examples involving attributive adjectives, namely those just referred to, as well as some examples involving predicative adjectives, as in (17), involve agreement with the furthest element.

(17) Niezbędnych było dobre parę metrów
indispensable.GEN was3.SG.N good.ACC couple.ACC meters.GEN
kwadratowych wykładziny.
square.GEN carpet.GEN
'A good couple square meters of carpet were indispensable.'

However, cases of furthest conjunct agreement may be readily found in the National Corpus of Polish (see (18)), so we conclude that Polish syntax makes this option independently available.[9]

[8] We crucially assume here that CASE is not a distributive feature; cf. Przepiórkowski and Patejuk 2012.

[9] On the other hand, it should still be explained why this option is much more readily available in case of agreement with numeral phrases than in case of agreement with true coordinate structures; we leave this for future research.

(18) Ewentualna porażka lub remis kosztowałaby ich utratę
prospective.SG.F defeat.SG.F or draw.SG.M would cost.SG.F them loss
żółtej koszulki lidera.
yellow jersey leader
'A prospective defeat or draw would cost them the leader's yellow jersey.'

4.3 Technical details

4.3.1 Independent assumptions

For the LFG analysis of Polish case assignment (Patejuk and Przepiórkowski 2011), largely carried over from the previous HPSG analysis (Przepiórkowski 1999), we assume the distinction between structural and lexical case assignment. Lexical case assignment happens in the lexicon; for example, the verb POMAGAĆ 'help' lexically specifies its complement to be dative, and it will remain dative regardless of the structural configuration or the form of this verb. On the other hand, the verb WSPIERAĆ 'support' specifies its object as structurally case assigned; it will normally be accusative, but it will be realised as genitive in the syntactic scope of negation or when the form of the verb is gerundial.

For structurally case assigned ("SC = +") subjects, we assume a simple statement which says that the case of such subjects depends on their category: normally it is nominative, but when the subject is a numeral phrase headed by a governing numeral ("ACM = REC"; cf. the *accommodability* category introduced in Section 2 above), it is accusative.

Technically, we formalise this statement as two implications:[10]

(19) a. $(\uparrow \text{SUBJ})$: $(\text{SC} =_c + \ \wedge \ \text{ACM} \neq \text{REC} \ \ \rightarrow \ \ \text{CASE NOM} = +)$

b. $(\uparrow \text{SUBJ})$: $(\text{SC} =_c + \ \wedge \ \text{ACM} =_c \text{REC} \ \ \rightarrow \ \ \text{CASE ACC} = +)$

Note that here and henceforth, unlike in the schematic (15) above, we assume the representation of case proposed in Dalrymple et al. 2009. According to this representation, and given the 7 morphological cases in Polish, an unambiguously nominative noun phrase such as *ewentualna porażka* in (18) will have the CASE value as shown in (20) below (instead of the atomic NOM):

(20) $\begin{bmatrix} \text{NOM} & + \\ \text{ACC} & - \\ \text{GEN} & - \\ \text{DAT} & - \\ \text{INST} & - \\ \text{LOC} & - \\ \text{VOC} & - \end{bmatrix}$

[10]See Przepiórkowski and Patejuk 2012 for details and justification.

More importantly, as independently justified at length in Przepiórkowski and Patejuk 2012 (in these proceedings), we assume a more subtle approach to the distributivity of features, where it is not features that are distributive, but statements.[11] In particular, each of the two implications in (19) must hold for each element in a hybrid feature structure separately.

In case of the numeral element, (19a) applies vacuously, because the antecedent is false ("$\text{ACM} \neq \text{REC}$" is false, as the numeral *is* governing), and (19b) applies non-vacuously and assigns the accusative case.

In case of the nominal element, both clauses apply only vacuously, because the noun is not structurally case marked (so "$\text{SC} =_c +$" is false) – instead, the genitive is checked by the relevant numeral phrase rule (see (21a) in § 4.3.2 below).

Note that, since the numeral element is structure-shared with the whole hybrid feature structure, the numeral phrase as such is unambiguously accusative. But since we assume that CASE is *not* a distributive feature (no feature is distributive by itself), this accusative case does not distribute to the nominal element, so there is no feature clash.

4.3.2 Structure of numeral phrases

After introducing these independently needed assumptions, the only special part of the analysis is the c-structure rule (21a), which gives rise to cyclic f-structures such as (21b), headed by the governing numeral, containing a genitive NP and occurring in structurally case marked positions.

(21) a.

$$\begin{array}{lccc} \text{NumP} & \rightarrow & \text{Num} & \text{NP} \\ & & (\downarrow \text{ACM}) =_c \text{REC} & (\downarrow \text{CASE GEN}) =_c + \\ & & (\downarrow \text{SC}) =_c + & \downarrow \in \uparrow \\ & & \uparrow = \downarrow & \\ & & \downarrow \in \uparrow & \end{array}$$

b.

$$\boxed{1}\begin{bmatrix} \text{ACM} & \text{REC} \\ \text{SC} & + \\ \multicolumn{2}{l}{\left\{ \boxed{1}, \boxed{2}\begin{bmatrix} \text{CASE} & \begin{bmatrix} \text{GEN} & + \end{bmatrix} \end{bmatrix} \right\}} \end{bmatrix}$$

When such a structure is the value of SUBJ, the statement (19b) ensures that the numeral phrase is in the accusative case, as explained in § 4.3.1.

Let us note that cyclicity is not a frequent feature of LFG analyses, but a very similar structure is proposed in Fang and Sells 2007, p. 209, to account for the Chinese verb copy construction:[12]

[11] By default, all statements are distributive; non-distributive ones are marked explicitly.

[12] Fang and Sells 2007, fn. 6, attribute this solution to Ron Kaplan, whom we would like to thank for pointing out to us the similarity between the two analyses.

(22) $$\begin{array}{llll} \text{VP(VCC)} & \rightarrow & \text{VP} & \text{VP+} \\ & & \downarrow\in\uparrow & \downarrow\in\uparrow \\ & & \downarrow=\uparrow & \end{array}$$

While *cyclicity* is never mentioned in Fang and Sells 2007, rule (22) gives rise to cyclic structures in the same way as (21a) does. Fang and Sells (2007), who assume the usual approach to the distributivity of features, use such cyclic structures to ensure that all non-initial VPs in the construction at hand "inherit" all relevant features from the first VP.

5 Previous attempts

It might be tempting to analyse numeral phrases as bi-headed or as structurally ambiguous, i.e., alternatively headed by the numeral and the noun. Such accounts are considered and rejected on various grounds in Przepiórkowski 2001 and we will not repeat this discussion here.

The alternative analysis proposed there assumes instead that the genitive noun is the SUBJECT of the numeral and that subjects are "visible" outside of their phrases.[13] Given this assumption, case agreement between a predicative adjective and its subject (structure-shared with the numeral subject of the copula) is formalised via a disjunctive constraint stating that the adjective agrees either with its subject (the accusative numeral phrase) or its subject's subject (the genitive noun).

Formally, case agreement is invoked in the two principles (23)–(24), and it is defined in (25).

(23) Attributive case agreement:

$$\begin{bmatrix} \textit{head} \\ \text{CASE } \boxed{1} \\ \text{MOD}|\text{LOC } \boxed{2}\begin{bmatrix}\text{CAT}|\text{HEAD}|\text{CASE } \boxed{0}\end{bmatrix} \end{bmatrix} \rightarrow \texttt{case-agreement}(\boxed{1}\texttt{, }\boxed{2})$$

(24) Predicative case agreement:

$$\begin{bmatrix} \textit{category} \\ \text{HEAD } \begin{bmatrix}\text{CASE } \boxed{1} \\ \text{PRD } + \end{bmatrix} \\ \text{SUBJ } \left\langle\begin{bmatrix}\text{LOC } \boxed{2}\begin{bmatrix}\text{CAT}|\text{HEAD}|\text{CASE } \boxed{0}\end{bmatrix}\end{bmatrix}\right\rangle \end{bmatrix} \rightarrow \texttt{case-agreement}(\boxed{1}\texttt{, }\boxed{2})$$

(25) Definition of case agreement:

$$\texttt{case-agreement}(\boxed{1}\,\textit{case}\texttt{, }\boxed{2}\,\textit{local}) \leftrightarrow$$
$$(\boxed{2} = \begin{bmatrix}\text{CAT}|\text{HEAD}|\text{CASE } \boxed{1}\end{bmatrix} \vee$$
$$\boxed{2} = \begin{bmatrix} \text{CAT}|\text{ARG-ST } \left\langle \begin{bmatrix}\text{CASE } \boxed{1} \\ \text{INDEX } \boxed{3}\end{bmatrix}, \ldots \right\rangle \\ \text{CONT}|\text{INDEX } \boxed{3} \end{bmatrix})$$

[13]While this assumption is natural in LFG, it is controversial on the strong view of locality held in HPSG, but see Sag 2007 for discussion.

In particular, according to (24)–(25), case agreement between a case-bearing phrase and its predicative head means the structure sharing of case values of the predicative element on one hand and *either* the phrase's case *or* the case of the phrase's subject (initial element on its ARG-ST list) on the other, the latter taking place only when the phrase and its subject have the same INDEX value (as arguably the numeral and its noun complement do).

Although this HPSG analysis may be carried over to LFG, and it still seems empirically adequate, it is rather unsatisfactory in the sense that, in order to explain a very specific construction involving subject numeral phrases, it proposes a more complicated (and disjunctive) general mechanism of case agreement. The current analysis seems to be theoretically more satisfying, as it provides an equally empirically adequate account in terms of a rather special structure of relevant numeral phrases, and leaves general agreement mechanisms untouched.

6 Conclusion

Although this paper deals with a very special parochial construction of Polish, the analysis is based on a couple of ideas that may be of a broader theoretical interest. First of all, we propose to extend the use of hybrid structures, previously employed in analyses of coordination, to the representation of a class of numeral phrases in Polish. Second, the analysis relies on the possibility of CASE being non-distributive. Third, we show how relevant case assignment and agreement facts can be dealt with by the assumption that hybrid feature structures representing numeral phrases are cyclic, i.e., that one of the elements of the set represented by the hybrid structure is the structure itself.

It needs to be noted, however, that these three aspects of the analysis have a very different standing. The assimilation of a class of numeral phrases to coordinate structures is crucial for the current analysis, as it makes it possible to recycle the standard agreement mechanisms, including the single conjunct agreement. What is also crucial is the possibility that a hybrid feature structure may bear a CASE value different from that of one of the elements in the set represented by the feature structure, i.e., we assume that CASE is not a distributive feature – perhaps it is not features that are distributive, but statements, as argued in Przepiórkowski and Patejuk 2012 (in these proceedings).

On the other hand, cyclicity is not a necessary feature of the current account: what is crucial is that the morphosyntactic features of the numeral, especially its CASE, be the same as those of the complete numeral phrase, and this can be ensured by equating just the values of the relevant features, without equating the whole feature structures. Nevertheless, we believe that Polish numeral phrases are rather naturally modelled as cyclic feature structures and that the proposed analysis opens the question of the place of such cyclic structures in LFG.

Acknowledgements

This research is supported by the POIG.01.01.02-14-013/09 project which is co-financed by the European Union under the European Regional Development Fund. The publication of the resulting implemented grammar is supported by the CESAR project (European project CIP ICT-PSP 271022, part of META-NET).

The authors are grateful to Tracy Holloway King for valuable comments on the previous version of this paper.

References

Andrews, Avery D. 1982. The Representation of Case in Modern Icelandic. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, MIT Press Series on Cognitive Theory and Mental Representation, pages 427–503, Cambridge, MA: The MIT Press.

Bień, Janusz S. and Saloni, Zygmunt. 1982. Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna). *Prace Filologiczne* XXXI, 31–45.

Butt, Miriam and King, Tracy Holloway (eds.). 2007. *The Proceedings of the LFG'07 Conference*, Stanford, CA.

Dalrymple, Mary and Kaplan, Ronald M. 2000. Feature Indeterminacy and Feature Resolution. *Language* 76(4), 759–798.

Dalrymple, Mary, King, Tracy Holloway and Sadler, Louisa. 2009. Indeterminacy by underspecification. *Journal of Linguistics* 45, 31–68.

Fang, Ji and Sells, Peter. 2007. A formal analysis of the verb copy construction in Chinese. In Butt and King (2007), pages 198–213.

Kallas, Krystyna. 1974. O zdaniach *Pachniał wiatr i morze.*, *Andrzej i Amelia milczeli. Studia z Filologii Polskiej i Słowiańskiej* XIV, 57–71.

Kuhn, Jonas and Sadler, Louisa. 2007. Single Conjunct Agreement and the Formal Treatment of Coordination in LFG. In Butt and King (2007), pages 302–322.

Małecki, Antoni. 1863. Gramatyka języka polskiego. Większa. Lwów.

Patejuk, Agnieszka and Przepiórkowski, Adam. 2011. Przypadki strukturalne w polskiej gramatyce LFG. In Mirosław Bańko and Dorota Kopcińska (eds.), *Różne formy, różne treści: Tom ofiarowany Profesorowi Markowi Świdzińskiemu*, pages 183–193, Warsaw: Uniwersytet Warszawski, Wydział Polonistyki.

Patejuk, Agnieszka and Przepiórkowski, Adam. 2012. Towards an LFG parser for Polish: An exercise in parasitic grammar development. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 3849–3852, ELRA, Istanbul, Turkey.

Przepiórkowski, Adam. 1999. *Case Assignment and the Complement-Adjunct Di-*

chotomy: A Non-Configurational Constraint-Based Approach. Ph.D. dissertation, Universität Tübingen, Tübingen.

Przepiórkowski, Adam. 2001. ARG-ST on Phrases: Evidence from Polish. In Dan Flickinger and Andreas Kathol (eds.), *Proceedings of the 7th International Conference on Head-Driven Phrase Structure Grammar*, pages 267–284, Stanford, CA: CSLI Publications.

Przepiórkowski, Adam. 2004. O wartości przypadka podmiotów liczebnikowych. *Biuletyn Polskiego Towarzystwa Językoznawczego* LX, 133–143.

Przepiórkowski, Adam and Patejuk, Agnieszka. 2012. On case assignment and the coordination of unlikes: The limits of distributive features. In Miriam Butt and Tracy Holloway King (eds.), *The Proceedings of the LFG'12 Conference*, pages 479–489, Stanford, CA: CSLI Publications.

Sag, Ivan A. 2007. Remarks on Locality. In Stefan Müller (ed.), *Proceedings of the HPSG 2007 Conference*, pages 394–414, Stanford, CA: CSLI Publications.

ON THE SYNTAX OF SOME APPARENT SPATIAL PARTICLES IN ITALIAN

Stefano Quaglia
Universität Konstanz

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

The paper deals with Italian Particle Verb Constructions that seem to display a different Grammatical Function assignment from the one of the base verb. I first demonstrate that the f-structures of these sentences are actually the same as the ones otherwise licensed by the verb. Then, I argue that the apparent spatial particles at stake are better analyzed as a particular class of prepositions that can realize their objects in non-adjacent c-structure nodes. Finally, I show how this discontinuous mapping from c- to f-structure (which obtains in other, unrelated constructions too) is licensed. As a consequence of the present account, a more restrictive and precise characterization of "Particle Verbs" for Italian is provided.

1 Introduction

Particle Verbs (henceforth, PVs) in English and in Germanic languages have been one major topic in generative linguistics for several decades (Emonds 1972; den Dikken 1995; Stiebels 1996; Dehé et al. 2002). The last years have seen an increasing interest in similar constructions in Italian and Romance, too, and many studies have been devoted to the topic, from different theoretical perspectives (Cini 2008; Cordin 2011; Iacobini & Masini 2007; Mateu & Rigau 2010, to name a few). Leaving aside a comparison of Italian and Germanic PVs, the present paper concentrates on Italian PVs that apparently exhibit a Grammatical Function assignment that is different from the one of their base verbs (cf. Cordin (2011:17); Iacobini & Masini (2007:159); Schwarze (2008:216)):

(1) a. Stefano è corso **alla fermata dell' autobus**
S. is run to-the stop of-the bus
'Stefano ran to the bus stop.'

b. Stefano **gli** è corso dietro
S. DAT.3SG.M is run behind
'Stefano ran after him.'

In (1a), the unaccusative verb *correre* 'to run' calls for a SUBJ (*Stefano*) and a spatial OBL (*alla fermata dell'autobus*). The PV *correre dietro* in (1b), on the other hand, seems to subcategorize for a SUBJ (*Stefano*) and an OBJ$_\theta$ (realized through the dative clitic pronoun *gli*). Notably, the verb *correre* alone does not normally take any OBJ$_\theta$:

(2) *Stefano **gli** è corso
S. DAT.3SG.M is run
'Stefano ran to him.'

[†]I am indebted to Christoph Schwarze and Miriam Butt for helpful discussion about the phenomena presented here. Moreover, I thank the participants to the LFG conference 2012 for pointing out interesting problems during my presentation.

In recent LFG literature, similar cases have been pointed out by Forst et al. (2010) for German and by Laczkó & Rákosi (2011) for Hungarian. The authors argue that the constructions at hand involve complex predication: verb and particle combine syntactically, and the new PRED features a GF-assignment that is different from the one of the verb.

The present paper aims first at demonstrating that a change in the GF-assignment is not what is going on in the Italian cases. By means of three syntactic tests, evidence is provided that sentences like (1b) feature a discontinuous OBL_θ, and not an OBJ_θ. I show that elements like *dietro* are not "true", but just "apparent" particles. They are better analyzed as a special class of prepositions that may govern their OBJs either in their c-structural complement position, or in non-adjacent nodes (like CL), provided that their CASE-requirements are met. It is precisely the last c-structural configuration (the same as in (1b)) that gives the double illusion of particle-syntax and change of GF-assignment. The c- to f-structure mappings displayed by the constructions at stake are then formalized (in terms of XLE-compatible annotated c-structure rules). Beyond giving a more restrictive and accurate characterization of spatial particles in Italian, the present account offers a window on how this language employs CASE as a means for the retrieval of GFs.

The paper is structured as follows: in section 2, an overview on Italian PVs is given; in section 3, I present three tests for the inspection of f-structure, in order to isolate the actual make-up of the f-structure; in section 4, I introduce the system I adopt for representing case; I then present the analysis of the spatial elements at stake, and I describe the c- to f-structure mappings displayed. In section 5, I summarize and make some concluding remarks.

2 Italian Particle verbs

Particle Verbs are commonly thought of as a linguistic phenomenon typical of Germanic languages, but absent in Romance ones (e.g. Snyder 2001). This generalization can be viewed as a corollary of Leonard Talmy's typology of motion events (Talmy 1985, 1991). Whereas Germanic languages, like English and German, lexicalize the "MANNER" meaning component in the verb root and "PATH" thrugh an adositional phrase or a particle, Romance languages, like Spanish and Italian, behave in the opposite way: "PATH" is lexicalized in the verb root, whereas "MANNER" is provided by a separate lexical item, such as a gerund. Accordingly, Germanic languages should be prone to constructions where a spatial particle encodes aspects of "PATH", like PVs: cf. English *to fly in*, German *hinein-//hinaus-fliegen*, Swedish *flyger in*. Although Talmy himself (1985) specified that his typology should not be interpreted as a sharp distinction without exception possibilities, the first linguist who pointed out the existence of Italian structures resembling Germanic PVs was Schwarze (1985), then followed by Simone (1997). Schwarze (1985) noticed that Italian features not only the typically Romance, expected pattern, but also the more Germanic-like one: the spatial particle encodes (aspects of) “PATH”, while the verb lexicalizes “MANNER”. Thus, beside the Romance type *uscire correndo* ‘to go out (while) running’, Italian features the Germanic-like *correre fuori* ‘to run out’, too.

The structure of Italian PVs can be descriptively characterized as follows: the combination of a verb and a spatial particle. One main issue in works on PVs (both Germanic (Booji 2002) and Italian (Iacobini & Masini 2007; Iacobini 2009)) is that of the locus of composition of these constructions: lexicon or syntax. Even within the LFG literature, one finds scholars defending opposite analyses: thus, as regards Hungarian PVs, Ackerman (1983) argues for a lexical account, whereas Laczkó & Rákosi (2011) prefer a syntactic one. Since, in Italian PVs, verb and particle can be separated at c-structure (cf. Masini 2008), a syntactic analysis would be the simplest assumption, and I will adopt it in this paper.

In the present work, I assume that lexical items that syntactically behave as particles belong to the major lexical category of P(repositions) (keeping to generalizations discussed in Emonds (1972) and Svenonius (2003; 2007)). On the contrary, particles are often classified as “adverbs” or “locative adverbs” in the literature on Italian PVs (cf. e.g. Cordin 2011; Iacobini & Masini 2007). This is because some of these elements need not take a complement (e.g. *fuori* ‘out(side)’, *dentro* ‘in(side)’, *sopra* ‘on, above’, *sotto* ‘under(neath)’), and some cannot take a complement altogether (e.g. *avanti* ‘ahead’, *indietro* ‘back(wards)’). Nonetheless, both their meaning and the distribution of the phrase the build set them together with “canonical” Ps: in some way, claiming that these items are not Ps would let us miss some important generalizations. Moreover, facts about complementation pose no problems for the approach defended here, if one adopts Emonds’ (1972) and Jackendoff’s (1983:57-60) view that the category P owns both transitive, and optionally transitive, and intransitive members – just like the category V.

As regards the meaning of Italian PVs, I will conform to Iacobini & Masini’s (2007:162) tripartite classification:

(3) a. **locative meanings**, as in *sbattere fuori* ‘to slap out’
b. **idiomatic meanings**, as in *fare fuori* ‘to kill’ (lit.: ‘to do out(side)’)
c. **aspectual and/or actional meanings**, as in *raschiare via* ‘to (successfully) scrape something away’

In this paper, I focus on PVs encoding locative meanings, since these are the ones where the phenomena at stake here can be appreciated at best.

3 Apparent changes in GF-assignment

3.1 The constructions at stake

The class of PVs I am going to focus on features, beside verb and spatial particle, the “Ground”-argument of the particle. This can be realized either (i) as a PP ((4a), (4b)) or (ii) as a case-marked clitic pronoun ((5a), (5b)):

(4) a. il difensore è corso dietro **all' attaccante**
the defender has run behind to-the attacker
'the defender ran after the attacker.'
b. il bandito salta dentro **al treno**
the bandit jumps inside to-the train
'the bandit jumps in the train.'

(5) a. il difensore **gli** è corso dietro
the defender DAT.3SG.M has run behind
'the defender ran after him.'
b. il bandito **ci** salta dentro
the bandit LOC jumps inside
'the bandit jumps in there.'

In the full-phrasal realization, one always gets PPs headed by *a* 'to'. On the other hand, if the "Ground" is encoded through a clitic, the animacy of the referent imposes a certain value for the attribute CASE: one gets dative clitics in case of [+animate] ((5a)), but locative clitics in case of [−animate] ((5b))[1]. Note that the same paradigm is exhibited by most other spatial particles (*addosso* 'on', *sotto* 'under(neath)', *sopra* 'upon, above', *vicino* 'near(by)', *contro* 'against', *intorno*, *attorno* '(a)round', *davanti* 'in front of', *accanto* 'beside', *incontro* 'towards', *appresso* 'by'), and with transitive verbs as well. In what follows, I am going to examine the structures involving PPs first, and the ones involving clitics later.
As Iacobini & Masini (2007:159) note, sentences like the ones in (4a) and (4b) are structurally ambiguous. On the one hand, the PP headed by *a* 'to' could be governed by the particle (yielding a complex PP):

(6) a. il difensore è corso [dietro [all' attaccante]$_{PP}$]$_{PP}$
b. il bandito salta [dentro [al treno]$_{PP}$]$_{PP}$

This obtains e.g. in sentences like the following:

(7) a. il difensore era **dietro all' attaccante**
the defender was behind to-the attacker
'the defender was behind the attacker.'
b. la pistola era **dentro alla borsa**
the gun was inside to-the bag
'the gun was inside the bag.'

On the other hand, the *a*-PP could be governed by the PV directly:

(8) a. il difensore è [corso dietro]$_{PV}$ [all' attaccante]$_{PP}$
b. il bandito [salta dentro]$_{PV}$ [al treno]$_{PP}$

PPs headed by *a* are indeed possible c-structural realizations of two clause-level GFs: OBJ$_\theta$ ((9a)) and OBL$_\theta$ ((9b)) respectively:

[1] In Italian, as in French, dative clitics bear PERS/NUM/GEND features, locative clitics do not.

(9) a. il difensore passa il pallone **al portiere**
the defender passes the ball to-the goalkeeper
'the defender passes the ball to the goalkeeper.'

b. il bandito abita **a Torino**
the bandit lives to Torino
'the bandit lives in Turin.'

Let us consider the implications of each hypothesis for the f-structures of the sentences in (4a) and (4b). If the first hypothesis were the case (i.e., the *a*-PP builds a unit together with the particle), the extra *a*-PP would bear a grammatical function subcategorized for by the PRED contributed by the spatial particle. For the time being, I won't make any claims about the precise identity of this function, and I will call it simply GF. *a*-PP and particle would then together correspond to a complex OBL_{loc}. This, in turn, would be subcategorized for by the verb. Verb and particle would correspond to separate predicates, at the level of f-structure. I will call this "Hypothesis (i)". In Figure 1, an underspecified f-structure consistent with Hypothesis (i) is given:

$$\begin{bmatrix} \text{SUBJ} & [...] \\ \text{PRED} & '\textbf{verb}<(\uparrow\text{SUBJ})\ ...\ (\uparrow\text{OBL}_{loc})>' \\ \cdots & [...] \\ \text{OBL}_{loc} & \begin{bmatrix} \text{PRED} & '\textbf{particle}<(\uparrow\text{GF})>' \\ \text{GF} & [...] \end{bmatrix} \end{bmatrix}$$

Figure 1: Hypothesis (i) (underspecified f-structure)

On the other hand, if the second hypothesis were the case (i.e., the *a*-PP is governed by the whole PV), the f-structure of the sentences would be deeply different: the PP headed by *a* would bear a clause-level GF on its own. Let us provisorily call this GF_{loc}. This grammatical function would be subcategorized for by a complex PRED, corresponding to the whole PV. Verb and particle would then build a single predicative unit – which is usually the case either (i) in case of Complex Predication, or (ii) in case of Applicatives[2]. I will call this "Hypothesis (ii)". In Figure 2, an underspecified f-structure consistent with Hypothesis (ii) is given:

$$\begin{bmatrix} \text{SUBJ} & [...] \\ \text{PRED} & '\textbf{verb}+\textbf{particle}<(\uparrow\ \text{SUBJ})\ ...\ (\uparrow\ \text{GF}_{loc})>' \\ \cdots & [...] \\ \text{GF}_{loc} & [...] \end{bmatrix}$$

Figure 2: Hypothesis (ii) (underspecified f-structure)

[2]I thank Miriam Butt for suggesting this last possibility to me with respect to these cases.

Let us now turn to the sentences where the "Ground"-argument is encoded by a clitic pronoun ((5a), (5b)). In Italian, both dative and locative clitics can realize clause-level GFs:

(10) a. il difensore **gli** passa il pallone
the defender DAT.3SG.M passes the ball
'the defender passes him the ball.'

b. il bandito **ci** abita
the bandit LOC lives
'the bandit lives there.'

gli in (10a) is an OBJ_{goal}, while *ci* in (10b) is an OBL_{loc}. In the literature, PVs appearing with "Ground"-clitics have not been investigated systematically. Iacobini (2008:113-5) considers structures involving dative clitics, and concludes that these are extra-arguments licensed by the PV (in line with our Hypothesis (ii)). Basing on evidence like (10a), dative clitics are thus considered 'bona fide' Indirect Objects. Masini (2008:86-7), on the other hand, argues that sentences like (10a) feature prepositions taking a clitic complement $-$ which corresponds to our Hypothesis (i). It should be noted that the data in (10a)-(10b) enable us to refine Hypothesis (ii). Since, in these sentences (as in many others), dative clitics encode OBJ_{θ} but locative clitics encode OBL_{θ}, it can be argued that [$+$animate] "Grounds" are OBJ_{loc}s, whereas [$-$animate] ones are OBL_{loc}s. As a matter of fact, the former alternate with dative clitics, the latter with locative clitics.

At this point of the paper, both possible analyses of the structures at hand have been sketched. In (3.2) I provide pieces of evidence that Hypothesis (ii) is untenable, whereas Hypothesis (i) correctly predicts the data. In section (4) I will describe the c- to f-structure mappings licensing the structures at stake.

3.2 Inspecting f-structure

3.2.1 Missing realization possibilities

In structures featuring PVs that lack a "Ground"-argument, the OBJ-NP $-$which encodes the "Figure"-argument, and clearly has to be analyzed as a clause-level GF[3]$-$ can appear either to the right ((11a)) or to the left ((11b)) of the particle, yielding something similar to the typically Germanic "particle shift":[4]

(11) a. il barista porta fuori le sedie
the barman brings out the chairs
'the barman brings out the chairs.'

[3]For example, it can be passivized: *le sedie vengono portate fuori* 'the chairs are brought out.'.

[4]This phenomenon appears to be more constrained in Italian than e.g. in English, even if it is driven by the same information-structural reasons (cf. Masini 2008). Nonetheless, these structures are licit, provided that certain lexical (PVs with locative meanings are preferred, cf. Masini (2008:92)) and prosodic (the interposed NP must not exceed one phonological phrase, cf. Schwarze (2008:220-1)) conditions are met.

b. il barista porta le sedie fuori
the barman brings the chairs out
'the barman brings the chairs out.'

If the "Ground"-PPs under scrutiny were to be analyzed as clause-level GFs, one would predict that they could be interposed between verb and particle as well. Though, just the linear order [...V − Prt − PP...] is grammatical ((12a)), whereas the order [...V − PP − Prt] is ungrammatical ((12b)):

(12) a. il cane salta addosso al ladro
the dog jumps on to-the thief
'the dog jumps up at the thief.'
b. *il cane salta al ladro addosso
the dog jumps to-the thief on
'the dog jumps up at the thief.'

It could be objected that "shiftability" is an idiosyncratic property of every single particle: *fuori* is shiftable, *addosso* is not. But it's easy to provide an immediate counterexample:

(13) il manifestante gli lancia delle pietre addosso
the protester DAT.3SG.M throws of-the stones on
'the protester throws stones at him.'

In (13), *addosso* appears in the shifted position: the OBJ-NP is now placed between it and the verb.
Under Hypothesis (i), these facts are easily explained. PPs like *al poliziotto* in (12a) are OBJs of the particle, which here actually behaves as a normal preposition. Subsequently, its complement must be realized on its right, as usual in Italian. Trying to place the PP to the right of the verb, in the c-structural position of OBJ_θ, fails, for the verb *saltare* in (12b) does not take any OBJ_θ.

3.2.2 Resumptive clitic pronouns in Clitic Left Dislocation

Clitic Left Dislocation (henceforth, CLLD) is a typically Romance structure where a phrase XP (it may be NP, PP, AP, VP, CP) is placed at the beginning of the sentence, and the GF it bears is indexed by means of a clitic pronoun, which functions as a resumptive element. The dislocated phrase XP is interpreted as the sentence Topic[5]. In (14a)-(14d), examples are provided:

(14) a. **Mario**, **lo** amiamo tutti
M. ACC.3SG.M love-1PL all
'Mario, we all love him.'

[5]For a survey of Italian and Romance CLLD, see Cinque (1990:56-97).

b. **a Mario**, Giorgio **gli** ha regalato un libro
to M. G. DAT.3SG.M has presented a book
'Mario, Giorgio gave a book to him as present.'

c. **in Russia**, Mario **ci** vuole andare
in Russia M. LOC wants to-go
'to Russia, Mario wants to go there.'

d. **che Mario è bravissimo**, **lo** abbiamo sempre saputo
that M. is good-SUPERL ACC.3SG.M have-1PL always known
'that Mario is very good, we always knew that.'

As can be seen in the examples, matching of GFs occurs by means of CASE and, where possible, PERS/NUM/GEND features. Thus, NPs functioning as OBJ must be resumed by accusative clitics matching agreement features ((14a)), PPs functioning as OBJ$_\theta$ must be resumed by dative ones ((14b)), whereas OBL$_\theta$ and COMP must be resumed by the locative clitic ((14c)) and by the accusative singular masculine clitic *lo* ((14d)) respectively.
Now, if Hypothesis (ii) were the case, in ambiguous structures like (15a) and (15b) it should not be possible to dislocate the spatial particle together with the PP, while getting the resumptive clitic indexing OBL$_\theta$:

(15) a. il difensore è corso dietro all' attaccante
the defender has run behind to-the attacker
'the defender ran after the attacker.'

b. l' allenatore piazza dietro all' attaccante un difensore
the coach places behind to-the attacker a defender
'the coach puts a defender behind the attacker.'

However, this is possible, indicating that Hypothesis (ii) makes wrong predictions:

(16) a. **dietro all' attaccante**, **ci** è corso il difensore
behind to-the attacker, LOC has run the defender
'after the attacker, the defender ran there.'

b. **dietro all' attaccante**, l' allenatore **ci** piazza un difensore
behind to-the attacker, the coach LOC places a defender
'behind the attacker, the coach puts a defender there.'

This structures are grammatical precisely because spatial particle and PP form a unit together, both at c-structure (a complex PP), and at f-structure (a complex OBL$_\theta$).
Turning to sentences where the "Ground"-argument is realized through a case-marked clitic pronoun ((17a), (17b)), we find out that CLLD can apply to the spatial particle alone, indexing it as an OBL$_{loc}$ ((18a), (18b)):

(17) a. il difensore gli è corso dietro
the defender DAT.3SG.M runs behind
'the defender runs after him.'

b. l' allenatore gli piazza dietro un difensore
the coach DAT.3SG.M places behind a defender
'the coach puts a defender behind him.'

(18) a. **dietro**, gli **ci** è corso il difensore
behind, DAT.3SG.M LOC has run the defender
'after him, the defender ran there.'

b. **dietro**, l' allenatore gli **ci** piazza un difensore
behind the coach DAT.3SG.M LOC places a defender
'behind him, the coach puts a defender there.'

The sentences in (18a)-(18b) might seem to contradict the data in (16a)-(16b), for the particle only is fronted, leaving the clitics in place. Though, this is consistent with Hypothesis (i): (17a) and (17b) display discontinuous OBL$_{loc}$s, where dative clitics contribute the OBJ attribute, while spatial particles contribute PRED. Of both, the only element feasible to be placed in the c-structure node hosting left-dislocated phrases (an XP-node adjoined to IP) is the particle, because clitic pronouns have to be attached either as sisters to I^0 or as sisters to V^0. But the sentences in (18a) and (18b) don't display just spatial particles and dative clitics. The resumptive locative clitic pronoun *ci* is present, too. On Hypothesis (i), this is predicted: since an OBL$_{loc}$ function is topicalized, it must be resumed within the clause by means of a locative clitic. On the contrary, the presence of the resumptive *ci* is not expected under Hypothesis (ii). According to the refined version of Hypothesis (ii) in section 3.1, in structures like (17a)-(17b) the f-structure would contain an OBJ$_{loc}$, but no OBL$_{loc}$ function at all. Therefore, indexation of OBL in CLLD would remain unexplained (and unpredicted).

In light of these facts, the test involving CLLD provides a crucial piece of evidence that only Hypothesis (i) is sustainable.

3.2.3 Binding of *proprio*

Binding data regarding the adjective *proprio* 'own' also suggest that the ambiguous sentences actually contain a complex OBL$_{loc}$, and not an OBJ$_{loc}$. Giorgi (1984, 1991) dubs *proprio* a "possessive anaphor": while it owns typical adjectival morphology (it must agree in NUM and GEND with a noun), it must be bound, like anaphors. Giorgi (1991:186) claims that this element can behave in two ways: it can be either clause-bound, or long-distance-bound. In the first case, both SUBJ and OBJ may be legitimate antecedents ((19a), taken from Giorgi (1984:314)); in the second, *proprio* is subject-oriented ((19b), taken from Giorgi (1991:186)):

(19) a. Gianni$_j$ ha ricondotto Maria$_i$ alla propria$_{i/j}$ famiglia
G. has taken-back M. to-the own family
'Gianni brought back Maria to his/her own family.'

b. Gianni$_j$ ha aizzato Maria$_i$ contro coloro che disprezzano il
G. has turned M. against those who despise the
proprio$_{j/*i}$ figlio
own son
'Gianni turned Maria against those who despise his/*her own son.'

In both cases, it seems that binding of *proprio* is constrained by a general f-command condition, as can be appreciated from the following examples:

(20) a. il presidente$_i$ ha ringraziato i propri$_i$ sostenitori
the president has thanked the own supporters
'the president thanked his own supporters.'
b. gli amici$_i$ di Gianni$_j$ apprezzano le proprie$_{i}/_{*j}$ poesie
the friends of G. appreciate the own poems
'Gianni's friends appreciate their/*his own poems.'
c. che i propri$_i$ ospiti siano arrivati in ritardo non
that the own guests have-SUBJUNCTIVE arrived in delay not
ha stupito Mario$_i$
has surprised M.
'that his own guests arrived late did not surprise Mario.'
d. *che Mario$_i$ sia arrivato in ritardo non ha stupito
that M. has-SUBJUNCTIVE arrived in delay not has surprised
i propri$_i$ amici
the own friends
'*that Mario arrived late did not surprise his own friends.'

Now, recalling that f-command is defined as follows (Bresnan (1982:334)):

(21) *F-command*:
For any occurrences of the functions α, β in an f-structure F, α *f-commands* β if and only if α does not contain β and every f-structure of F that contains α contains β

it is easy to see that in (20a)-(20d), the anaphor *proprio* can be bound only by those GFs that f-command it. Thus, in (20a) *il presidente* (value of SUBJ) f-commands the f-structure corresponding to the OBJ, and also *propri*, which is contained within it. In (20b), *Gianni* cannot be a binder, for the first f-structure containing it (the f-structure corresponding to the SUBJ) does not contain *proprie*. Similar arguments apply to (20c) and (20d).
Binding of *proprio* provides us with a probe into the f-structure of the ambiguous sentences: if the "Ground"-PPs really were clause-level OBJ$_{loc}$s, they should be possible binders. However, this is not the case, as the following examples show:

(22) a. Paolo$_i$ mette dietro a Maria$_j$ il proprio$_{i/*j}$ ritratto
P. puts behind to M. the own portrait
'Paolo$_i$ puts his/*her own portrait behind Maria.'

b. il ninja$_i$ lancia contro al samurai$_j$ la propria$_{i/*j}$ spada
the ninja throws against to-the samurai the own sword
'the ninja$_i$ throws his$_{i/*j}$ own sword against the samurai$_j$.'

Under Hypothesis (i), these facts are predicted: the PPs would be GFs embedded in a clause-level OBL$_{loc}$s, and from their structural position they could not f-command the anaphor, which is embedded in the clause-level OBJs.
Interestingly, however, the sentences featuring clitic pronouns behave in the opposite way:

(23) a. Paolo$_i$ le$_j$ mette dietro il proprio$_{i/j}$ ritratto
P. DAT.3SG.F puts behind the own portrait
'Paolo$_i$ puts his/her own portrait behind her.'
b. il ninja$_i$ gli$_j$ lancia contro la propria$_{i/j}$ spada
the ninja DAT.3SG.M throws against the own sword
'the ninja$_i$ throws his$_{i/j}$ own sword against him$_j$.'

The clitics *le* and *gli* have the same f-structural position as the PPs *a Maria* and *al samurai* in (22a) and (22b) respectively. Subsequently, it is predicted that they should not be able to bind the anaphor, for they do not f-command it. Indeed, surprisingly, they are able to bind *proprio*. These facts can be explained by appealing to the information-structural status of clitic pronouns. In Italian, clitic pronouns are topical: as Berretta (1986:71) points out, they convey "de-emphatic old information". In sentences like (23a) and (23b), they receive an i(nformation)-structural representation that is different from the one of the particle (probably, TOPIC). Accordingly, at i-structure they are separate from the rest of the OBL$_{loc}$, and they therefore regain a prominence they do not have at f-structure[6]. Thus, I tentatively argue that this kind of prominence relaxes the f-command condition, enabling the clitics at stake to bind the anaphor[7]. In sum, data concerning binding of the anaphoric adjective *proprio* are also compatible with Hypothesis (i).

The three tests I have presented so far provide evidence that Hypothesis (i), and not Hypothesis (ii), is a sustainable representation for the examined constructions. In the course of the discussion, it may already have become clear to the reader why the spatial elements at stake only display an apparent particle-syntax: the tests suggest that these elements syntactically behave like prepositions. As a matter of fact, they constantly keep a dependency relation to an OBJ. This is evident in case they govern it on their right, but might seem bizarre when the OBJ is encoded as a clitic. In section 4, I concentrate on the last kind of mapping, showing that it is not peculiar to this class of lexical items.

[6]An alternative solution consists in resorting to c-command. Under a c-structural analysis of clitics as non-projecting nodes adjoined to V^0/I^0 (e.g. Toivonen (2001)), the dative clitics would c-command the XP containing the anaphor.

[7]Indeed, there is evidence from unrelated constructions that prominence at i-structure plays an important role with respect to grammatical processes, in Italian (cf. Salvi 1986).

4 Apparent spatial particles and their c- to f-structure mapping

In this section, I explain how the mapping from c-structure to f-structure takes place in the constructions discussed in section 3. The analysis consists of two major premises, to be introduced in turn, and a presentation of the c-structure rules and the functional annotations licensing the correspondence.

4.1 Case in Italian

A framework like LFG makes it possible to formally represent the acknowledged generalization that different categories (e.g. P and CL) can contribute an identical grammatical information (e.g. CASE). As regards Romance languages, much work has been done on the role of case and its representation (Grimshaw (1982), Frank (1996) and Schwarze (1996) on French; Alsina (1996) on Catalan; Schwarze (2012) on Italian). The representation of case I propose in this paper is in line with the one worked out in Schwarze (1996; 2012).
Italian features a "janus-faced" case-marking system: on the one hand, CASE can be expressed syntactically, namely through Ps devoid of a PRED attribute. On the other, clitics encode CASE-oppositions morphologically (although syncretic forms often neutralize such oppositions, like *ne*, as is expected in lexical paradigms):

	ACC	DAT	LOC	GEN	ABL
P	–	*a*	*a*	*di*	*da*
CL	*lo, la, le, li*	*gli, le*	*ci*	*ne*	*ne*

Table 1: Sketch of Standard Contemporary Italian case system

In the present system, CASE is assigned to a given f-structure only in presence of an overt marker. A consequence of this is that SUBJs and OBJs, if encoded by NPs, do not contain a CASE attribute. In these cases, the encoding of Grammatical Functions obtains configurationally. However, if one wants to keep to generalizations about case-assignment, it is possible to assign NOMINATIVE and ACCUSATIVE structurally, i.e. by means of additional functional annotations on the c-structure nodes where SUBJ and OBJ can be realized.
In what follows, I illustrate how the system of case interacts with the class of spatial particles I have been focussing on.

4.2 "True" and "apparent" particles

The main claim of this paper is that the P-elements involved in the constructions under scrutiny syntactically behave as prepositions, and not as particles –as they are usually analyzed. Precisely, they belong to a special sub-class of Italian Ps that can lexically impose CASE requirements on their governed GFs.
Consider the following sentences:

(24) a. l' allenatore piazza dietro **all' attaccante** un difensore
the coach places behind to-the attacker a defender
'the coach puts a defender behind the attacker.'

b. l' allenatore **gli** piazza dietro un difensore
the coach DAT.3SG.M places behind a defender
'the coach puts a defender behind him.'

(25) a. il bambino mette dietro **al muretto** i giochi
the kid puts behind to-the wall-DIMINUTIVE the toys
'the kid puts the toys behind the little wall.'

b. il bambino **ci** mette dietro i giochi
the kid LOC puts behind the toys
'the kid puts the toys behind there.'

In these structures, the P *dietro* 'behind' requires its OBJ to be either DATIVE or LOCATIVE, depending on the value for the attribute ANIMATE. The grammatical P *a* 'to' is ambiguous: it can contribute either ($\uparrow$ CASE) $=$ DATIVE or ($\uparrow$ CASE) $=$ LOCATIVE. Subsequently, the opposition between the two values is superficially neutralized in (24a) and (25a). But, as soon as the subcategorized OBJ is encoded as a CL at c-structure, the opposition comes to the surface, as can be seen in (24b) and (25b). Interestingly, *dietro* (like *contro* 'against', *dentro* 'in(side)', *sopra* 'on; above', *sotto* 'under(neath)') can also take an NP as OBJ, without a "mediating" grammatical P:

(26) a. l' allenatore piazza dietro **l' attaccante** un difensore
the coach places behind the attacker a defender
'the coach puts a defender behind the attacker.'

b. il bambino mette dietro **il muretto** i giochi
the kid puts behind the wall-DIMINUTIVE the toys
'the kid puts the toys behind the little wall.'

As argued in section 4.1 with respect to sentential OBJs encoded by NPs, also for the OBJs of *dietro* in (26a) and (26b) two treatments are possible: they can be either not marked for CASE, or structurally marked as ACCUSATIVE. In either case, they are not lexically marked by the governing P. According to this analysis, prepositions like *dietro* have two government patterns: they can either (i) lexically impose a certain value for CASE, or (ii) not impose any. However, the inventory of Italian predicative prepositions also contains classes that behave in a more restrictive way, allowing only one of the two strategies. Ps like *addosso* 'on', *davanti* 'in front of', *incontro* 'towards' exhibit (i), but not (ii):

(27) a. andavo incontro *(a) Maria
go-IMPF-1SG towards to M.
'I was going towards Maria.'

b. le andavo incontro
DAT.3SG.F go-IMPF-1SG towards
'I was going towards her.'

Ps like *lungo* 'along', *verso* 'towards', *oltre* 'beyond' (in its spatial meaning) behave in the opposite way, exhibiting (ii), but not (i). Moreover, P-elements of this class do not tolerate clitic OBJs:

(28) a. andavo verso (*a) Maria
go-IMPF-1SG towards M.
'I was going towards Maria.'
b. *la/le andavo verso
ACC.3SG.F/DAT.3SG.F go-IMPF-1SG towards'
'I was going towards her.'

The generalization thus appears to be that only Ps that can lexically impose a CASE value on their OBJs can realize them on separate c-structure nodes.
Under the analysis presented here structures like (24a), (24b) and (25a), (25b) do not involve particles, but prepositions. This is a welcome conclusion: if these elements were analyzed as particles, this would argue against Svenonius' (2003:434) generalization that particles tend to introduce a "Figure" only, and no "Ground". The author himself points out that this statement should be interpreted as the typical case rather than as a strict generalization (Svenonius (2007:81)), and refers to cases where a particle does introduce a "Ground" as a syntactic argument of the verb (after demotion of the "Figure", cf. Svenonius (2003:437-8)). Nonetheless, it seems safe to assume that a P-element that *directly governs* a "Ground" is a preposition, and not a particle (as Svenonius (2003:434) proposes). In the constructions described so far, a "Ground" is always there, and it is always governed by the P-elements, be it realized as an adjacent PP or as a non-adjacent CL node. Now, in the latter realization option, c-structure rules produce a deceiving linear order, which closely resembles the typical one featured by "true" particles:

(29) a. [NP_{figure} ...$-$ $CL_{ground}V - P$ ...] (intransitive Vs)
b. [... $CL_{ground}V - \{P\} - NP_{figure} - \{P\}$ ...] (transitive Vs)

Italian does have "bona fide" spatial particles, as can be seen in the following sentences:

(30) a. il ladro saltò dentro
the thief jumped in(side)
'the thief jumped in.'
b. Luca ha buttato giù i birilli
L. has thrown down the skittles
'Luca threw down the skittles.'

But the crucial difference between these structures and the ones investigated in this paper lies in the absence vs. presence of a “Ground”.

4.3 Mapping c-structure to f-structure

The two structure types I have been considering feature a similar f-structure, but differ with respect to the ϕ-projection. Though, this is expected, given that they also differ as regards c-structure. Sentences where the locative P take a PP as c-structure complement, present the standard mapping of locative PPs (depicted in Figure 3). On the other hand, sentences where a clitic pronoun encodes the “Ground” involve a discontinuous mapping (depicted in Figure 4).

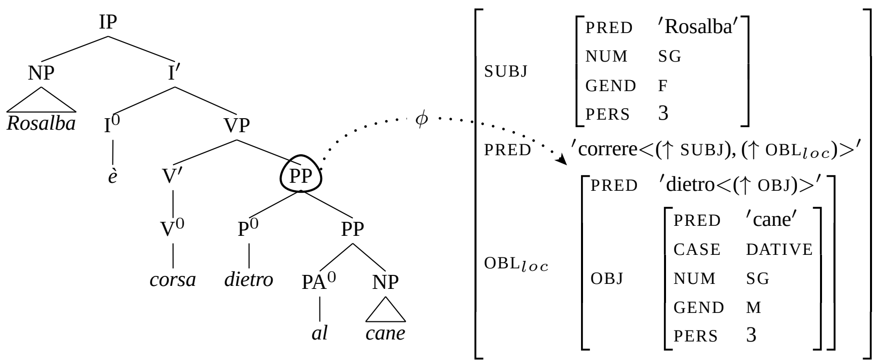


Figure 3: *Rosalba è corsa dietro al cane* ‘Rosalba ran after the dog.’

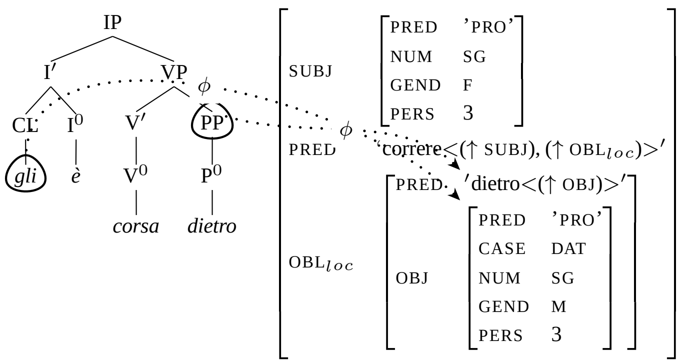


Figure 4: *gli è corsa dietro* ‘She ran after him.’

The first type of mapping is effected by means of the following functional annotations on c-structure rules:

(31)
$$\begin{array}{cccccc} \text{VP} & \longrightarrow & ... & \text{V}' & ... & (\text{PP}) \\ & & & \uparrow = \downarrow & & (\uparrow \text{OBL}_{loc}) = \downarrow \end{array}$$

(32) $\begin{array}{llll} \text{PP} & \longrightarrow \quad ... & \text{P}^0 & (\text{PP}) \\ & & \uparrow = \downarrow & (\uparrow \textsc{obj}) = \downarrow \end{array}$

The second type of mapping is more complex. Whereas the annotation on the PP in (31) applies here too, the OBJ of OBL_{loc} is contributed by the clitic pronoun, attached as a sister of either V^0 or I^0[8]. If the CASE feature provided by the clitic is consistent with the requirements imposed by the PRED of the locative P, the partial f-structures will correctly unify as a complex OBL_{loc}.
This kind of ϕ-projection is not only found in the sentences examined here, but it instantiates a general mapping mechanism available for Italian clitic pronouns. Consider the following data:

(33) a. Paolo è fedele ad Anna
P. is loyal to A.
'Paolo is loyal to Anna.'
b. Paolo le è fedele
P. DAT.3SG.F loyal
'Paolo is loyal to her.'

(34) a. Nerone desidera la distruzione di Roma
N. wishes the destruction of R.
'Nero wishes the destruction of Rome.'
b. Nerone ne desidera la distruzione
N. GEN wishes the destruction
'Nerone wishes its destruction.'

In (33a)-(33b), the adjective *fedele* 'loyal' subcategorizes for a GF (it might be an OBJ, or an OBJ_θ): this is encoded either through an adjacent PP, or through a clitic. Similarly, in (34a)-(34b) the event noun *distruzione* 'destruction' calls for a POSS, which is realized either through a PP or through a clitic pronoun. Importantly, in both cases –like in the sentences involving locative Ps– CL nodes are mapped onto a GF that is governed by a PRED embedded in a clause-level GF (PREDLINK in (33b), OBJ in (34b)). Obviously, CL nodes can be mapped onto clause-level GFs, too:

(35) a. Lucio parla di film horror
L. talks of movie horror
'Lucio talks about horror movies.'
b. Lucio ne parla
L. GEN talks
'Lucio talks about it.'

[8]Toivonen (2001) argues that Romance clitic pronouns are non-projecting nodes. Accordingly, clitics are adjoined to V^0/I^0, resulting in another V^0/I^0. This seems a very interesting proposal to me, but its implementation in XLE easily runs into overgeneration problems. These can be avoided resorting to more complicated c-structure rules (involving disjunction) and to additional constraints, but for the purposes of this paper I keep to the more "traditional" c-structure rules (as proposed first by Grimshaw (1982) for French), which represent clitics as sisters of V^0/I^0 and daughters of V'/I'.

In (35b), the genitive clitic *ne* encodes the OBL_θ of the verb *parlare* ‘to talk’, which must bear GENITIVE as value for CASE. Nonetheless, the “search space” where CL nodes can retrieve their GFs has to be constrained. As a matter of fact, GFs contained in COMP ((36a)-(36b)) and XCOMP ((37a)-(37b)) seem to be unavailable:

(36) a. Marcolino promette che farà i compiti
M. promises that do-FUT.3SG the homeworks
‘Marcolino promises that he will do his homework.’

b. *Marcolino li promette che farà
M. ACC.3PL.M promises that do-FUT.3SG
‘Marcolino promises that he will do it.’

(37) a. Matteo vede Stefano dare un regalo a Susanna
M. sees S. give-INF a present to S.
‘Matteo sees Stefano give a present to Susanna.’

b. *Matteo le vede Stefano dare un regalo
M. DAT.3SG.F sees S. give-INF a present
‘Matteo sees Stefano give her a present.’

Also GFs realized as clauses have this “island”-effect:

(38) a. che i deputati non vadano in parlamento è una vergogna
that the deputies not go-SUBJUNCTIVE-3PL in parliament is a shame
‘that deputies don’t go to the parliament is a shame.’

b. *che i deputati non vadano ci è una vergogna
that the deputies not go-SUBJUNCTIVE-3PL LOC is a shame
‘that deputies don’t go there is a shame.’

The right generalization to be captured thus seems to be that the “search space” cannot cross a GF that contains a SUBJ. This can be easily represented by means of a functional uncertainty path, restricted by an off-path constraint:

(39) $$(\uparrow \underset{\neg(\rightarrow\ \text{SUBJ})}{\text{GF}^*}\ \text{GF}) = \downarrow$$

Moreover, GFs contained in a SUBJ are excluded as well[9]:

(40) a. il trailer del documentario è molto bello
the trailer of-the documentary is very nice
‘the trailer of the documentary is very nice.’

9 Rizzi (2001:540-1) claims that also adjuncts are “islands” for this kind of mapping. He provides sentences involving copular verbs, like **Gianni le è felice accanto* ‘Gianni is happy beside her’. Nonetheless, grammatical sentences can be easily found where a clitic pronoun encodes the OBJ of an ADJ function: *i bambini ci giocano sopra* ‘the kids play (while being) on it’. The ungrammaticality of the sentences provided by Rizzi seems thus to depend on the verb type, or on its lexical semantics.

b. *il trailer ne è molto bello
the trailer GEN is very nice
'its trailer is very nice.'

The expression in (39) must be then further constrained. The final version of the functional uncertainty (to be annotated on CL nodes) is then:

(41)
$$\begin{array}{c}(\uparrow \text{GF}^* - \text{SUBJ GF}) = \downarrow \\ \neg(\rightarrow \text{SUBJ})\end{array}$$

The annotated c-structure rules (already implemented in an XLE-grammar fragment for Italian) would look like as follows:

(42)
$$\begin{array}{cccccc} \text{I}' & \longrightarrow & ... & \text{(CL)} & \text{(CL)} & ... & \text{I}^0 \\ & & & (\uparrow \text{GF}^* - \text{SUBJ GF}) = \downarrow & (\uparrow \text{GF}^* - \text{SUBJ GF}) = \downarrow & & \uparrow = \downarrow \\ & & & \neg(\rightarrow \text{SUBJ}) & \neg(\rightarrow \text{SUBJ}) & & \\ & & & & & & \\ & & & ((\downarrow \text{CASE}) =_c \text{DAT}) \vee & ((\downarrow \text{CASE}) =_c \text{ACC}) \vee & & \\ & & & ((\downarrow \text{CASE}) =_c \text{LOC}) & ((\downarrow \text{CASE}) =_c \text{GEN}) \vee & & \\ & & & & ((\downarrow \text{CASE}) =_c \text{ABL}) & & \end{array}$$

(43)
$$\begin{array}{cccccc} \text{V}' & \longrightarrow & ... & \text{(CL)} & \text{(CL)} & ... & \text{V}^0 \\ & & & (\uparrow \text{GF}^* - \text{SUBJ GF}) = \downarrow & (\uparrow \text{GF}^* - \text{SUBJ GF}) = \downarrow & & \uparrow = \downarrow \\ & & & \neg(\rightarrow \text{SUBJ}) & \neg(\rightarrow \text{SUBJ}) & & \\ & & & & & & \\ & & & ((\downarrow \text{CASE}) =_c \text{DAT}) \vee & ((\downarrow \text{CASE}) =_c \text{ACC}) \vee & & \\ & & & ((\downarrow \text{CASE}) =_c \text{LOC}) & ((\downarrow \text{CASE}) =_c \text{GEN}) \vee & & \\ & & & & ((\downarrow \text{CASE}) =_c \text{ABL}) & & \end{array}$$

Whereas the annotation in (41) will be associated to every CL node, linear order constraints exhibited in clitic clusters (i.e., DATIVE $>$ ACCUSATIVE) can be easily represented by means of additional constraining equations, as can be seen in (42) and (43). Accordingly, DATIVE and LOCATIVE clitic pronouns are forced to be associated to the first CL-slot, ACCUSATIVE, GENITIVE and ABLATIVE ones to the second.
These c-structure rules, together with their respective functional annotations, will license the second type of mapping discussed above, which appears not only in structures involving the locative Ps examined in this paper, but also in other, unrelated constructions.

5 Conclusion

In this paper, I considered Italian Particle-Verb Constructions where the "Ground" argument of the spatial particle is realized, either as a PP or as a case-marked clitic pronoun. Resorting to three different tests (licit c-structural realization possibilities, resumption in Clitic-Left-Dislocation contexts, Binding of anaphoric adjective *proprio*), I showed that the "Ground" cannot be represented as a clause-level GF at f-structure: subsequently, it cannot be maintained that the construction features a GF-assignment different from that of the verb (e.g. as a result of either Complex Predication or Applicativization). Whereas the mapping from c-structure to f-structure involved in constructions featuring complex PPs can be viewed as a "trivial" one, the one exhibited by sentences with clitic pronouns is more complex, and instantiates as general ϕ-projection mechanism (feasible to be represented by means of a functional uncertainty), available for CL nodes in many other unrelated constructions. CASE-properties of both clitic pronouns and GF-taking lexical items were showed to be crucial for this last mechanism to apply successfully. Moreover, in the analysis presented here the "apparent" spatial particles under scrutiny were showed to be actually a particular sub-class of P-elements displaying prepositional (and not particle-like) syntax. Their distinctive property is the ability to lexically impose CASE-requirements on their governed OBJs. These results may contribute to a better understanding of the phenomenon of "Particle Verbs" in Italian.

References

Ackerman, Farrell. 1983. *Miscreant morphemes: Phrasal Predicates in Ugric*. Ph.D. Dissertation, UC Berkeley.

Alsina, Alex. 1996. *The role of argument structure in grammar. Evidence from Romance* Stanford: CSLI Publications.

Berman, Judith & Anette Frank. 1996. *Deutsche und französische Syntax im Formalismus der LFG* (Linguistische Arbeiten 344). Tübingen: Niemeyer.

Berretta, Monica. 1986. Struttura informativa e sintassi dei pronomi atoni: condizioni che favoriscono la "risalita". In Stammerjohann (1986: 71-84).

Booji, Geert. 2002. Separable complex verbs in Dutch: A case of periphrastic word formation. In Dehé et al. (2002: 21-41).

Bresnan, Joan. 1982. Control and Complementation. In Bresnan & Kaplan (1982: 282-390).

Bresnan, Joan & Ron Kaplan (eds.). 1982. *The mental representation of grammatical relations*. Cambridge: MIT Press.

Cini, Monica (ed.). 2008. *I verbi sintagmatici in italiano e nelle varietà dialettali. Stato dell'arte e prospettive di ricerca* (Kommunikative Räume 3). Frankfurt am Main: Peter Lang.

Cinque, Guglielmo. 1990. *Types of A′-Dependencies* (*Linguistic Inquiry* Monographs). Cambridge (MA) & London: The MIT Press.

Cordin, Patrizia. 2011. *Le costruzioni verbo-locativo in area romanza. Dallo spazio all' aspetto* (Beihefte zur Zeitschrift für romanische Philologie 365). Berlin, New York: De Gruyter.

Dehé, Nicole, Ray Jackendoff, Andrew McIntyre & Silke Urban (eds.). 2002. *Verb-Particle Explorations* (Interface Explorations 1). Berlin/New York: Mouton de Gruyter.

den Dikken, Marcel. 1995. *Particles: On the syntax of verb-particle, triadic, and causative constructions*. New York: Oxford University Press.

Emonds, Joseph. 1972. Evidence that indirect object movement is a structure-preserving rule. *Foundations of Language* 8(4). 546-561.

Forst, Martin, Tracy H. King & Tibor Laczkó. 2010. Particle verbs in computational LFG: Issues from English, German and Hungarian. In Miriam Butt & Tracy H. King (eds.) *Proceedings of the LFG'10 Conference*, 228-248. Stanford: CSLI Publications.

Giorgi, Alessandra. 1984. Toward a theory of long distance anaphors: a GB approach. *The Linguistic Review* 3. 307-361.

Giorgi, Alessandra. 1991. Prepositions, binding and θ-theory. In Jan Koster & Eric Reuland (eds.), *Long-distance anaphora*, 184-203. Cambridge: Cambridge University Press.

Grimshaw, Jane. 1982. On the lexical representation of romance reflexive clitics. In Bresnan & Kaplan (1982: 87-148).

Iacobini, Claudio. 2008. Presenza e uso dei verbi sintagmatici nel parlato dell'Italiano. In Cini (2008: 103-20).

Iacobini, Claudio. 2009. Phrasal verbs between syntax and lexicon. *Rivista di Linguistica* 21(1). 97-117.

Iacobini, Claudio & Francesca Masini. 2007. The emergence of verb-particle constructions in Italian: locative and actional meanings. *Morphology* 16(2). 155-188.

Jackendoff, Ray. 1983. *Semantics and Cognition*. Cambridge & London: The MIT Press.

Laczkó, Tibor and Rákosi, György. 2011. On particularly predicative particles in Hungarian. In Miriam Butt, & Tracy H. King (eds.) *Proceedings of the LFG'11 Conference*, 299-319. Stanford: CSLI Publications.

Masini, Francesca. 2008. Verbi sintagmatici e ordine delle parole. In Cini (2008: 83-102).

Mateu, Jaume & Gemma Rigau. 2010. Verb-particle constructions in Romance: A lexical-syntactic account. *Probus* 23. 241-269.

Rizzi, Luigi. 2001. Il sintagma preposizionale. In Lorenzo Renzi, Giampaolo Salvi & Anna Cardinaletti (eds.), *Grande grammatica italiana di consultazione*, Volume 1: 521-545. Bologna: Il Mulino.

Salvi, Giampaolo. 1986. Asimmetrie soggetto/tema in italiano. In Stammerjohann (1986: 37-54).

Schwarze, Christoph. 1985. "Uscire" e "andare fuori": struttura sintattica e semantica lessicale. In Annalisa Franchi de Bellis & Leonardo M. Savoia (eds.) *Sintassi e morfologia della lingua italiana d'uso*, 355-371. Roma: Bulzoni.

Schwarze, Christoph. 1996. Die farblosen Präpositionen des Französischen: vage Prädikate oder Kasusmarker? *Romanische Forschungen* 108. 1-22.

Schwarze, Christoph. 2008. I verbi sintagmatici: prospettive di ricerca. In Cini (2008: 209-224).

Schwarze, Christoph. 2012. Romance clitic pronouns in lexical paradigms. In Sascha Gaglia & Marc-Olivier Hinzelin (eds.), *Inflection and Word Formation in Romance Languages*, 119-140. Amsterdam: Benjamins.

Simone, Raffaele. 1997. Esistono verbi sintagmatici in italiano? In Tullio De Mauro & Vincenzo Lo Cascio (eds.), *Lessico e grammatica*, 155-170. Roma: Bulzoni.

Snyder, William. 2001. On the nature of syntactic variation: evidence from complex predicates and complex word-formation. *Language* 77. 324-342.

Stammerjohann, Harro (ed.). 1986. *Tema-Rema in Italiano/Theme-Rheme in Italian/Thema-Rhema im Italienischen*. Tübingen: Narr.

Stiebels, Barbara.1996. *Lexikalische Argumente und Adjunkte: zum semantischen Beitrag verbaler Präfixe und Partikeln* (Studia Grammatica 39). Berlin: Akademie Verlag.

Svenonius, Peter. 2003. Limits on P: Filling in holes vs. falling in holes. *Nordlyd* 31(2). 431-445.

Svenonius, Peter. 2007. Adpositions, Particles, and the Arguments they Introduce. In Eric Reuland, Tanmoy Bhattacharya, and Giorgos Spathas (eds.), *Argument Structure*, 63-103. Amsterdam: John Benjamins.

Talmy, Leonard. 1985. Lexicalization Patterns: Semantic Structure in Lexical Forms. In Timothy Shopen (ed.), *Language Typology and Syntactic Description*, Vol. 3: 57-149. Cambridge: Cambridge University Press.

Talmy, Leonard. 1991. Path to realization: A typology of event conflation. *Berkeley Linguistics Society (BLS)* 17. 480–519.

Toivonen, Ida. 2001. *The Phrase-Structure of Non-Projecting Words*. Ph.D. dissertation, Stanford University.

NON-CORE PARTICIPANT PPs ARE ADJUNCTS

György Rákosi

University of Debrecen

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper revisits the question of whether optional, non-core participant PPs are to be treated as arguments or as adjuncts in linguistic theory in general and in LFG grammars in particular. I argue that a number of considerations converge on pointing towards the latter option.

1. Introduction

The need to be able to track arguments across verbal alternations has been an important motivation behind the use of thematic roles in linguistic theory. By classifying the noun phrase *the window* as the patient argument of the verb in both (1a) and (1b), a convenient tool is made available for the linguist to relate the object of the transitive construction to the subject of the intransitive construction.

(1)	a.	*The heat of the fire broke the window*.	**patient**
	b.	*The window broke*.	**patient**

At a pure descriptive level, (1) illustrates the fact that the expression of certain distinguished event participant types is not restricted to a specific syntactic function or position in a particular construction type.

While the exact nature of the relation between (1a) and (1b) is an issue in linguistic theory, the fact that *there is* a relation is unquestionable. Consider now the representative set of examples in (2) and (3) from the perspective (1) provides.

(2)	a.	*John shook hands with Kate*.	**comitative**
	b.	*John cut the meat with my knife*.	**instrument**
	c.	*John doesn't appeal to Kate*.	**experiencer**
	d.	*John baked Kate a loaf of bread*.	**benefactive**
	e.	*The heat of the fire broke the window*.	**cause**
(3)	a.	*John cleaned the room with Kate*.	**comitative**
	b.	*John broke the window with a hammer*.	**instrument**
	c.	*John seemed to Kate to be happy*.	**experiencer**
	d.	*John baked a loaf of bread for Kate*.	**benefactive**
	e.	*The window broke from the heat of the fire*.	**cause**

There is a good consensus in the pertinent literature that the underlined expressions in (2) are syntactic arguments of their verbs, with the type of participation that the descriptive labels on the right spell out. The underlined expressions in (3), which I will be referring to as *non-core participant PPs*, refer to event-internal participants whose type of participation is roughly describable by the same labels as in the corresponding cases in (2). An important question for linguistic theory is whether the two sets of expressions are to be related to each other, and if yes, then what exactly is the level where the correspondence is to be drawn. Most importantly, do non-core participant PPs populate a(rgument-)structure, and if they do, are they indistinguishable at this level from the underlined expressions in (2)?

This issue has received some fresh attention within the LFG literature in recent years. In their programmatic paper, Needham and Toivonen (2011) argue that non-core participant PPs form a subset of expressions that they analyse as derived arguments. Zaenen and Crouch (2009), on the other hand, argue that all semantically marked obliques should be treated as adjuncts. In earlier work on dative experiencers (Rákosi 2006a,b), I proposed an account in which non-core participant PPs are thematic adjuncts.

The goal of this paper is to revisit this question and provide further arguments for the adjunct analysis of non-core participant PPs. Assuming that the non-core participant PPs in (3) belong to the same broad semantic types as the respective arguments in (2), I embrace a view of grammar that by default allows for varying syntactic instantiations of participant types in the absence of constraints to the contrary. For example, comitatives can be on the argument list of verbs of social interaction and thus realized as complements, but they can also be licensed in a much larger set of contexts as adjuncts. This assumption can be viewed as a stronger version of the *dual analysis* of Dowty (2003), who argues that every complement can be analysed as an adjunct in the default case, and vice versa.

The paper does not discuss VP-internal directional, locative, source, manner, temporal or purpose PPs. I restrict my attention to the types listed above in (3), and I use the term non-core participant PPs to cover this subset of what have been called elsewhere circumstantial phrases (see, for example, Cinque 2006). Furthermore, I focus on English and Hungarian data, and I do not discuss applicative-marking languages, where the grammar of non-core participant PPs is markedly different.

The structure of the paper is as follows. After this introduction, I give an overview of previous LFG proposals in Section 2. In Section 3, I list a number of primary arguments in favour of the adjunct analysis. In Section 4, I take a bird’s eye view of conventional syntactic tests of argumenthood, only to conclude with other authors that these tests do not identify non-core participant PPs unambiguously either as arguments or as adjuncts. In Section 5, I add further conceptual arguments for the adjunct analysis, occasioned by two case studies that I briefly discuss. I conclude in Section 6.

2. An overview of the LFG literature

To my knowledge, the earliest discussion of the proper LFG treatment of non-core participant PPs appears in Bresnan (1982). Since the following passage gives a concise statement of the argument analysis, I quote it in full:

> "It is possible to define a lexical rule of *Instrumentalization* (analogous to lexical rules of *Causativization*) which converts an *n*-adic predicate argument structure *P* to an *n*+1-adic predicate argument structure *P-with* whose *n*+1st argument is assigned the grammatical function INSTR OBJ [instrumental object]. For motivation, note that Instrumentalization alters the inherent semantic properties of a predicate as illustrated in (54-55). …
>
> (54) a. John killed Harry.
>   b. John killed Harry with dynamite.
> (55) a. An explosion killed Harry.
>   b. #An explosion killed Harry with dynamite."
>
> (Bresnan 1982: 165)

This analysis is in the spirit of early approaches to argument structure in the 1960s, where non-core participant PPs were generally treated as arguments. In essence, Bresnan argues here that the lexical process in question creates an agentive predicate (55) from a basic lexical entry with an agent or cause subject argument (54). An obvious alternative is to assume a single lexical entry with a subject argument underspecified for agentivity, and to let the agent reading license the instrument. I discuss this alternative in some detail in Section 3. What is directly relevant is that as far as the treatment of non-core participant PPs is concerned, Bresnan (1982) can be regarded as the argument analysis par excellence in LFG.

Needham and Toivonen (2011) give renewed impetus to this analysis. They, nevertheless, do not treat what I call here non-core participant PPs as fully-fledged arguments, but as *derived arguments*. The following extracts from their paper are illustrative:

> "Bresnan (2001, 310) notes: 'The lexical stock of a-structures in a language can be extended by morphological means.' This implies that there is a basic lexical stock that is a subset of the entire lexical stock. The analysis that we propose assumes that arguments listed in the basic a-structure of verbs have a different status than arguments listed in the manipulated a-structure. …
> Our treatment of instrumentalization differs slightly from Bresnan's … but the basic idea is the same. … Under this analysis, *with*-

instruments are arguments. However, they are not listed as part of the basic argument structure of verbs, but optionally added … ."

That is, they assume that the lexicon contains core and derived entries that are distinguishable from each other, and, concomitantly, so are core and derived arguments. However, they do not formalize this difference beyond the proposal that lexical rules introduce derived arguments into what essentially is a regular argument structure.

Webb (2008) takes the underlying intuition a step further in his analysis of instruments. While discussing non-core instrument PPs as *adjuncts*, he does introduce them on what he calls the second tier of *argument structure*. Consider (4) for illustration:

(4) *Jack opened the door with the key.*

1st tier < p-a p-p >

2nd tier < p-a >

The thematic content of the two tiers is described in terms of Dowtian (1991) proto-roles, with the respective mappings being informally indicated here with the arrows. The instrument is treated as a sort of a secondary (proto-) agent, mapped onto the thematic oblique phrase. Hurst (2010) presents a very similar analysis of comitatives. Comitatives are briefly discussed here in Section 5.

My earlier proposal (Rákosi 2006a) is based on a rather similar core of background assumptions. A significant difference is that I explicitly treat non-core participant PPs as adjuncts that receive thematic specification in terms of Reinhart's (2002) Theta System. In Rákosi (2006b), I describe a possible LFG-theoretic implementation of this analysis. Consider the case of optional dative experiencers, which I claim to be thematic adjuncts:

(5) *This doesn't much matter to/for me.*

'matter	< [-m] >'	([-c])
	stimulus	(affected) experiencer / undergoer
	SUBJ	ADJ_{Θ}

A thematic adjunct is treated as a regular adjunct, but it is indexed by a thematic role label ([-m] stands for a mentally non-involved participant in Reinhart's system, and [-c] is a participant that is not related causally to the event). Such an adjunct is not introduced on the argument list, and the round bracket notation used in (5) is essentially only a reminder that a thematic adjunct of the given type can be licensed in the context of the argument structure to its left.

Finally, in what we can call the *pure adjunct analysis*, non-core participant PPs are treated as essentially regular VP-internal modifiers without any thematic specification. Asudeh and Toivonen (2007, 2012) argue, for example, that verbs can assign (non-thematic) semantic roles to adjuncts that are not on their argument list. In their analysis, the experiencer PP in (5) is a PGOAL ('goal of perception'). Driven mostly by the exigencies of computational implementation, Zaenen and Crouch (2009) make a proposal to treat *all* semantically marked PPs as adjuncts. I discuss their implementational concerns in Section 5. At this point it should suffice to note that there are proponents of the pure adjunct analysis within the LFG framework.

Thus the overall picture is a relatively varied scene stretching from the strong argument analysis to the strong adjunct analysis. This analytical spectrum seems to reflect an underlying variation in how strongly non-core participant PPs are assumed to be associated with the licensing verb. The strong argument analysis postulates a strong association, whereas the strong adjunct analysis stems from an increased emphasis on the independence of such PPs. If the above sample of analyses is representative, then LFG seems to have been moving towards the assumption a weaker association. I note here without further comment that this move parallels recent developments elsewhere in generative grammar, cf. especially the *generalized theory of applicatives* (Pylkkänen 2002, Cuervo 2003) and the *cartographic approach to circumstantials* (Schweikert 2004, Cinque 2006).

3. Primary arguments for the adjunct analysis

I repeat examples (3) as (6) to illustrate the forthcoming discussion.

(6) a. *John cleaned the room* (<u>*with Kate*</u>). **comitative**
b. *John broke the window* (<u>*with a hammer*</u>). **instrument**
c. *John seemed* (<u>*to Kate*</u>) *to be happy*. **experiencer**
d. *John baked a loaf of bread* (<u>*for Kate*</u>). **benefactive**
e. *The window broke* (<u>*from the heat of the fire*</u>). **cause**

The underlined PPs are all classified here as non-core participant PPs. As such, they all show a level of independence from the governing verb that is not characteristic of arguments. My aim in this section is to substantiate this essential fact of the grammar of non-core participant PPs.

It is a defining property of these PPs that they are syntactically optional. Each of the underlined phrases can be dropped in (6), and the remaining structure stays grammatical. And though certain types of arguments may also be optional, optionality is a characteristic property of adjuncts (see Asudeh and Toivonen 2012 for a discussion).

Whether non-core participant PPs are also optional semantically, i.e., whether they are entailed by the predicate or not, is a more contentious issue. The comitative (6a), the instrument (6b) and the benefactive (6d) are clearly not entailed. Cleaning, breaking or baking events do not need to involve either a participant who accompanies the agent, or an instrument, or someone who benefits from the results. As I briefly argue in Section 5, anticausative verbs are also non-causal in nature and, consequently, they do not entail the presence of a cause (6e). It follows then that the PP in (6e) genuinely introduces a cause, rather than spells out or modifies one that is present in the semantics of the verb. The existence of the participant denoted by the PP experiencer in (6c) does appear to be entailed – see Asudeh and Toivonen (2007, 2012) for an in-depth discussion of this issue. With other experiencer predicates, however, the presence of such an entailment relation is not so obvious, cf.:

(7) a. *This doesn't much matter.*
b. *This situation is unpleasant.*

Matter or *unpleasant* fairly frequently occur without the experiencer PP. It is often not trivially clear in these cases whether we are dealing with the lack of an entailed experiencer or with the presence of an entailed indefinite implicit argument. I have argued for the former position in Rákosi (2006a), and I refer the reader to Jackendoff (2007) for more on this issue. Here I simply conclude that non-core participant PPs are dominantly non-entailed, and dative experiencers represent a more complex case.

Given that these PPs are not subcategorized by the verb, their morphological form is not fixed, but is subject to variation as is allowed by the semantics of the given participant type. The following sentences illustrate how non-core PPs differ from true arguments in this respect:

(8) a. *This has never appealed to/*for me.*
b. *This has never occurred to/*for me.*

(9) a. *This doesn't matter to/for me.*
b. *This doesn't seem the best option to/for me.*

(10) a. *John shook hands with Kate.*
b. **John shook hands without/together with Kate.*

(11) a. *John cleaned the room with Kate.*
b. *John cleaned the room without/together with Kate.*

Experiencer and comitative arguments (8 and 10) are coded via designated markers, unlike the corresponding non-core PPs (9 and 11), whose morphosyntactic coding is subject to variation.

The assumption that non-core participant PPs are adjuncts explains why they do not change the semantic or the grammatical properties of the verb they combine with. Consider the following examples:

(12) a. *Peter works for Kate*.
b. *Peter works with Kate*.

(13) a. *Peter goes for Kate*.
b. *I like the salary that goes with the job*.

(14) a. *Peter walked for 10 minutes*/**in 10 minutes*.
b. *Peter walked to the bank *for ten minutes/in 10 minutes*.

(15) a. *This mattered only for 10 minutes*/**in 10 minutes*.
b. *This mattered to Peter only for 10 minutes*/**in 10 minutes*.

While *goes for* and *goes with* describe different types of events (13), the addition of a benefactive or a comitative PP in (12) does not change the semantics of the verb and the construction refers to the same type of *working* event as one without any non-core PP. (14) illustrates the well-known fact that directional PPs have the force of creating a telic predicate and thus change the inherent aspectual profile of the verb (the stars with the adverbials are relative to the intended, default aspectual interpretations of the predicates). Notice that this fact indicates that directionals are indeed more argument-like than the non-core participant PPs we investigate here. These latter do not change the inherent aspectual specification of the verb, as (15) shows. Thus non-core participant PPs seem to be modifiers, rather than arguments of the verb. This is only to be expected under the adjunct analysis.

Under this view, the data discussed by Bresnan (1982), which we have seen earlier in Section 2, receive a different explanation. If the instrumental in (6) is not an argument, but an adjunct modifier, the question is not what its insertion does to the base verb. Instead, the question is what properties of the base structure license the insertion of the instrument.

(16) a. *John killed Harry with a dynamite*.
b. #*An explosion killed Harry with a dynamite*.

As many have argued in the literature, instruments are licensed in the presence of an agent argument (see, a.o., Reinhart 2002 and Needham and Toivonen 2011). The subject argument of *kill* can be either an agent or a cause, but since only the former licenses the instrument, (16b) is not well-formed.

So non-core instruments at first appear to be licensed by the argument structure of the verb. But benefactives, for example, are known to be licensed by properties of the event denoted by the verb, rather by its argument structure (see Marelj 2005). The presence of an agent is required, but this agent can be only implied (17b,c) rather than be explicitly present (17a).

(17) a. *I did it for you*.
b. *He died for you*.
c. *I'll be there for you*.

On closer inspection, it turns out that instruments are also subject to somewhat weaker licensing conditions. As (18a) from Schütze (1995: 127) shows, they can be licensed in the presence of an implied agent; and Schäfer (2008: 99) argues that animacy in itself is enough to license an instrument even in the absence of volition (18b):

(18) a. *The nail came away from the wall with the back of a hammer*.
b. *John unintentionally broke the vase with the hammer*.

Space limitations prevent me from discussing further examples, but this behaviour is characteristic of each non-core participant PP type discussed in this paper. While their licensing is primarily dependent on the argument structure of the base verb, they do seem to be accommodated at the level of the event denoted. This property they share with agent-oriented adverbials, which are subject to similar, weak licensing conditions, cf.:

(19) a. *I am here deliberately*.
b. *I like you on purpose*.

We can conclude that as far as their licensing is concerned, non-core participant PPs pattern with certain types of adjuncts, rather than arguments.

I must also mention two facts that at first appear to render non-core participant PP similar to arguments. First, their semantics is not conditioned by the c-structure position that they occupy (20) – the *for*-PP has the same semantic type in its usual position (20a) as it has sentence-initially as a topic. This is a property they share with arguments (21).

(20) a. *John didn't bake a loaf of bread for Kate*.
b. *For Kate, John didn't bake a loaf of bread*.

(21) a. *John didn't appeal to Kate*.
b. *To Kate, John didn't appeal*.

Adverbial adjuncts fall into two groups in this respect. Light adverbials often have a position-sensitive interpretation. The pair in (22) is from Morzycki (2005: 8). In (22a), the adverb *happily* describes the manner of playing, but in (22b) it describes the speaker's attitude towards the event. Heavy adverbials, however, have invariable semantics that c-structure variation will not affect (23). The PP is a manner adverbial in both (23a) and (23b).

(22) a. *Clyde would play the tuba happily*.
b. *Happily, Clyde would play the tuba*.

(23) a. *In a happy manner, Clyde would play the tuba*.
b. *Clyde would play the tuba in a happy manner*.

Therefore non-core participant PPs, *qua* adjuncts, pattern with heavy adverbials, as is expected.

The second fundamental argument-like property of non-core PPs is that they are generally non-iterable, a point already raised by Bresnan (1982: 165). Schütze (1995: 130-131) points out nevertheless, that sometimes this constraint can be violated - compare (24a) with (24b).

(24) a. **I wrote this paper with my computer with my Macintosh Quadra*.
b. *I wrote this paper with my computer with Microsoft Word*.

Whether (24b) tells us something deep about the grammar of non-core instruments is an issue that I leave open here (cf. also Zaenen and Crouch (2009: 646) on how our linguistic stand may influence our interpretation of such data). What underlies the lack of iteration is the uniqueness constraint, and a more substantial question is whether this constraint regulates only true arguments. Asudeh and Toivonen (2012) argue on independent grounds that the domain of uniqueness includes not only arguments, but also a subset of non-thematic semantic roles (see also Carlson 1998 for an important discussion of this issue). If that is a legitimate extension, then non-iterativity is not a sufficient condition for argumenthood. What we have seen in this section then is a sort of behaviour which is fully consistent with and is explained by the assumed adjunct status of non-core participant PPs.

4. A quick look at the syntactic scene

Needham and Toivonen (2011) catalogue a number of syntactic tests that have been discussed in the pertinent literature as argument diagnostics, including preposition stranding, VP anaphora, VP-focussed pseudo-clefting, VP-preposing and *wh*-word conjunction. In the LFG literature, Bresnan (1982) and Dalrymple (2001) provide further overviews of the adjunct/argument distinction and its grammatical correlates. There is obviously a much wider literature on this fundamental issue, which I cannot give due credit to here. The single point I want to stress here, together with Needham and Toivonen (2001), is that non-core participant PPs show mixed behaviour with respect to traditional syntactic tests of argumenthood.

The data that I present here to illustrate this point concern preposition stranding and VP-preposing. Extraction of the NP-complement of a preposition is possible if the PP is an argument (25a). If the PP is a non-core participant phrase, then such extraction is sometimes possible (25c), sometimes not (25b). (25b) is taken from Needham and Toivonen (2011: 411).

(25) a. *Who did you tell it to that it is going to rain*?
b. **Who does it look to like it's going to rain*?
c. *Who did you cook the dinner with*?

On the basis of this test, non-core comitatives are argument-like, but non-core experiencers are not. Does this warrant the conclusion that the former is an argument but the latter is not?

First of all, preposition stranding is subject to many constraints that may influence the result of the testing. Consider the following data quoted from Hornstein and Weinberg (1981) in Cinque (2006: 150):

(26) a. *Who did John talk to about Harry yesterday*?
b. ??*Who did John talk about Harry to yesterday*?

(27) a. *Who did John talk to Harry about*?
b. ??*Who did John talk about to Harry*?

The basic order of the two PPs after the verb *talk* is *to*-PP > *about*-PP. If the P-object is extracted from its basic position (26a and 27a), then the result is fully acceptable for both PPs. If we flip the two, which is an otherwise attested and grammatical order, extraction and the concomitant preposition stranding becomes hardly acceptable (26b and 27b). Notice that this constraint has nothing to do with the argument status of the PP per se: it is a constraint that bans preposition stranding if the PP does not occupy its basic/neutral c-structure position. So at best, argumenthood is a precondition for preposition stranding (assuming that the *about*-PP in 26-27 is an argument, which is not obvious).

Still, let us suppose that the *with*-PP in (25c) is an argument on the basis of the success of preposition stranding. Consider then what happens if we try to apply the VP-preposing test (see Needham and Toivonen 2011). Only adjuncts can be left behind if the VP is preposed, but arguments must be included in the preposed VP.

(28) a. **I wanted to meet with him, and meet I did with him at 10pm.*
b. *I wanted to cook with him, and cook I did with him at 10pm.*

(28a) includes the agentive ('appointment') *meet* predicate, which takes an oblique *with*-PP. Since the PP is not included in the preposed VP, the result is ungrammatical. The *with*-PP in (28b) is, however, a non-core comitative phrase and it can stay behind. This tells us that the non-core comitative *with*-PP is an adjunct, contrary to our earlier conclusion reached on the basis of (25c).

If this brief example convinces the reader as representative, then we can conclude that non-core PPs show mixed behaviour with respect to traditional syntactic tests of argumenthood. Needham and Toivonen (2011) explain this by pointing out that some of these tests do not test for argumenthood *per se*, but may be conditioned by further functional or configurational factors. Furthermore, they also argue that some tests may distinguish between arguments and what they call *derived arguments*. Recall that non-core participant PPs form a subset of their category of derived arguments.

But there is another way of looking at this situation. Given that the test results are not consistent in this domain, there is no a priori advantage in classifying non-core PPs either as adjuncts or as arguments of any sort. And in fact, our grammar can stay more constrained if we allow for a strict and relatively well-behaving category of arguments, one which does not include non-core PPs. Classifying non-core PPs as adjuncts is no better or worse explanation for their mixed behaviour than classifying them as arguments. Given that we have seen a number of arguments supporting the adjunct analysis in Section 3, I conclude this section by maintaining this analysis in the face of the data discussed here.

5. Three further arguments for the adjunct analysis

Finally, I want to add three further conceptual arguments supporting the adjunct analysis that I have tried to substantiate in the previous sections. They support the adjunct analysis by offering theoretical and implementational advantages.

Recall that I started this paper with the assumption that languages, by default, allow for variable syntactic realizations of the same semantic supertypes of participants. So in principle, nothing precludes the possibility that a comitative or a cause can be realized as an argument in certain cases and as a (VP-internal) adjunct in certain others. The former happens in the case of designated predicate classes, and the latter happens, as we have seen, if certain properties of the event license the adjunct. I briefly discuss here two case studies as the first argument for the adjunct analysis. The logic of the argument is the same in the two cases: if we lift what I claim to be an adjunct participant PP to become an argument, then we lose the ability to properly account for why and how these phrases differ from true arguments.

Consider comitative arguments first, discussed in more detail in, among others, Dimitriadis (2008), Hurst (2010), Rákosi (2003, 2008) and Siloni (2008, 2011). Verbs of social interaction consistently take comitative arguments in Hungarian (29a), whereas non-core comitative adjuncts are generally licensed if an agentive participant is present (29b).

(29) a. *János* *csókolóz-ott* (*Kati-val*).
John kiss-PAST.3SG Kate-with
'John was involved in a mutual kissing activity (with Kate).'

b. *János fut-ott* (*Kati-val*).
John run-PAST.3SG Kate-with
'John ran with Kate.'

Both types are syntactically optional in Hungarian, but notice that argument comitatives are always entailed, as the English translation tries to show.

I mention here two further facts that differentiate the two types. First, argument comitatives have fixed coding, unlike adjunct comitatives, which can be modified. For expository purposes, I use English examples, but the same facts carry over to Hungarian.

(30) a. *John shook hands* (**together*) *with Kate*.
b. *John ran* (*together*) *with Kate*.

Second, only comitative arguments license anaphors in Hungarian, comitative adjuncts do not, cf.:

(31) a. *János és* *Kati* *egymás-sal* *csókolóz-t-ak*.
John and Kate each.other-with kiss-PAST-3PL
'John and Kate were involved in a mutual kissing activity with each other.'

b. **János és* *Kati* *egymás-sal* *futot-t-ak*.
John and Kate each.other-with run-PAST-3PL
'John and Kate ran with each other.'

Further differences between the two types are discussed in Rákosi (2003). These differences are substantial enough to claim that non-core comitatives are adjuncts, and not oblique arguments of the verb.

For a second quick thought experiment, consider the issue of the anticausative alternation:

(33) a. *The heat of the fire broke the window*.
b. *The window was broken by the heat of the fire*.
c. *The window broke* (*from the heat of the fire*).

The transitive *break* is obviously a semantically dyadic predicate, and so is the passive verb. Analyses diverge in whether they treat the passive *by*-phrase as an adjunct parasitic on the underlying but supressed argument position (see Grimshaw 1991), or as an oblique argument (see Kibort 2001). In either case, a cause argument is present in the semantic representation of the verb.

It is this causal component that is missing from the basic anticausative verb in (33c). As discussed by Piñón (2001), Reinhart (2002), Giorgolo and Asudeh (this volume) and Rákosi (2012), among others, there is no straightforward evidence for the presence of a causal component in either English or Hungarian anticausatives (but see Alexiadou et al. 2006, and Koontz-Garboden 2009 for claims to the contrary). The following example is from Giorgolo and Asudeh, and it illustrates that no external causer is entailed in the structure:

(34) *Yesterday, at three, the door closed. Nothing closed it*.

Thus the *from*-PP in (33c) must be a genuine introducer of a cause. That this is so is indicated by the fact that *from*-PPs can introduce causes even in the context of stative predicates, cf.:

(35) *She was somewhat tired from the journey*.

Given these considerations, it seems motivated to treat the *from*-PP in (33c) as an adjunct, not as an argument. Notice that if we did not do so, and treated this PP as an argument, then the difference between passive and anticausative structures would be somewhat mysterious. Also, such a move would be highly unnatural, since it would involve the deletion of a cause argument during anticausative formation and the subsequent introduction of another one via the insertion of the *from*-PP. Much more motivated is to assume that these *from*-causes are adjuncts, which is what I aimed to demonstrate.

The next (and the last) two arguments for the adjunct-analysis of non-core participant PPs are closely related, as they respectively target the same underlying issue from the perspective of the theory and that of the computational implementation. For the sake of argument, let us assume that the adjunct analysis is not on the right track, and the two bracketed PPs in (36) are optional or derived arguments.

(36) *I painted a picture* (*with Mary*) (*for her father*).

(37) a. $paint_1$ < agent, patient >
b. $paint_2$ < agent, patient, comitative >
c. $paint_3$ < agent, patient, benefactive >
d. $paint_4$ < agent, patient, comitative, benefactive >

Under the argument analysis of non-core participant PPs, we need at least the 4 lexical entries in (37) for the verb *paint* to be able to describe the data in (36).

Obviously, such a consequence is not alien to the spirit of LFG, given that it is designed to have a large lexicon. Appropriate lexical rules can derive (37b-d) from (37a), as both Bresnan (1982) and Needham and Toivonen (2011) show. The result will potentially be an exponential increase in the number of verbal lexical entries, many of which will come with heavy argument structures of a relatively large size. However, research on argument structure seems to have been going in a different direction. To be able to handle heavy argument structures, and, especially, to be able to distinguish a larger number of oblique argument types, we need a larger set of thematic roles (or features) than what most would like to see (see Carlson 1998 on this). And some theories of argument structure have been designed explicitly not to allow for more than 4 arguments in any given argument structure. Reinhart's (2002) Theta System, for example, is one such framework (see especially Marelj 2005).

The analysis in (37) also raises some issues for the computational implementation of LFG grammars. This is the major concern that Zaenen and Crouch (2009) have against the argument analysis: it creates oblique/adjunct ambiguities in parsing unless we are able to constrain it properly. But that task is not easy. Since core (or non-derived) argument structures like (37a) exists, any non-core PP can be analysed by default as an adjunct or as an oblique argument. If we, however, assume the adjunct analysis of non-core participant PPs, then this kind of parsing ambiguity disappears.

These last two considerations are not decisive in and of themselves. It is possible to maintain the argument analysis that (37) represents both from a theoretical and an implementational perspective. However, taken together with the rest of the argumentation that I have presented in this paper, I believe these considerations give further support to the adjunct analysis rather than weaken it.

6. Conclusions

In this paper, I have argued for the adjunct analysis of non-core participant PPs on the grounds of the following considerations. First, this analysis captures certain salient grammatical properties of non-core PPs in an obvious way. Second, acknowledging the mixed syntactic properties of non-core PPs, the adjunct analysis allows for a stricter and more constrained treatment of true arguments. Third, there are comparable constructions with participants belonging to the same broader semantic type where it is clearly motivated empirically to maintain an argument/adjunct distinction to be able to capture the facts. Fourth, the adjunct analysis has implementational advantages (as discussed by Zaenen and Crouch 2009). Five, no heavy argument structures are generated under the adjunct approach, which may be seen as an advantage given certain theoretical and pre-theoretical assumptions.

A question that I have not discussed here is whether non-core participant PPs, qua adjuncts, are to be distinguished formally from regular adjuncts. As I have briefly noted in Section 2, I proposed in earlier work that they may be indexed by thematic features (Rákosi 2006a,b). The resulting system is summarized in Table 1.

	+thematic	**-thematic**
+argument	ARG_{Θ}	ARG
-argument	ADJ_{Θ}	ADJ

Table 1. *Feature decomposition of argument and adjunct expressions*

Semantic arguments are thematic, but the type inventory of LFG also includes non-semantic (non-thematic) syntactic arguments. Expletives or “raised” arguments are treated as syntactic arguments of the matrix predicate that are not listed on the semantic argument list. Adjuncts do not receive a thematic role. If non-core participant PPs receive thematic specification, then the above feature-based inventory becomes complete.

This move raises a number of issues. Most importantly for our purposes, now we need to handle two types of adjuncts, rather than distinguish between two types of arguments, as happens in the system proposed by Needham and Toivonen (2011). Under either approach, we enrich the inventory of our grammar, which may have unwelcome consequences in both cases, some of which have been discussed in this article. This issue, however, is largely orthogonal to the primary claim that I wanted to defend in this paper: the adjunct analysis of non-core participant PPs may offer more advantages for LFG grammars than the argument analysis.

Acknowledgements

I thank the LFG12 reviewers and the audience of the Bali conference for useful comments and discussions. I also thank the editors of this volume for their assistance. Any errors that remain are solely mine.

I acknowledge that the research reported here is supported, in part, by OTKA (*Hungarian Scientific Research Fund*), grant number K 72983, and by the TÁMOP-4.2.1/B-09/1/KONV-2010-0007 project, which is co-financed by the European Union and the European Social Fund. I also thank the support of the Research Group for Theoretical Linguistics of the Hungarian Academy of Sciences at the University of Debrecen.

Bibliography

Alexiadou, Artemis, Anagnostopoulou, Elena and Schäfer, Florian 2006. The properties of anticausatives cross-linguistically. In Mara Frascarelli ed. *Phases of Interpretation*. Berlin: Mouton. 187-212.

Asudeh, Ash and Toivonen, Ida. 2007. Copy raising and its consequences for perception reports. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling and Christopher Manning eds. *Architectures, Rules, and Preferences: Variations on Themes by Joan W Bresnan*. Stanford: CSLI Publications. 49-67.

Asudeh, Ash and Toivonen, Ida. 2012. Copy raising and perception. *Natural Language and Linguistic Theory* 30 (2). 321-380.

Bresnan, Joan. 1982. Polyadicity. In Joan Bresnan ed. *The Mental Representation of Grammar*. Cambridge, MA: The MIT Press. 149-172.

Bresnan, Joan. 2001. *Lexical-functional Syntax*. Oxford: Blackwell Publishers.

Carlson, Greg N. 1998. Thematic roles and the individuation of events. In Rothstein, Suzanne eds. *Events and Grammar*. Dordrecht: Kluwer Academic Publishers. 35-51.

Cinque, Guglielmo. 2006. Complement and adverbial PPs: implications for clause structure. In Guglielmo Cinque. *Restructuring and Functional Heads. The Cartography of Syntactic Structures, Volume 4*. Oxford: OUP. 145-166.

Cuervo, María C. 2003. *Datives at large*. PhD dissertation. Cambridge, MA.: MIT.

Dalrymple, Mary. 2001. *Lexical Functional Grammar. Volume 34 of Syntax and Semantics*. New York: Academic Press.

Dimitriadis, Alexis. 2008. Irreducible symmetry in reciprocal constructions. In Ekkehard König and Volker Gast eds. *Reciprocals and reflexives. Theoretical and typological explorations*. Trends in Linguistics Studies and Monographs 192. Berlin/New York: Mouton de Gruyter. 375-409.

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67 (3). 547-619.

Dowty, David. 2003. The dual analysis of adjuncts/complements in Categorial Grammar. In Ewald Lang, Claudia Maienborn, Cathrine Fabricius-Hansen eds. *Modifying Adjuncts*. Berlin/New York: Mouton de Gruyter. 33-66.

Grimshaw, Jane. 1990. *Argument Structure*. Cambridge, MA: The MIT Press.

Hurst, Peter. 2010. The syntax of reciprocal constructions. In Miriam Butt and Tracy Holloway King eds. *The proceedings of the LFG10 conference*. Stanford: CSLI. 290-310.

Jackendoff, Ray. 2007. *Language, consciousness, culture – Essays on mental structure*. Cambridge, MA: The MIT Press.

Kibort, Anna. 2001. The Polish passive and impersonal in Lexical Mapping Theory. In Miriam Butt and Tracy Holloway King eds. *Proceedings of the LFG01 Conference*. Stanford, CA: CSLI Publications. 163-183.

Koontz-Garboden, Andrew. 2009. Anticausativization. *Natural Language and Linguistic Theory* 27 (1). 77-138.

Marelj, Marijana. 2004. *Middles and Argument Structure across Languages*. PhD dissertation. University of Utrecht.

Marelj, Marijana. 2005. From *with* to *for*. Ms. Utrecht: Uil-OTS.

Morzycki, Marcin. 2005. *Mediated Modification: Functional Structure and the Interpretation of Modifier Position*. PhD dissertation. University of Masschusetts Amherst.

Needham, Stephanie and Toivonen, Ida. 2011. Derived arguments. In Miriam Butt and Tracy Holloway King eds. *Proceedings of the LFG11 Conference*. Standford, CA: CSLI Publications. 401-421.

Pylkkänen, Liina. 2002. *Introducing arguments*. PhD dissertation. Cambridge, MA: MIT.

Piñón, Christopher. 2001. A finer look at the causative-inchoative alternation. In Rachel Hastings, Brendan Jackson, Zsófia Zvolenszky eds. *Proceedings of Semantics and Linguistic Theory* 11. Ithaca, NY: CLC Publications.

Rákosi György. 2003. Comitative arguments in Hungarian. In Willemijn Heeren, Dimitra Papangeli and Evangelia Vlachou eds. *Uil-OTS Yearbook* 2003. Utrecht Institute of Linguistics OTS. 47-57.

Rákosi, György 2006a. *Dative Experiencer Predicates in Hungarian*. PhD. Thesis. University of Utrecht.

Rákosi, György 2006b. On the need for a more refined approach to the argument-adjunct distinction: The case of dative experiencers. In Miriam Butt and Tracy Holloway King eds. *Proceedings of the LFG00 Conference*. Stanford, CA: CSLI Publications. 416-436.

Rákosi György. 2008. The inherently reflexive and the inherently reciprocal predicate in Hungarian: each to their own argument structure. In Ekkehard König and Volker Gast eds. *Reciprocals and reflexives*. *Theoretical and typological explorations. Trends in Linguistics Studies and Monographs 192*. Berlin/New York: Mouton de Gruyter. 411-450.

Rákosi, György. 2012. In defence of the non-causative analysis of anticausatives. Martin Everaert, Marijan Marelj and Tal Siloni eds. *The Theta System: Argument Structure at the Interface*. Oxford: OUP. 177-199.

Reinhart, Tanya. 2002. The Theta System - an Overview. *Theoretical Linguistics* 28 (3). 229-290.

Schäfer, Florian. 2008. *The Syntax of (Anti-)Causatives*. Amsterdam/Philadelphia: John Benjamins.

Schweikert, Walter. 2004. *The Order of Prepositional Phrases in the Structure of the Clause*. PhD dissertation. University of Venice.

Schütze, Carson. 1995. PP attachment and argumenthood. In Carson T. Schütze, Jennifer B. Ganger, and Kevin Broihier eds. *Papers on language processing and acquisition. MIT Working Papers in Linguistics Volume 267*. Cambridge MA: MIT. 95-151.

Siloni, Tali. 2008. The syntax of reciprocal verbs: an overview. In Ekkehard König and Volker Gast eds. *Reciprocals and Reflexives*. *Theoretical and typological explorations*. Trends in Linguistics Studies and Monographs 192. Berlin/New York: Mouton de Gruyter. 451-498.

Siloni, Tal. 2012. Reciprocal verbs and symmetry. *Natural Language and Linguistic Theory* 30 (1). 261-320.

Webb, James. 2008. *Instruments in LFG's Argument Structure*. Master's thesis. University of Oxford.

Zaenen, Annie and Dick Crouch. 2009. OBLs hobble computations. In Miriam Butt and Tracy Holloway King eds. *Proceedings of the LFG09 Conference*. Standford, CA: CSLI Publications. 644-654.

COMBINATORY POSSIBILITIES IN MURRINH-PATHA COMPLEX PREDICATES: A TYPE-DRIVEN APPROACH

Melanie Seiss
Universität Konstanz

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper is concerned with complex predicates in Murrinh-Patha, a Northern Australian language. In Murrinh-Patha, a verb usually has a bipartite structure, i.e., the lexical meaning of a word is determined together by two different parts. This paper looks at the combinatory possibilities of these two parts, establishes some factors which may play a role in the selection process and proposes a formal modeling of the data using a fine-grained semantic type hierarchy. The paper then compares this approach to accounts utilizing the idea of lexical conceptual structures (LCSs) (Jackendoff, 1990) such as e.g., Butt (1995) and Wilson (1999).

1 Introduction

This paper is concerned with verbal structure in Murrinh-Patha, a North Australian language. In Murrinh-Patha, a verb usually has a bipartite structure, i.e., the lexical meaning of a word is determined together by two different parts. These bipartite structures can be considered complex predicates in the sense of Alsina et al. (1997) as the two parts together determine the argument structure of the phrase.

Such bipartite verbal complexes have also been treated as instances of event classification (McGregor, 2002; Schultze-Berndt, 2000), as one of the parts, the so-called classifier stem, functions to classify the event. In (1a), the classifier stem HANDS(8) classifies the event as an event involving hands, while in (1b) the classifier stem FEET(7) classifies the event as involving feet. In both examples, the lexical stem *rirda* 'push' is used.[1]

(1)	a. *marntirda*	b. *nungarntirda*
	mam-rirda	nungam-rirda
	3sgS.HANDS(8).nFut-push	3sgS.FEET(7).nFut-push
	'He pushed him (with hands).'	'He kicked him.'

(Nordlinger, 2008)

As has been pointed out by Nordlinger (2008), sometimes it is obvious why a specific classifier stem is chosen (as in (1)), but sometimes it is not obvious how

[†]Many heartful thanks go to Rachel Nordlinger for extensive discussions about Murrinh-Patha and for testing some of the data in the field. I would also like to thank Joe Blythe for providing me with Murrinh-Patha data, Patrick Caudal for discussions about Type Composition Logic (Asher, 2011), my supervisor Miriam Butt and finally the audience of the LFG 2012 conference.

[1]Traditionally, the Murrinh-Patha classifier stems have been glossed with a number. In more recent publications it has become common to use small capitals to account for the generic meaning of the classifier stem and to keep the number to ensure compatibility with earlier publications. Murrinh-Patha classifier stems are inflected for subject person and number as well as for tense in portmanteau forms. The same classifier stem can thus have very different surface forms. The following abbreviations have been used in the glosses: sg = singular, pl = plural, S = subject, DO = direct object, RDP = reduplicated, nFut = non-future tense, Fut = future, PImpf = past imperfective, Pres = present tense, Asp = unmarked for aspect, Foc = focus marker, NC = noun class, DEM = demonstrative.

the meaning of the classifier and lexical stem is composed to form the meaning of the complex predicate. For example, in (2) it is not clear why the classifier stem HANDS(8) is used.

(2) *mam-pun-mardaraki*
3sgS.HANDS(8).nFut-3sgDO-disappoint
'He disappointed them.' (Nordlinger, 2008)

This paper investigates some of the factors which are involved in Murrinh-Patha event classification. McGregor (2002) claims that three different factors may be important in event classification in Australian languages generally: valency, aspect/Aktionsart and vectorial configuration. Seiss and Nordlinger (2010) were mainly concerned with the factors valency and aspect for Murrinh-Patha complex predicates. However, these two factors are not enough to explain the combinatory possibilities. Vectorial configuration, i.e., the lexical semantic content of the classifier and lexical stems, also plays an important role, which is the main focus of this paper.

In Murrinh-Patha, the same classifier stem can be used in a range of different complex predicates. For example, the classifier stem POKE(19) can be used in events in which contact is made with the tip of a long object, in events of linear movement and in certain mouth-associated events, among others. This variety can be nicely modeled with a type-driven approach in which a lexical item has a simple lexical entry with multiple typing restrictions. For this purpose, the paper makes use of Asher's (2011) idea of fine-grained semantic type hierarchies and his Type Composition Logic (TCL).

For the formal modeling of the lexical semantics of complex predicates in other languages, many approaches have utilized the idea of lexical conceptual structures (LCSs) (Jackendoff, 1990), e.g., Butt (1995), Andrews and Manning (1999), Broadwell (2000) or Wilson (1999). The paper compares the two approaches and shows that applying such approaches to the Murrinh-Patha data is difficult.

The paper is structured as follows. In section 2, a very brief introduction to Murrinh-Patha is presented. Section 3 is concerned with the argument structure of Murrinh-Patha complex predicates, showing that an account which builds purely on argument structure alone does not suffice. Section 4 then introduces data with the classifier stems POKE(19), BASH(14) and SLASH(23) which are used in the case study of the formal approach using types in section 5. Section 6 compares the type-driven approach to the more established LCS account and section 7 concludes the paper.

2 A brief introduction to Murrinh-Patha

Murrinh-Patha is a non-Pama-Nyungan language spoken in and around Wadeye in the Daly river region, approximately 400 kilometers south-west of Darwin. Green (2003) showed that Murrinh-Patha is related to Ngan'gitymerri (e.g. Reid, 2011), forming the Southern Daly language family. In contrast to most other Australian languages, it is still spoken in everyday life and still acquired by children, with about 2500 current speakers (Nordlinger, 2008).

Murrinh-Patha is a highly polysynthetic language with a complex verbal morphology and optional case and discourse marking on nouns. Besides the bipartite verbal structure of classifier and lexical stems which are the focus of this work, tense, reflexivity/reciprocality as well as subject and object markers can be part of the verbal complex. Additionally, body parts as well as adverbials and particles can be incorporated. For a detailed overview over the verbal template see Nordlinger (2010c). Further descriptions of the language include, among others, Street (1987); Walsh (1976); Nordlinger (2010a, 2011); Nordlinger and Caudal (2012) and Blythe (2009).

3 Argument structure

Seiss and Nordlinger (2010) provide a basic overview over the behavior of the argument structure in Murrinh-Patha complex predicates. They claim that in general, the classifier stem provides the number of arguments while the lexical stem fills in the thematic role specifications. However, they also report on (more or less common) exceptions from this general tendency.

This section provides an overview over the findings discussed by Seiss and Nordlinger (2010) and shows that valency is one factor in determining the selectional restrictions on classifier and lexical stems, but not the only one, with a complex interaction of the selectional factors.

Murrinh-Patha has 38 paradigms of classifier stems which can be roughly divided into intransitive, transitive and reflexive/reciprocal classifier stems. The classifier stems 1 to 6 are posture and motion classifier stems and have been glossed SIT(1), LIE(2), STAND(3), BE(4), PERCH(5) and MOVE(6). These classifier stems can also function as the sole verbal predicate, i.e., without an accompanying lexical stem, and are intransitive in these cases.

These classifier stems mostly form intransitive complex predicates with lexical stems. Two examples are provided in (3). In (3a), the lexical stem *karrk* 'cry' combines with the classifier stem SIT(1) to form a complex predicate. The lexical stem *karrk* 'cry' can be considered an intransitive lexical stem, as it never occurs in a transitive complex predicate and it refers to a semantically monovalent activity. Seiss and Nordlinger (2010) thus assume that *karrk* 'cry' contributes an agent to the complex predicate formation.

(3) a. *dim-karrk*
3sgS.SIT(1).nFut-cry
'He's crying.'

b. *dim-lerrkperrk*
3sgS.SIT(1).nFut-crush
'It's smashed.'
(Seiss and Nordlinger, 2010)

(3b) is another intransitive complex predicate formed with SIT(1). However, in this case the lexical stem involved is *lerrkperrk* 'crush', which is considered to be a transitive lexical stem as it otherwise combines with different transitive classifier stems in transitive complex predicates. *lerrkperrk* is hence treated as involving two arguments, an agent and a patient, by Seiss and Nordlinger (2010). The construction itself is considered to be an anticausative/resultative construction as the classifier stem SIT(1) only provides one argument slot, resulting in an intransitive complex predicate with just a theme argument.

The intransitive classifier stems 1 to 6 may also convey aspectual meaning and form transitive complex predicates such as in (4). Looking at the combinations for intransitive classifier stems with lexical stems thus reveals that argument structure plays a role in the selectional process, but that other factors are involved, too.

(4) a. *ku* *ngurlmirl* *wurran-ku*
NC$_{anim}$ fish 3sgS.MOVE(6).nFut-fish
'He continually catches fish.' (Seiss and Nordlinger, 2010)

b. *ngani-nan-part-nu-warda* *ngurru-warda*
1sgS.BE(4).Fut-2plDO-leave-Fut-now 1sS.GO(6).Fut-now
'I've got to leave you behind, I'm going.' (Seiss and Nordlinger, 2010)

Similar findings can be reported for transitive classifier stems and their combinatory possibilities with lexical stems. (5) provides an example of a typical transitive complex predicate formed with the classifier stem HANDS(8).

(5) *mam-kurrk*
1sgS.HANDS(8).nFut-scratch
'I scratched it.' (Seiss and Nordlinger, 2010)

Barone-Nugent (2008, 53) claims that the prototypical use of the classifier stem HANDS(8) is the following: "x is in physical contact over a period of time with y using hands". In this use, the classifier stem only ever occurs in transitive complex predicate constructions. However, the classifier stem HANDS(8) can also be used (although rarely) in intransitive complex predicates such as in (6). Barone-Nugent (2008) argues that HANDS(8) can be used in combination with the lexical stem *wel* 'glide' (and other similar lexical stems) because a wing can be seen as similar to a hand, and that the classifier stem HANDS(8) can be used in events involving hands more generally.[2]

(6) *ku* *murrirrbe* *mampel=kanam*
ku murrirrbe mam-wel=kanam
NC$_{anim}$ bird 3sgS.HANDS(8).nFut-glide=3sgS.BE(4).nFut
'The bird is gliding.' (Street, 1989)

[2]Boban Arsenijevic pointed out that the classifier stem in (6) could be understood as applying an action carried out with the hands to *yourself* and that the intransitive behavior results from this kind of reflexive reading. However, I would then expect the complex predicate to be formed with the reflexive/reciprocal version HANDS:RR(10).

As with the case of intransitive classifier stems, transitive classifier stems do not determine the argument structure of the complex predicate in all cases either. Semantic concepts such as hand-like body parts may play a role in the selection process of the classifier and lexical stem combination, overruling the valency requirements.

The aim of the remainder of this paper is to determine some semantic concepts which may play a role in the selection process and to present an account which explains the combinatory possibilities.

4 Case study: BASH(14), POKE(19) and SLASH(23)

This section introduces the core data for which the formal approach is tested in section 5. As a case study, the classifier stems BASH(14), POKE(19) and SLASH(23) are discussed, paying special attention to the similarities and differences in the combinatory possibilities. For POKE(19), Barone-Nugent (2008) has already presented a detailed study, cast within cognitive semantics. His findings are used here to contrast them with the behavior of the classifier stems BASH(14) and SLASH(23), for which Street (1989) as well as field notes by Rachel Nordlinger and Joe Blythe have been used as a data base.

Barone-Nugent (2008) states that the basic meaning of POKE(19) is that of events in which contact is made with a pointed end of an instrument, such as a stick or spear. This prototypical use of the classifier stem is illustrated in (7).

(7) *ku* *thithay* *nganthak=ngem*
ku thithay ngam-thak=ngem
NC_{anim} honey 1sgS.POKE(19).nFut-dip=1sgS.SIT(1).nFut
'I'm dipping into the honey.' (Street, 1989)

In this use of caused contact, the classifier stem POKE(19) contrasts with the classifier stems BASH(14) and SLASH(23). BASH(14) is used to denote events in which flat, solid objects such as stones, hammers etc. play a role. In contrast, SLASH(23) denotes events in which the long side of an object such as a knife etc. figures prominently. In this reading, a range of lexical stems combines with all three classifier stems which illustrates the difference in meaning nicely. An example is provided in (8) in which the lexical stem *wirntay* 'miss' is used with all three classifier stems. According to Street (1989), *wirntay* 'miss' in combination with POKE(19) means "miss with a spear", while it means "miss with a stone or short spear" with the classifier BASH(14) and "miss with a stick" with the classifier stem SLASH(23).

(8) a. *nga-wirntay-nu*
1sgS.POKE(19).Fut-miss-Fut
'I will miss (with a spear).' (Street, 1989)

b. *bangam-na-wirntay*
1sgS.BASH(14).nFut-3sgIOm-miss
'I missed him (with a stone).' (Street, 1989)

c. *thu* *thay* *pan-na-wirntay*
NC$_{weapon}$ stick 1sgS.SLASH(23).nFut-3sgIOm-miss
'I missed (hitting) him with the stick.' (Street, 1989)

However, not all lexical stems need to combine with all three different classifier stems to form complex predicates in which caused contact plays a role. The lexical stem *rtal* 'cut off', for example, only combines with the classifier stems BASH(14) and SLASH(23), denoting an action of cutting something down with an axe and of cutting something with a knife respectively. An action in which something is cut with the tip of an instrument probably does not exist or is very rare, so that the combination of *rtal* 'cut off' and POKE(19) does not exist (at least in my database).

(9) a. *thay* *wakal* *bangarntal*
thay wakal bangam-rtal
tree little 1sgS.BASH(14).nFut-cut.off
'I cut down the little tree with an axe.' (Street, 1989)

b. *nanthi* *terert* *pana* *ngu-rartal-nu*
NC$_{residue}$ many DEM 1sgS.SLASH(23).Fut-cut.off(RDP)-Fut
'I'll cut those things there with a knife.' (Street, 1989)

Beyond these basic meanings of the classifier stems BASH(14), POKE(19) and SLASH(23), each classifier stem also has further meanings. All three classifier stems can be used in events involving movement. Barone-Nugent (2008) showed that POKE(19) is used in linear movement, both in horizontal movement along the x axis as in (10a) and in vertical movement along the y axis such as in (10b).

(10) b. *nga-riwak-nu*
1sgS.POKE(19).Fut-follow-Fut
'I will follow him.' (Street, 1989)

a. *nga-wintigat-nu*
1sgS.POKE(19).Fut-descend-Fut
'I'm going down.' (Fieldnotes R. Nordlinger)

In contrast, BASH(14) and SLASH(23) can be found in the database with lexical stems of non-linear movement, such as circular or undirected movement as in (11a,b). They cannot be used with lexical stems of linear movement, as can be seen in (11c).

(11) a. *ba-rikat-nu*
1sgS.BASH(14).Fut-circuit-Fut
'I'll go around.' (Fieldnotes R. Nordlinger)

b. *ngu-rikat-nu*
1sgS.SLASH(23).Fut-circuit-Fut
'I will go around.' (Street, 1989)

c. **ba-wintigat-nu*
1sgS.BASH(14).Fut-descend-Fut (Fieldnotes R. Nordlinger)

A third meaning range for all three classifier stems can be identified as "mouth-associated" actions. Barone-Nugent (2008) points out that POKE(19) can be used in actions in which the mouth or mouth-associated body parts such as lips, teeth etc. play a role. He assumes that the perception of the tongue or the teeth as pointed ends may have enabled the use of the classifier stem POKE(19) in actions in which the mouth plays a role.

The subgroup of mouth-associated actions is quite large; it comprises, among others, blowing, licking and chewing actions. For the purpose of illustrating the different combinatory possibilities of the three classifier stems POKE(19), BASH(14) and SLASH, only two subgroups, speech actions and ingesting, is considered.

The classifier stem POKE(19) can be used both in complex predicates denoting speech actions and ingestion. Two examples are provided in (12).

(12) a. *nga-nhi-dharrpu-nu*
1sgS.POKE(19).Fut-2sgDO-ask-Fut
'I'll ask you.' (Street, 1989)

b. *kura* *parranthap*
kura parram-thap
NC_{aqua} 3plS.POKE(19).nFut-taste
'They tasted the water.' (Street, 1989)

In contrast, the classifier stem BASH(14) is not used in complex predicates denoting speech actions. It can be used in complex predicates of ingesting, such as in (13). In this meaning range, it can combine with some lexical stems which also combine with POKE(19), e.g., with *thap* 'taste'.

(13) *ku* *ngen* *ba-gatkat-nu*
NC_{anim} meat 1sgS.BASH(14).Fut-eat.until.satisfied
'I'll eat meat until I'm satisfied.' (Street, 1989)

SLASH(23) behaves the other way around, i.e., it can be readily used in speech actions ((14a)), but usually not in complex predicates of ingesting. The only two examples which can be found in my data base of SLASH(23) being used in a complex predicate of ingesting are given in (14b,c). In these examples it can probably be argued that it is not the ingesting that is important for the selection of the classifier stem but rather the action that leads to the food being brought close to the mouth.

(14) a. *pan-ngi-rerda=kanam*
3sgS.SLASH(23).nFut-1sgDO-blame=3sgS.BE(4).nFut
'He continually blames me' (Street, 1989)

b. *ku* *ngalek*
ku ngalek
NC_{anim} mosquito
puninkatattha=dini
puni-nkatat-dha=dini
3sgS.SLASH(23).PImpf-catch-PImpf=1sgS.SIT(1).PImpf
'He was catching mosquitoes (with his tongue).' (Street, 1989)

c. *ku* *lapi* *pan-purl*
NC_{anim} rib membrane 3sgS.SLASH(23).nFut
'He dragged the membrane from the rib bone with his teeth.'
(Street, 1989)

What becomes clear from this discussion of the data is that the same classifier stem can be used in a wide variety of different complex predicates. For some complex predicates, it is quite obvious why a specific classifier stem is used, e.g., in the examples involving caused contact above. Similarly, even for quite different complex predicates such as the caused contact and movement complex predicates involving POKE(19) it is understandable why the same classifier stem can be used. The core meaning of the classifier stem POKE(19) seems to be something like "moving in a pointy direction", which can account both for the caused contact complex predicates and for the movement complex predicates.

In contrast, for some complex predicates it is not clear why the same classifier stem is used. For example, Barone-Nugent (2008) suggests that POKE(19) is licensed in complex predicates in which the mouth plays a role because the teeth and the tongue can be perceived as the pointed end of a long object. However, this extension is quite difficult to accommodate, especially because the explanation does not extend to the cases in which SLASH(23) is used in speech actions. If POKE(19) is used in mouth-associated actions because the teeth and the tongue are received as pointed ends of long objects, it is not clear why at the same time the side of a long object should play a role in speech actions.

It is thus questionable whether all combinatory possibilities rely on cognitive, perceptual factors. In some cases, it may be pure morphological coincidences that the same classifier stem is used. Barone-Nugent (2008) argues for such an explanation for the classifier stem HANDS(8) which is used in actions performed by the hands but which is also used in speech acts. Barone-Nugent (2008) points out that the paradigm for the classifier stem HANDS(8) is very similar to the paradigm of the classifier stem SAY/DO(34) and that this similarity together with the similarity in meaning for the non-speech acts may have licensed the use of HANDS(8) with speech acts.

The approach taken here aims at looking at the combinatory possibilities and focusses on the factors which play a role in the selection of the classifier stems (such as caused contact, movement, etc.) and the subtypes which are determined by different classifier stems (linear vs. non-linear movement etc.). In the following section, an approach using a fine-grained semantic type hierarchy is proposed to model the various combinatory possibilities.

5 An account using types

The previous section discussed the classifier stems POKE(19), BASH(14) and SLASH(23) and their similarities and differences in the combination with lexical stems. The discussion showed that all three classifier stems can combine with some lexical stems to form complex predicates of caused contact, but also that subgroups of lexical stems exist which can only combine with one or two of these classifiers.

This section aims at a formalization of the findings of the previous section. It discusses the requirements that such an approach needs, introduces Asher's (2011) Type Composition Logic (TCL) and discusses how TCL can be applied to the Murrinh-Patha data.

From the discussion it should be clear that a simple enumeration approach, in which each possible classifier and lexical stem is listed, is not satisfying. What is needed instead is an approach which enables a flexible grouping of the classifier stems and the lexical stems into different subclasses to describe which combinations are possible.

While similar classifier stems have been discussed more or less closely together in a range of works, among them Schultze-Berndt (2000) and Reid (2011), this paper proposes a formal modeling of the grouping of classifier stems and lexical stems. The formal modeling should account for the flexible subgroupings of the classifier stems, i.e., it should model the fact that one classifier stem can belong to one or more subclasses, combining with various subgroups of lexical stems. Such an approach is offered by multi-dimensional type hierarchies.

For the purpose of modeling the type hierarchies, I use concepts adopted from Asher's (2011) Type Composition Logic (TCL). Asher (2011) proposes a very fine-grained semantic type hierarchy, in which very specific types as well as very general types can be assumed. He combines this type hierarchy with simple lexical entries which can come with a whole range of defeasible typing restrictions. This view of the lexicon makes it possible to account for the very general meaning of the Murrinh-Patha classifier stems and the flexibility in the combinatory possibilities with lexical stems in an elegant manner. TCL has also been used to account for verb-formation patterns in the Australian language Panyjima by Caudal et al. (2012).

Asher (2011) is mainly concerned with cases of coercion such as those given in (15). In (15a), people usually assume that Mary either started *writing* or *reading* the book. How people come to this understanding has been the matter of extensive research, with Pustejovsky's (1995) Generative Lexicon as a seminal work.

(15) a. *Mary started the book.*
b. *Mary enjoyed the book.*

c. *The goat enjoyed the book.*

Asher (2011) extends the Generative Lexicon approach to handle data more flexibly. He proposes that it is not only the object of the verb which triggers different event readings but that the choices can also be restricted by the subject. For example, if the subject is an *author* who started the book it is most likely that the event was one of starting to *write* a book. Alternatively, if the subject is a goat, then the event is most probably an *eating* event ((15c)).

Asher (2011) proposes that such specifications are part of the lexicon and are modeled as defeasible specifications such as in (16). For example, (16a) specifies that if there is a subject α which is a human and there is an object β which has a "physical" and an "informational" aspect ($\textsc{p} \bullet \textsc{i}$) such as a book, then it follows for a statement involving coercion ($\textsc{ev}(\alpha, \epsilon(\alpha, \beta))$) that the event ($\epsilon$) in which the subject and the object are involved is most probably a *reading* event. The $>$ is used as a weak conditional operator accounting for the defeasibility of the rule.

The defeasible rule in (16b) specifies that if the subject is not only human but also an author, i.e., a more specific sub-type, then the specification from (16a) can be overruled and the event is most probably a *writing* event. (16c) accounts for the fact that the event is most probably an *eating* event if the subject is a goat.

(16) Defeasible specifications (Asher, 2011, 228):

a. $(\alpha \sqsubseteq \textsc{human} \wedge \beta \sqsubseteq \textsc{p} \bullet \textsc{i}) \rightarrow (\textsc{ev}(\alpha, \epsilon(\alpha, \beta)) > \epsilon(\alpha, \beta) = \textsc{read}(\alpha, \beta)$

b. $(\alpha \sqsubseteq \textsc{author} \wedge \beta \sqsubseteq \textsc{p} \bullet \textsc{i}) \rightarrow (\textsc{ev}(\alpha, \epsilon(\alpha, \beta)) > \epsilon(\alpha, \beta) = \textsc{write}(\alpha, \beta)$

c. $(\alpha \sqsubseteq \textsc{goat} \wedge \beta \sqsubseteq \textsc{p} \bullet \textsc{i}) \rightarrow (\textsc{ev}(\alpha, \epsilon(\alpha, \beta)) > \epsilon(\alpha, \beta) = \textsc{eat}(\alpha, \beta)$

Asher (2011) offers a very detailed formal account of the mathematics of TCL which cannot be discussed here. But the examples given show the main properties of the TCL approach: there are simple types such as HUMAN or more complex types with multiple aspects such as $\textsc{p} \bullet \textsc{i}$. A type can be very generic such as physical property (P) or it can be very specific such as GOAT. These types in combination with a type hierarchy and the defeasible specifications account for the cases of coercions such as exemplified in (15).

The Murrinh-Patha data as illustrated and understood so far seems to need a detailed type hierarchy in which more specific types block the combination of lexical stems with classifier stems of less specific types similarly to blocking principles in morphology. For modeling this kind of type hierarchy, other accounts of subtyping could be used. However, it seems that for complex predicates crosslinguistically, the defeasibility of the specifications and the modeling of coercions which is build into TCL is needed as well.

The defeasibility of the specifications accounts for the fact that classifier and lexical stem combinations can be used in novel contexts denoting new meanings.

Coercion may be involved in cases in which the resulting complex predicate actually involves more than the pure sum of the meanings of the classifier and lexical stem. Butt and Geuder (2001) and Butt and Tantos (2004) discuss this issue for complex predicates in Urdu. An example of such a phenomenon in a Northern Australian language is found in Jaminjung. As Schultze-Berndt (2000) states, the inflecting verb HIT is normally used in cases in which impact is made in a non-specified way. It is thus similar to the Murrinh-Patha classifier stems BASH(14), POKE(19) and SLASH(23) but does not specify the shape of the instrument. However, HIT in certain complex predicate combinations "encodes complete affectedness" (Schultze-Berndt, 2000, 314). Schultze-Berndt (2000) illustrates this with complex predicates of 'encircling'. She states that a lexical stem such as *walig* 'move around' can combine with motion classifier stems, but in combination with HIT, a sense of complete encircling arises. An example is provided in (17). This can be modeled with the specification in (18) which states that if the classifier stem HIT (α) combines with lexical stems of encircling (β), the resulting complex predicate (*cp*(α, β)) is one of complete encircling.

(17) *walig* *gani-ma-m* *gurrurrij*
around 3sg:3sg-HIT-Pres car
'He walks around the car.' (Jaminjung, Schultze-Berndt, 2000, 314)

(18) ($\alpha \sqsubseteq$ HIT $\wedge\ \beta \sqsubseteq$ ENCIRCLE) $>$ *cp*(α, β) = COMPLETE ENCIRCLING

As no lexical stems and only few classifier stems can occur on their own in Murrinh-Patha, it is difficult to determine whether some meaning parts just evolve from the combination or whether these meaning parts are part of the classifier or lexical stem. However, a more refined understanding will probably reveal situations very similar to the Jaminjung case. For this reason, TCL is adopted as a formalism.

The TCL approach is now applied to the Murrinh-Patha data which was discussed in section 4. To model the similar behavior of POKE(19), BASH(14) and SLASH(23) with lexical stems, one can assume that all three classifier stems are of a rather general type CAUSED CONTACT. This is formalized in (19). The formula in (19c) accounts for the data in (8). It states that if α is a classifier stem of caused contact, and if β is a lexical stem of type MISS, e.g. the lexical stem *wirntay* 'miss', then most likely they can combine in a complex predicate and the resulting complex predicate is one of missed caused contact.

(19) a. POKE(19), BASH(14), SLASH(23) $\sqsubseteq$ CAUSED CONTACT(x,y)
b. *wirntay* 'miss' $\sqsubseteq$ MISS(x,y)
c. ($\alpha \sqsubseteq$ CAUSED CONTACT(x,y) $\wedge\ \beta \sqsubseteq$ MISS(x,y))
$>$ $cp(\alpha, \beta)$ = MISSED CAUSED CONTACT(x,y)

The lexical stem *rtal* 'cut off' from the examples in (9) is only listed with BASH(14) and SLASH(23) in Street (1989), not with POKE(19). To account for this combination, BASH(14) and SLASH(23) form a subgroup of the classifier stems of CAUSED CONTACT: they form a subtype CUTTING. The specifications in (20) account for the combinatory possibilities.

(20) a. BASH(14), SLASH(23) $\sqsubseteq$ CUTTING(x,y)
b. *rtal* 'cut off' $\sqsubseteq$ CUTTING(x,y)
c. ($\alpha \sqsubseteq$ CUTTING(x,y) $\wedge$ $\beta \sqsubseteq$ CUTTING(x,y))
$> cp(\alpha, \beta)$ = CUTTING(x,y)

As was also discussed above, the classifier stems show a different behavior when combined with lexical stems of movement. This means that the classifier stems belong to different types with respect to movement, i.e., POKE(19) has the type LINEAR MOVEMENT while BASH(14) and SLASH have the type NON-LINEAR MOVEMENT. The specifications in (21b) display different lexical stems with the subclasses LINEAR MOVEMENT and NON-LINEAR MOVEMENT.

(21) a. POKE(19) $\sqsubseteq$ LINEAR MOVEMENT
BASH(14), SLASH(23) $\sqsubseteq$ NON-LINEAR MOVEMENT
b. *riwak* 'follow', *wintigat* 'descend' $\sqsubseteq$ LINEAR MOVEMENT
rikat 'go around', *rdertpart* 'skirt' $\sqsubseteq$ NON-LINEAR MOVEMENT
c. ($\alpha \sqsubseteq$ LINEAR MOVEMENT $\wedge$ $\beta \sqsubseteq$ LINEAR MOVEMENT)
$> cp(\alpha, \beta)$ = LINEAR MOVEMENT(x,y)
($\alpha \sqsubseteq$ NON-LINEAR MOVEMENT $\wedge$ $\beta \sqsubseteq$ NON-LINEAR MOVEMENT)
$> cp(\alpha, \beta)$ = NON-LINEAR MOVEMENT(x,y)

The different typing restrictions are summarized in Figure 1 as a multi-dimensional type hierarchy. The hierarchy only comprises the data which has been discussed in this paper. A more elaborate structure is needed to account for all different combinatory possibilities in Murrinh-Patha complex predicate formation. The multi-dimensional hierarchy again shows the idea that a classifier stem can belong to more than one type and this enables the modeling of the fact that a classifier stem may pattern with other classifier stems for one type but does not need to pattern with these classifier stems for other types.

That the classifier stems indeed have multiple types along different dimensions can be seen in combinations of classifier stems with certain lexical stems such as *rikerdek* 'finish' or *wirntay* 'miss'. The resulting complex predicate carries the meaning of 'finish an event specified by the classifier stem' or 'miss an event specified by the classifier stem' respectively. That is, the combination of POKE(19) plus *rikerdek* can refer to either finishing a writing event or finishing an eating event. Similarly, POKE(19) plus *wirntay* can be used to refer to an event in which someone was missed with a spear or to an event in which a message was missed, referring to the 'talking' aspect of POKE(19). The meaning of this kind of complex predicate is underspecified if used in isolation but receives a specialized interpretation from the context.

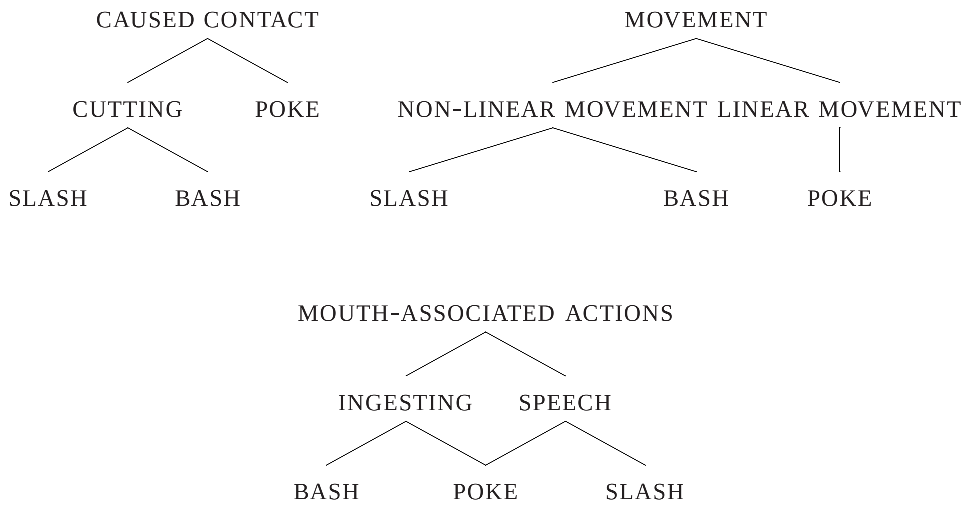


Figure 1: Multi-dimensional type hierarchy: for ease of readability, the hierarchies have been displayed in separate trees. On the left: a simplified type hierarchy for the type CAUSED CONTACT. On the right: a simplified type hierarchy for the type MOVEMENT. On the bottom: a simplified type hierarchy for the type MOUTH-ASSOCIATED ACTIONS.

To sum up, this section discussed a formal modeling of the combinatory possibilities using a multidimensional, fine-grained semantic type hierarchy. The following section compares this approach to the more established approaches using Lexical Conceptual Structures (Jackendoff, 1990).

6 Lexical Conceptual Structures

Many approaches to complex predicate formation use Jackendoff's (1990) LCSs to model the compatibility of the complex predicate constituents and to exclude illformed combinations. These analyses were inspired by Butt's (1995) analysis of complex predicates in Urdu. For Australian languages, LCSs have been used in the analysis of complex predicates in e.g., Wagiman (Wilson 1999, Wilson 2006), Wambaya (Nordlinger, 2010b) or across languages (Baker and Harvey, 2010; Andrews and Manning, 1999). In this section I compare the LCS approaches with the type-driven approach proposed in the previous section.

Butt (1995) is concerned with complex predicates in Urdu in which so-called light verbs, which roughly correspond to Murrinh-Patha classifier stems, combine with another verb, noun or adjective. To model the semantic contribution of each part of the complex predicate, Butt (1995) uses LCSs for each of the parts and different mechanisms of how these LCSs can combine. She accounts for the fact that light verbs are semantically bleached by proposing an LCS with a *transparent event* for the light verb. In complex predicate formation, the LCS of the full verb is inserted into the transparent event and, depending on the light verb, either event fusion or argument fusion takes place.

Wilson (1999) uses the idea of LCSs for complex predicate formation in Wagiman, a non-Pama-Nyungan language of the Northern Territory of Australia. However, he shows that a different approach is needed in Wagiman, as the two parts which form a complex predicate in Wagiman can combine in more diverse ways than in Urdu. This is also the case for complex predicates in Murrinh-Patha and other Australian languages. For this reason, his approach has also been adopted by others, e.g., Baker and Harvey (2010), and will therefore be discussed in more detail here.

In Wagiman, the complex predicate consists of two morphologically distinct words, and the terminology therefore differs from what is used for Murrinh-Patha. Wilson (1999) (among others) uses the term *inflecting verb* for the corresponding Murrinh-Patha classifier stem and the term *coverb* for the equivalent to the Murrinh-Patha lexical stem.

In contrast to Butt's (1995) approach to Urdu complex predicates, the LCS of the coverb is merged into the LCS of the inflecting verb **wherever it fits**. This accounts for the fact that coverbs and inflecting verbs can combine in various ways in Wagiman, and for the observation that inflecting verbs in Wagiman do not necessarily have to be semantically light. Ungrammatical combinations are ruled out if merging the LCS of the coverb into the LCS of the inflecting verb is impossible. This is best explained by an example. In Wagiman, coverbs of state can combine with stative inflecting verbs ((22a)) while coverbs of change of state cannot ((22b)).

(22) a. *ga-yu* *guk-ga* *gahan* *labingan*
3sg-be.Pres sleep-Asp that baby
'That baby is asleep.' (Wagiman, Wilson 1999, 150)

b. **bort-da* *ga-yu*
die-Asp 3sg-be.Pres
'He is dead.' (Wagiman, Wilson 1999, 150)

Wilson (1999) explains this with the fact that one can merge the LCS of *guk* 'sleep' into the LCS of -*yu*- 'be', but this is not the case for the LCS of *bort* 'die'. This contrast can be observed in (23a) and (23b). The LCS of *guk* 'sleep' can combine with the LCS of -*yu*- 'be', because both are states and the more detailed information for place in the LCS of *guk* 'sleep' can fill in the underspecified place in the LCS of -*yu*- 'be'. The LCS of the complex predicate *guk*-*yu*- ((23d)), thus, is the same as the LCS of *guk* 'sleep' in (23a).

(23) a. *guk* 'sleep'
$[_{State}\ \mathrm{BE}_{Ident}\ ([_{Thing}\ \ ]_A\ ,\ [_{Place}\ \mathrm{AT}_{Ident}\ (\ [_{Property}\ \textit{asleep}])])]$

b. *bort* ‘die’

$[_{\text{Event}}\ \text{BECOME}\ ([_{\text{State}}\ \text{BE}_{\text{Ident}}\ ([_{\text{Thing}}\ \ ]_{\text{A}}, [_{\text{Place}}\ \text{AT}_{\text{Ident}}\ ([_{\text{Property}}\ \textit{dead}\ ])])])]$

c. -*yu*- ‘be’

$[_{\text{State}}\ \text{BE}\ ([_{\text{Thing}}\ \ ]_{\text{A}}, [_{\text{Place}}\ \text{—}])]$

d. *guk*-*yu*- ‘sleep-be’

$[_{\text{State}}\ \text{BE}_{\text{Ident}}\ ([_{\text{Thing}}\ \ ]_{\text{A}}\ , [_{\text{Place}}\ \text{AT}_{\text{Ident}}\ (\ [_{\text{Property}}\ \textit{asleep}])])]$

The change of state coverb *bort* ‘die’, however, cannot combine with the inflecting verb -*yu*- ‘be’ because the two LCSs cannot be merged in an appropriate way. In Wilson’s (1999) account, the inflecting verb determines the general shape of the LCS of the complex predicates, which means that only the LCS of the coverb can be merged into the LCS of the inflecting verb, not vice versa. The combination **bort*-*yu*- ‘die-be’ is ungrammatical because the LCS of the coverb cannot be merged into the LCS of the inflecting verb -*yu*- ‘be’.

To summarize Wilson’s (1999) account of Wagiman complex predicates, he uses the compatibility of the LCSs of the inflecting verb and the coverb to explain grammatical and ungrammatical combinations. The rule he uses for the compatibility is very simple: a complex predicate is only grammatical if the LCS of the coverb can be fused into the LCS of the inflecting verb.

To apply this account to the Murrinh-Patha data is difficult. Although some patterns of combinations of inflecting verbs and coverbs are similar in Wagiman to the combinations of classifier and lexical stems in Murrinh-Patha, the differences that do exist result in a more complicated system. As a consequence, more rules for possible combinations would have to be defined to account for the Murrinh-Patha data. Additionally, different LCSs for the same classifier stem would be needed to account for the different combinations. LCSs thus do not serve to restrict the combinatory possibilities in Murrinh-Patha complex predicates as they do in Wagiman complex predicates. Consequently, an account building on the compatibility of the LCSs does not have explanatory power for Murrinh-Patha. This is not to say that LCSs cannot be helpful in establishing the meaning of a certain range of classifier and lexical stems and their combinations, but that an account in which the LCSs themselves account for the combinatory possibilities is not helpful. The remainder of this section discusses these claims in more detail.

One difference between complex predicate formation in Wagiman and complex predicate formation in Murrinh-Patha seems to be that in Murrinh-Patha, the number of arguments of the lexical stem can be reduced. This is the case for Murrinh-Patha anticausative/resultative constructions with the classifier stem SIT(1), discussed above and repeated in (24) for convenience.

(24) *dim-lerrkperrk*
3sgS.SIT(1).nFut-crush
‘It is smashed.’ (Seiss and Nordlinger, 2010)

In this anticausative/resultative construction, it seems that the single argument of the classifier stem SIT(1) picks out the theme object of the lexical stem and thus reduces the number of arguments the lexical stem takes. This cannot be accounted for by merging the LCSs of the lexical stem in terms of Wilson's (1999) proposal.

(25) a. SIT(1)
[$_{\text{State}}$ BE ([$_{\text{Thing}}$]$_{\text{A}}$, [$_{\text{Place}}$ —])]
b. *lerrkperrk* 'crush'
[$_{\text{Event}}$ CAUSE ([$_{\text{Thing}}$]$_{\text{A}}$, [$_{\text{Event}}$ BECOME
([$_{\text{State}}$ BE ([$_{\text{Thing}}$]$_{\text{A}}$, [$_{\text{Place}}$ AT ([$_{\text{Property}}$ *crushed*])])])])])]
c. SIT(1)-*lerrkperrk* 'be crushed'
[$_{\text{State}}$ BE ([$_{\text{Thing}}$]$_{\text{A}}$, [$_{\text{Place}}$ AT ([$_{\text{Property}}$ *crushed*])])]

The problem is that the LCS of the lexical stem should always be merged into the LCS of the classifier stem, which is not possible in this case. What happens intuitively is that the LCS of the lexical stem is reduced, i.e., the two events CAUSE and BECOME in the LCS of the lexical stem are deleted because they do not match with the LCS of the classifier stem. While the process of picking out a patient argument of a lexical stem can be explained in terms of LCSs, the process itself changes the algorithm put forward by Wilson (1999).

For other combinations of classifier and lexical stems it is not obvious how rules should be defined to combine the prototypical LCS of the classifier stem with an LCS of the lexical stem. The only way of accounting for these combinations seems to be to assume a different LCS for the classifier stem in different combinations. This can be illustrated with the different uses of the classifier stem POKE(19) discussed above. As we have seen, POKE(19) is used in constructions in which contact is made with the pointed end of a long object. This is considered the prototypical use of POKE(19) by Barone-Nugent (2008). An example was provided in (7) in which the lexical stem *thak* 'dip (into liquid)' is combined with POKE(19).

One could define an LCS for the prototypical use of POKE(19) as in (26a). This LCS uses the basic LCS proposed by Jackendoff (1990) for verbs of contact, in which the slot for the instrument is already filled by an object with a pointed end. Similarly, for the LCS of the lexical stem, the slot for the place has already been specified, i.e. liquid.

(26) a. POKE(19)
[$_{\text{Event}}$ CAUSE ([$_{\text{Thing}}$]$_{\text{A}}$, [$_{\text{Event}}$ BECOME ([$_{\text{State}}$ BE,
([POINTED_END_OBJECT], [$_{\text{Place}}$ —])])])])]
b. *thak* 'dip in liquid'
[$_{\text{Event}}$ CAUSE ([$_{\text{Thing}}$]$_{\text{A}}$, [$_{\text{Event}}$ BECOME ([$_{\text{State}}$ BE,
([$_{\text{thing}}$]$_{\text{A}}$, [$_{\text{Place}}$ IN ([LIQUID])])])])])]

Because the LCSs of the lexical stem and the classifier stem share most of their structure, they can be combined as in (27) to form a coherent complex predicate.

(27) POKE(19) + *thak* 'dip in liquid':

$[_{\text{Event}}$ CAUSE $([_{\text{Thing}}\ ]_{\text{A}}$, $[_{\text{Event}}$ BECOME $([_{\text{State}}$ BE,
([POINTED_END_OBJECT], $[_{\text{Place}}$ IN ([LIQUID])])])])]

The LCS account thus seems to work nicely for this kind of classifier and lexical stem combination. However, as was also discussed above, the classifier stem POKE(19) can also be used in complex predicates of movement, and actually in both transitive and intransitive ones. Examples were provided in (10) for the lexical stems *riwak* 'follow' and *wintigat* 'descend'. Another example is provided in (28) for the lexical stem *dhadumnum* which is paraphrased as 'bob/poke one's head up and down or in and out to look around' by Street (1989).

(28) *ku* *pangkuy* *pana-ka*
ku pangkuy pana-ka
NC$_{anim}$ snake that-Foc
danthadumnum=wurran
dam-dhadumnum=wurran
3sgS.POKE(19).nFut-poke.head(RDP)=3sgS.MOVE(6).nFut
'That snake is poking his head in and out looking around.' (Street, 1989)

To account for these combinations involving movement in an LCS account, different LCSs would be needed as different valencies as well as different path requirements are involved even for the linear movement usage of POKE(19). Because the range of combinations of lexical stems with POKE(19) is large, using different LCSs for all the minor differences is not feasible. The basic problem for an LCS account is that LCSs cannot capture the core semantic meaning of POKE(19) in these different combinations. This is also the case for other classifier stems.

Finally, it is not clear how to incorporate lexicalized combinations of classifier and lexical stems, i.e., combinations in which the selecting factors are not detectable, into an LCS account. A lexicalized combination has been given in (2); (29) provides another example involving the classifier stem POKE(19). Although the meaning range of the classifier stem POKE(19) is well studied thanks to Barone-Nugent (2008), so far no determining factor could be established which licenses the use of the lexical stem *riwiye* 'pollute' with it.

(29) *kura* *nga-riwiye-nu*
NC$_{water}$ 1sgS.POKE(19).Fut-pollute-Fut
'I will pollute the water.' (Street, 1989)

If such lexicalized combinations were to be incorporated into the LCS system, an LCS for the classifier stem and an LCS for the lexical stem would have to be stipulated to account for the combination. In contrast, TCL is more suited to incorporating lexicalized combinations as part of the fine-grained semantic type hierarchy.

To sum up, the combinatorial possibilities of the semantics of classifier and lexical stems is much higher in Murrinh-Patha than what has been described by Wilson (1999) (and the additions in Wilson (2006)) for Wagiman. This is due to the wide range of meanings which are associated with one classifier stem. In many cases a LCS decomposition is too detailed to account for the combinatory possibilities. That is, LCSs require the specification of the valency, the path requirements etc., while probably all that the classifier stem POKE(19), for example, denotes is that it has something to do with linear movement and a pointed end of a long object. In contrast to the LCS account, the type account is especially targeted at defining such classes of types and defining the behavior of the combinations accordingly. That is, the explanatory power of the type-driven approach lies in the possibility of grouping the lexical items into various types so that statements can be expressed for more than one lexical item.

7 Conclusion

This paper discussed complex predicate formation in Murrinh-Patha and proposed an analysis which uses a fine-grained semantic type hierarchy. This system models the fact that classifier stems have a rather general meaning which allows them to combine with a wide range of different lexical stems. In this system, classifier stems can have multiple different types while lexical stems usually have only one type. The grouping into types allows us to define possible combinations of classifier and lexical stems according to the defined subtypes.

The paper further looked at approaches for complex predicate formation involving lexical conceptual structures and discussed Wilson's (1999) approach for Wagiman in detail. It was discussed that such an approach using the compatibility of the LCSs of classifier and lexical stems does not yield the required explanatory power for Murrinh-Patha complex predicate formation, as many different fusion rules and LCSs would have to be defined to account for the variety in the data.

This is not to say that LCSs are not useful in determining the meaning contributions of some of the classifier and lexical stem combinations. But they cannot be used elegantly to determine whether classifier and lexical stem combinations are grammatical or ungrammatical as the combinatory rules for Murrinh-Patha are much more diverse than the rules for Wagiman.

However, the type-driven and LCS accounts can probably be combined to provide more insight into the process of complex predicate formation in Murrinh-Patha. That is, the type hierarchy could be used to define templates of LCSs for the classifier and lexical stems and how they combine. For example, a template LCS for the caused contact complex predicates could be defined in which the classifier stems fill in the slot for the instrument and the lexical stems fill in the result state. Further research is needed to pursue this approach. Future research will also include establishing what other semantic concepts play a role in Murrinh-Patha complex predicate formation.

References

Alsina, Alex, Bresnan, Joan and Sells, Peter (eds.). 1997. *Complex predicates*. Stanford: CSLI Publications.

Andrews, Avery D. and Manning, Christopher D. 1999. *Complex predicates and information spreading in LFG*. Stanford: CSLI Publications.

Asher, Nicholas. 2011. *Lexical meaning in context: A web of words*. Cambridge: Cambridge University Press.

Baker, Brett and Harvey, Mark. 2010. Complex predicate formation. In Mengistu Amberber, Brett Baker and Mark Harvey (eds.), *Complex predicates: A cross-linguistic perspective*, Cambridge: Cambridge University Press.

Barone-Nugent, Ben. 2008. Event classification: a cognitive semantic analysis of two Murrinh-Patha event categories, Honours thesis, University of Melbourne.

Blythe, Joe. 2009. *Doing referring in Murriny Patha conversation*. Ph. D.thesis, University of Sydney.

Broadwell, George Aaron. 2000. Choctaw Directionals and the Syntax of Complex Predicates. In Miriam Butt and Tracy Holloway King (eds.), *Argument Realization*, pages 111–134, Stanford: CSLI Publications.

Butt, Miriam. 1995. *The structure of complex predicates in Urdu*. Stanford: CSLI Publications.

Butt, Miriam and Geuder, Wilhelm. 2001. On the (Semi)Lexical Status of Light Verbs. In Norbert Corver and Henk van Riemsdijk (eds.), *Semi-lexical Categories: On the Content of Function Words and the Function of Content Words*, pages 323–370, Berlin: Mouton de Gruyter.

Butt, Miriam and Tantos, Alexandros. 2004. Verbal Semantics via Petri Nets. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG04 Conference, University of Canterbury, Christchurch*.

Caudal, Patrick, Dench, Alan and Roussarie, Laurent. 2012. A semantic type-driven cccount of verb-formation patterns in Panyjima. *Australian Journal of Linguistics* 32(1), 115–155.

Green, Ian. 2003. The genetic status of Murrinh-Patha. In Nicholas Evans (ed.), *The non-Pama-Nyungan languages of Northern Australia*, Canberra: Pacific Linguistics.

Jackendoff, Ray. 1990. *Semantic structures*. Cambridge: The MIT Press.

McGregor, William B. 2002. *Verb classification in Australian Languages*. Berlin: Mouton de Gruyter.

Nordlinger, Rachel. 2008. Reciprocals in Murrinh-Patha, Manuscript, University of Melbourne.

Nordlinger, Rachel. 2010a. Agreement in Murrinh-Patha serial verbs. In Yvonne Treis and Rik De Busser (eds.), *Selected Papers from the 2009 Conference of the Australian Linguistic Society*.

Nordlinger, Rachel. 2010b. Complex predicates in Wambaya. In Mengistu Amberber, Brett Baker and Mark Harvey (eds.), *Complex predicates: A cross-linguistic perspective*, Cambridge: Cambridge University Press.

Nordlinger, Rachel. 2010c. Verbal morphology in Murrinh-Patha: Evidence for templates. *Morphology* 20(2), 321–341.

Nordlinger, Rachel. 2011. Transitivity in Murrinh-Patha. *Studies in Language* 35(3), 702–734.

Nordlinger, Rachel and Caudal, Patrick. 2012. The tense, aspect and modality system in Murrinh-Patha. *Australian Journal of Linguistics* 32(1), 73–113.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

Reid, Nicholas. 2011. *Ngan'gityemerri: A language of the Daly River region, Northern Territory of Australia*. Outstanding Grammars from Australia, Munich: Lincom.

Schultze-Berndt, Eva. 2000. *Simple and complex predicates in Jaminjung*. Ph. D.thesis, University of Nijmegen.

Seiss, Melanie and Nordlinger, Rachel. 2010. Applicativizing complex predicates: A case study from Murrinh-Patha. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG10 Conference, University of Ottawa, Canada*, pages 416 – 436.

Street, Chester. 1987. *An introduction to the language and culture of the Murrinh-Patha*. Port Keats: Pacific Linguistics.

Street, Chester. 1989. Murrinh-Patha vocabulary, electronic MS Word file, ms, Darwin N.T.

Walsh, Michael. 1976. *The Murinypatha language of north-west Australia*. Ph. D.thesis, Australian National University.

Wilson, Aidan. 2006. Negative evidence in linguistics: The case of Wagiman complex predicates, BA Honours thesis, University of Sydney.

Wilson, Stephen. 1999. *Coverbs and complex predicates in Wagiman*. Stanford: CSLI Publications.

ISSUES CONCERNING CONSTRAINTS ON DISCONTINUOUS NPS IN LATIN

Liselotte Snijders
University of Oxford

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

In this paper, two main issues concerning discontinuity of Latin noun phrases (NPs) will be discussed from a lexical-functional grammar (LFG) perspective. It is often assumed that Latin word order is free and that therefore discontinuity is not subject to any constraints. However, it appears that the discontinuity of Latin NPs is in fact constrained. Two constraints on this type of discontinuity form problems for LFG as a theory of syntax. For the first constraint, one on discontinuous prepositional phrases (PPs), an exception to the principle of Economy of Expression is proposed. This paper claims that c-structure should be more constrained in LFG. A second constraint on Latin discontinuous adjuncts shows that the way in which LFG treats adjuncts (with a set notation) is problematic for a proper account of discontinuous adjuncts. This paper provides an initial discussion of this issue.

1 Introduction

Latin is well known to be a 'free word order language', a language with no fixed word order patterns. This variation in word order is allowed because the specification of grammatical relations in the sentence is not determined by the ordering of constituents, but rather by morphology. This means that Latin also allows a great deal of discontinuity in its phrase structure (Panhuis, 1982; Spevak, 2010). It is often believed that Latin allows discontinuity of phrases in an unlimited way. However, as shown by Bolkestein (2001), discontinuity in Latin is in fact constrained. Bolkestein found three absolute constraints on the discontinuity of NPs in Latin, two of which will be the focus of this paper.[1] Discontinuity (and Latin word order in general) is highly determined by information structure. In this paper, the focus will not be on these information structural tendencies, but only on strict constraints. LFG is generally capable of accounting for discontinuous phrases, as is shown for example in the work by Simpson (1991) on Warlpiri. However, two of the three constraints found by Bolkestein pose problems for LFG. These two constraints will be addressed in this paper. For the first constraint, involving prepositional phrases, this paper suggests an exception to Economy of Expression. It appears that Economy of Expression allows for too much freedom in word order when it comes to discontinuity of NPs, and this freedom needs to be restricted. A second constraint, involving discontinuous adjuncts, brings forth another problem; it shows that the way in which LFG treats adjuncts is not capable of accounting for discontinuous adjuncts at all in any language. This is highly problematic for LFG as a theory of syntax, and it appears that there is no proper solution at the moment. This paper will discuss the two main issues connected to these two constraints; it will illustrate how they are problematic for LFG and how they can potentially be accounted for.

Before turning to the real issues at hand, a note needs to be made of two issues concerning the use of Latin in linguistic research. Latin is a dead language and therefore only limited data is available. Due to the status of the (classical) Latin that is available to us (literary Latin), we can only draw conclusions on this type of Latin, not on Latin as a whole. Spoken Latin might have been somewhat different from the variety of Latin which is available to us and being used for academic research. However, this type of Latin is interesting in its own right. It is merely important to keep in mind that we are dealing with a specific variety of Latin. Secondly, assignment of grammaticality

[†] I would like to thank Mary Dalrymple for her help and her supervision in the writing of my master's dissertation (Snijders, 2012), which has led to this paper. Also her book (Dalrymple, 2001) has helped a great deal in the formation of my dissertation and this paper.

[1]For a full overview of the constraints and their LFG analysis, please see my master's dissertation (Snijders, 2012). This paper is an outgrowth of this dissertation. One of the constraints is not discussed here because its analysis is relatively straightforward in LFG and does not present any problems to the theory.

in a dead language is different than assignment of grammaticality in languages spoken today. Any grammaticality judgement of a dead language is based on whether a specific sentence is attested in the data or not. Unattested sentences are assumed to be ungrammatical (only if the data set is large enough). Examples in this paper which are marked as ungrammatical are thus assumed to be ungrammatical because they were not found in the data set.[2]

2 Discontinuity of NPs in Latin

There are two different types of discontinuous NPs in Latin, as defined by Spevak (2010).[3] The first kind is an 'obligatory' discontinuous NP in which the intervening element occurs in an obligatory position in the sentence. This happens with certain connective particles (e.g. *autem*, meaning 'indeed'), which always occur in the second position in the sentence. These particles can therefore split up an NP, if this NP occurs at the beginning of the sentence. The phenomenon of clitics appearing obligatorily in second position is very interesting, especially in relation to how it is involved in discontinuity. An LFG account of second position clitics in another dead language, Ṛgvedic Sanskrit, can be found in Lowe (2011). Because the type of discontinuity caused by an intervening second position clitic is obligatory and appears to be a specific rule for Latin, this is not relevant for the discussion at hand. The other type defined by Spevak, and the one relevant for this paper is the type of discontinuous NP with an 'alien' element intervening, something which is neither related to the NP nor appearing in its position obligatorily. This type of discontinuity is strongly determined by information structure. This type of discontinuous NP will be the only type considered in this paper. From here on, when talking about 'discontinuous NP's, the term will refer to discontinuous NPs with intervening alien elements.

Discontinuous NPs are quite common in Latin; according to Pinkster (2005), 12% of all NPs in Latin are discontinuous. Here a discontinuous NP refers to an (f-structure) NP with two or more parts which are separated in phrase structure (c-structure). As mentioned, the intervening element is not part of this (f-structure) NP; it is an 'alien' element (for example a verb). An example of this in Latin is the following:[4]

(1)	*...a*	*qua*	*ego*	***nullum***	*confiteor*	*aetatis*	*meae*
	of+ABL	which.ABL	I.NOM	no.ACC	admit.1SG	lifetime.GEN	my.GEN
	tempus	*abhorruisse...*					
	time.ACC	deter.INF.PASS					

'... of which I admit that at no point in my lifetime I have been deterred...'

(Cic. *Arch.* 1)

[2]Grammaticality in a language such as Latin (classical Latin in this case) is also dependent on the specific author, since some authors will allow certain constructions and ordering of constituents, whereas others may not. Therefore it is always important to keep in mind which data set is used for research of the kind in this paper. Most canonical texts of Classical Latin (often used for linguistic research) are available in the Perseus Digital Library, an open-access corpus which at the moment contains 5.5 million words of Classical Latin (http://www.perseus.tufts.edu/).

[3]Spevak (2010) defines three types of discontinuous NPs, but one of them is one in which the intervening element is a modifier of the larger NP. This is not a real type of discontinuity, and was therefore not taken into account in this paper.

[4]The glossing 'of+ABL' means that the preposition *a* has the meaning 'of' and takes a complement with ablative case. In Latin prepositions specify the case of their objects.

In this example, the two parts of the NP *nullum tempus* are separated in phrase structure.[5] The intervening elements are a genitive modifier (which is technically part of the larger NP) and the main verb of the sentence, which is unrelated to the NP. The two parts of the NP can be interpreted to be part of the same (f-structure) NP, shown by the fact that they have the same case marking (accusative in this case). Latin's system of morphology thus allows for discontinuity to occur. In this case the two separated parts of the NP are a determiner and a noun, but they can be two different elements, for example an adjective and a noun.

3 Discontinuity in LFG

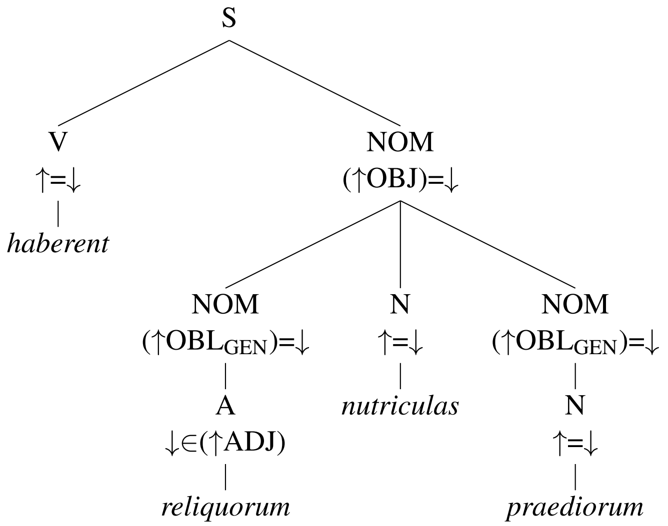
In this paper, the approach by Simpson (1991) accounting for discontinuous NPs in Warlpiri is followed, in order to account for discontinuity of phrases in Latin. In her account, Simpson simply assigns all elements of the discontinuous phrase the same category (N' in her case), and annotates these multiple occurrences of this category with the same grammatical function. An example of this for Latin is the tree structure in (3) for the sentence in (2):[6]

(2)	*...haberent*	***reliquorum***	*nutriculas*	***praediorum***.
	have.3PL.IMPF.CONJ	other.GEN.PL	foster-mothers.ACC	farms.GEN

'...they might have foster mothers for their other farms.'

(Cic. *Phil.* 11.12, taken from Bolkestein (2001, p. 253))

(3)



As one can see, the discontinuous genitive phrase *reliquorum...praediorum* ('for their other farms') is analyzed as being an argument of the noun *nutriculas* ('foster mothers') because a foster mother always has someone/something which she is taking care of. In the structure in (3), the adjective (A) and the noun (N) each form a separate NOM projection, both of which carry $(\uparrow \mathrm{OBL}_{\mathrm{GEN}}){=}\downarrow$ specification. If a nominal phrase is contiguous, it will form one NOM constituent, which is annotated once for its grammatical function. If it is discontinuous, every separate element

[5]Note that the case of *tempus* is accusative. Most commonly the ending *-us* is used for the nominative alone (as a matter of fact *tempus* is also used for the nominative), but the word *tempus* is of the third declension in Latin, meaning that the nominative and accusitive case have the same form, *-us* in this case.

[6]The type of nominal constituent is called 'NOM' here, a term introduced in Snijders (2012) to refer to any type of nominal constituent. This annotation was chosen in order to distinguish it from the classical NP: the structure of Latin nominals appears to be flat, and NPs are generally hierarchical constituents. For the purposes of this paper, the exact annotation is not very important, and 'NOM' can therefore be read as 'NP' or 'nominal constituent'.

or group of elements (if the NOM consists of more than two elements) forms its own NOM. These NOMs, part of the 'larger' nominal phrase (one could say), are all annotated with the same grammatical function in c-structure. Here the concept of *unification* is crucial. The annotation of separate elements with the same function ensures that the separate pieces of information of the 'larger' NOM appear in the same f-structure matrix: the information from both parts *unifies* in f-structure.

4 Problem 1: Constraint on Prepositional Phrases

4.1 Constraint and Data

One of the three absolute constraints on Latin discontinuous NPs found by Bolkestein (2001) is the following on prepositional phrases:[7]

(4) **Constraint 1 on Latin discontinuous NPs:**
No discontinuity between a P and the NP it governs (yet the NP may be internally discontinuous, meaning that part of the NP may be separated from the P)

This constraint states that there is no discontinuity between a preposition and the noun phrase which it governs. This constraint is similar to saying that preposition stranding is not allowed. Preposition stranding is allowed in some languages, for example in English:[8]

(5) *This is* ***the man*** *I told you* ***about****.*

This type of separation does not occur in Latin. Coming back to the Latin data, an example of a grammatical (attested) sentence is the following:

(6)

In	***quo***	*ego*	***accusatore,***	*iudices,*	*primum*	*illud*
in+ABL	which.ABL	I.NOM	prosecutor.ABL	judges.VOC	firstly	this.ACC

deprecabor,	*ne*	*quid*	*L.*	*Murenae*	*dignitas*	*illius...*
pray.1SG	so-that	any.ACC	L.	Murena.DAT	dignity.NOM	he.GEN

obsit...
injure.3SG.CONJ

'In the case of this prosecutor, gentlemen, I pray this first, that his dignity may not injure L. Murena in any way...'

(Cic. *Mur.* 58, taken from Bolkestein (2001, p. 251), but extended with original text)

In this example, the word *quo* has to appear adjacent to the preposition *in* (alternatively, *accusatore* would have to appear adjacent to *in* if *quo* was not in that position). Cases in which the preposition is completely separate from its NP are not found (in Bolkestein (2001)'s data), and it will therefore be assumed that they are ungrammatical. The NP in this example is internally discontinuous. The order within the NP is irrelevant: either dependent or head (noun) may come first. In (6), the order is P - D - N, and the noun is separated from the preposition. The determiner is also allowed to be separated from the preposition, and it may even occur before the preposition. This is shown in the next example:

[7] Bolkestein (2001) covers data from Cicero, Pliny the Elder and Petronius. In total she used 59 examples of discontinuous NPs from Cicero and 83 examples from the later works of Pliny the Elder and Petronius; 142 examples in total.

[8] Preposition stranding is allowed in English, but the place where the preposition occurs is very restricted.

(7) *Deinde, si* ***qua*** *ego* ***in*** ***re*** *fratri tuo*
secondly if any.ABL I.NOM in+ABL matter.ABL brother.DAT your.DAT
rei publicae causa restiterim...
good.GEN public.GEN for-the-sake-of oppose.1SG.PERF.CONJ
'Secondly, if in any matter I have opposed your brother for the sake of the public good...'
(Cic. *Fam.* 5.2.6, taken from Spevak (2010, p. 25))

Even a three-way split of the NP is allowed, but still part of it occurs adjacent to the preposition:

(8) *Ac ne* ***in*** ***hoc*** *quidem* ***tam*** ***molesto*** *tacebant*
and not in+ABL this.ABL even so unpleasant.ABL be-silent.3PL.IMPF
officio
duty.ABL
'And they were not even silent by this unpleasant duty.'
(Petr. *Sat.* 31, taken from Bolkestein (2001, p. 255))

The constraint in (4) implies that examples of the following kind are not grammatical (ungrammatical version of example (6)):

(9) ****In*** *ego* ***quo*** ***accusatore,*** *iudices, primum illud*
in+ABL I.NOM which.ABL prosecutor.ABL judges.VOC firstly this.ACC
deprecabor, ne quid L. Murenae dignitas illius...
pray.1SG so-that any.ACC L. Murena.DAT dignity.NOM he.GEN
obsit...
injure.3SG.CONJ

This type of example is indeed not attested in the data. It appears that preposition needs to occur to at least part of its object NP. In the following section, an analysis for this constraint in LFG will be presented.

4.2 Problem for Economy of Expression and Solution

This constraint poses a problem for the LFG notion of optionality of nodes, or in a formalized way, for Economy of Expression. Economy of Expression may be defined in the following way (Bresnan, 2001):

(10) **Economy of Expression:**
All syntactic phrase structure nodes are optional and are not used unless required by independent principles (completeness, coherence, semantic expressitivity).

This principle makes all nodes, complements and heads, optional, giving the c-structure a great degree of freedom. Economy of Expression works well for cases in which transformational theories have problems, for example when a maximal XP phrase does not dominate a corresponding X head, so-called headless constructions as occur for example in Russian. For examples of this see King (1995).

It appears that Economy of Expression is problematic for the PP-NP adjacency constraint in Latin (described in (4)). Economy of Expression ensures that there are no 'messy' empty categories, which is favorable for multiple reasons (Bresnan, 2001). One of these reasons is that empty nodes are problematic for natural language processing, for which syntax should be a model. However, it also means that there will be an overgeneration, or overacceptance of sentences, meaning that certain sentences are incorrectly classified as grammatical.[9] In other words, Economy of Expression predicts discontinuity in an unconstrained way (not only in Latin), which is contradictory to the data.

In order to overcome this problem, I propose to allow for exceptions to Economy of Expression, making it possible to make certain nodes obligatory. This will be explained in more detail later. Firstly, it must be made clear that the solution to the problem needs to be sought in the formulation of phrase structure rules, since this is a c-structure problem. Therefore I define a set of phrase structure rules for Latin which account for the constraint. These rules are presented below in (11). This set of rules accounts for the constraint on prepositional phrases, but not yet for the constraint on discontinuous adjuncts which will be discussed in the next section. In order to account for this constraint, the rule for S needs to be altered, which will be shown later. For the moment the following set of rules is proposed:[10]

(11) Non-final rules:

$$\text{GF} \equiv \{\text{SUBJ} \mid \text{OBJ} \mid \text{OBJ}_\theta \mid \text{OBL}_\theta \mid \text{ADJ} \in \mid \text{OBL}_\theta\ \text{OBJ} \mid \text{ADJ} \in \text{OBJ}\}$$

$$\text{S} \rightarrow \{ \underset{\uparrow=\downarrow}{\text{V}} \mid \underset{(\uparrow\ \text{GF}) =\downarrow}{\text{NOM}} \mid \underset{(\uparrow\ \{\text{OBL}_\theta|\text{ADJ} \in\}) =\downarrow}{\text{PP}} \}^*$$

$$\text{NOM} \rightarrow \{ \underset{\uparrow=\downarrow}{\text{D}} \mid \underset{\downarrow\in (\uparrow\ \text{ADJ})}{\text{A}} \mid \underset{\uparrow=\downarrow}{\text{N}} \mid \underset{(\uparrow\ \{\text{OBL}_\theta|\text{ADJ} \in\}) =\downarrow}{\text{PP}} \}^*$$

$$\text{PP} \rightarrow \underset{\uparrow=\downarrow}{\text{P}}\ ,\ \underset{(\uparrow\ \text{OBJ}) =\downarrow}{\text{NOM}}$$

Before discussing these rules, it is important to note that the P here may be either a preposition or a postposition, which is shown by the fact that there is a comma between P and NOM. The comma (or shuffle operator, this will be explained later) shows that the order of P and NOM inside the PP does not matter, and therefore that the P can be either a preposition or a postposition. In most cases the P will be a preposition: Latin has predominantly prepositions, and only two postpositions. Unfortunately Bolkestein (2001) does not explicitly mention postpositions; she only mentions prepositions. In order to keep the analysis uniform, it will be assumed that postpositions function in the same way as prepositions, but more research needs to be done on this issue. From this point onward, however, PPs will be referred to as prepositional phrases, but keep in mind that the rule in (11) also takes potential postpositional phrases into account.

As one can see in (11), three phrase structure rules are posited for Latin: one for the sentence S, one for the nominal constituents (NOM or NP) and one for the prepositional phrase PP. For

[9]Within LFG, with phrase structure rules being node admissability conditions, it is better to say overacceptance than overgeneration.

[10]GF refers to 'grammatical function'.

the purpose of this paper, other constituents such as adverbs (or adverbial phrases), auxiliaries, XCOMP and COMP are not directly relevant; therefore they are not included in these rules. The first rule shows that the structure of the sentence S is flat; this is indicated by the curly brackets, which denote disjunction (and not a set). The Kleene star (*) ensures that any number of constituents (as long as they are in the rule) can be daughters of S.[11] Verbs, nominal constituents and prepositional phrases (with the function of either oblique argument or adjunct) may occur in any order. As mentioned above, it is important to note that this ordering is largely determined by the sentence's information structure, and that it is not completely random.[12] However, in this paper we are concerned with absolute constraints and not tendencies, and therefore the information structural aspect of these issues will not be elaborated upon. The second rule shows that the structure of the nominal constituent (NOM) is also flat, but it may include hierarchically structured PPs. Determiner and noun may be c-structural co-heads, as signified by the annotation on D and N, $\uparrow=\downarrow$ for both.

The rule relevant for the constraint discussed in this section is the last rule: the one for the PP constituent. Firstly, it needs to be made clear why a PP constituent was posited to start with. Technically it is possible to not posit a PP and say that the S directly dominates prepositions and nominal complements of prepositions. This is possible because nominal complements can be marked as being objects of a preposition, either with $(\uparrow \text{OBL}_\theta \text{ OBJ})=\downarrow$ (for PP arguments) or $(\uparrow \text{ADJ} \in \text{OBJ})=\downarrow$ (for PP adjuncts).[13] The preposition would simply be marked with its grammatical function. This is possible in Latin because prepositions specify the case marking of their complement. However, it seems a little odd to have a preposition on its own be marked with a specific grammatical function. Also, the fact that discontinuity is relatively unconstrained, but that one of the constraints is that the preposition needs to be adjacent to at least part of its complement, is strong evidence in favor of the existence of the PP category. For this reason, the PP is posited as a constituent. However, if the PP is discontinuous, part of the f-structure prepositional phrase will still need to occur separately as a NOM marked as being the object of the preposition, with the notation $(\uparrow \text{OBL}_\theta \text{ OBJ})=\downarrow$ or $(\uparrow \text{ADJ} \in \text{OBJ})=\downarrow$, as mentioned above. A structural representation of this will be presented in the next section. A similar reasoning as for the PP constituent is used for the existence of the NOM constituent: 12% of all nominal phrases are discontinuous, meaning that contiguity of nominal phrases seems to be the norm. It would be strange to assume that this is a coincidence, and therefore it is assumed that there is a nominal constituent in c-structure (and not just separate adjectives, nouns and determiners annotated to be part of the same f-structure nominal constituent). For a more elaborate explanation of these considerations, see Snijders (2012).

Now let us turn back to the PP rule. This rule, unlike the others, does not involve a notation with Kleene star (and thus optionality) but rather requires that both the P and NOM are present inside the PP. This means that a PP node will always dominate a preposition and a nominal dependent, which is marked as the object of the preposition. By making the PP rule a separate rule, and by saying that in this rule the P and NOM are both obligatory (this nicely contrasts with the Kleene star notation in the other two rules), we make an exception to Economy of Expression. In general Economy of Expression works well (as mentioned, in specific constructions, see King (1995)) and the principle is one of the strong points of a constraint-based theory of syntax. In

[11] Kleene star traditionally means that zero or more of the constituents are present. The S-rule thus says that an S may contain any number of Vs, NOMs or PPs, in any order. One could also use a Kleene plus ($^+$) here, meaning that at least one of these constituents is present as a daughter of S. However, f-structure constraints should take care of the fact that in a sentence at least one constituent is present: therefore Kleene star suffices.

[12] Whether Latin has an underlying standard word order is debated, see Ledgeway (2011). It is in any case definitely true that Latin allows any ordering of subject, object and verb (Pinkster, 1990).

[13] Here, complement of a preposition and object of a preposition refer to the same thing.

transformational theories, phrase structure is less flexible, meaning that empty categories can be present. Also, making certain categories and levels of structure obligatory means that in the syntax there will be redundant structure. It seems implausible, particularly from the perspective of natural language processing, to posit that empty structure is present. Economy of Expression is a strong principle prohibiting this redundant structure. However, it does cause problems for an analysis of the constraint at hand.

One could alternatively propose to abandon Economy of Expression altogether. It is a crucial concept within LFG, however, and one of the clear points in which LFG differs from transformational accounts, as mentioned. Proposing to abandon Economy of Expression would mean a complete change in theory. Nonetheless, the constraint discussed in this paper has shown that Economy of Expression has weak points, and that we might want to reconsider at least part of it, by positing exceptions when necessary. It appears that c-structure is in some cases not constrained enough and that it will overadmit sentences; it allows for grammatical sentences, but also for certain ungrammatical ones. As explained, f-structure cannot rule these sentences out because the problem is not completeness/coherence of information. In terms of f-structure, all the information is present; it is the ordering that is the problem. Therefore I believe that c-structure should be more strongly constrained than it currently is in LFG.

As a final note, it should be mentioned that one could maintain a principle of economy without positing it explicitly. One simply needs to define the phrase structure rules in such a way that it is clear that in general, all nodes are optional. One would then need to find a proper notation to show that in some cases, nodes are in fact obligatory. As long as the notation is explained properly, this is possible. This is certainly a valid way of going about the issue, but it might become a little messy. Also, the fact that there is a principle such as Economy of Expression makes it explicit that in general, one assumes that there is no redundant structure present in the syntax. Therefore the proposal in this paper is to maintain Economy of Expression, but make exceptions possible.

4.3 Analysis of Discontinuous PPs: Structure

In order to make clear what a Latin prepositional phrase structurally looks like, and to give an illustration of the rules in (11), the structure of a Latin sentence including a PP will be presented here. In this paper it is assumed that an f-structure prepositional phrase which is discontinuous in c-structure consists of two (or more) parts: one PP node with part of the nominal object inside and one (or more) separate nominal object(s). These separate nominal objects are annotated as being the object of the preposition with either $(\uparrow \mathrm{ADJ} \in \mathrm{OBJ})=\downarrow$ or $(\uparrow \mathrm{OBL}_{\theta}\ \mathrm{OBJ})=\downarrow$ (depending on the grammatical function of the PP), with the value θ dependent on the grammatical function of the phrase, often dependent on the specific preposition. Example (6) is repeated below in (12):

(12)

__In__	*__quo__*	*ego*	*__accusatore,__*	*iudices,*	*primum*	*illud*
in+ABL	which.ABL	I.NOM	prosecutor.ABL	judges.VOC	firstly	this.ACC

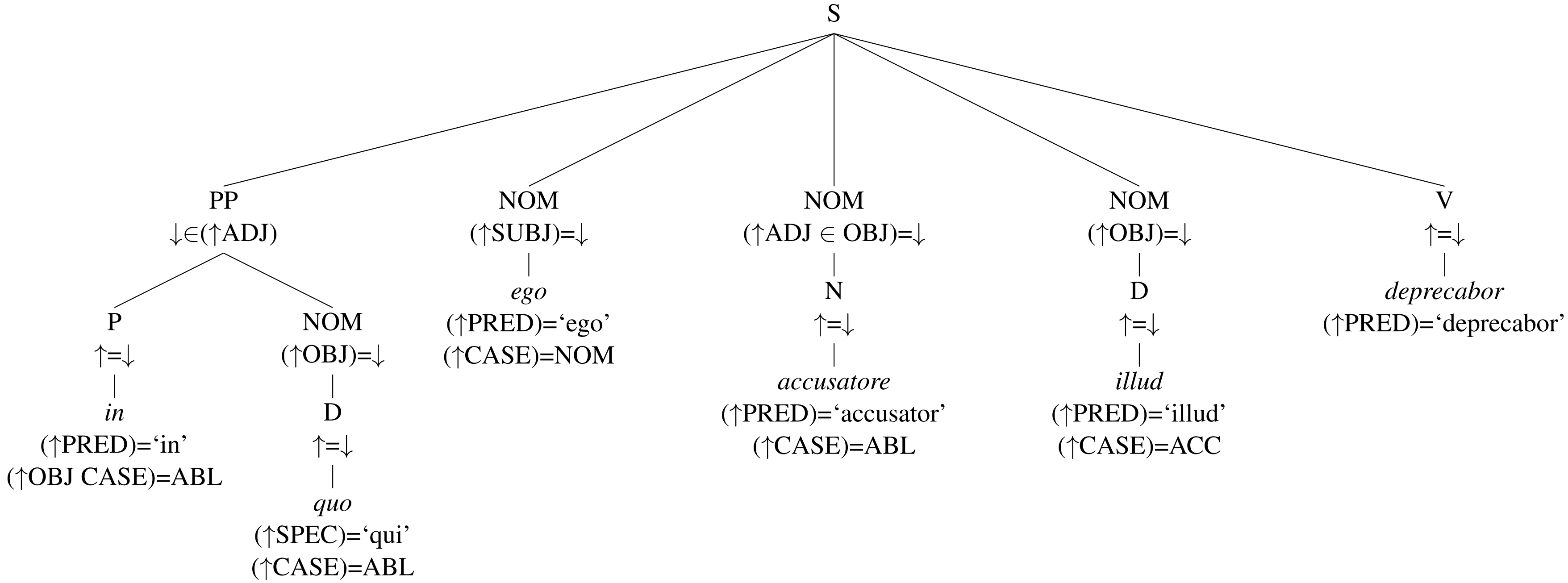
deprecabor,	*ne*	*quid*	*L.*	*Murenae*	*dignitas*	*illius...*
pray.1SG	so-that	any.ACC	L.	Murena.DAT	dignity.NOM	he.GEN

obsit...
injure.3SG.CONJ

'In the case of this prosecutor, gentlemen, I pray this first, that his dignity may not injure L. Murena in any way...'

The f-structure and c-structure of a simplified version of this sentence are the following (the sentence is simplified to make the analysis presented in this paper clearer):

(13)


S
PP
↓∈(↑ADJ)
P
↑=↓
in
(↑PRED)=‘in’
(↑OBJ CASE)=ABL
NOM
(↑OBJ)=↓
D
↑=↓
quo
(↑SPEC)=‘qui’
(↑CASE)=ABL
NOM
(↑SUBJ)=↓
ego
(↑PRED)=‘ego’
(↑CASE)=NOM
NOM
(↑ADJ ∈ OBJ)=↓
N
↑=↓
accusatore
(↑PRED)=‘accusator’
(↑CASE)=ABL
NOM
(↑OBJ)=↓
D
↑=↓
illud
(↑PRED)=‘illud’
(↑CASE)=ACC
V
↑=↓
deprecabor
(↑PRED)=‘deprecabor’

$$
(14)\quad \begin{bmatrix}
\text{PRED} & \text{‘deprecabor}\langle\text{SUBJ,OBJ}\rangle\text{’} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{‘ego’} \\ \text{CASE} & \text{ACC} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{‘illud’} \\ \text{CASE} & \text{ACC} \end{bmatrix} \\
\text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{‘in’} \\ \text{OBJ CASE} & \text{ABL} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{‘accusator’} \\ \text{SPEC} & \text{‘qui’} \\ \text{CASE} & \text{ABL} \end{bmatrix} \end{bmatrix} \right\}
\end{bmatrix}
$$

In this simplified c-structure and f-structure, we can see that the verb subcategorizes for a subject and an object, and that both are present in c-structure (and f-structure). The prepositional phrase is an adjunct, and therefore ends up in the set of adjuncts. In c-structure, the prepositional phrase consists of one PP node marked as an adjunct, and one NOM node marked as object of an adjunct. The information from both nodes ends up in the same f-structure, namely that of the adjunct in the sentence. In this sentence this notation works fine because there is only one adjunct, but note that problems might arise when the sentence has more than one discontinuous adjunct. This is a serious issue, which will be addressed in the next section. For the purpose of illustration, however, these two structures neatly show how Latin discontinuous PPs can be analyzed.

5 Problem 2: Constraint on Discontinous Adjuncts

5.1 Constraint and Data

A second constraint on discontinuous nominal phrases which is problematic for LFG is the following, as proposed by Bolkestein (2001):

(15) **Constraint 2 on Latin discontinuous NPs:**
No discontinuous adjuncts when the intervening element is itself an adjunct

This constraint states that in Latin, there are no discontinuous nominal phrases which are adjuncts with the intervening element being an adjunct. This ties in with two findings by Bolkestein (2001) that discontinuous adjuncts are rarer than discontinuous arguments, and that adjuncts as intervening elements are less common than arguments as intervening elements. In general, adjuncts are not involved in discontinuous nominal phrases as commonly as arguments are.

As an illustration of this constraint, see the following two examples:[14]

[14] The first example contains a so-called *Accusativus cum Infinitivo*-construction (*AcI*) in which a complement clause is formed by a subject in accusative case and verb in infinitive form. As one can see, it is translated with the accusative phrase being the subject and the verb in finite form in English.

(16)

ternis	*expeditionem*	*eam*	***mensibus***	*confici*
three.ABL	expedition.ACC	this.ACC	months.ABL	accomplish.INF.PASS

'that this expedition was accomplished in three months'

(Plin. *Nat.* 7.26, taken from Bolkestein (2001, p. 254))

(17)

***ternis**	*magna*	*celeritate*	***mensibus***	*expeditionem*
three.ABL	great.ABL	speed.ABL	months.ABL	expedition.ACC
confecit				
accomplish.3SG.PERF				

'?? he finished the expedition very quickly in three months'

(Taken from Bolkestein (2001, p. 255))

In the first example, the ablative adjunct phrase *ternis...mensibus* ('in three months') is discontinuous, and the intervening element is the subject of the complement clause (it has accusative case because of the *AcI*-construction of which it is part, see footnote 14). In the second example, which is of a type not occuring in the data, a discontinuous adjunct is split by another adjunct. This type of split is not attested in the data and we therefore assume that it is not grammatical.

The examples in the literature of adjuncts that are discontinuous only include adjuncts on the clausal level; adjuncts on the nominal level are not mentioned. The constraint found by Bolkestein discussed in this chapter seems to be true only for adjuncts on the clausal level, and because of lack of data we will assume this is the case and provide an analysis only for discontinuous adjuncts on the clausal level.

This constraint is highly problematic for LFG, as will be shown in the next sections. It turns out that not only the constraint itself is problematic, but more importantly it appears that an analysis for any type of discontinuous adjunct (no matter what element is intervening) is problematic. Before turning to the actual discussion of a potential analysis, it will be shown what type of elements may in fact intervene in discontinuous adjuncts, in order to give a complete overview of the data, and therefore to aid in finding a proper solution to this problem.

5.2 Intervening Elements

In finding an analysis for this constraint, an investigation into the type and number of elements splitting up a discontinuous nominal phrase was made. As mentioned earlier, discontinuity of NPs may be caused by obligatory intervening elements, when a specific element needs to have a set position in the sentence (e.g. second position). This is one type of intervening element. More interesting and relevant to the discussion in this paper is the type of discontinuous NP in which the intervening element is 'alien' and can be anything which is not directly associated with the discontinuous phrase. In discontinuity caused by alien elements only one element intervenes in the majority of cases. In Bolkestein's [2001] data from Cicero (containing 59 instances of discontinuous nominal phrases) there are only nine instances (15%) in which there is more than one constituent intervening. In these cases, they are very often only two elements, either the predicate and one of its arguments or two arguments, as in the following example:[15]

(18)

putares...	***aliquo***	*te*	*cum*	*hoc*	***rei***	***publicae***
think.2SG.CONJ	some.ABL	you.ACC	by	this.ABL	affair.GEN	public.GEN
vinculo	*esse*	*coniunctum*				
responsibility.ABL	be.INF	bind.PTPC				

[15]This example is another case of an *AcI*-construction.

‘you would have thought... that you were bound to him by some responsibility for the state’
(Cic. *Mur.* 64, taken from Bolkestein (2001, p. 252))

The intervening elements in this example are the subject of the verb, *te* (‘you’, in accusative case in the *AcI*-construction), and the oblique argument of the verb, *cum hoc* (‘to him’). In Bolkestein’s data set of Cicero there are no occurrences of adjuncts as intervening elements in any of the discontinuous nominal phrases (even when the discontinuous phrase is an argument).

In the works by Pliny the Elder and Petronius the number of intervening elements can be even larger, and adjuncts may intervene:

(19)	***nostram***	*scilicet*	*de*	*more*	*ridebant*	***invidiam***
	our.ACC	namely	out+ABL	habit.ABL	laugh.3PL.IMPF	envy.ACC

‘they namely laughed at our envy out of habit’
(Petr. *Sat.* 14, taken from Bolkestein (2001, p. 254))

In this example, an adverb, an adjunct PP and the main verb all intervene between the two parts of the object of the sentence. This shows that adjuncts may in fact occur as intervening elements when the discontinuous NP is an argument. Bolkestein (2001) only gives examples of intervening adjuncts which are prepositional phrases, not of, for example, bare ablative phrases (not governed by a preposition).[16] She notes that this needs further investigation. In any case we assume that adjuncts can intervene if the discontinuous nominal phrase is an argument. From the example in (19) it appears that any number of constituents may intervene between the parts of the discontinuous nominal phrase; the two parts of the NOM are located at the edges of the phrase. Since there are examples of this kind, we may conclude that there is no real constraint on the number of intervening elements.

From the work on this issue by Bolkestein (2001) (see also Spevak (2010)) we conclude that in general there may be more than one intervening element, but that in the case of discontinuous adjuncts, rare as they are, only one constituent may intervene, and this is never another adjunct. The intervening element may be either an argument of the verb, or the verb itself. This insight is important for a potential analysis, since one only has to constrain one element (the intervening element). If more intervening elements were allowed, it would be quite difficult to formalize how one of these intervening constituents is not allowed to be an adjunct. Even under the assumption that only one element may intervene, the analysis of this constraint in LFG will not be straightforward. In fact, this is an incredibly difficult, even seemingly impossible, constraint to analyze within LFG, as will be shown in the next section.

5.3 Search for an Analysis of Discontinuous Adjuncts

The constraint on discontinuous adjuncts and the number of intervening elements pose some problems for a potential analysis within LFG. First of all, the S rule which was posited in (11) is problematic. It is repeated below in (20):

(20) $\text{GF} \equiv \{\text{SUBJ} \mid \text{OBJ} \mid \text{OBJ}_\theta \mid \text{OBL}_\theta \mid \text{ADJ} \in \mid \text{OBL}_\theta\ \text{OBJ} \mid \text{ADJ} \in \text{OBJ}\ \}$

$$\text{S} \rightarrow \left\{ \begin{array}{ccccc} \text{V} & \mid & \text{NOM} & \mid & \text{PP} \\ \uparrow=\downarrow & & (\uparrow \text{GF}) =\downarrow & & (\uparrow \{\text{OBL}_\theta | \text{ADJ} \in\}) =\downarrow \end{array} \right\}^*$$

[16]Bolkestein does not even mention genitive phrases. Ablative phrases are nearly always adjuncts because of their nature of providing extra information not specified by the predicate, such as place, manner, accompaniment.

As explained in Section 3, this rule allows NOMs to be discontinuous (by positing more than one NOM in c-structure, the information of which unifies in f-structure), but there is no way to constrain what type of elements may intervene between the different parts of the nominal phrase. Anything is allowed to intervene according to the rule in (20). As explained in the previous section, it may be assumed that in most cases this is fine and that in general any number and kind of constituents are allowed to intervene. However, the rule needs to be changed in order to account for the constraint on discontinuous adjuncts.

The closest approximation of a properly working S rule is displayed below in (21). Unfortunately it is not completely capable of accounting for the constraint, or in fact any type of discontinuous adjunct, as will be explained later. However, it is displayed here in order to show what the problem is. The close approximation rule is the following:

(21) Closest approximation but faulty S rule:[17]

$\text{GF} \equiv \{\text{ARG-GF} \mid \text{ADJ-GF}\}$

$\text{ARG-GF} \equiv \{\text{SUBJ} \mid \text{OBJ} \mid \text{OBJ}_\theta \mid \text{OBL}_\theta \mid \text{OBL}_\theta\ \text{OBJ} \mid \text{ADJ} \in \text{OBJ}\ \}$

$\text{ADJ-GF} \equiv \text{ADJ} \in$

$$\begin{array}{cccccccccc} \text{S} & \rightarrow & [\{ & \text{V} & \mid & \text{NOM} & \mid & \text{PP} & \}^*] & , \\ & & & \uparrow=\downarrow & & (\uparrow \text{ARG-GF}) =\downarrow & & (\uparrow \{\text{OBL}_\theta|\text{ADJ-GF}\}) =\downarrow & & \end{array}$$

$$\begin{array}{cccccccc} [(\text{NOM}) & (\{ & \text{V} & \mid & \text{NOM} & \}) & (\text{NOM})]^* \\ (\uparrow \text{ADJ-GF}) =\downarrow & & \uparrow=\downarrow & & (\uparrow \text{ARG-GF}) =\downarrow & & (\uparrow \text{ADJ-GF}) =\downarrow \end{array}$$

In this rule, the shuffle operator (comma) is used and the constraint on the discontinuous adjunct is expressed in the bracketed group after the operator. The shuffle operator ensures that anything occuring before it in a PS-rule can be freely ordered (‘shuffled’) in relation to everything occuring after it in a PS-rule. In the rule below, the shuffle operator ensures that anything occuring after it in the rule (for example, an adjunct NOM) can occur anywhere within the first part of the rule before the operator. This means that an adjunct NOM can, for example, occur in between an argument NOM and a verb (these last two are freely ordered as well, shown by the disjunction notation).

The first part of the rule (before the shuffle operator) is very similar to the S rule proposed in (11), except that the NOMs inside this part of the rule are only allowed to be arguments. Adjunct NOMs are represented in the bracketed group after the shuffle operator. This part of the rule ensures that, first of all, adjuncts are optional (denoted by the parentheses, although Economy of Expression technically takes care of this). Secondly, it makes sure that the only type of intervening element in a discontinuous adjunct is either a verb or an argument of the verb.[18] Thirdly, the Kleene star ensures that more than one discontinuous adjunct is allowed per sentence (which should technically be possible).

[17] Note that GF is now divided into argument functions (ARG-GF) and adjunct functions (ADJ- GF); this distinction is used to show that only arguments may intervene in discontinuous adjuncts.

[18] There appear to be no examples in the data of argument PPs occuring as intervening elements of discontinuous adjuncts. If it is the case that argument PPs are allowed to occur in this position, the group V | NOM can be extended to include PP (with annotation $(\uparrow \text{OBL}_\theta)=\downarrow$).

This rule unfortunately fails to account for the fact that the two adjunct NOMs in the second part of the rule can form a unit in f-structure. With the current annotation, all adjuncts end up in the same set, and there is no proper distinction between possibly different (f-structure) adjuncts. For example, if there are two adjuncts with the same case, number and gender in c-structure, there is no real way to distinguish between the situation in which they form separate f-structure adjuncts (separate elements in the set of adjuncts) or the situation in which they form two different subparts of the same f-structure adjunct (in a discontinuous phrase, as one element in the set of adjuncts). The only thing constraining this is the concept of PRED clash. One adjunct cannot have two PRED values. This is illustrated by the example f-structure below:

(22)
$$\left[\begin{array}{ll} \textsc{pred} & \text{`VERB}\langle\textsc{subj}\rangle\text{'} \\ \textsc{subj} & \left[\textsc{pred}\quad \text{`SUBJ'}\right] \\ \textsc{adj} & \left\{ \left[\textsc{pred}\quad \text{`ADJ1'}\right], \left[\textsc{pred}\quad \text{`ADJ2'}\right] \right\} \end{array}\right]$$

This example f-structure shows that all adjuncts are in a set, and each adjunct has its own PRED value. If a sentence has two adjuncts and they both provide a PRED value, they will therefore be two separate adjuncts in the set of adjuncts in f-structure. If there is a discontinuous adjunct, consisting of for example a determiner and a noun, the information from both parts of the discontinuous adjunct will unify in f-structure in the same element in the set of adjuncts. The determiner does not provide a PRED value (rather it provides a SPEC value), and it will therefore need to merge with an element that does provide a PRED value (because every separate f-structure adjunct has a PRED value). This unification will only happen if case/number/gender of the determiner and the noun agree. In most cases this will work fine, but a serious issue arises when there is more than one determiner present in the sentence with the same agreement values (for case, number and gender), annotated as being an adjunct. There is in principle no way to link the correct determiners to the correct nouns in this case. In actual language comprehension, this would most likely not be a problem because the reader can infer from the semantics of the words what is going on. However, the core issue remains that in LFG there is no way to mark two parts of a discontinuous adjunct in such a way that it shows that they are part of the same f-structure adjunct. One cannot exclusively link two different c-structure adjunct nodes together in one f-structure element, which does happen with grammatical functions. For example, a sentence exclusively has one subject only, and every node annotated with $(\uparrow\text{SUBJ})=\downarrow$ will end up in the f-structure of the SUBJ. This is not possible for adjuncts. There is no way to control how different nodes annotated with $\downarrow\in(\uparrow\text{ ADJ})$ are assigned a place in the set of adjuncts. This is a major problem in the treatment of (discontinuous) adjuncts. This means not only that the constraint discussed in this section cannot be analyzed properly, but also that in general discontinuous adjuncts are a major problem for LFG.[19]

It appears that LFG cannot account for the constraint that a discontinuous adjunct cannot have another adjunct as its intervening element. With the use of the shuffle operator it is possible to constrain what types of intervening elements may occur between the two parts of the discontinuous NOM, but the real problem is the annotation on adjuncts. In LFG currently, adjuncts are marked in

[19] As mentioned by Dag Haug at the LFG12 conference, there might be a way to solve this issue in a computational way, by assigning one of the parts of the discontinuous phrase an annotation with functional uncertainty. This is very briefly mentioned in Haug (2011, pg. 8). This can then be implemented as such in XLE, for example. This is not part of my work, but will be left for future research.

such a way that they all end up in the same f-structure set, which is problematic. The problem here is that merging of adjuncts is (relatively) uncontrolled. In languages such as English this is fine because in English one adjunct appears in the c-structure as one node, because they are not allowed to be discontinuous. Each adjunct will become a separate member in the set of adjuncts (with its own PRED value). In the case of discontinuous adjuncts in Latin, however, one cannot ensure that the two different nodes marked as being adjuncts form a unit in f-structure. Merging (unification) of adjuncts is uncontrolled (apart from when there is a PRED clash). This is a serious problem in LFG, for which there is not yet a solution. If there was a way to ensure that the two adjunct NOMs were part of the same adjunct in f-structure, the rule would work fine. Finding a solution might involve marking the two parts of the adjunct in such a way that they 'belong together', but an important question here is how this would affect the computational power of the system. This is an important consideration to take into account. Instead of finding a solution for this in LFG, one could also find a different way to account for discontinuity of nominal phrases, other than the one proposed by Simpson (1991). However, it not clear what an analysis of this would look like. For the moment it may be concludde that LFG's treatment of adjuncts is problematic and that it cannot account for discontinuous adjuncts in Latin (or in other languages).

6 Conclusion and Future Work

In conclusion, it has been shown that constraints on discontinuity of Latin nominal phrases bring forth two main problems for LFG as a theory of syntax. The first constraint, which states that in discontinuous PPs the preposition needs to occur adjacent to at least part of its nominal object, shows that c-structure in LFG is not sufficiently constrained. The principle of Economy of Expression overgenerates sentences. This paper has shown that c-structure needs to be more strongly constrained in LFG than it is at the moment; much of LFG's attention has been focused on constraints on f-structure. In this paper, it is proposed to make an exception to Economy of Expression in order to account for the Latin data. The second issue discussed in this paper came to light with a closer investigation into another constraint on discontinuous NPs in Latin, namely the one on discontinuous adjuncts. As shown in this paper, the problem goes beyond the constraint because it turns out that in general, discontinuous adjuncts cannot be accounted for properly within LFG, due to LFG's set notation for adjuncts. No proper analysis has been found for this problem at the moment. This is an important issue which merits future research.

Future work will involve the issue of adjuncts in LFG. Also, an increased data set would allow for further conclusions to be drawn; for this paper only 142 examples of discontinuous NPs (from three different authors) were used.[20] It would be interesting to see how genitive adjuncts come into this, since they were not discussed in the data; adjuncts in Latin are most often bare ablative phrases. In order to acquire a better overview of discontinuity in general, other languages will need to be looked at. Discontinuous nominal phrases are not only found in Latin; they also occur in languages such as ancient Greek, Polish and Russian (Siewierska, 1988). Discontinuity and its limits say something about what is ultimately possible in a language, to the extent that it is still possible for listeners to understand the speaker (or readers understand the writer in the case of classical literary Latin). A cross-linguistic investigation would help give a better insight into the limits of discontinuity, and thereby into the limits of the cognitive capacities of language users in perception and production.

[20]Considering that according to Pinkster (2005) 12% of all NPs are discontinuous, this is still a reasonable amount of data.

7 References

Bolkestein, A. Machteld. Random Scrambling? Constraints on Discontinuity in Latin Noun Phrases. In Moussy, C., editor, *De lingua Latina novae quaestiones*, pages 245–258. Peeters, Louvain, 2001.

Bresnan, Joan. *Lexical-Functional Syntax*. Blackwell Publishers Inc., Malden, MA; Oxford, 2001.

Dalrymple, Mary. *Lexical-Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press, San Diego; London, 2001.

Haug, Dag. Backward Control in Ancient Greek - Subsumption or Linearization? In Butt, Miriam and King, Tracy Holloway, editors, *Proceedings of the LFG11 Conference*. CSLI Publications, 2011.

King, Tracy Holloway. *Configuring Topic and Focus in Russian*. Dissertations in Linguistics. CSLI, Stanford, CA, 1995.

Ledgeway, Adam. Syntactic and morphosyntactic change. In Maiden, M.; Smith, J.C., and Ledgeway, A., editors, *The Cambridge History of the Romance Languages, Volume 1: Structures*, pages 382–471. Cambridge University Press, Cambridge, 2011.

Lowe, John J. Ṛgvedic clitics and 'prosodic movement'. In Butt, Miriam and King, Tracy Holloway, editors, *Proceedings of the LFG11 Conference*. CSLI Publications, 2011.

Panhuis, Dirk. *The Communicative Perspective in the Sentence: A Study of Latin Word Order*. Benjamins, Amsterdam/Philadelphia, 1982.

Pinkster, Harm. *Latin Syntax and Semantics*. Routledge, London, 1990.

Pinkster, Harm. The language of Pliny the Elder. In Reinhardt, T.; Lapidge, M., and Adams, J.N., editors, *The Language of Latin Prose*, pages 239–256. Oxford University Press, Oxford, 2005.

Siewierska, Anna. *Word Order Rules*. Croom Helm, London, 1988.

Simpson, Jane. *Warlpiri Morpho-Syntax: a Lexicalist Approach*, volume 23 of *Studies in Natural Language and Linguistic Theory*. Kluwer Academic Publishers, Dordrecht, 1991.

Snijders, Liselotte. Constraints on Discontinuous Nominal Phrases in Latin from a Lexical-Functional Grammar Perspective. Unpublished Master's dissertation, University of Oxford, 2012.

Spevak, Olga. *Constituent order in Classical Latin prose*. Studies in Language Companion Series. John Benjamins Publishing Co., Amsterdam, 2010.

NOMINAL ARGUMENT STRUCTURE AND THE STAGE-/INDIVIDUAL-LEVEL CONTRAST IN HINDI-URDU

Sebastian Sulger
University of Konstanz

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper focuses on a very specific aspect of a construction in Hindi-Urdu involving the verb *ho* 'be'. The construction was previously described as displaying a pattern of differential case marking (DCM) in the literature, expressing specific semantic effects. These effects were previously compared to the contrast known as the stage-/individual-level contrast. The paper will show, however, that this view does not take into account various syntactic and semantic facts about this construction, and argues for a more differentiated view: what has been regarded as a single construction to be differentiated only by the case marking, should rather be treated as two separate constructions with differences in the nominal argument structure, case marking and semantics.

1 Introduction

This paper focuses on a very specific aspect of a construction in Hindi-Urdu involving the verb *ho* 'be'. The construction was previously described as displaying a pattern of differential case marking (DCM) in the literature, resulting in specific semantic effects. The paper will show, however, that this view does not take into account various syntactic and semantic facts about this construction, and argues for a more differentiated view: what has been regarded as a single construction to be differentiated only by the case marking, should rather be treated as two separate constructions with differences in the nominal argument structure, case marking and semantics.

2 The Data

In this section, I give a brief overview of the data. The construction which is the focus of this paper is as follows. An oblique subject marked by either the dative case marker *ko* ((1a), (2a)) or the locative case marker *mẽ* 'in' ((1b), (2b)) is followed by a noun and the verb *ho* 'be'. As noted by e.g. Mohanan (1994), the interpretation depends on the choice of the case marker on the subject. While the sentences refer to a temporary state when the subject is marked with the dative case, they denote more permanent properties with locative subjects. Mohanan (1994) concludes that these cases exhibit a pattern of differential case marking (DCM), arguing that the case markers encode different configurations in a semantic field, resulting in the different semantics. She compares the semantic effects to those encoded by the stage-/individual-level contrast (Carlson, 1977; Kratzer, 1995); see the quote in (3).[1]

[†]Thanks to the audience of the LFG12 conference for comments and suggestions, as well as to my informants: Qaiser Abbas, Tafseer Ahmed, Rajesh Bhatt, Miriam Butt, Ghulam Raza. Special thanks to Rajesh Bhatt and Miriam Butt for many comments on earlier versions of this paper.

[1]The abbreviations in the glosses are as follows: 1/2/3 – 1st/2nd/3rd person, Acc – accusative, Comp – complementizer, Dat – dative, Erg – ergative, exp – experiencer argument, F – feminine, Inst – instrumental, Loc – locative, loc – locative argument, M – masculine, Nom – nominative, Obl – oblique, Perf – perfect, Pl – plural, Pres – present, Sg – singular, src – source argument, th – theme argument.

(1) a. nina=ko bʰay hɛ
Nina.F.Sg=Dat fear.M.Sg=Nom be.Pres.3.Sg
'Nina is afraid.' Mohanan (1994, p. 172)

b. nina=mẽ bʰay hɛ
Nina.F.Sg=Loc$_{in}$ fear.M.Sg=Nom be.Pres.3.Sg
'Nina is fearful.' (lit. 'There is fear in Nina.') Mohanan (1994, p. 172)

(2) a. nina=ko pyar hɛ
Nina.F.Sg=Dat love.F.Sg=Nom be.Pres.3.Sg
'Nina is in love.'

b. nina=mẽ pyar hɛ
Nina.F.Sg=Loc$_{in}$ love.F.Sg=Nom be.Pres.3.Sg
'Nina is in love.' (lit. 'There is love in Nina.')

(3) While *-ko* encodes the abstract location of a temporary state, such as happiness or worry, or a temporary fear [...], *-mã* [*-mẽ*] expresses the location of a characteristic attribute that is relatively permanent, such as a fearful disposition [...]. When the state is inherently temporary, as in the event of a cough or a fever, the use of *-mã* [*-mẽ*] is disallowed, perhaps because abstract containment cannot be extended to temporary states. Mohanan (1994, p. 172)

However, serious problems for this analysis are presented by examples as in (4). Here, an additional argument marked by the instrumental case marker *se* is introduced. Notice that the sentences are grammatical only with the *ko*-marked subjects, but ungrammatical with the *mẽ*-marked subjects. The original assumptions by Mohanan (1994) do not predict this; if the only difference were in the choice of the case marker, we would simply predict a different interpretation, but not ungrammaticality. For example, we would expect that (4b) expresses a more permanent fear relation towards *yasin* than (4a), but not the ungrammaticality of (4b). The only way to introduce the object of the fear/love relation in sentences with a locative subject is by inserting it with *ke liye* 'for', which is an adjunct marker in Hindi-Urdu.

(4) a. nina=ko yasin=se bɑhʊt bʰay hɛ
Nina.F.Sg=Dat Yassin.M.Sg=Inst much fear.M.Sg=Nom be.Pres.3.Sg
'Nina is afraid of Yassin.'

b. *nina=mẽ yasin=se bɑhʊt bʰay
Nina.F.Sg=Loc$_{in}$ Yassin.M.Sg=Inst much fear.M.Sg=Nom
hɛ
be.Pres.3.Sg

c. nina=ko yasin=se bɑhʊt pyar hɛ
Nina.F.Sg=Dat Yassin.M.Sg=Inst much love.F.Sg=Nom be.Pres.3.Sg
'Nina is in love with Yassin.'

d. *nina=mẽ yasin=se bɑhʊt pyar
Nina.F.Sg=Loc$_{in}$ Yassin.M.Sg=Inst much love.F.Sg=Nom
hɛ
be.Pres.3.Sg

e. nina=mẽ yasin=ke liye bɑhʊt pyar hɛ
Nina.F.Sg=Loc$_{in}$ Yassin.M.Sg=for much love.F.Sg=Nom be.Pres.3.Sg
'Nina carries much love (in her) for Yassin.' $\sim$ 'Nina is in love with Yassin.'

Similar problems emerge when we vary the other noun involved in this construction: the noun describing the relation/the feeling. Consider (5), for example. In contrast to the examples above, even without specifying an additional argument, a locative subject is ruled out.

(5) a. nina=ko tɑlaš hɛ
Nina.F.Sg=Dat search.F.Sg=Nom be.Pres.3.Sg
'Nina is searching.'

b. *nina=mẽ tɑlaš hɛ
Nina.F.Sg=Loc$_{in}$ search.F.Sg=Nom be.Pres.3.Sg

Again, assuming solely a pattern of DCM resulting in a semantic contrast does not give the full picture: we would simply predict a different interpretation (something along the lines of (6b) expressing a more permanent state of "being in search" than (6a)), but not ungrammaticality of (6b). To explain these effects, we are in need of a new analysis.

2.1 Some Generalizations and Open Questions

At this point, I lay out some basic generalizations about the data. We can observe the following:

- Introducing source arguments is only felicitous with dative subjects.
- Certain abstract nouns are only felicitous with one kind of subject (e.g., *tɑlaš* 'search' is only acceptable with a dative subject).
- The examples we have looked at so far all seem to involve a specific category of relational nouns (e.g., *love, fear, hate, search, regret,* etc.) and animate subjects.
- The stage-/individual-level contrast does not suffice to explain the observed effects.

I also formulate some questions to be answered in the remainder of this paper:

- If the stage-/individual-level contrast is indeed not primarily responsible for choosing among the case markers — what is?

- If there are indeed different *constructions* at the backend of the *ko* vs. *mẽ* examples — how do they differ? And what can all of this tell us about the syntax and semantics of Hindi-Urdu?

- What should a formal treatment of this look like?

The verb *ho* 'be' in Hindi-Urdu may either function as a copula verb connecting a subject to its predicate, or as a light verb taking part in complex predicate (CP) formation.[2] Thus, a good starting point for taking a closer look at the data (and, eventually, analyzing the structure within) seems to be the theory of CP formation and how it contrasts with copula constructions (Section 3).

The remainder of this paper is structured as follows. In Section 3, I discuss complex predicates in Hindi-Urdu, setting them apart from copula constructions; the differences in these two types of constructions are essential for the purpose of this paper. Section 4 provides an overview of locative copula constructions in Hindi-Urdu, showing that these have several features in common with the data we have seen above. In Section 5, I take a detailed look at different classes of Hindi-Urdu nouns and argue that the differences in these classes with respect to argument selection ultimately account for the data above. A novel analysis using Mapping Theory is then presented in Section 6. Furthermore, I discuss the semantics of the different constructions in Section 7. I conclude in Section 8.

3 Complex Predicates and Copula Constructions

Hindi-Urdu has about 700 simple verbs (Humayoun, 2006). As is the case in other South Asian languages, Hindi-Urdu uses a variety of different types of complex predicates (CPs) to express its full range of verbal predication. These CPs may be formed using different combinations of parts of speech: noun-verb, verb-verb, adjective-verb, preposition-verb. The verbs involved in the CPs have often been referred to as "light verbs" since they neither retain their full semantic predicational content, nor are they semantically empty; they seem to work like a licenser for the other, semantically more fundamental part of the CP, nevertheless retaining some semantics of their own (Butt, 2010). CPs in Hindi-Urdu have been thoroughly examined and analyzed in a bulk of work, for example Hook (1974); Singh (1990); Mohanan (1994); Butt (1995, 2010); Ahmed and Butt (2011) and references in all of these.

A major step in analyzing the data reviewed above is to determine their syntactic status: whether they form CPs or not. A starting point is the definition of a CP given in Butt (1995, p. 2), repeated below:

- The argument structure is complex (two or more semantic heads contribute arguments).

- The grammatical functional structure is that of a simple predicate. It is flat: there is only a single predicate (a nuclear pred) and a single subject.

[2] *ho* 'be' may also function as an auxiliary in Hindi-Urdu, but this use is not of immediate importance for this paper.

- The phrase structure may be either simple or complex. It does not necessarily determine the status of the complex predicate.

An example is given in (6) and in the functional structure in Figure 1. Here, we have a noun-verb CP; the argument structure is complex in that the light verb *lɑg* ‘attach’ selects two arguments (the “attachee” and the thing the attachee is attached to), and the noun *ḍar* ‘fear’ selects one argument (the thing being feared). This results in the complex argument structure of (1), where the main PRED is composed of the light verb *lɑg* and the noun *ḍar*. The grammatical functional structure of the sentence, though, is flat in that there is only a single main predicate and a single subject and there are no embeddings.[3]

(6) nadya=ko haṭ^{h}i=se ḍɑr lɑg-a
Nadya.F.Sg=Dat elephant.M.Sg=Instr fear.M.Sg=Nom attach-Perf.M.Sg
‘Nadya was frightened by the elephant.’

$$\begin{bmatrix} \text{PRED} & \text{‘lag<(}\uparrow\text{ SUBJ), ‘ḍɑr<(}\uparrow\text{ OBL)>’>’} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED ‘nadya’}\\ \text{CASE dat}\end{bmatrix} \\ \text{OBJ} & \begin{bmatrix}\text{PRED ‘ḍɑr’}\\ \text{CASE nom}\end{bmatrix} \\ \text{OBL} & \begin{bmatrix}\text{PRED ‘haṭ}^{\text{h}}\text{i’}\\ \text{CASE inst}\end{bmatrix} \end{bmatrix}$$

Figure 1: F-Structure for (6)

Recall that the verb *ho* ‘be’ in Hindi-Urdu may be used in different ways: as a light verb in CP constructions or as a copula verb. An extensive discussion of the different usages of *ho* ‘be’ in Urdu is given in Raza (2011). The issue which arises in connection to this paper is: how can we distinguish between the light verb and copula usages?

3.1 Tests for Complex Predicatehood

Butt (1995) provides several diagnostics for CPs based on agreement, control and anaphora. The tests are designed so as to distinguish monoclausal, non-embedding CP structures from polyclausal, embedding, non-CP structures. However, the tests identified by Butt do not give an answer to the open question whether the sentences constitute copula constructions (XCOMP/PREDLINK in LFG terms, depending on how you think about copula predication (Butt et al., 1999; Dalrymple et al., 2004;

[3] In fact, a recent dependency banking effort for complex predicates suggests that CP predicates be rewritten (i.e., as *ḍar_lag* in the case of (6)) when banking the structures for further processing to reflect their syntactic and semantic behavior as unities (Ahmed et al., 2012).

Attia, 2008)) or CPs. This is because the constructions in (2)–(5) are unmistakably monoclausal in nature, e.g., there is only a single verbal element and they have only a single subject. A more promising syntactic test for distinguishing copula constructions and CPs concerns coordination.

The Coordination Test When *ho* ‘be’ acts as a light verb in a noun-verb CP, the noun is in itself a predicator that introduces an argument. According to Raza (2011), the nominal predicators in noun-verb CPs may not be coordinated; this is especially clear in cases where the nominal introduces a clausal argument. (7) shows such examples. While (7a) is grammatical, the coordination of the nominal predicator as in (7b) is not felicitous.

(7) a. ɑli=ko xɑbɑr hɛ [kɪh ...
Ali.M.Sg=Dat news.M.Sg=Nom be.Pres.3.Sg Comp ...
‘Ali knows that ...’

b. *ɑli=ko xɑbɑr ya xʊsa hɛ
Ali.M.Sg=Dat news.M.Sg=Nom or anger.M.Sg=Nom be.Pres.3.Sg
[kɪh ...
Comp ...

Coordination, however, is possible in copula constructions. Crucially, coordination is also possible in (8c) and (8d), where the subject bears locative case. Coordination thus serves as a test for distinguishing noun-verb CPs from copula constructions.

(8) a. nina g^{h}ɑr=mẽ hɛ
Nina.F.Sg=Nom house.M.Sg=Loc$_{in}$ be.Pres.3.Sg
‘Nina is in the house.’

b. nina g^{h}ɑr=mẽ ya bag=mẽ hɛ
Nina.F.Sg=Nom house.M.Sg=Loc$_{in}$ or garden.M.Sg=Loc$_{in}$ be.Pres.3.Sg
‘Nina is in the house or in the garden.’

c. g^{h}ɑr=mẽ (ek) cuha ya (ek) kʊṭṭa
house.M.Sg=Loc$_{in}$ (one) rat.M.Sg=Nom or (one) dog.M.Sg=Nom
hɛ
be.Pres.3.Sg
‘A rat or a dog is in the house.’ (lit. ‘There is a rat or a dog in the house.’)

d. nina=mẽ pyar ya b^{h}ay hɛ
Nina=Loc$_{in}$ love.F.Sg=Nom ya fear.M.Sg=Nom be.Pres.3.Sg
‘Nina is in love or fear.’ (lit.: ‘There is love or fear in Nina.’)

The coordination facts suggest a structural difference between the constructions examined. A CP analysis seems right for the part of the data that exhibits complex argument structures — so distinguishing between CP and copula constructions essentially boils down to the question: do all of the data exhibit complex argument structures?

4 Interlude: Locatives in Hindi-Urdu

Let us briefly review what is known about locative constructions in Hindi-Urdu. Locative predication in Hindi-Urdu is achieved via the frame in (9). I assume the copula *ho* 'be' may select a theme and a location; this is a cross-linguistically valid assumption (Bresnan and Kanerva, 1989; Curnow, 1999; Pustet, 2003). A linking analysis using Lexical Mapping Theory (LMT) (Bresnan and Kanerva, 1989; Bresnan and Zaenen, 1990; Bresnan, 2001) is provided in Figure 2.

(9) ho $<$ th loc $>$

(10) nina gʰɑr=mẽ hɛ
Nina.F.Sg=Nom house.M.Sg=Loc$_{in}$ be.Pres.3.Sg
'Nina is in the house.'

ho	<	th	loc	>
		|	|	
intrinsic		[-r]	[-o]	
defaults			[+r]	
		OBJ/SUBJ	OBL$_{loc}$	
well-formedness		SUBJ	OBL$_{loc}$	
		|	|	
case		nom	loc	

Figure 2: Linking analysis for predicative locatives

In Hindi-Urdu the locative case-marked phrase (i.e., the location) may also be realized as the sentence's subject. Compare (11a) to the inverted example in (11b); in (11a), the nominative theme is realized as the subject, while in (11b) the location is realized as the subject. Mohanan (1994) presents evidence that the locative in (11b) is in fact the subject of the sentence and I adopt this view.

(11) a. kʊt̤t̤a gʰɑr=mẽ hɛ
dog.M.Sg=Nom house.M.Sg=Loc$_{in}$ be.Pres.3.Sg
'The dog is in the house.'

b. gʰɑr=mẽ (ek) kʊt̤t̤a hɛ
house.M.Sg=Loc$_{in}$ (one) dog.M.Sg=Nom be.Pres.3.Sg
'A dog is in the house.' (lit. 'There is a dog in the house.')

Looking at sentences such as (11a) and (11b) more closely, we notice a difference concerning discourse structure. If the theme is realized as the subject, it must be a definite referent (i.e., a referent already given in the discourse) as in (11a). On the other hand, if the location is realized as the subject, the theme must not be a definite referent, but must be an indefinite one, as in (11b). (11b) can not mean *The dog is in the house*. Consider the dialogues in (12) and (13). The answers marked by '???' are not felicitous in the course of the dialogue.

(12) a. kʊṭṭa ɑhan hɛ
dog.M.Sg=Nom where be.Pres.3.Sg
'Where is the dog?'

b. kʊṭṭa g^{h}ɑr=mẽ hɛ
dog.M.Sg=Nom house.M.Sg=Loc$_{in}$ be.Pres.3.Sg
'The dog is in the house.'

c. ??? g^{h}ɑr=mẽ kʊṭṭa hɛ
house.M.Sg=Loc$_{in}$ dog.M.Sg=Nom be.Pres.3.Sg
'A dog is in the house.' (lit. 'There is a dog in the house.')

(13) a. g^{h}ɑr=mẽ kɪya hɛ
house.M.Sg=Loc$_{in}$ what be.Pres.3.Sg
'What is in the house?'

b. g^{h}ɑr=mẽ (ek) kʊṭṭa hɛ
house.M.Sg=Loc$_{in}$ (one) dog.M.Sg=Nom be.Pres.3.Sg
'A dog is in the house.' (lit. 'There is a dog in the house.')

c. ??? kʊṭṭa g^{h}ɑr=mẽ hɛ
dog.M.Sg=Nom house.M.Sg=Loc$_{in}$ be.Pres.3.Sg
'The dog is in the house.'

These examples point strongly to a difference in discourse structure between (12c) and (13c). Discourse structure in Hindi-Urdu is reflected by word order; while topics occur in clause-initial position, the focus position in Hindi-Urdu is generally immediately preverbal (Butt and King, 1997). This generalization is borne out by the data in (12) and (13). The sentences marked by '???' are not felicitous as they focus the wrong part of the clause in response to the question. Furthermore, the definiteness effects are predicted under this analysis — topics are referents given in discourse (and may therefore be definite) while focused constituents are new information (and may therefore not be definite).

Locative Inversion I make the following proposal. Hindi-Urdu has locative inversion, cf. Bresnan and Kanerva (1989): in cases of locative inversion, the theme role is optionally classified as objective (the reason for which is discussed below), thus rendering the locative as a subject and the theme as an object. Bresnan and Kanerva (1989) motivate this optional assignment in terms of discourse functions: inverted locatives have a presentational function whereby the theme is focussed, thus the locative role must be realized as the subject/topic. By well-formedness conditions, the theme is classified as the object/focus. Kibort (2007) argues instead that the theme must receive [+o] and is realized as an object ('demotion of subject to an object'), leaving the locative to become the subject by well-formedness conditions. Although the two solutions result in the same GF assignment, it seems more intuitive to further specify the theme argument, as this is the one being focussed (compared to the solution put forward by Bresnan and Kanerva (1989)). I adapt Kibort's approach to locative inversion, shown in Figures 3 and 4.

	<	th	loc	>
loc. inv.		[+o]		

Figure 3: Optional classification for locative inversion (Kibort, 2007)

ho	<	th	loc	>
intrinsic		[-r]	[-o]	
loc. inv.		**[+o]**		
		OBJ	OBL/SUBJ	
well-formedness		OBJ	SUBJ	
case		nom	loc	

Figure 4: Linking analysis for inverted locatives

Now, note the following similarities between examples like (1b) and (2b), repeated here in (15) and (16). My argument is that they in fact represent the same construction. The claim that the argument structure of (15) is in fact simple (and not complex as in CPs) will receive further reasoning in the next section.

(14)
1. locative (not dative) case marking;
2. existential interpretation/indefinite theme;
3. simple argument structure;
4. verb *ho* 'be'.

(15) a. nina=mẽ b^hay hɛ
Nina.F.Sg=Loc_{in} fear.M.Sg=Nom be.Pres.3.Sg
'Nina is fearful.' (lit. 'There is fear in Nina.') Mohanan (1994, p. 172)

b. nina=mẽ pyar hɛ
Nina.F.Sg=Loc_{in} love.F.Sg=Nom be.Pres.3.Sg
'Nina is in love.' (lit. 'There is love in Nina.')

(16) g^hɑr=mẽ (ek) kʊṭṭa hɛ
house.M.Sg=Loc_{in} (one) dog.M.Sg=Nom be.Pres.3.Sg
'A dog is in the house.' (lit. 'There is a dog in the house.')

5 Nominal Argument Structure

It has long been known that nouns across languages may take arguments (Chomsky, 1970; Higginbotham, 1983; Grimshaw, 1990, among others). The crucial point here for our purposes is that many nouns are in fact ambiguous between interpretations in which they realize arguments and other interpretations in which they do not (Grimshaw, 1990). Other nouns are not ambiguous in this respect; some nouns never allow arguments, while some nouns always require arguments. In this section, I argue that there is evidence from semantics that Hindi-Urdu has all of these, and that ultimately, it is this diversity in nominal argument structure that makes for the differences discussed above.

5.1 Ambiguous Nouns

5.1.1 Argument-Taking Uses

In certain contexts, relational nouns in Hindi-Urdu such as *nɑfrat* 'hate'/*pyar* 'love' allow oblique arguments marked by the instrumental case marker *se*.

(17) mʊjhe (bɪllɪyõ=se) nɑfrɑt hɛ
I.Obl.Dat (cat.F.Pl.Obl=Inst) hate.F.Sg=Nom be.Pres.3.Sg
∼ 'I hate (cats).'

(18) nina=ko (yasin=se) pyar hɛ
Nina.F.Sg=Dat (Yassin.M.Sg=Inst) love.F.Sg=Nom be.Pres.3.Sg
'Nina is in love (with Yassin).'

Native speakers inform me that in (17a)/(18a), it is always understood that Nina's love/hate is *directed at someone/something specific*. Hindi-Urdu makes use of pro-drop (all arguments may in principle be dropped) (Butt, 1995; Butt and King, 2007), which explains why the *se*-marked nominal may be absent. Notice that we have dative case marking on the subject in all these cases; since the copula does not license dative case on its arguments, we must assume the dative (experiencer) case is licensed by the relational noun (*nɑfrat/pyar*).

5.1.2 Non-Argument-Taking Uses

In other contexts, the same abstract relational nouns never allow any oblique arguments. These are exactly the cases where we have locative case marking on the subject.

(19) a. mʊjh=mẽ nɑfrɑt hɛ
I.Obl=Loc$_{in}$ hate.F.Sg=Nom be.Pres.3.Sg
∼ 'I hate.'

b. *mʊjh=mẽ bɪllɪyõ=se nɑfrɑt hɛ
I.Obl=Loc$_{in}$ cat.F.Pl.Obl=Inst hate.F.Sg=Nom be.Pres.3.Sg

(20) a. nina=mẽ pyar hɛ
Nina.F.Sg=Loc_{in} love.F.Sg=Nom be.Pres.3.Sg
'Nina is in love.'

b. *nina=mẽ yasin=se pyar hɛ
Nina.F.Sg=Loc_{in} Yassin.M.Sg=Inst love.F.Sg=Nom be.Pres.3.Sg

According to native speaker judgement, in (19a)/(20a), the emphasis in the utterance is on the *feeling by itself*; crucially, it is not immediately understood that the *hate/love* relations have objects in these sentences. In these sentences, *pyar* 'love' expresses a more detached and somewhat more concrete reading than in the sentences with a dative subject. The difference can be compared to the one between R-nouns and Ev-nouns put forward by Grimshaw (1990). According to Grimshaw's account, two types of nominalizations may be distinguished: complex event nouns (Ev-nouns) that retain the properties of their verbal base, and result nouns (R-nouns) in which those properties are no longer transparent. I conclude that these nouns have an concrete reading where they do not realize the object argument (i.e., the argument expressing the direction of the feeling).

5.2 Unambiguous Nouns

5.2.1 Obligatorily Argument-Taking Nouns

Nouns such as *tɑlaš* 'search' seem to obligatorily select arguments. They are not allowed to appear with locative subjects as in (21b), but only with dative subjects as in (21a). They seem to be inherently relational, selecting for an experiencer and an (optionally expressed) source. Even when the source argument marked by *se* is not realized, it is always understood that the event expressed by the noun is directed at someone/something.

(21) a. nina=ko (yasin=se) tɑlaš hɛ
Nina.F.Sg=Dat Yassin.M.Sg=Inst search.F.Sg=Nom be.Pres.3.Sg
'Nina is searching (for Yassin).'

b. *nina=mẽ tɑlaš hɛ
Nina.F.Sg=Loc_{in} search.F.Sg=Nom be.Pres.3.Sg

In Hindi-Urdu, a search is not a search without being experienced by someone and being directed at something. This explains the ungrammaticality of (21b): the experiencer argument licensed by the nominal *tɑlaš* 'search' cannot be assigned locative case, since there is no locative role. This points to a distinction between locatives and experiencers and presents evidence against a view that unites locations and experiencers (e.g., Landau, 2010). I return to this issue in Section 7.

5.2.2 Obligatorily Non-Argument-Taking Nouns

Other nouns such as *acc^{h}ai* 'goodness' may only appear with a single locative argument realized as the subject; these nouns may never appear with dative subjects as in (22b), nor with additional source arguments as in (22c).

(22) a. nina=mẽ ɑcc^{h}i hɛ
Nina.F.Sg=Loc$_{in}$ goodness.F.Sg=Nom be.Pres.3.Sg
'Nina is good/a good person.' (lit. 'There is goodness in Nina.')

b. *nina=ko ɑcc^{h}i hɛ
Nina.F.Sg=Dat goodness.F.Sg=Nom be.Pres.3.Sg

c. *nina=mẽ yasin=se ɑcc^{h}i hɛ
Nina.F.Sg=Loc$_{in}$ Yassin.M.Sg=Inst goodness.F.Sg=Nom be.Pres.3.Sg

I assume that inherently non-relational nouns such as *acc^{h}ai* 'goodness' do not select for arguments, since they are not directed at anyone; they are, of course, abstract in nature, which is a lexical property setting them apart from concrete nouns such as *kɪtab* 'book', but syntactically (i.e. regarding argument selection) these two kinds of nouns work alike.

5.3 Intermediate Summary

We have identified four different classes of nouns with respect to argument selection for Hindi-Urdu. The classes are depicted in Table 4. We also have identified two different patterns of experiencer constructions. In the first pattern, which I will call the "dative experiencer construction" (DEC), the subject is dative, the noun is relational, licensing both an experiencer and a source argument, and the argument structure is complex; *ho* 'be' in these cases is a light verb, forming a CP with the predicative noun. The second pattern, which I will call the "locative experiencer construction" (LEC), is entirely different in that the subject is locative in an inverted locative construction and the argument structure is simple; *ho* 'be' here is a copula verb, selecting for a theme and a location.

nouns	ambiguous	realize arguments (relational *pyar* 'love', *b^{h}ay* 'fear')
		do not realize arguments (non-relational *pyar*, *b^{h}ay*)
	not ambiguous	always realize arguments (*talaš* 'search')
		never realize arguments (*acc^{h}ai* 'goodness', *kɪtab* 'book')

Table 1: Overview of Hindi-Urdu noun classes wrt. argument selection

6 A Novel Analysis Using Mapping Theory

In this section, I present an analysis of the two patterns identified above (DECs and LECs) in terms of Mapping Theory as described in e.g. Butt (1995); Alsina (1996); Butt et al. (1997); Butt (1998). That is, I adopt amendments to original Lexical Mapping Theory, which was reformulated as Mapping Theory by e.g. Butt (1995) and Alsina (1996) to account for the formation of complex predicates. In particular, I assume argument fusion as triggered by a pertinent characteristic of light verbs, namely the transparent event argument ev_T. It is this argument that models the "semantically bleached" nature of light verbs. Essentially, CP formation must take place if ev_T is present (Butt, 1995). Case in this framework is seen as a separate system interacting with linking principles and clausal semantics, but not wholly determining them (Butt, 1998).

My analysis makes use of two different frames for the copula *ho* 'be': a locative copula frame and a light verb frame used for CP formation. My assumptions about case are as follows. An ev_T argument never receives case marking: it always bears nominative case (Butt, 1995). Relational nouns, on the other hand, may license case depending on their argument structure: experiencers receive dative case (Butt et al., 2006), sources take instrumental case, locations receive locative case.

6.1 Predicative Locatives

First, let's review predicative locatives. An example is given in (23). As discussed above, the theme must be definite for this particular linking to be realized. The linking analysis is given in Figure 5.

(23) admi kɑmre=mẽ hɛ
man.M.Sg=Nom room.M.Sg.Obl=Loc$_{in}$ be.Pres.3.Sg
'The man is in the room.'

ho	<	th	loc	>
intrinsic		[-r]	[-o]	
defaults			[+r]	
		OBJ/SUBJ	OBL$_{loc}$	
well-formedness		SUBJ	OBL$_{loc}$	
case		nom	loc	

Figure 5: Linking analysis for predicative locatives

6.2 Inverted Locatives, Locative Experiencer Constructions

The linking for inverted locatives such as (24) is given in Figure 6. This is also the frame used for the LECs as in (25). As we have seen above, the theme argument in both (24) and (25) must be indefinite and receives the [+o] feature in the linking process, which causes it to be realized as an object. If the theme is a relational noun such as *b^{h}ay*, it may not realize its arguments in the clause, as discussed in Section 5.1.2. The linking for the LEC in (25) is given in Figure 7.

(24)

kɑmre=mẽ	(ek)	admi	hɛ
room.M.Sg.Obl=Loc_{in}	(one)	man.M.Sg=Nom	be.Pres.3.Sg

‘There is a man in the room.’

(25)

nina=mẽ	b^{h}ay	hɛ
Nina.F.Sg=Loc_{in}	fear.M.Sg=Nom	be.Pres.3.Sg

‘Nina is fearful.’ (lit.: ‘There is fear in Nina.’) Mohanan (1994:172)

ho	<	th	loc	>
		|	|	
intrinsic		[-r]	[-o]	
loc. inv.		**[+o]**		
		OBJ	OBL/SUBJ	
well-formedness		OBJ	SUBJ	
		|	|	
case		nom	loc	

Figure 6: Linking analysis for inverted locatives

ho	<	th	loc	>				
pyar	<	exp	src	>				
ho	<	pyar	<	exp	src	>	loc	>
		|					|	
intrinsic		[-r]					[-o]	
loc.inv.		**[+o]**						
		OBJ					OBL/SUBJ	
well-formedness		OBJ					SUBJ	
		|					|	
case		nom					loc	

Figure 7: Linking analysis for LEC

6.3 Dative Experiencer Constructions

Relational nouns such as *pyar* ‘love’ supply two arguments: an experiencer and a source. The resulting argument structure is complex, and complex predicate formation takes place. The highest argument of the embedded predicate is fused with the lowest argument of the matrix predicate (Butt, 1995, 1998). The resulting frame and the linking is depicted in Figure 8.

(26) nina=ko yasin=se bɑhʊt pyar hɛ
Nina.F.Sg=Dat Yassin.M.Sg=Inst much love.F.Sg=Nom be.Pres.3.Sg
'Nina carries much love (in her) for Yassin.' ∼ 'Nina is in love with Yassin.'

ho	$<$	th	ev_t	$>$				
pyar	$<$	exp	src	$>$				
ho	$<$	th	pyar	$<$	exp	src	$>$	$>$
intrinsic			[-r]					
defaults		[-r]				[+r]		
$\hat{\theta}$		[-o]						
		SUBJ	OBJ/SUBJ			OBJ_θ/OBL_θ		
well-formedness		SUBJ	OBJ			OBL_θ		
case		dat	nom			inst		

Figure 8: Linking analysis for experiencer complex predicate (I)

The matrix frame *ho* $< th \quad ev_t >$ is also selected for the "illness" examples such as (27) (linking analysis in Figure 9). The difference between these cases and the data involving relational nouns is obvious: in the "illness" examples, the predicative nominal selects a single experiencer argument, while in the examples involving relational nominals, the nominal selects two arguments: an experiencer and a source.

(27) nina=ko bɑhʊt k^{h}ãsi hɛ
Nina.F.Sg=Dat much cough.F.Sg=Nom be.Pres.3.Sg
'Nina has a severe cough.'

6.4 Copula and Light Verb *ho* 'be'

The present analysis thus gives us an idea of how the two different readings of Hindi-Urdu *ho* 'be' work. In one reading, *ho* is a copula verb selecting for a theme and a locative. The theme may be simple (Figure 6) or complex (Figure 7) regarding its argument structure, but even when it's complex, its own arguments may not be realized in the clause, since the theme is not a transparent event (ev_t), and argument fusion cannot take place. This explains why a source argument (marked with *se*) may not be licensed in the clause.

In the other reading, *ho* is a light verb selecting for a theme and a transparent event (ev_t). In this frame, argument fusion must take place. The transparent

ho	$<$	th	ev_t	$>$			
k^hãsi	$<$	exp	$>$				
ho	$<$	th	k^hãsi	$<$	exp	$>$	$>$
intrinsic			[-r]				
defaults		[-r]					
$\hat{\theta}$		[-o]					
		SUBJ	OBJ/SUBJ				
well-formedness		SUBJ	OBJ				
case		dat	nom				

Figure 9: Linking analysis for experiencer complex predicate (II)

event may contribute one (Figure 9) or two arguments (Figure 8). The valency of the transparent event is determined by its lexical entry; relational nouns such as *pyar* 'love' supply two arguments, "illness" nouns such as *k^hãsi* supply a single argument. In both cases, the dative case on the subject is required by the experiencer argument of the noun. As argument fusion takes place, the additional source argument from the relational noun can be realized.

7 The Semantics of Sentient Locations

As argued above, under the present analysis of the constructions' syntax, the semantics observed in the data are expected. Comparing my own analysis to the approaches of Mohanan (1994) and Landau (2010), the present analysis involves a strict distinction between locations (abstract or concrete) on the one hand versus experiencers on the other hand. While locations (whether sentient or not) get locative case, experiencers receive dative case. Emotional experiencers always also have a source at argument structure, which may be pro-dropped, while "illness" experiencers don't involve a source. The different syntactic analyses in general, and the ambiguities observed in the argument-structure of Hindi-Urdu nominals in particular, give rise to differing semantic interpretations. The relational nouns may only realize their source argument in the CP construction, but in the copula construction, they never do so, hence the reading is a rather concrete one (as described in Section 5.1.2). While an individual-level vs. stage-level (or permanent vs. temporary) distinction as described by Mohanan (1994) suggests itself, it is the source argument which is not realized which renders the semantics of the overall clause more concrete and gives the sentence a less time-dependent flavor.

To formally describe the semantics of the LEC is not straightforward. While Landau (2010) suggests that all experiencers are nothing but syntactic locations, this turns out to be only partly true for Hindi-Urdu. Experiencers are encoded using two separate syntactic constructions, namely a locative frame with locative case marking and a complex predicate frame with dative case marking, and they are used to convey quite different meanings. The next section expands on this.

7.1 A Scenario

To illustrate the differences in meaning, imagine you were part of an experiment on animals, and while you would not consider yourself averse to cats, the experimental setup would prove otherwise. The proposition expressed by (28a) would therefore be false, since by your internal judgment you would not subscribe to the fact that you hate cats; the statement in (28b) would still be true as shown by the experiment. That is, the sentences with dative subjects seem to describe more direct relations than the sentences with locative subjects.

(28) a. #mʊjhe bɪllɪyõ=se nɑfrɑt hɛ
I.Obl.Dat cat.F.Pl.Obl=Inst hate.F.Sg=Nom be.Pres.3.Sg
'I hate cats.'

b. mujh=mẽ bɪllɪyõ=ke liye nɑfrɑt hɛ
I.Obl=Loc$_{in}$ cat.F.Pl.Obl=for hate.F.Sg=Nom be.Pres.3.Sg
'I hate cats.' (lit. 'There is hate in me against cats.')

The examples show that the locatives with sentient subjects do not necessarily express more permanent states than the dative experiencer cases (as put forward by Mohanan (1994)), but rather more independent, objective and concrete descriptions of states. I maintain that the contrasts observed in the data cannot be analyzed simply by calling upon the stage-/individual-level contrast.

7.2 A Formal Semantic Treatment?

A formal semantic treatment, e.g., in terms of Glue logic (Dalrymple et al., 1993; Dalrymple, 1999; Asudeh, 2012), is in need of more research and outside of the scope of this paper. The LFG/Glue architecture in principle allows for a deliberate number of inferences for the different constructions. Asudeh and Giorgolo (this volume), for example, present an LFG/Glue analysis for optional arguments (29) and derived arguments (e.g., passive *by*-phrases (30), instrumental *with*-phrases (31)).[4] Using flexible semantic composition in combination with generalizations over descriptions, Asudeh and Giorgolo enforce implications of optional and derived arguments at the syntax-semantics interface. For example, for the semantically transitive verb *drink*, they enforce the implication that the (missing) object is an alcoholic drink.

(29) a. Any child of Kim's is unfortunately likely to drink ___ .

b. Kim ate ___ at 10 o'clock.

[4]The examples are due to Asudeh and Giorgolo (this volume).

(30) The hole was plugged by Kim.

(31) Kim plugged the hole with a cork.

The approach of Asudeh and Giorgolo (this volume) is relevant to the work presented in this paper inasmuch as this paper also deals with arguments that may or may not be realized in a sentence, and when they are not realized, the sentence carries certain connotations, as shown above. The exact nature of the implications, however, must be left for future work.

8 Conclusion

This paper has presented a treatment of Hindi-Urdu relational nouns at the syntax-semantics interface. Some of these nouns feature a complex argument structure, resulting in different syntactic and semantic behavior, depending on whether they occur with copular *ho* 'be' or light verb *ho* 'be'; the difference is essentially one of copula constructions versus complex predicate formation (i.e., no argument fusion vs. argument fusion). The paper discussed the relevant data and presented an analysis using Mapping Theory. It showed that assuming a simple temporary/permanent distinction, triggered by the case markers involved, does not do full justice to the data, as the syntax and semantics are more detailed than it was assumed before. What must be left for future work is the exact makeup of the semantic restriction on the LECs (see Section 7).

References

Ahmed, Tafseer and Butt, Miriam. 2011. Discovering Semantic Classes for Urdu N-V Complex Predicates. In *Proceedings of the International Conference on Computational Semantics (IWCS 2011)*.

Ahmed, Tafseer, Butt, Miriam, Hautli, Annette and Sulger, Sebastian. 2012. A Reference Dependency Bank for Analyzing Complex Predicates. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA).

Alsina, Alex. 1996. *The Role of Argument Structure in Grammar: Evidence from Romance*. CSLI Publications.

Asudeh, Ash. 2012. *The Logic of Pronominal Resumption*. Oxford University Press.

Asudeh, Ash and Giorgolo, Gianluca. this volume. Flexible Composition for Optional and Derived Arguments. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG12 Conference*, CSLI Publications.

Attia, Mohammed. 2008. A Unified Analysis of Copula Constructions in LFG. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG08 Conference*, CSLI Publications.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Blackwell Publishing.

Bresnan, Joan and Kanerva, Jonni M. 1989. Locative Inversion in Chicheŵa: A Case Study of Factorization in Grammar. *Linguistic Inquiry* 20(1).

Bresnan, Joan and Zaenen, Annie. 1990. Deep Unaccusativity in LFG. In Katarzyna Dziwirek, Patrick Farrell and Errapel Meijas-Bikandi (eds.), *Grammatical Relations: A Cross-Theoretical Perspective*, CSLI Publications.

Butt, Miriam. 1995. *The Structure of Complex Predicates in Urdu*. CSLI Publications.

Butt, Miriam. 1998. Constraining Argument Merger through Aspect. In E. Hinrichs, A. Kathol and T. Nakazawa (eds.), *Complex Predicates in Nonderivational Syntax*, Academic Press.

Butt, Miriam. 2010. The Light Verb Jungle: Still Hacking Away. In Mengistu Amberber, Brett Baker and Mark Harvey (eds.), *Complex Predicates in Cross-Linguistic Perspective*, Cambridge University Press.

Butt, Miriam, Dalrymple, Mary and Frank, Anette. 1997. An Architecture for Linking Theory in LFG. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG97 Conference*, CSLI Publications.

Butt, Miriam, Grimm, Scott and Ahmed, Tafseer. 2006. Dative Subjects. Talk given at the *NWO/DFG Workshop on Optimal Sentence Processing*.

Butt, Miriam and King, Tracy Holloway. 1997. Null Elements in Discourse Structure. Manuscript.

Butt, Miriam and King, Tracy Holloway. 2007. Urdu in a Parallel Grammar Development Environment. *Language Resources and Evaluation: Special Issue on Asian Language Processing: State of the Art Resources and Processing* 41.

Butt, Miriam, King, Tracy Holloway, Niño, María-Eugenia and Segond, Frédérique. 1999. *A Grammar Writer's Cookbook*. CSLI Publications.

Carlson, Gregory. 1977. *Reference to Kinds in English*. Ph. D.thesis, University of Massachusetts at Amherst.

Chomsky, Noam. 1970. Remarks on Nominalization. In R. A. Jacobs and P. S. Rosenbaum (eds.), *Readings in English Transformational Grammar*, Ginn.

Curnow, Timothy. 1999. Towards a Cross-linguistic Typology of Copula Constructions. In *Proceedings of the 1999 Conference of the Australian Linguistic Society*.

Dalrymple, Mary (ed.). 1999. *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*. The MIT Press.

Dalrymple, Mary, Dyvik, Helge and King, Tracy Holloway. 2004. Copular Complements: Closed or Open? In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG04 Conference*, CSLI Publications.

Dalrymple, Mary, Lamping, John and Saraswat, Vijay. 1993. LFG Semantics via Constraints. In *Proceedings of EACL 93*.

Grimshaw, Jane. 1990. *Argument Structure*. The MIT Press.

Higginbotham, James. 1983. Logical Form, Binding, and Nominals. *Linguistic Inquiry* 14.

Hook, Peter. 1974. *The Compound Verb in Hindi*. Center for South and Southeast Asian Studies, University of Michigan.

Humayoun, Muhammad. 2006. *Urdu Morphology, Orthography and Lexicon Extraction*. Masters Thesis, Department of Computing Science, Chalmers University of Technology.

Kibort, Anna. 2007. Extending the Applicability of Lexical Mapping Theory. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG07 Conference*, CSLI Publications.

Kratzer, Angelika. 1995. Stage-Level and Individual-Level Predicates. In Gregory Carlson and Francis Pelletier (eds.), *The Generic Book*, Chicago University Press.

Landau, Idan. 2010. *The Locative Syntax of Experiencers*, volume 53 of *Linguistic Inquiry Monographs*. The MIT Press.

Mohanan, Tara. 1994. *Argument Structure in Hindi*. CSLI Publications.

Pustet, Regina. 2003. *Copulas — Universals in the Categorization of the Lexicon*. Oxford University Press.

Raza, Ghulam. 2011. *Subcategorization Acquisition and Classes of Predication in Urdu*. Ph. D.thesis, University of Konstanz.

Singh, Mona. 1990. The Aspectual Content of Compound Verbs. In Y. No and M. Libucha (eds.), *Proceedings of the Seventh Eastern States Conference on Linguistics*.

TWO FOR THE PRICE OF ONE: AN LFG TREATMENT OF SENTENCE INITIAL OBJECT *ES* IN GERMAN

Nadine Theiler and Gerlof Bouma
Universität Potsdam   Göteborgs Universitet

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King (Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

We present an analysis of sentence initial object *es* ‘it’ in German. The weak pronoun *es* may only realize such an object under specific information structural conditions. We follow recent work suggesting these conditions are exactly those that licence the use of the presentational construction, marked by a sentence initial dummy *es*. We propose that the initial objects are an example of function amalgamation, show that only objects that may also appear in the clause-internal postverbal domain can participate in this fusion and make this precise in LFG. We end the paper with a contrastive discussion.

1 Introduction

In this paper, we will be concerned with German sentences of the following kind:

(1) a. [E]s haben keine Spinner angerufen.[1]
it have no idiots called
‘No idiots have called.’

b. [[B]itte im Sitzen pinkeln]$_1$, es$_1$ sieht keiner![2]
it sees nobody
‘Please sit down to pee, nobody can see you do it!’

In (a) we have an example of the, rather common, presentational construction. It is flagged by the preverbal dummy *es* ‘it’ and licensed by information structural properties of the subject. The sentence in (b) is more remarkable: an object pronoun *es* occupies the preverbal position. The theoretical literature has commonly claimed that non-subject *es* cannot fill this position. As seen from the example, and as discussed below, this is too strong a statement. Instead, we will argue that German allows amalgamation of the presentational marking function and the object realization function into one occurrence of *es*. In these cases, we may observe an object pronoun *es* in the otherwise forbidden preverbal position.

Discussion of German clausal syntax is facilitated by the *topological model*, in which we divide the sentence into fields. In a declarative main clause, the finite verb is in second position and any other verbs are in the clause-final verb cluster. These fixed positions delineate the topological fields *Vorfeld*, one constituent in front of the finite verb, and *Mittelfeld*, between the finite verb and the verb cluster.[3]

[†]The reported work was carried out in the context of the Collaborative Research Centre SFB 632 “Information Structure” at the University of Potsdam and the Centre for Language Technology of the University of Gothenburg. The authors gratefully acknowledge financial support from these research centers. We would also like to thank the audience at LFG12 in Denpasar and the members of *Grammatikseminariet* of the Dept. of Swedish in Gothenburg for discussion. In particular, we thank Adam Przepiórkowski and Elisabet Engdahl for drawing our attention to the Polish and Swedish data, and to Elizabeth Coppock for discussion of repeated *es*-es and economy.

[1]`http://www.talkteria.de/forum/topic-181017.html`

[2]First observed on a sign in the men’s bathroom at the linguistics department of the University of Potsdam. See also `http://www.shirtfactory24.de/products/de/Fun-Schilder/Tuer-Tor/Blechschild-Kloordnung-17x22cm.html`

(2)

Vorfeld	V_{fin}	*Mittelfeld*	*Verb cluster*
Es	haben	keine Spinner	angerufen.

In Section 2, we discuss the distributional properties of the Vorfeld-bound presentational *es* and the (normally) Mittelfeld-bound object *es*, and show how in examples like (1b), properties of these *es*-es are combined. Section 3 then introduces our LFG analysis, making precise the intuitive notion of combining characteristics of Vorfeld and Mittelfeld *es*. Section 4 ends the paper with a discussion of open issues and function amalgamation in other languages.

2 The multiple uses of *es*

The form *es* is the third person singular neuter pronoun. It has a wide range of uses: it can act as a regular, thematic subject or object, as the correlative pronoun for extraposed subordinate clauses, as the subject of weather verbs and certain existential and psych verbs, and as a dummy argument in some idiomatic constructions. In addition, there is a placeholder *es*, which occurs in presentational constructions and impersonal passives. Here, we will consider this placeholder *es* (Section 2.1) and relate it to the role of *es* as an object pronoun (Section 2.2). Where relevant, subscripts will be used to distinguish different functions of *es*.

2.1 Placeholder *es*

The German presentational construction is characterised by a placeholder *es*$_{\text{vf}}$, which appears directly before the verb in main clauses (3a), in the *Vorfeld*. The same *es* can also occur in impersonal passives (3b).

(3) a. Es sind immer noch keine Briefe für mich angekommen.[4]
es$_{\text{vf}}$ are still no letters for me arrived
'There have still been no letters for me so far.'

b. Es wurde jeden Abend gesungen und getanzt.
es$_{\text{vf}}$ was every night sung and danced
'People were singing and dancing every night.'

Note that placeholder *es*$_{\text{vf}}$ never determines agreement on the verb. In presentational constructions, the verb agrees with the (logical) subject of the sentence, *Briefe* 'letters' in (3a). Impersonal passives on the other hand are thematic-subject-less sentences, with the finite verb always in the third singular form.

Placeholder *es*$_{\text{vf}}$ may be called an *expletive*, but note that it is not an expletive in the sense of a non-referential pronoun used to fill an otherwise empty argument slot. This is already evident from (3a), where no argument is missing. In addition, *es*$_{\text{vf}}$ is set apart from expletive arguments by being strictly bound to the Vorfeld.

[3]The field after he verb cluster, the *Nachfeld*, will not be relevant in the present discussion.

[4]`http://www.bym.de/forum/bym-your-haushalt/439379-neue-bym-your-haushalt-cafe-hereinspaziert-49.html`

While the expletive subject of a weather verb (4a) can also appear after the finite verb (4b) or after a complementizer like *weil* 'because' in a subordinate clause (4c), *es*$_{\mathrm{vf}}$ is excluded from these positions (5a,b).

(4) a. Es regnet nun schon seit Stunden.
es$_{\mathrm{vf}}$ rains now already for hours
'It has been raining for hours now.'
b. Seit Stunden regnet es nun schon.
c. . . . weil es nun schon seit Stunden regnet.

(5) a. Jeden Abend wurde (*es) gesungen und getanzt.
b. . . . weil (*es) immer noch keine Briefe für mich angekommen sind.

Instead, *es*$_{\mathrm{vf}}$ is commonly analysed as not being selected by the verb, but fulfilling a purely structural function (e. g., Berman, 2003): it serves to uphold V2.

Crucially for what follows in the rest of the paper, German presentational sentences are not limited to intransitive verbs, but also allow transitives (6).

(6) Aber es wird sie leider keiner knuddeln.[5]
but es$_{\mathrm{vf}}$ will them sadly nobody cuddle
'Sadly, nobody will cuddle them.'

Even though there are consequently no clear conditions on which type of *verb* may appear in German presentational constructions, the class of possible *subjects* is information structurally restricted: *es*$_{\mathrm{vf}}$ may only occur if the subject is not an (aboutness) topic. For impersonal passives, this condition holds vacuously since they do not have a thematic subject. For presentational sentences, the restriction is illustrated by the mini-discourse in (7). After *Marie* is explicitly made an aboutness topic, it is possible to continue with canonical (7a), but not presentational (7b).

(7) Let me tell you something about Marie.
a. Marie kommt morgen zu Besuch.
Marie comes tomorrow for a visit
'Marie will come for a visit tomorrow.'
b. #Es kommt morgen Marie zu Besuch.

According to Frey (2004), German is a discourse configurational language. He divides the Mittelfeld into a topic region preceding modal and speaker-oriented adverbs, and a comment region following them. In (8), the position of the subject is ambiguous; it could sit either in the topic or in the comment region. The presence of *leider* 'unfortunately' in (9) resolves this uncertainty. Followed by *leider*, the subject is topical (9a), while preceded by the adverb it is not (9b). We have tried to convey the perceived meaning difference through accent placement in the English translations in (9).[6]

[5] `http://www.kaninchenforum.de/kaninchenhaltung-allgemeines/11824-urlaub-hilfe.html`

(8) Dieses Wochenende ist die Brücke gesperrt.
this weekend is the bridge closed
'The bridge is closed this weekend.'

(9) a. Dieses Wochenende ist [$_{\mathrm{topic}}$ die Brücke] **leider** gesperrt.
(About the bridge:)
'This weekend, the bridge will be clósed, unfortunately.'

b. Dieses Wochenende ist **leider** [$_{\mathrm{comment}}$ die Brücke] gesperrt.
(About what is going on in general:)
'Unfortunately, this weekend the brídge will be closed.'

As a result of the information structural properties of the presentational construction, *es*$_{\mathrm{vf}}$ cannot cooccur with a subject in the topic region (10, cf. Frey's ex. 19).

(10) Es spielt (erfreulicherweise) Max Greger (*erfreulicherweise) für
es$_{\mathrm{vf}}$ plays fortunately Max Greger fortunately for
unsere Gäste.
our guests
'I am delighted that Max Greger will play for our guests.'

We add that this non-topicality constraint only applies to the subject. As evidenced by the position of the object pronoun *ihn* 'him' in (12), material other than the subject *is* allowed to be topical.

(11) Es hat ihn zum Glück keiner nach seinem Ausweis gefragt.
es$_{\mathrm{vf}}$ has him luckily nobody for his passport asked
'Luckily, nobody asked for his passport.'

All in all, we take *es*$_{\mathrm{vf}}$ to be a flag for clauses without a topical subject (Sells, 2005, for a similar view of Icelandic *það*). This is also in line with the predictions made by *topicality hierarchies* that have appeared in the literature in various forms (see e.g., Dunbar, 1983). Indefinite and quantified expressions, which rank very low on such hierarchies, can be considered inherently non-topical phrases. As such, they make very natural subjects in presentational sentences (12). On the other extreme, an unstressed personal pronoun, which we may consider to be inherently topical, cannot appear as a subject in such constructions (14a). Focussed pronouns can, however, as in (14b). Definite expressions are situated relatively high on the topicality hierarchy. They are possible, but marked subjects of presentational constructions (13a) and work best in thetic statements like (13b).

(12) Es bleibt niemand zuhause.
es$_{\mathrm{vf}}$ stays nobody at home
'Nobody is staying at home.'

[6] The adverb sitting between the topic and comment region can sometimes induce narrow focus on an accompanying constituent. For all sentences in this paper, the reader is asked to consider only readings in which the adverb takes sentential scope.

(13) a. ?Es bleibt Otto zuhause.
b. Es kam die Polizei.
es$_{\mathrm{vf}}$ came the police
'The police showed up.'

(14) a. *Es bleibt er zuhause.
b. Es bleibt nur ER zuhause.
es$_{\mathrm{vf}}$ stays only he at home
'It is only he who is staying at home.'

2.2 Object *es*

The direct object pronoun *es*$_{\mathrm{obj}}$ forms an exception in the German pronominal system in more than one respect (Cardinaletti and Starke, 1996). Here, we will focus only on its limited positional distribution. At first sight, *es*$_{\mathrm{obj}}$ is in complementary distribution with the placeholder *es*$_{\mathrm{vf}}$: While *es*$_{\mathrm{vf}}$ only appears in the Vorfeld, *es*$_{\mathrm{obj}}$ seems to be excluded from just this position (15), from Travis (1984, p121), (16a). In contrast to *es*$_{\mathrm{obj}}$, the masculine and feminine object pronouns are not subject to the same restriction (16b).

(15) Er hat es / *Es hat er gegessen.
he has *es*$_{\mathrm{obj}}$ / *es*$_{\mathrm{obj}}$ has he eaten
'He ate it.'

(16) a. Im Zoo haben sie jetzt [ein Eisbärenjunges]$_1$.
in the zoo have they now a polar bear cub
*Es$_1$ hat Otto schon gesehen.
es$_{\mathrm{obj}}$ has Otto already seen
'At the zoo, they now have a polar bear cub. Otto has already seen it.'
b. Im Zoo haben sie jetzt [einen Pandabären]$_1$.
in the zoo have they now a panda bear
Ihn$_1$ hat Otto schon gesehen.
him has Otto already seen
'At the zoo, they now have a panda bear. Otto has already seen it.'

Classification of *es*$_{\mathrm{obj}}$ as a weak pronoun, which cannot be topicalized, explains these and other facts about *es*$_{\mathrm{obj}}$ (Cardinaletti and Starke, 1996).

However, more recently it has come to be accepted that there are in fact valid, non-marginal instances of *es*$_{\mathrm{obj}}$ in the Vorfeld (Meinunger, 2007, and references therein). The probably most famous example (17) is from Lenerz (1994).

(17) Ihr Geld$_1$ ist nicht weg, meine Damen und Herren.
your money has not gone my ladies and gentlemen
Es$_1$ haben jetzt nur andere.
es$_{\mathrm{vf/obj}}$ have now only others
'Your money has not disappeared, ladies and gentlemen. It just belongs to others now.'

As illustrated by the following examples, $es_{\mathrm{vf/obj}}$ can also cooccur with a dative object (18a), and can occupy the Vorfeld of subordinate V2 clauses (18b).

(18) a. [Er] hat das ganze Geld$_1$ irgendwo verloren
he has the whole money somewhere lost
oder es$_1$ hat ihm jemand geklaut.[7]
or $es_{\mathrm{vf/obj}}$ has him.DAT someone stolen
'He lost the entire sum somewhere, or someone stole it from him.'

b. wir [könnten dir helfen]$_1$.
we could help you
aber: ich denke, es$_1$ wird keiner machen.[8]
but I think $es_{\mathrm{vf/obj}}$ will nobody do
'We could help you, but I think nobody will.'

Meinunger (2007) notes that the conditions under which $es_{\mathrm{vf/obj}}$ is licensed correspond to the conditions under which we can observe es_{vf}: the subject has to be non-topical. This explains why examples like (15) and (16a) are marginal or even ungrammatical: they contain subjects high on the topicality hierarchy. In contrast, the acceptable sentences (17) and (18) have inherently non-topical subjects. This effect is systematically illustrated in the minimal quadruple in (19): as the subject's topicality increases from (a) through (d), acceptability dwindles.

(19) a. Es hat leider niemand gehört.
$es_{\mathrm{vf/obj}}$ has unfortunately nobody heard
'Unfortunately, nobody has heard it.'

b. ?Es hat leider Otto gehört.

c. ?*Es hat Otto leider gehört.

d. *Es hat er leider gehört.

Why does $es_{\mathrm{vf/obj}}$ place this information structural constraint on the subject? After all, the object pronoun es_{obj} does not impose any such restrictions. Then again, normal es_{obj} cannot occupy the Vorfeld, either. The placeholder es_{vf}, in turn, can appear in the Vorfeld and does introduce the non-topicality constraint; but it does not fill a slot in the verb's argument frame. The fact that $es_{\mathrm{vf/obj}}$ both brings along the information structural restriction and fills an argument slot suggests an analysis in which the two separate functions es_{vf} and es_{obj} are amalgamated into a single occurrence of *es*. The following section describes how an analysis like this can be implemented in LFG. The topic of function amalgamation is taken up again from a comparative perspective in Section 4.3.

[7] www.witzplanet.de/ddr.htm

[8] www.bodybuildingforum.at/threads/57831-Hilfe-bei-Ernahrungsplan

3 LFG Analysis

We have argued that $es_{\mathrm{vf/obj}}$ is an amalgamation of es_{vf} and es_{obj}. Before we show how this idea can be implemented in LFG, we will discuss our background assumptions about German clausal syntax.

3.1 The German clause

We follow the work of Berman (2003) in assuming that the German clause is a CP. SpecCP corresponds to the Vorfeld of the topological tradition. In a V2 sentence, the finite matrix verb resides in C. We do not distinguish between subject-initial and other V2 clauses in terms of c-structure, as f-structure annotations alone are enough to allow a differential treatment of subjects and non-subjects in the Vorfeld. The basic rules for the C projection are as follows:

(20) a. $\mathrm{CP} \rightarrow \underset{(\uparrow \text{ SUB}|\text{UDF}) = \downarrow}{\mathrm{XP}} \quad \mathrm{C}'$

b. $\mathrm{C}' \rightarrow \mathrm{C} \quad \mathrm{VP}$

The constituent in SpecCP is either the matrix subject or involved in a long-distance dependency. Instead of using the traditional grammaticalized discourse functions TOPIC and FOCUS for the latter, we assume there is one unbounded dependency function UDF at f-structure (we use Asudeh's, 2011, terminology; a similar proposal was made in Alsina, 2008). Information structural distinctions are better handled separately. We will take this to happen at s-structure (see below).

Constituents in the German Mittelfeld may be either the subject of the highest verb or local dependents of any of the verbs in the same *coherence domain*, roughly, the verb in second position plus the verbs in the verb cluster (see Müller, 2002, Ch 2, for an overview and references). Verbs that allow embedded verbs to share the verb cluster and Mittelfeld are said to *construct coherently*. As a result of coherent construction, we cannot tell which verb an argument belongs to from position alone.

We deviate from Berman (2003) in assuming that the Mittelfeld and verb cluster are in a flat VP in CompCP (Kaplan and Zaenen, 2003; Forst and Rohrer, 2009). Following Kaplan and Zaenen (2003), dependents relate to their verb through a functional uncertainty equation. We introduce the label EXTENDED COMPLEMENT FUNCTION for this equation and will investigate its definition shortly. Note that ECF is merely a convenience label and is not an f-structure feature. The VP contains two stretches of (nominal) arguments, divided by the speaker-oriented sentence adverbs that mark the boundary between the topic and comment regions (Frey, 2004). Information structural information is located at s-structure, accessible through the σ-projection from f-structure (Dalrymple and Nikolaeva, 2011).[9] The s-structure

[9]Our proposal does not depend on this exact implementation of information structure. Any setup that allows us to combine information structural hints from different sources (position in the sentence, lexical specification, construction type, etc) should do. We will keep the information structure

feature DISCOURSE FUNCTION allows us to specify whether material belongs to the topic or comment.

(21) a. VP $\rightarrow$ XP* XP* XP* V′

$(\uparrow \text{SUB}|\text{ECF}) = \downarrow$ $\downarrow \in (\uparrow \text{ADJ})$ $(\uparrow \text{SUB}|\text{ECF}) = \downarrow$

$(\downarrow_\sigma\ \text{DF}) = \text{TOPIC}$ $(\downarrow \text{ADV-TYPE}) = \text{SPOR}$ $(\downarrow_\sigma\ \text{DF}) \neq \text{TOPIC}$

b. V′ $\rightarrow$ V′ V

$(\uparrow (\text{XCOMP})) = \downarrow$

For reasons of exposition, we do not consider any adjuncts in our grammar fragment other than the speaker oriented sentence adverbs.

Final in the VP is the verb cluster as a V′. The V′ contains verbs that either project to the same f-structure as the c-structurally higher verb, e.g., when this is a perfective auxiliary, or to an XCOMP of other coherently constructing verbs.[10]

3.2 The Extended Complement Function

When Mittelfeld constituents are not subjects, they relate to their selecting verbs through the ECF equation, which we will define as follows:

(22) ECF = XCOMP* (PREDLINK) GF−SUB

This definition is motivated by the data below. For space considerations, we focus on es_{obj} in our examples. However, bar the linear position of the argument in the Mittelfeld, the results carry over to non-pronominal objects and to obliques. The ECF only concerns non-subjects, as subjects are always attached to the highest verb. If a Mittelfeld constituent is the (understood) subject of an embedded verb, this is the result of control. Non-nominal complements, like CP COMPs, normally do not appear in the Mittelfeld. This fact is not captured by the current grammar fragment, but may be modelled in the c-structure rules in (21a).

The simplest Mittelfeld argument-verb dependency is when a constituent is the argument of the matrix verb or any of its co-heads. We give projection annotations as subscripts to indicate the relation between the whole clause and the verb-argument pair of interest.

(23) a. Leider $\text{findet}_{\uparrow}$ $\text{es}_{(\uparrow \text{OBJ})}$ keiner wieder.
sadly finds es_{obj} nobody again
'Unfortunately, nobody will find es_{obj} again.'

annotations to a minimum and refer the reader to the cited book for details of the architecture.

[10]We are aware that our flat VP with mixed subjects, complements and adjuncts does not adhere to the endocentric mapping principles (Bresnan, 2001). To assume instead that the postverbal material is an S and the verb cluster a VP would equally well fit our needs. However, Berman's (2003) structure, in which arguments *adjoin*, one by one, to the VP containing the verbal cluster is less attractive than a flat structure containing all arguments, as this makes it hard to introduce the split between topical and non-topical regions of the Mittelfeld without introducing a new phrase label

b. Leider hat es$_{(\uparrow \text{OBJ})}$ keiner wiedergefunden$_{\uparrow}$.
sadly has *es*$_{\text{obj}}$ nobody found again
'Unfortunately, nobody has found it again.'

c. Jeder hat$_{\uparrow}$ es$_{(\uparrow \text{OBJ})}$ mal eilig.
everyone has *es*$_{\text{obj}}$ PART hurried
'Everyone is in a hurry at some time or another.'

Note that (23c) is an expletive object pronoun in a verbal idiom headed by the verb *haben* 'to have'.

A Mittelfeld object cannot come from a finite subordinate clause, whether complement or adjunct:

(24) a. *Keiner hat es$_{(\uparrow \text{COMP OBJ})}$ bemerkt, dass ich gefunden$_{(\uparrow \text{COMP})}$ habe.
nobody has *es*$_{\text{obj}}$ noticed, that I found have
Intended: 'Nobody noticed that I have found it.'

b. *Jeder hat es$_{(\uparrow \text{ADJ} \in \text{OBJ})}$ sich gefreut,
everyone has *es*$_{\text{obj}}$ been glad,
weil ich wiedergefunden$_{(\uparrow \text{ADJ} \in)}$ habe.
because I found again have
Intended: 'Everyone was happy, because I found it again.'

However, as mentioned above, it can be the object of an embedded verb in the same coherence domain, which are XCOMPs of any depth:

(25) a. Leider konnte es$_{(\uparrow \text{XCOMP OBJ})}$ keiner wiederfinden$_{(\uparrow \text{XCOMP})}$.
sadly could *es*$_{\text{obj}}$ nobody find again
'Unfortunately, nobody could find it again.'

b. Wahrscheinlich möchte es$_{(\uparrow \text{XCOMP XCOMP OBJ})}$ keiner
probably wants *es*$_{\text{obj}}$ nobody
wiederfinden$_{(\uparrow \text{XCOMP XCOMP})}$ können.
find again can.INF
'Probably, nobody wants to be able to find it again.'

c. Leider haben es$_{(\uparrow \text{XCOMP OBJ})}$ ihn auch alle spüren$_{(\uparrow \text{XCOMP})}$ lassen.
sadly have *es*$_{\text{obj}}$ him too all feel let
'Unfortunately, everyone let him know it, too.'

Example (25c) is an *accusative and infinitive*. The matrix object *ihn* 'him.ACC' controls the embedded subject; *es*$_{\text{obj}}$ is the embedded object.

As for objects of non-verbal heads, we see that (predicative) adjectives can have an object in the Mittelfeld (see also Forst, 2006; Vincent and Börjars, 2010).

(26) Da waren es$_{(\uparrow \text{PREDLINK OBJ})}$ alle schon ziemlich leid$_{(\uparrow \text{PREDLINK})}$.
then were *es*$_{\text{obj}}$ all PART rather fed up
'By that time, everyone was pretty fed up with it.'

In contrast, objects of prepositions need to be realized in situ and cannot move out of the PP into the Mittelfeld.

(27) a. *Keiner kommt es$_{(\uparrow \text{OBL}_{ohne}\ \text{OBJ})}$ noch [$_{\text{PP}}$ ohne$_{(\uparrow \text{OBL}_{ohne})}$] klar.
nobody manages *es*$_{\text{obj}}$ PART without VPART
Intended: 'Nobody can manage without it anymore.'

b. *Keiner möchte es$_{(\uparrow \text{ADJ} \in \text{OBJ})}$ noch [$_{\text{PP}}$ ohne$_{(\uparrow \text{ADJ} \in)}$] leben.
nobody want *es*$_{\text{obj}}$ PART without live
'Nobody wants to live without it anymore.'

These latter two examples now fall out from not allowing ADJs and OBLs in any expansion of the ECF equation. For the future, we might wish to explain them from German PP syntax in general, which does not allow preposition stranding.

3.3 The presentational construction

With our grammatical sketch of the German clause in place, we turn to the constructions at hand. The lexical entry for *es*, in all its uses, is as follows:

(28) *es* NP

$$(\uparrow \text{FORM}) = \text{ES}$$
$$\{(\uparrow \text{PRED}) = \text{'pro'} \mid \neg(\uparrow \text{PRED})\}$$
$$\neg(\text{UDF} \uparrow)$$
$$(\uparrow \text{CASE}) \in \{\text{NOM}, \text{ACC}\}$$
$$(\uparrow \text{AGR}) = 3\text{SG}$$

The disjunction for the PRED value allows its use as a referential pronoun or an expletive pronoun. Subject and object uses are allowed by the underdetermined case specification. The negative inside-out constraint $\neg(\text{UDF} \uparrow)$ lexically prevents *es* from entering such a long-distance dependency. Thus, in the CP in (20a), *es* can only appear as subject. As arguments in the VP are not assigned to UDF, *es* is free to realize subject or object there, in line with the observations of Travis (1984).

We model the presentational construction with a c-structure rule (Asudeh et al., 2008) that explicitly selects *es* in SpecCP. The constraint against topical subjects is also introduced here.

(29) CP $\rightarrow$ XP C′ *(to be revised)*

$$\begin{array}{c} \text{XP} \\ (\downarrow \text{FORM}) =_c \text{ES} \\ ((\uparrow \text{SUB})_\sigma\ \text{DF}) \neq \text{TOPIC} \end{array} \quad \text{C}'$$

To see how this c-structure rule interacts with the rest of the sentence, consider the ungrammatical combination of presentational *es*$_{\text{vf}}$ and a subject in the Mittelfeld topic region:

(30) *Es spielt Max Greger erfreulicherweise für unsere Gäste.
it plays Max Greger luckily for our guests
Intended: 'I am delighted that Max Greger will play for our guests.'

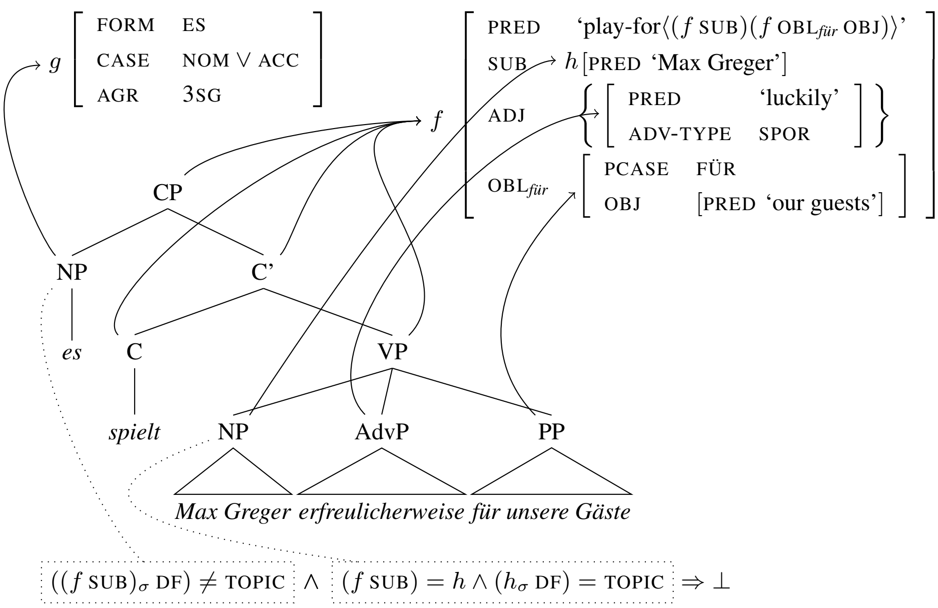


Figure 1: Structure and conflicting constraints for example (30).

The associated f- and c-structures are in Fig. 1. Note that the f-structure projected from es_{vf}, labelled g, is not integrated in the clausal f-structure f. The dummy es_{vf} is not selected by anything, is not assigned a grammatical function, and does not introduce a PRED. Thus, f and g are well-formed. Even valid uses of es_{vf} will project to a non-integrated f-structure. The information structural constraint against a topical subject is introduced in SpecCP, as per (29). The constraint talks directly about the subject of f, without referring to g at all. In Fig. 1, this constraint is given in the leftmost dotted box. At the same, the left periphery of the VP introduces two constraints, as per (21a): h is the subject of f and h is topical. These constraints are in the rightmost dotted box in the figure. The resulting conflict at s-structure is what rules out (30). In the absence of a speaker oriented sentence adverb, the topicality constraint on the subject in the VP can be avoided. S-structure information about the subject may also come from the lexicon or discourse to determine the felicity of the presentational construction, though (cf. Dalrymple and Nikolaeva, 2011).

3.4 Function amalgamation in the presentational construction

Finally, we turn to the task of modelling $es_{\mathrm{vf/obj}}$. As we have shown in Section 2.2, this *es* combines the properties of flagging the presentational construction and realizing an argument. One question we have not dealt with, however, is exactly which heads may realize their objects as $es_{\mathrm{vf/obj}}$. It turns out that the range of possibilities matches that of a Mittelfeld object. That is, a Vorfeld object *es* can be the object of anything within the same coherence domain. This is demonstrated in the sentences in (31)–(35), counterparts of the sentences in (23)–(27).

(31) a. Es$_{(\uparrow \text{OBJ})}$ findet$_{\uparrow}$ leider keiner wieder.
es$_{\text{vf/obj}}$ finds sadly nobody again
'Unfortunately, nobody will find it again.'

b. Es$_{(\uparrow \text{OBJ})}$ hat leider keiner wiedergefunden$_{\uparrow}$.
es$_{\text{vf/obj}}$ has sadly nobody found again
'Unfortunately, nobody has found it again.'

c. [E]s$_{(\uparrow \text{OBJ})}$ hat$_{\uparrow}$ jeder mal eilig[.][11]
es$_{\text{vf/obj}}$ has everyone PART hurried
'Everyone is in a hurry at some time or another.'

(32) a. *Es$_{(\uparrow \text{COMP OBJ})}$ hat keiner bemerkt, dass ich gefunden$_{(\uparrow \text{COMP})}$ habe.
es$_{\text{vf/obj}}$ has nobody noticed, that I found have
Intended: 'Nobody noticed that I have found it.'

b. *Es$_{(\uparrow \text{ADJ} \in \text{OBJ})}$ hat jeder sich gefreut,
es$_{\text{vf/obj}}$ has everyone been glad,
weil ich wiedergefunden$_{(\uparrow \text{ADJ} \in)}$ habe.
because I found again have
Intended: 'Everyone was happy, because I found it again.'

(33) a. Es$_{(\uparrow \text{XCOMP OBJ})}$ konnte leider keiner wiederfinden$_{(\uparrow \text{XCOMP})}$.
es$_{\text{vf/obj}}$ could sadly nobody find again
'Unfortunately, nobody could find it again.'

b. Es$_{(\uparrow \text{XCOMP XCOMP OBJ})}$ möchte wahrscheinlich keiner
es$_{\text{vf/obj}}$ wants probably nobody
wiederfinden$_{(\uparrow \text{XCOMP XCOMP})}$ können.
find again can.INF
'Probably nobody wants to be able to find it again.'

c. Es$_{(\uparrow \text{XCOMP OBJ})}$ haben ihn leider auch alle spüren$_{(\uparrow \text{XCOMP})}$ lassen.
es$_{\text{vf/obj}}$ have him sadly too all feel let
'Unfortunately, everyone let him know it, too.'

(34) Es$_{(\uparrow \text{PREDLINK OBJ})}$ waren alle schon ziemlich leid$_{(\uparrow \text{PREDLINK})}$.
es$_{\text{vf/obj}}$ were all PART rather fed up
'Everyone was pretty fed up with it.'

(35) a. Es kommt keiner noch ohne$_{(\uparrow \text{OBL}_{ohne})}$ klar.
es$_{\text{vf}}$ manages nobody PART without VPART
'Nobody can manage without it anymore.'

b. Es möchte keiner noch ohne$_{(\uparrow \text{ADJ} \in)}$ leben.
es$_{\text{vf}}$ wants nobody PART without live
'Nobody wants to live without it anymore.'

As remarked for its non-presentational counterpart, the *es* in (31c) is an expletive pronoun. However, this expletive is selected by the verb, and is not just a flag for the presentational construction. The grammaticality of the examples in (35)

[11] http://m.urbia.de/archiv/forum/th-1365744/-ueber-rote-Ampeln-laufen-obwohl-Kinder-daneben-stehen.html

is explained by the fact that elision of *ohne* 'without' is particularly easy. These sentences are thus regular presentational sentences and, as indicated in the gloss, do not involve $es_{\mathrm{vf/obj}}$.

Regarding head-argument dependencies, an $es_{\mathrm{vf/obj}}$ behaves in the Vorfeld just as an es_{obj} in the Mittelfeld. Regarding the information structural constraints, we are looking at a presentational construction. An optional ECF assignment in the c-structure rule for the presentational construction models this amalgamation.

(36) $$\mathrm{CP} \rightarrow \underset{\substack{(\downarrow \text{FORM}) =_c \text{ES} \\ ((\uparrow \text{SUB})_\sigma \text{DF}) \neq \text{TOPIC} \\ \big((\uparrow \text{ECF}) = \downarrow\big)}}{\mathrm{XP}} \quad \mathrm{C}'$$ *(replaces 29)*

4 Discussion: open issues and comparitive perspective

We have claimed that German allows amalgamation of functional properties of different *es*-es into one Vorfeld occurrence of *es* and showed how this notion of combining properties can be made precise in a constructional LFG analysis. We end this paper by discussing open problems in the analysis of the German data and data from three other languages that potentially could be analysed in a similar fashion.

4.1 Incoherently constructing verbs

In our discussion of non-finite complements in the ECF definition, we have only considered bare infinitives and participles. Such complements are always constructed coherently. A third type of non-finite complement, *zu*-marked infinitives, are sometimes constructed coherently. When they are, they allow $es_{\mathrm{vf/obj}}$ for their object, as in (37) with *scheinen* 'to seem', which takes a coherent *zu*-infinitive.

(37) [Ich lächele ein bisschen zu viel]$_1$, es$_1$ scheint niemand zu bemerken.
$es_{\mathrm{vf/obj}}$ seems nobody to notice
'I smile a bit too much, nobody seems to notice that.'[12]

Depending on the selecting verb, *zu*-infinitivals may also be constructed incoherently, that is, the embedded verb and its dependents are realized apart from the selecting verb, as in (38):

(38) Keiner hat dich gezwungen [es zu sagen].
nobody has you forced es_{obj} to say
'Nobody forced you to say it.'

On the basis of Dutch, Kaplan and Zaenen (2003) argue that incoherently constructed dependents are COMPs. If we assume this, the ECF equation predicts that an

[12] http://www.theaterzeitung-koeln.de/archiv/akt29-januar-2012/aus-der-koelner-theaterszene/facebook-mit-anfassen/

object of an incoherently combined verb cannot appear as $es_{\mathrm{vf/obj}}$. This prediction appears to be correct, although the data is not as solid as with finite COMP.

(39) ??Es hat dich keiner gezwungen zu sagen.
$es_{\mathrm{vf/obj}}$ has you nobody force to say
'Nobody forced you to say it.'

A further complication with *zu*-infinitivals is the so called 'third construction', a (marked) construction in which one of the the dependents of an incoherently combined verb appears in the Mittelfeld of the selecting verb.

(40) Keiner hat dich das gezwungen zu sagen.
nobody has you that forced to say
'Nobody forced you to say that.'

More work establishing the data and investigating the third construction is needed to see how *zu*-infinitivals interact with $es_{\mathrm{vf/obj}}$.

4.2 Multiple occurrences of *es*

Thus far we have argued that the constraints the presentational construction introduces are on the topicality of its subject. However, there are further coocurrence restrictions on the rest of the sentence that do not fall under this characterization. First, consider the following data with an (optional) expletive subject.

(41) a. Es regnet vielleicht.
it rains perhaps
'Maybe it rains.'
b. Vielleicht regnet es.
c. *Es regnet es vielleicht.

(42) a. Es graut mir vor dem Abend.
it horrifies me for the evening
'I fear the night.'
b. Mir graut (es) vor dem Abend.
c. *Es graut es mir vor dem Abend.

The (a) sentences show the expletive subject in the Vorfeld. The (b) sentences show that *es* is selected and not just a Vorfeld *es*, as it can appear in the Mittelfeld. Note that in the case of *grauen* 'horrify', this subject is optional. As we take these subjects to be non-thematic, we would also have to assume that they are non-topical. However, if they are non-topical, it should be possible to use them in a presentational construction. The (c) examples above show that this is not the case: one cannot use unselected es_{vf} with these verbs.

It is unclear how to handle this properly at this point. A possible solution would be to constrain the subject to have a PRED value. This could either be done directly or by changing the constraint forbidding topical subjects into a constraint enforcing the subject to be part of the comment. Unfortunately, this solution would cause problems for impersonal passives, which we also assume to lack a thematic subject, but which can be used in a presentational construction.

The challenges these expletive subjects pose are not limited to the presentational construction. Expletive subjects seem to pattern like other unstressed personal pronouns in the Mittelfeld, too: they appear in the topic region (43a). This is unexpected under the topic-comment division of the Mittelfeld. This appears to be specific to unstressed *personal* pronouns, and not any unstressed pronouns, as (43b) shows – the unstressed indefinite pronoun *wer* 'someone' is in the Mittelfeld comment region and occurs in a presentational construction.

(43) a. Morgen wird es wohl / *wohl es regnen.
tomorrow shall it PART rain
'I guess it will rain tomorrow.'

b. … und es wird wohl wer vorbei geschickt.
and es_{vf} is PART someone.NOM over sent.PRFPRT
'… and I guess they will send someone over.'[13]

Given the data in (43a), the impossibility of (41c) is perhaps less surprising: these expletive subjects follow the general correlation we saw between Mittelfeld positioning and the possibility of occurring in a presentational construction. Spelling out what the information structural properties of these subjects is and how this relates to our observations will have to remain a topic for future work.

A related but slightly different observation can be made about the object in a presentational construction. In the attested example below, we find a Mitteld es_{obj} together with presentational es_{vf}.

(44) Sie sehen ja, wie er es hervorbringt. Bei dem Dichter sehen Sie es aber nicht.
Of course, you see how he [the painter or sculptor] creates it [the admired artwork]. But in the poet's case, you don't see it.

(*)Es hat es keiner gesehen.[14]
es_{vf} has es_{obj} nobody seen.
'Nobody has seen it.'

The example is starred within parentheses to indicate that there is variation in the grammaticality judgements. On the one hand, the example is from edited text and other examples like it can easily be found. One the other, informal polling of native speaker informants suggests a low degree of acceptability for some. Note that es_{vf}… es_{obj} is possible in our grammar fragment, so we would predict (44) to be grammatical. It is the variation in its grammaticality that is problematic under our account.

[13] http://www.mietrecht-hilfe.de/mietrecht-forum/kuendigung/1031-schimmel-nach-3-wochen-aber-3-jahres-klausel.html

[14] From page 172 of Theodor Adorno, 1967, *Ohne Leitbild. Parva Aesthetica*, Suhrkamp.

The (for some) obligatory amalgamation of two *es*-es might be explained as a kind of syntactic haplology (Neeleman and Van de Koot, 2006), under which repetition of material is avoided for (prosodic) wellformedness reasons. Haplology is commonly restricted to functional material, and indeed, there is nothing remarkable about multiple *es*-es in a German sentence in general.

(45) Es hat es nicht verdient.
it has it not deserved
'It/he/she hasn't deserved it.'

However, haplology also normally applies to material that is adjacent. Our data thus poses a problem for an explanation from haplology. Not only are the two *es*-es not adjacent in (44), removing any pressure to leave one unexpressed, they *cannot* be adjacent, since that would violate V2. There are well-known (apparent) exceptions to V2 in German, but none is of the type needed for a deletion rule of the type ... *es* ~~*es*~~ ... to apply. For instance, es_{vf} plus another accusative pronoun in the Vorfeld is also ruled out:

(46) *Es ihn hat keiner gesehen.
es_{vf} him has nobody seen
Intended: 'Nobody has seen him.'

An appeal to economy would be a more promising direction: if function amalgamation gives us the option to express with one *es* exactly as much as we would express with two *es*-es, pronouncing both would be uneconomic. The strength of such a theory relies, however, on finding the right notions of 'economy' and 'expressing'. In LFG, the Principle of Economy (Bresnan, 2001) would appear to suggest itself. However, c-structure terminals and preterminals are exempt from this principle, so it would not have anything to say about repeated material in a sentence, at all. If we were to broaden the application of the Principle of Economy, the whole presentational construction would become suspect: The construction involves 'extra' material (i.e., es_{vf}) to express exactly the same f-structure as its non-presentational counterpart. Even the information about the information structural properties of the subject is not contributed as such by the presentational construction, as this is information already supplied by the lexical, discursive or positional properties of the subject. The fact that the presentational works best with subjects that are clearly non-topical, such as negatively quantified subjects, underlines this.

The economy intuition might be better captured in an Optimality Theoretic setting, where constraints that promote flagging a marked situation (such as lacking a topical subject) and keeping separate information separate (and thus not using one *es* for two purposes) interact with economy constraints that give rise to function amalgamation. Under such a model, we would expect there to be a correlation between liking $es_{\mathrm{vf/obj}}$ and disliking $es_{\mathrm{vf}} \ldots es_{\mathrm{obj}}$ for a native speaker, which is a testable hypothesis.

4.3 Comparison to other languages

We end with a brief discussion of similar phenomena in Dutch, Swedish and Polish, and the potential for carrying over our analysis to these languages.

Dutch The concept of function amalgamation of the kind that we have argued for in this paper was already introduced for the Dutch existential construction by Bech (1952). Dutch has an existential construction marked by the weak adverbial pronoun *er* 'there' appearing in the Vorfeld or preceding the subject in the Mittelfeld. This construction is licensed with impersonal passives or when the subject is indefinite. In one of its other uses, *er* may realize the object of an adposition, when this object has floated leftward of its head into the Mittelfeld. Since *er* is a weak pronoun, er_{obj} is normally banned from the Vorfeld, just like es_{obj} in German. A further parallel with the German data is that when the conditions for an existential construction are met, er_{obj} can appear in the Vorfeld. This is illustrated in (47). The grammaticality of the sentence varies with the definiteness of the subject, indicating that the Vorfeld *er* not only realizes the object of *op* 'on', but also flags the existential construction.

(47) Er staat een handtekening van mij / *mijn handtekening op.
$\text{er}_{\text{ex/obj}}$ stands a signature of me my signature on
'My signature is on it.'

Further discussion and analysis of this construction can be found in Odijk (1993), Bouma (2000), and Neeleman and Van de Koot (2006). Our LFG model should carry over to a great extent. Aside from the mentioned fact that er_{ex} is not Vorfeld-bound, a difference with the German data is that three different *er*-s (objective, locative, partitive) may be combined with er_{ex}. Also, because er_{ex} may appear in the Mittelfeld, repeated mentions of *er* could in principle be adjacent, so that haplogogy might be an explanatory factor in some of these data (Odijk, 1993).

Swedish Clausal subjects in Swedish can either be realized clause initially (in-situ) as in (48a) or clause finally (extraposed) as in (48b). In the latter case, a correlative pronoun *det* 'it' is used as a preliminary subject. However, as shown in example (48b), from Engdahl (2007), realizing the object of extraposed clause *det* 'it' now becomes optional, even though in (a) it was obligatory (see also Engdahl, To appear).

(48) a. [Att du sa *(det)] var dumt.
that you said det_{obj} was stupid
'It was stupid (of you) to say it/so.'

b. Det var dumt att du sa (det)
det_{corr} was stupid that you said det_{obj}

c. Därför var *(det) dumt att du sa *(det)
therefore was det_{corr} stupid that you said det_{obj}
'That is why it was stupid of you to say so.'

It appears to be possible to long-distance bind the embedded object *det* from the front of the clause and have it act as a preliminary subject for the matrix clause at the same time. Example (48c) shows that this is restricted to the sentence initial position for unbounded dependencies, as the sentence internal subject position after the verb cannot participate in such function amalgamation – both *det*-s must be realized in this case.

The Swedish data are different from the German and Dutch in that there is no (obvious) information structural marking involved. However, given the strong positional restrictions of the effect, it should be possible to analyze the Swedish data with the help of a dedicated c-structure rule or annotation, too.

Polish Finally, we will consider the reflexive clitic *się* in Polish, which can be involved in rather elaborate function amalgams. A basic example is in (49), from Kupść (1999). The sentence contains two inherently reflexive verbs requiring *się*. However, only one reflexive clitic is needed to meet the needs of both verbs.

(49) Boję się głośno roześmiać.
fear.1SG.INH *się*$_{\text{inh}}$ loudly laugh.INF.INH
'I'm afraid to laugh loudly.'

Since Polish has clitic climbing, it is hard to pin this sharing effect down to a specific c-structure position. A lexical treatment, where reflexive verbs carry optional control equations passing *się* on to embedded verbs may be more appropriate for these data than a construction-based implementation. See Kupść (1999) for an HPSG-based lexical account.

The example in (50), idem, shows two further challenging aspects of the Polish reflexive clitic. First, Polish has a reflexive impersonal construction that is marked by *się* (Kibort, 2008, for an LFG analysis). As the example shows, one *się* for *powinno* 'should' both marks it as impersonal and supplies the required inherent reflexive. Thus, *się* can not only fulfil the same role for different verbs, but also different roles for one verb. In fact, multiple *się*-s for one verb are not allowed.

(50) Po tych lekach powinno mu się zacząć
after these pills should.INH.IMP him.DAT się$_{\text{imp/inh}}$ begin.INF
udawać mniej obawiać spotkać ze znajomymi
succeed.INF.INH less fear.INF.INH meet.INF.INH with friends
sprzed wojny.
from before war
'As a result of these pills, he should begin to succeed to be less afraid of meeting with friends from before the war.'

Secondly, *się* in (50) is the reflexive marker for a total of four inherently reflexive verbs. This is also a reason a constructional account is not attractive for the Polish data. We would need an arbitrary number of c-structure annotations to distribute the one reflexive clitic over all embedded inherently reflexive verbs. A lexical control analysis would probably fare better in this respect, too, as it is possible to chain the reflexive from one verb to another, irrespective of where and how often the reflexive was realized.

References

Alsina, Alex. 2008. A Theory of Structure-sharing: Focusing on Long-distance Dependencies and Parasitic Gaps. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG08*, pages 5–25, CSLI Online Proceedings.

Asudeh, Ash. 2011. Towards a Unified Theory of Resumption. In Alain Rouveret (ed.), *Resumptive Pronouns at the Interfaces*, pages 121–187, Amsterdam: John Benjamins.

Asudeh, Ash, Dalrymple, Mary and Toivonen, Ida. 2008. Constructions with Lexical Integrity: Templates as the lexicon-syntax interface. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG08*, pages 68–88, CSLI Online Publications.

Bech, Gunnar. 1952. Über das niederländische Adverbialpronomen *er*. *Traveaux du Cercle Linguistique de Copenhague* 8, 5–32, `http://www.dbnl.org/tekst/bech001uber01_01/bech001uber01_01-x.pdf`.

Berman, Judith. 2003. *Clausal Syntax of German*. CSLI Publications.

Bouma, Gosse. 2000. Argument Realization and Dutch R-Pronouns: Solving Bech's problem without movement or deletion. In Ronnie Cann, Claire Grover and Philip Miller (eds.), *Grammatical Interfaces in Head-driven Phrase Structure Grammar*, CSLI Publications.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Blackwell.

Cardinaletti, Anna and Starke, Michael. 1996. Deficient Pronouns: A view from Germanic. In Samual Epstein and Steve Peter (eds.), *Studies in Comparative Germanic Syntax*, volume 2, pages 21–65, Kluwer Academic Publishing.

Dalrymple, Mary and Nikolaeva, Irina. 2011. *Objects and Information Structure*. Cambridge University Press.

Dunbar, Ronald. 1983. The Topicality Hierarchy in German Syntax. *Monatshefte* 75(2), 141–146.

Engdahl, Elisabet. 2007. Flerfunktionelt *det*. In Karin Milles and Anna Vogel (eds.), *Språkets roll och räckvidd*, Stockholm Studies in Scandinavian Philology, New Series, No. 42, pages 97–106, University of Stockholm.

Engdahl, Elisabet. To appear. Optional explitive subjects in Swedish. *Nordic Journal of Linguistics*.

Forst, Martin. 2006. COMP in (Parallel) Grammar Writing. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG06*, CSLI Online Proceedings.

Forst, Martin and Rohrer, Christian. 2009. Problems of German VP Coordination. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG09*, CSLI Online Proceedings.

Frey, Werner. 2004. A Medial Topic Position for German. *Linguistische Berichte* 198, 153–190.

Kaplan, Ron and Zaenen, Annie. 2003. West-Germanic Verb Clusters in LFG. In Pieter Seuren and Gerard Kempen (eds.), *Verb Constructions in German and Dutch*, pages 127–150, John Benjamins Publishing Company.

Kibort, Anna. 2008. Impersonals in Polish - an LFG perspective. *Transactions of the Philological Society* 106(2), 246–289.

Kupść, Anna. 1999. Haplology of the Polish Reflexive Marker. In Robert Dorsley and Adam Przepiórkowski (eds.), *Slavic in Head-Driven Phrase Structure Grammar*, Studies in Constraint-Based Lexicalism, pages 91–125, CSLI.

Lenerz, Jürgen. 1994. Pronomenprobleme. In Brigitta Haftka (ed.), *Was determiniert Wortstellungsvariation? Studien zu einem Interaktionsfeld von Grammatik, Pragmatik und Sprachtypologie*, pages 161–175, Westdeutscher Verlag.

Meinunger, André. 2007. About Object *es* in the German Vorfeld. *Linguistic Inquiry* 38(3), 553–563.

Müller, Stefan. 2002. *Complex Predicates: Verbal complexes, resultative constructions, and particle verbs in German*. Studies in Constraint-Based Lexicalism, CSLI Publications.

Neeleman, Ad and Van de Koot, Hans. 2006. Syntactic Haplology. In Martin Everaert and Henk van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, pages 685–710, Blackwell.

Odijk, Jan. 1993. *Compositionality and Syntactic Generalizations*. Ph. D.thesis, Katholieke Universiteit Brabant, Tilburg.

Sells, Peter. 2005. The peripherality of the Icelandic expletive. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG05*, pages 408–428, CSLI Online Proceedings.

Travis, Lisa. 1984. *Parameters and Effects of Word Order Variation*. Ph. D.thesis, Massachussetts Institute of Technology.

Vincent, Nigel and Börjars, Kersti. 2010. Complements of Adjectives: A Diachronic Approach. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG10*, pages 458–478, CSLI Online Proceedings.