Proceedings of LFG10

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Contents

1 Editors' Note

The program committee for LFG10 were Martin Forst and Louisa Sadler. We would like to thank them again for 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. The local organizing committee consisted of Ash Asudeh and Ida Toivonen. We would like to thank Mary Dalrymple for helping with the light editing and Dikran Karagueuzian for his and CSLI's unfailing support.

The html table of contents lists all the papers presented at the conference. Some papers were not submitted to the proceedings. For these papers, we suggest contacting the authors directly.

THE CATALAN DEFINITE ARTICLE AS LEXICAL SHARING

Alex Alsina

Pompeu Fabra University

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

This paper argues for the claim that the definite article (d-article) in Catalan is a bound morpheme that adds its information about grammatical category to that of the host it forms a compound with. In this way, the derived compound satisfies two X^0 categories in the c-structure, in an instance of lexical sharing (Wescoat 2002). The present proposal is shown to be superior to alternatives that, while assuming bound morpheme status for the d-article, assume a single category for the word that contains the d-article (as in Miller 1992 for French). The crucial evidence is provided by potentially boundless sequences in which the d-article and weak prepositions are adjacent to each other. Only a theory that recognizes the status of these elements as bound morphemes that add a grammatical category to the word they are part of correctly accounts for the relevant evidence.

1 Introduction

The goal of this paper is to argue for the claim that the definite article (d-article) in Catalan is a bound morpheme and that the word that the d-article is part of corresponds to two X^0 categories in the c-structure. This is an instance of the theoretical construct of *lexical sharing*, proposed by Wescoat (2002), whereby a single word (or vocabulary item) instantiates, or corresponds to, more than one X^0 category. The paper likewise argues for the claim that weak prepositions in Catalan are also lexical sharing morphemes. A consequence of these claims is that, since the d-article and the weak preposition *de* can be adjacent to each other in potentially boundless sequences, they give rise to instances of multiple lexical sharing (words corresponding to many X^0 categories). The evidence provided suggests that lexical sharing is the right way to handle "phrasal affixes" in general, over alternative analyses.[†]

The paper starts out by outlining the structure of the DP in Catalan and arguing for the existence of two classes of determiners. In the context of the Catalan determiner system, in section 3, the d-article is shown to have some properties common to all determiners, but also some special properties that set it aside from both classes of determiners. In section 4, it is proposed that the explanation for these properties involves assuming (a) that it is a bound morpheme (following similar proposals by Miller 1992 for French) and (b) that the word containing this morpheme has the grammatical categories of both its component elements—an instance of lexical sharing. Next, in section 5, it is shown that so-called weak prepositions in Catalan are also bound morphemes that attach to a following word yielding a word with the grammatical categories of its several component elements. Since the d-article can attach to the weak preposition *de* to its right and this item can attach to

[†] This paper has been supported in part by research grant HUM2007–61916FILO. I thank Joan Bresnan and participants in LFG10 for observations, Martin Forst for providing many examples from the web that have made me change and refine my analysis, and Michael Wescoat for detailed comments on earlier versions of the paper. All errors are due to me.

the d-article to its right, complex words consisting of sequences of d-articles and weak prepositions are formed: these words have a list of the categories contributed by their morphemes. It will be argued that this approach is superior to alternative approaches without lexical sharing. The present approach will be shown to provide a simple way of accounting for co-occurrence restrictions involving particular lexical items, in section 6. Finally, conclusions will be presented.

2 The structure of the Catalan DP

This section proposes an analysis of the Catalan DP: it includes the assumption of two classes of determiners (subsection 2.1), a proposal about its c-structure (2.2), and a proposal about its f-structure and the mapping between the two structures (2.3).

2.1 Two classes of determiners

If we use the working definition of determiner as a word belonging to a closed class that can only appear in the initial position of a noun phrase, we find that, in Catalan, the class of determiners splits into two classes on the basis of the syntactic behavior of these words. We shall refer to the two classes of determiners as *semipronominal* and *nonpronominal* determiners; some examples of words belonging to these two classes are given in (1):

(1) a. Semipronominal Ds: *aquest* 'this', *aquell* 'that', *quin* 'which', *quant* 'how much/many', *cap* 'no/none/any', *un* 'one/a/an', *algun* 'some/any', *tant* 'so much/many', *molt* 'many/much', etc.

b. Nonpronominal Ds: *cada* 'each', *sengles* 'each his own', *qualque* 'some', *cert* 'certain', *mon* 'my', *ton* 'your', *son* 'his/her', *llur* 'their', *tot* 'every' (as in *tot estudiant* 'every student', not as in *tots els estudiants* 'all the students'), etc.

The difference between the two classes is that, whereas nonpronominal Ds must head a DP with a noun heading its sister NP, semipronominal Ds may occur without this noun. In other words, semipronominal Ds may introduce headless noun phrases, whereas nonpronominal Ds may not. Taking *aquella* 'that(F,S)' and *cada* 'each' as examples of semipronominal and nonpronominal Ds respectively, we see that both can introduce noun phrases with overt nouns, but only *aquella* can introduce a noun phrase lacking an overt noun, that is, consisting only of the determiner, or of the D and a PP, or of the D and an AP, or of the D and a relative clause:

(2) a. Hem insistit en aquella situació (compromesa)
we-have insisted on that(F,S) situation(F,S) delicate(F,S)
(que es va presentar)
that arose
'We insisted on that (delicate) situation (that arose).'

b. Hem insistit en aquella.
we-have insisted on that(F,S) ('We insisted on that one.')

c. … en aquella compromesa.
… on that(F,S) delicate(F,S) ('… on that delicate one.')

d. … en aquella que es va presentar.
… on that(F,S) that arose ('… on that one that arose.')

(3) a. Hem insistit en cada situació (compromesa) (que es va presentar)
we-have insisted on each situation(F,S) delicate(F,S) that arose

'We insisted on each (delicate) situation (that arose).'

b. * Hem insistit en cada.
we-have insisted on each ('We insisted on each one.')

c. * … en cada compromesa.
… on each delicate(F,S) ('… on each delicate one.')

d. * … en cada que es va presentar.
… on each that arose ('… on each one that arose.')

2.2 The c-structure of the DP

We will assume that both classes of determiners belong to category D and that the category D projects a DP, adopting the DP hypothesis of Abney (1987) and others, which has been incorporated into the LFG framework by Bresnan (2001), among others. The structure of phrases, and specifically the structure of DP, conforms to X-bar Theory, as expressed in (4):

(4) X-bar Theory:

a. $X'' \rightarrow YP\ X'^{(')}$

b. $X' \rightarrow X^{(')}\ ZP$

We take XP to be equivalent to X″. The sister of D, if there is one, is an NP. Spec of DP is reserved for definite *tot/s* 'all' and adverbs or adverbial phrases such as *fins i tot* 'even', *exactament* 'exactly', *només* 'only', *almenys* 'at least', etc. Examples containing some of these specifiers of DP would be: *tots els llibres* 'all the books', *només aquelles notes* 'only those notes', etc.

Spec of NP is reserved for a special class of adjectives that we can call specificational adjectives, or SpecA, some examples of which would be: *altre* 'other', *dos* 'two', *tres* 'three', *primer* 'first', *segon* 'second', *últim* 'last', *seu* 'his', *nostre* 'our', etc.

In addition, phrase structures are subject to Bresnan's (2001) Economy of Expression, which allows any category licensed by (4) to be missing:

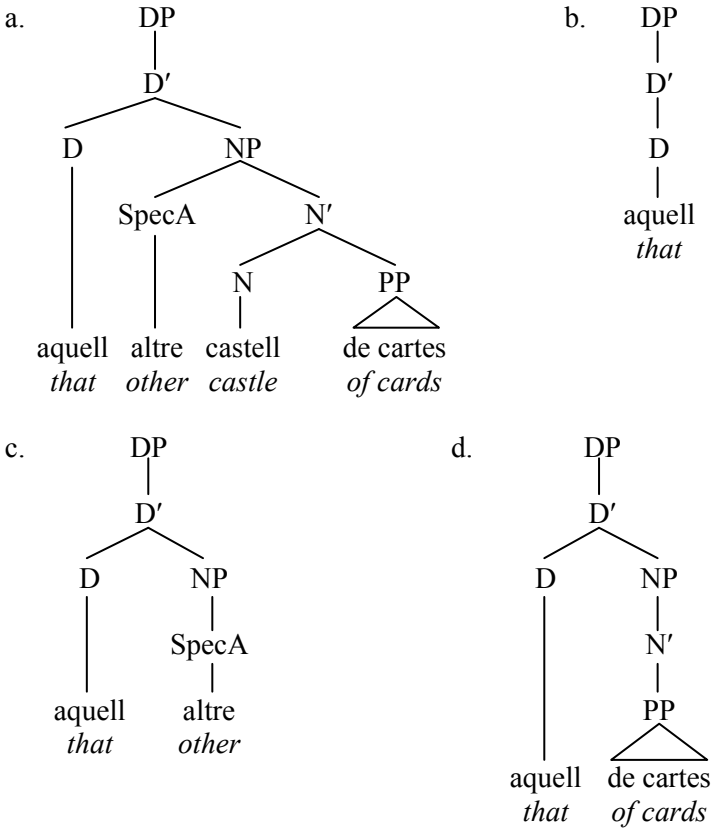
(5) Economy of Expression:

All c-structure nodes are optional and are only used if needed for semantic reasons or to satisfy well-formedness conditions.

According to this principle, there can be DPs that do not include an NP and NPs that do not include an N′ or an N. Nonhead categories can likewise be missing. There are noun phrases without a noun, but there are no null nouns. Economy of Expression rules out the null noun: if there were a null noun, there would have to be an N node in the c-structure that would not be needed for semantic reasons or to satisfy a well-formedness condition.

Following are some examples of c-structures of DPs:

(6)



2.3 The f-structure of the DP

The combination of (4) and (5) correctly allows headless noun phrases, that is, DPs that either lack an NP altogether, as in (6b), or include an NP that lacks the head N, as in (6c,d). Headless noun phrases are well-formed provided the determiner belongs to the class of semipronominal Ds. But the c-structure principles given so far do not distinguish between the two classes of Ds. So, by these principles we expect to find headless noun phrases with nonpronominal Ds. Headless noun phrases such as those given in (3b–d) are ungrammatical and yet are assigned a well-formed c-structure.

The reason why the noun phrases in (3b–d) are ungrammatical is because they violate a well-formedness condition on the f-structure,

specifically, Completeness. Completeness, as stated in Bresnan (2001), has two parts: the requirement that an f-structure have all the grammatical functions specified in its local PRED feature, with which we will not be concerned here, and the requirement in (7):

(7) Completeness:

Every f-structure with a thematic role must have a PRED feature.

Words of categories N, V, and A generally have a PRED feature. Therefore, a headed NP will normally satisfy condition (7) because the head noun provides the necessary PRED feature. Ds split into two groups: some do not have this feature—the nonpronominal Ds—and some have an optional [PRED 'pro'] feature—the semipronominal Ds.[1] See vocabulary entries of a D of each type:

(8) a. aquell: D $\begin{bmatrix} \text{SPEC} & [\text{DEM} +] \\ \text{NUM} & \text{sg} \\ \text{GEN} & \text{masc} \\ (\text{PRED} & \text{'pro'}) \end{bmatrix}$

b. cada D $\begin{bmatrix} \text{SPEC} & [\text{DISTR} +] \\ \text{NUM} & \text{sg} \end{bmatrix}$

In the mapping between f-structure and c-structure, the following conditions apply: (a) The f-structure of a phrasal category is that of its head, and (b) the f-structure of a functional category is that of its complement phrase. Thus, the phrasal categories DP and NP map onto the same f-structure as their head D and N respectively. Given that D is a functional category (like C and I), its f-structure must be that of its NP complement. Consequently, the f-structure of a DP is the unification of the f-structure of the head D and of the N head of the complement NP.

A semipronominal D can head a well-formed DP without an NP sister or with a headless NP, as it has the PRED feature needed to satisfy Completeness. A nonpronominal D cannot do so, as the DP would lack a PRED feature and so violate Completeness. The optionality of the feature [PRED 'pro'] in (8a) allows a semipronominal D such as *aquell* to co-occur with a headed NP complement, in which case that feature is not used, or with a headless NP complement, in which case that feature is used. The absence of that feature in (8b) forces a nonpronominal D such as *cada* to co-occur with a headed NP, as that is the only way for the DP to satisfy Completeness.

This provides the explanation for the contrast between a grammatical DP consisting only of the word *aquell* and an ungrammatical DP consisting only of the word *cada*. [$_{DP}$ aquell] maps onto a feature structure containing all the features in (8a), including the PRED feature, whereas [$_{DP}$ cada] maps onto the feature structure in (8b), which lacks this feature.

[1] There is a third type of D that we are not concerned with here: pronominal Ds, which always have the [PRED 'pro'] feature , such as *algú* 'someone', *qui* 'who', *res* 'nothing', *ell* 'he', etc.

3 Properties of the d-article

Having outlined the structure of the DP and having established the existence of two classes of determiners in Catalan, we will see why it is reasonable to consider d-articles determiners, in 3.1, and how they differ from other determiners syntactically, in 3.2, and morphophonologically, in 3.3.

3.1 The d-article as a determiner

The d-article is like a determiner, in that it is initial in the noun phrase:

(9) a. el pas ràpid ('the fast pace')
the(M,S) pace(M,S) fast(M,S)

b. *pas el ràpid / *pas ràpid el

(10) a. la primera veu ('the first voice')
the(F,S) first(F,S) voice(F,S)

b. *primera la veu / *primera veu la

(11) a. les peces de ferro ('the iron parts')
the(F,P) parts(F,P) of iron

b. *peces les de ferro / *peces de ferro les

Consequently, it cannot co-occur with another determiner in the same noun phrase, as the two would be competing for initial position: **cada la situació* 'each the situation', **la cada situació* 'the each situation', **l'aquella situació* 'the that situation', **aquella la situació* 'that the situation', etc.

The d-article is like a semipronominal determiner, in that it may introduce not only a noun phrase with a head noun, as in (9)–(11), but also a noun phrase without a head noun, as in (12) (the relevant phrase in boldface):

(12) a. D'aquestes peces, només vull **les de ferro**.
of these(F,P) parts(F,P), only I-want the(F,P) of iron
'Of those parts, I only want the iron ones.'

b. La segona pàgina és més interessant
the(F,S) second(F,S) page(F,S) is more interesting
que **la primera**.
than the(F,S) first(F,S)
'The second page is more interesting than the first one.'

c. Aquest cavall és **el que corre més**.
this(M,S) horse(M,S) is the(M,S) that runs the most
'This horse is the one that runs the most.'

In these examples, the d-article is followed by an NP consisting only of a PP, as in (12a), of a SpecA, as in (12b), or of a relative clause, as in (12c).

3.2 Special syntactic properties of the d-article

The d-article differs from ordinary semipronominal determiners in that it must be followed by some lexical element in its NP sister:

(13) a. Mentre esperava els amics, pensava en
while he-waited the(M,P) friends(M,P) he-thought in
els *(de la infantesa).
the(M,P) of the childhood
'While he waited for his friends, he thought about {*them/his childhood friends}.'

b. Parlant de llibres, ja he llegit el
talking of books(M,P) already I-have read the(M,S)
*(que recomanes).
that you recommend
'Talking about books, I have already read the one *(that you recommend).'

This indicates that the d-article needs a host: there must be some element in the NP it introduces that it can attach to. At this point, we could say: the d-article is a semipronominal determiner and what is special about it is that it cliticizes onto its NP sister (see Brucart and Gràcia 1984 for a similar claim). In other words, there has to be some lexical item in the NP sister of the d-article that the d-article can depend on phonologically.

However, not just anything in the NP serves this purpose. It is useful to distinguish between the *prenominal sector* of the NP—everything in the NP that precedes the head noun (or would precede it if it is absent)—and the *postnominal sector*—everything in the NP that follows the head noun (or would follow it if it is absent). The d-article can attach to the head noun, as in (9a) and (11a), and to anything in the prenominal sector: to a SpecA, whether the head noun is present, as in (10a), or not, as in (12b); to a prenominal adjective, or PreN-A, as in (14); to a modifier of a following adjective or noun, which can be an intensifier or an adverbial phrase, as in (15a−c), a PP headed by any preposition,[2] as in (15d−e), and even a clause, as in (15f).

(14) a. l' antiga fàbrica ('the old factory')
the old(F,S) factory(F,S)

b. el pobre gos ('the poor dog')
the poor(M,S) dog(M,S)

(15) a. la molt celebrada resposta del ministre
the very celebrated reply of-the minister

b. els increïblement intensos minuts de concert
the incredibly intense minutes of concert

c. l' encara president del govern
the as of yet president of-the government

[2] The importance of this class of examples was brought to my attention by Martin Forst, who culled many examples of d-articles followed by PPs from the web. The grammatical instances of PPs other than *de*-PPs were prenominal (see below for discussion of postnominal PPs).

d. el, des del mes passat, millor cuiner del món
the from the month past best cook of-the world

e. el {sens dubte / per descomptat} principal factor econòmic
the without doubt/ of course main factor economic

f. la, que jo sàpiga, única solució al problema[3]
the that I know only solution to-the problem

There does not seem to be anything in the prenominal sector that the d-article cannot attach to. Crucially, in all of the cases illustrated in (15), the word or phrase that immediately follows the d-article is a modifier of a following word or phrase, which may be an adjective, a noun, or an N′.

We consider next what can follow the d-article in the postnominal sector. It may seem that it may be followed by any phrase that can appear postnominally: a PP, as seen in (12a) and (13a), a clause (or CP), as shown in (12c) and (13b), or an AP of the postnominal variety. PPs and CPs such as those of (16) and (17) are always postnominal:

(16) a. la filla dels veïns
the daughter of-the neighbors
'the neighbors' daughter'

b. * la dels veïns filla

(17) a. el llibre que et vaig mostrar
the book that you I-PST show
'the book that I showed you'

b. * el que et vaig mostrar llibre

APs, as modifiers of a noun, can appear prenominally or postnominally. Some adjectives can appear in either position without a difference in meaning. However, some can only appear in one of the two positions or have a different meaning depending on the position they appear in. For example, *pobre* in prenominal position, as in (14b), does not mean 'lacking in financial resources', as it does in postnominal position, but signals an object of pity on the part of the speaker. The adjectives *francès*/*francesa* 'French' or *industrial* are postnominal adjectives (PostN-A) and can only appear postnominally:

(18) a. la muntanya francesa
the mountain(F,S) French(F,S)
'the French mountain'

b. * la francesa muntanya

(19) a. l' activitat industrial
the activity(F,S) industrial(S)
'industrial activity'

[3] When the modifier to which the d-article is adjacent is of some structural complexity, as in (15d,f), it is generally written between commas.

b. * la industrial activitat

The fact that the d-article can immediately precede these adjectives, as in (20), indicates that the d-article can be adjacent to a postnominal AP:

(20) a. M' agrada tant la muntanya italiana com **la francesa**.
me likes as much the mountain Italian as the French
'I like the Italian mountain as much as the French one.'

b. L' activitat comercial ha superat **la industrial**.
the activity comercial has exceeded the industrial
'Comercial activity has exceeded industrial activity.'

When we say that the d-article may be adjacent to a *postnominal* phrase (PP, CP, or AP), it does not mean that the phrase follows a noun. It only means that it is a phrase of the kind that follows a noun if a noun is present. The phrase is licensed by rule (4b): this rule licenses a maximal projection following a head X or X′. Since both nodes are optional by Economy of Expression, the rule may license only the nonhead constituent, which is the case in the boldfaced sequences in (20).

Whereas the d-article may be adjacent to any AP, there are significant restrictions on the postnominal PPs and CPs that may follow the d-article. A postnominal PP must be headed by the preposition *de*, if adjacent to the d-article (see Martí (2002: 1286); see also Lobeck (2006: 159) concerning the same restriction in Spanish).[4] Other semipronominal Ds can be followed by the same range of PPs as nouns, but not the d-article:

(21) No vull aquelles sabates, sinó
not I-want those(F,P) shoes(F,P), but
aquestes / *les amb cordons blancs.
these(F,P) / *the(F,P) with laces white

(22) a. L' abric / aquest / el de pell sense caputxa
the coat / this / the of leather without hood

[4] An anonymous reviewer indicated apparent counterexamples to this generalization:

(i) a. els sense sostre (lit. *the without ceiling*) 'the homeless'
b. el sense nom (lit. *the without name*) 'the nameless one'
c. els sense feina (lit. *the without job*) 'the jobless'

The sequences like *sense sostre* in these examples have to be analyzed as lexical units of category N. Unlike true headless NPs, the NPs in (i) are not interpreted as if there were a missing N whose semantics is filled by identity with some N in the discourse. Those sequences can be preceded by PreN-As like *mer* 'mere' or *antic* 'former', which is only possible with Ns: *un mer sense sostre* 'a mere homeless person'. They cannot be altered in any way: **els sense aquest nom* 'those without this name', **els sense cap sostre* 'those without any ceiling', etc. A few other PPs not headed by *de* are marginally acceptable after the d-article (e.g., *?els en vies de desenvolupament* (lit. *the in track of development*), where *en vies de desenvolupament* = 'developing'), but they have to be fixed phrases that are plausibly reanalyzed as single words, such as an adjective.

b. L' abric / aquest / *el sense caputxa de pell
the coat / this / *the without hood of leather
(Martí 2002: 1286)

The d-article can be immediately followed by a clause, only if the clause is introduced by *que* (see (12c)),[5] contrasting with other semipronominal Ds or with the d-article adjacent to a SpecA, which can be followed by the same range of relative clauses as nouns, as shown in (23):

(23) a. aquell/ *el de qui et vaig parlar
that(M,S)/ *the(M,S) of whom you I-PST speak
'that/*the one I spoke to you about'

b. aquests/ *els en els quals ens trobem
these(M,P)/ *the(M,P) in the which we find ourselves
'those/*the ones in which we find ourselves'

c. la *(primera) on vaig viure
the(F,S) *(first(F,S)) where I-PST live
'the *(first) one where I lived'

d. l'altre / *el per on passarem
the other / *the through where we will pass
'the *(other) one that we will pass through'

3.3 Special morphophonological properties of the d-article

There are some morphophonological properties that make the d-article unique among determiners:

(1) The d-article is stressless, unlike most other Ds.[6] Evidence for this is found in the schwa in these forms (*el* [əl], *la* [lə], *els* [əls], *les* [ləs]) and in that they form a phonological word with the following word, as shown by the homophony of single words with d-article-word combinations:

(24) a. [láβil]: *làbil* 'labile' / *l'hàbil* 'the skilled one'

b. [ləkáβə]: *lacava* 'lacquer (past)' / *la cava* 'the cellar'

c. [ləmén]: *lament* 'lament' / *la ment* 'the mind'

d. [lətén]: *latent* 'latent' / *l'atent* 'the attentive one'

(2) There is a syllabic/asyllabic alternation: the masculine singular form is *el* [əl] before a consonant and *l'* [l] before a vowel: *el noi* / **l' noi* 'the boy', *l'amic* / **el amic* 'the friend'. There is a similar alternation in the feminine singular form (*la/l'*); however, whereas this alternation could be accounted

[5] Some dialects and registers also allow a relative clause immediately following a d-article to begin with *qui*, signalling a human referent (e.g., *els qui vindran* 'those who will come'). I will ignore this option here for ease of exposition.

[6] The only other Ds that are stressless are the so-called weak possessives—*mon/ma* 'my', *ton/ta* 'your', *son/sa* 'his/her/its'—which are nonpronominal determiners and very restricted in their use.

for by a regular phonological rule of deletion of schwa adjacent to another vowel, no general rule can account for the *el/l'* alternation.

(3) The d-article patterns with verbal clitics morphophonologically. The syllabic/asyllabic alternation just noted is only found in so-called clitics, such as *em/m'* (1st p.sg.), *et/t'* (2nd p.sg.), *es/s'* (3rd refl.), *el/l'* (3rd p.masc.sg): *em veus* 'you see me', *m'ajudes* 'you help me', *el sentim* 'we hear him', *l'escoltem* 'we listen to him', etc. The d-article is homophonous with the 3rd person clitic, giving rise to VP/NP ambiguities such as: *la veu* 'the voice' / '(s)he sees her', *les cordes* 'the ropes' / 'you tie them', *el viu* 'the live one' / '(s)he lives it', etc.

3.4 Summary of properties of the d-article

The d-article:

- is initial in the noun phrase, like a D;
- may occur without the head noun, like a semipronominal D;
- must precede some element in its noun phrase, unlike a semipronominal D (i.e., it needs a host);
- may be adjacent to any element in the NP; however,
  - if adjacent to a postnominal PP, this PP must be headed by *de*;
  - if adjacent to a postnominal clause, it must be a *que*-clause;
- is morphophonologically unlike other Ds:
  - it is stressless: its vowel is schwa, if any, and it forms a phonological word with the following word;
  - it shows a syllabic/asyllabic alternation, only found in verbal clitics.

4 Explaining these properties

Simply assuming the d-article is a D, of the semipronominal kind, would explain some facts, but would leave the special properties of the d-article unexplained: the need for a host, the syntactic restrictions on the host, and the morphophonological properties. We will adopt the idea defended by Miller 1992 for French that the d-article is a bound morpheme, forming a word with the following word.[7] This is the treatment also given to verbal clitics in French by Miller 1992 and Miller and Sag 1997 and widely adopted for the other Romance languages. This analysis provides a way to explain those special properties of the d-article. If we assume that the d-article is the same type of morphological element as verbal clitics, it follows that they should all exhibit the same syllabic/asyllabic alternation, that they are all stressless and form a phonological word with the following word.

[7] I prefer to think of the d-article as a member of a compound, rather than an affix. The fact that it can take inflectional affixes itself (feminine and plural morphs: e.g., *les*: [[l][a][z]]) is inconsistent with its being an affix.

However, what is the syntactic category of the compound containing the d-article (the d-compound)? Is it a D, ignoring the category of the host, or is it the category of the host, obliterating the D? Both alternatives face serious problems or complications.

The analysis of the d-compound as a D explains the fact that it is initial in the DP, but does not explain: (a) why, if the host in the d-compound is a noun, it cannot be followed by a SpecA like *altre* 'other': **el-gat altre* 'the cat other'; (b) why, if the host is a PreN-A like *bon* 'good', it needs to be followed by a noun: **el-bon* 'the good', cf. *el bon amic* 'the good friend', **l'amic bon*; (c) why, if the host is a P, an NP must follow: *el de *(ferro)* 'the iron one'; (d) why, if the host is a C, a clause must follow: *el que *(surt)* 'the one that is coming out'.

On the other hand, the analysis that assigns the d-compound the category of the host does not have these problems, but cannot explain (without additional machinery) why the d-compound cannot occur with a D in the same DP and must be initial in the noun phrase. If the d-compound had the category of the host (whether N, PreN-A, PostN-A, etc.), it should be possible for it to occur with a D (**aquell l'amic* 'that the friend', **cada l'hora* 'each the hour', etc.) and for it to be preceded by a word that can precede the host (**bon l'amic* 'good the friend', **altre el gat* 'other the cat', etc.). Miller 1992 adopts a version of this analysis and enriches the theory with an EDGE feature and with principles, including Linear Precedence constraints that linearize constituents with the appropriate EDGE feature such as the d-compound to the left of other constituents (see also Tseng 2003).

Both analyses fail to recognize that the categories of both components of the d-compound are needed to account for its distribution. If the d-compound has the category of the d-article (D), it follows that it cannot co-occur with another D (only one D per DP) and cannot be preceded by a SpecA or any other category in the NP (the D precedes its sister NP). If it has the category of the second member of the compound, it has to appear in the position reserved for this category: If this category is N, it cannot be followed by a SpecA like *altre* (SpecA precedes N); if it is a PreN-A like *bon*, it must be followed by a noun; if it is the preposition *de*, it must be followed by a noun phrase; if it is the complementizer *que*, it must be followed by a tensed clause. The conclusion is that the d-compound has the categories of both components of the compound.

This is possible only if we allow two different X^0 categories in the c-structure to be instantiated by the same word. The theory of Lexical Sharing (Wescoat 2002, 2005, 2007) allows this possibility. All we need is to allow a single word to fill (or map onto) two or more X^0 nodes in the c-structure. If we assume that words are not part of the c-structure, but are part of a separate level—the word string (or lexical-structure, in Wescoat's terms)—where words are linearly ordered, there is a mapping, or correspondence, between terminal nodes (X^0 nodes) in the c-structure and words in the word string.

This mapping, by which all terminal nodes are instantiated by words and all words instantiate terminal nodes, is constrained by Wescoat's Order Preservation Axiom (Wescoat 2007: 444−445):

(25) Order Preservation Axiom:

For all c-structure nodes X and Y, if the word that instantiates X precedes the word that instantiates Y, then X precedes Y.

By this constraint, words with multiple categories can only appear in structures where the various linked categories are adjacent.

The lexical entry of all words includes a list of categories, which, in the simplest case, consists of only one member. The order of the categories in this list matches the order of the X^0 nodes in the c-structure that the word maps onto. The d-article, as a lexical item below the word level, carries information about morphology, c-structure, and f-structure.

(26) Lexical information of the d-article:

Morphology: the d-article is a bound stem and attaches to a word to its right to form a word (the d-compound).

C-structure: the list of categories of the d-compound is the concatenation of the category D (for the d-article) and the list of categories of the 2nd member of the compound.

F-structure: the f-structure of the 2nd member is included in the f-structure of the first member (where f-structure A includes f-structure B iff B = A or is contained in A).

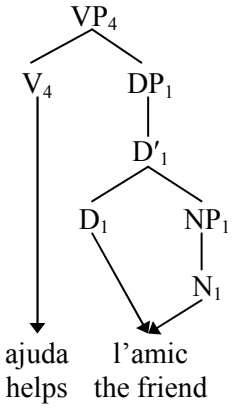
Restriction: if the category of the 2nd member is dominated by a PP or CP adjunct of the d-article, the 2nd member is *de* or *que* respectively.

The morphological status of the d-article allows us to explain the article's properties in 3.3. The c-structure information in (26) provides the explanation for the distribution of the d-compound: it depends on the category of both members of the compound. The f-structure information captures the observation that the d-article needs a host that belongs in the same DP. The restriction in (26) accounts for the observation that, if the d-article is immediately followed by a PP adjunct of the noun phrase, the head of this PP is the preposition *de*, and if the d-article is immediately followed by a CP adjunct of the noun phrase, this CP begins with the complementizer *que*.

As an illustration of this theory, consider the d-compound *l'amic*. (27a) shows the vocabulary entry of this compound, with a list of categories consisting of a D and an N and the f-structures corresponding to each category, with indices signaling the correspondence. (27b) shows the c-structure and f-structure of the sentence *ajuda l'amic*, where arrows, rather than lines, represent the correspondence between c-structure and the word string.

(27) a. l'amic: $\langle D_1, N_2 \rangle$ $\begin{bmatrix} \text{Def} & + \\ \text{Gend} & \text{m} \\ \text{Numb} & \text{sg} \end{bmatrix}_1 \begin{bmatrix} \text{Pred} & \text{'friend'} \\ \text{Gend} & \text{m} \\ \text{Numb} & \text{sg} \end{bmatrix}_2$

b.



$\begin{bmatrix} \text{Pred} & \text{'friend'} \\ \text{Def} & + \\ \text{Gend} & \text{m} \\ \text{Numb} & \text{sg} \end{bmatrix}_1$

Note: $f_1 = f_2$

Notice that these structures satisfy all the constraints proposed, specifically: the d-article is the first part of a word that has the two categories of its component elements in a list, and the f-structure of the second member is included in (here, is equal to) the f-structure of the first member. The restriction in (26) applies vacuously as there is no PP or CP adjunct involved.

The morphological and f-structure requirements in (26) ensure that the d-article is followed by some element, explaining contrasts such as those in (13), and that this following element is part of the noun phrase introduced by the d-article, ruling out examples such as the following:

(28) a. * Pensava en [$_{DP}$ els] sovint.
he-thought in the(M,P) often

b. * Consideren [$_{DP}$ la] una bona amiga.
they-consider the(F,S) a(F,S) good(F,S) friend(F,S)

The restriction in (26) explains the contrasts illustrated in (21) and (22): if the second member of the d-compound heads a PP adjunct of the d-article, it must be the preposition *de*, accounting for the contrast between *les de ferro* (*the of iron*) and **les amb cordons* (*the with laces*). This condition does not prevent the d-article from attaching to a preposition other than *de*, provided this P heads a PP that is not an adjunct of the d-article. This is what we see in (15d,e), where the P adjacent to the d-article heads an adjunct of the following adjective or noun. That restriction also explains the contrasts in (23): in these examples, the second member of the compound is part of a CP adjunct of the d-article, but it is not *que*, and therefore they are ruled out.[8]

[8] The fact that the restriction in (26) is a separate statement from the other statements about the d-article is not a negative feature of the analysis, but a positive one, as there are languages that have a d-article as a lexical sharing bound morpheme, but lack this restriction or have a different one. Spanish, from the 15th to the 17th centuries, allowed constructions very much

Whereas a d-article can be followed by a restrictive relative clause beginning with *que*, it cannot be followed by a nonrestrictive relative clause, even if it begins with *que*:

(29) M'agrada aquell,/ *el, que em vas ensenyar.
I-like that(M,S)/ *the(M,S) that me you-PST show
'I like that/*the, which you showed me.'

A nonrestrictive relative clause is distinguished from the word that immediately precedes it by an intonational break. The demonstrative *aquest* in (29) can be followed by a nonrestrictive relative clause and an intonational break separates *aquest* from *que*, which begins the relative clause. The d-article has to form a compound with the following word, but if this is the first element of a nonrestrictive relative clause, there would have to be an intonational break inside a word. An intonational break can separate a word from another word, but not a word part from another word part.[9] This is an additional argument for the claim that the d-article forms a word with a following word.

5 Weak prepositions and multiple lexical sharing

This section argues that weak prepositions—*per, de, en, a, amb*—are also lexical sharing bound morphemes. After presenting evidence for this claim, we will see that the preposition *de* and the d-article can alternate in indefinitely long sequences. The implications of these structures will be considered for choosing among alternative theories of the d-article.

Evidence that weak prepositions in Catalan are bound morphemes includes the following observations:

(1) Weak prepositions need a host. Other prepositions can function without a following NP, i.e, they can be the sole element in the PP. This is the case of *sense* 'without', *sota* 'under', *contra* 'against', among others: *Vols el te* ***amb*** *sucre o* ***sense****?* 'Do you want your tea with sugar or without?'; cf. **Vols el te* ***sense*** *sucre o* ***amb****?* 'Do you want your tea without sugar or with?'.

(2) The choice between the prepositions *a* and *en* is partly conditioned by phonology: e.g., *he caigut {****en*** *un forat /* ***al*** *forat /* ***en*** *aquell forat /* ***a*** *dins del forat}* 'I have fallen {in a hole / in the hole / in that hole / inside the hole}'. The rule governing this alternation can be stated as: use *en* if followed by a

like those with the d-article in (23). A search in CORDE—the historical corpus of Spanish texts—reveals many examples from that period such as *la por quien muero* (lit. the for whom I die) 'the one I die for'. (The modern Spanish corpus CRAE, on the other hand, has no example like these.) To deal with different grammars, the restriction in (26) can be redefined or eliminated altogether, without affecting the remaining statements about the d-article.

[9] The intonational break that occurs at the boundary of an appositional or restrictive clause is completely different from the intonation we find in expletive insertion cases (e.g., *fan-fucking-tastic*) or in coordination of prefixes in English and German (e.g., *pro- and anti-abortion activists*; *ein- und aussteign* 'to get on and off' from Forst, King, and Laczko 2010).

vowel; otherwise, use *a*. By the Principle of Phonology-Free Syntax (Zwicky 1969; Zwicky and Pullum 1986; Miller, Pullum, and Zwicky 1997), this cannot be an instance of a phonologically conditioned choice of lexical items. But, according to this principle, it can be an instance of phonologically conditioned allomorphy or prefix suppletion.

(3) The prepositions *de*, *per*, and *a* alternate with contracted forms involving the d-article: *del*, *dels*; *pel*, *pels*; *al*, *als*. The contracted forms are used when the d-article morph /l/ is not in a syllable onset; otherwise, the contracted forms are used: *per la noia* 'by the girl', *per l'amic* 'by the friend', *pel noi* 'by the boy', *pels amics* 'by the friends'.[10]

(4) Certain phonological properties of these prepositions also argue for their status as bound morphemes:

a) They are stressless: their vowel is schwa: [pər], [də], [ən], [ə], [əm].
b) *Amb* 'with' is not pronounced [əm] in all contexts (which would be expected if it were a word, given the rule of word-final stop deletion after a homorganic consonant), but is pronounced [əmb] when followed by a vowel: [əm lə má] 'with the hand', [əmb əl dít] 'with the finger'.
c) *Per* 'by, for' triggers the deletion of the following vowel if it is part of a demonstrative: *per (a)quest camí* 'by this path', *per (a)quella raó* 'for that reason', *per (ai)xò* 'because of this', *per (a)llà* 'that way' (lit. 'by there'), etc. This vowel does not drop in other contexts. And *per* does not cause the dropping of a following schwa in other cases. Again, it is a case of allomorphy or a morphologically conditioned phonological rule.

These facts support the claim that weak prepositions are bound morphemes.[11] The reasons for assuming that there is a P in the c-structure corresponding to the word they are part of include the following: (a) a phrase beginning with a weak preposition occupies the same positions as a phrase beginning with an independent word of category P; (b) the c-structure rule stating that a P is initial in the PP applies both to independent words of category P and to words containing a weak preposition, requiring them to be initial in their PP; (c) a weak preposition can scope over a coordinate NP, just as an independent P

[10] Lexical sharing arises not only in the contracted forms (as assumed in Wescoat (2007)), but in all occurrences of weak Ps. The choice between the contracted and uncontracted forms does not depend on syntactic conditions, but on phonological conditions. Assuming phonology-free syntax, the alternation cannot be a choice between different words. If weak Ps are always lexical sharing prefixes, the alternation is phonologically conditioned allomorphy. (Similar facts occur in French: *à l'homme* 'to the man' / *au bon homme* 'to the good man'.)

[11] Wescoat (personal communication) notes that the bound morpheme status of weak prepositions can be traced back to Latin, at least. The Latin so-called clitic *-que* 'and' attaches to the end of the first word of a conjunct, but monosyllabic prepositions frequently do not count for the purpose of *-que* placement (see Gildersleeve and Lodge 1895: 300), as the following examples (from Carlson 1983: 81) illustrate: *in forōque* 'and in the forum', *ob eāsque rēs* 'and because of these achievements'. The reason prepositions such as *in* and *ob* do not count as the first word of the conjunct is that they are not separate words, but part of a word together with the following stem.

can: *del pare i la mare* 'of-the father and the mother', *amb la flauta i el violí* 'with the flute and the violin', etc.

So, weak prepositions are bound morphemes that add the category P as the first member to the list of categories of the word they attach to. Since a preposition can be followed by a d-article and a d-article has to be followed by another word in its noun phrase, and the first two elements are bound morphemes of the lexical sharing kind, we often find words that are sequences of three categories, i.e., instances of multiple lexical sharing: *amb el dit* 'with the finger'⟨P, D, N⟩. Given that a d-article can also be followed by a weak P, we have recursivity (a d-article can attach to a following P *de* and a weak P can attach to a following d-article, and this can be repeated indefinitely):

(30)	a.	les	de	ferro			⟨D, P, N⟩
		the(F,P)	of	iron			
	b.	pels	de	la	primera		⟨P, D, P, D, SpecA⟩
		by-the(M,P)	of	the(F,S)	first(F,S)		
	c.	els	de	la	de	París	⟨D, P, D, P, N⟩
		the(M,P)	of	the(F,S)	of	Paris	
	d.	amb	les	del	de	la	becària
		with	the(F,P)	of-the(M,S)	of	the(F,S)	assistant(F,S)
							⟨P, D, P, D, P, D, N⟩

The limit in generating words such as these is human processing capabilities. Every DP consisting only of a d-article and a PP adjunct is interpreted with a semantic restrictor supplied by identity with a noun present in the discourse of the same gender as the nounless DP. If (30d) appears in a discourse where *dades* 'data(F,P)' and *treball* 'paper(M,S)' are mentioned, it can be interpreted like (31), where the crossed out words can be assumed to have a representation in the semantics but not in the syntax:

(31)	amb	les	~~dades~~	del	~~treball~~	de la	becària
	with	the(F,P)	data(F,P)	of-the(M,S)	paper(M,S)	of the(F,S)	asst.(F,S)

Miller (1992: 291) notes that an example like (32a), which he represents as (32b), is ungrammatical in French:

(32)	a.	* le	de	Paris	est	arrivé	hier
		the(M,P)	of	Paris	is	arrived	yesterday

b. * [$_{NP}$ le ε de Paris] est arrivé hier

Miller crucially assumes the presence of a null noun (ε) in all apparently headless NPs. The ungrammaticality of (32a) follows from this and from the natural assumption that an affix (such as the d-article) cannot be realized on a null stem. But these assumptions incorrectly predict the equivalent form in Catalan to be ungrammatical:

(33) el de París va arribar ahir
the(M,P) of Paris 3sg-PST arrive yesterday
'the one from Paris arrived yesterday'

Whereas Catalan and Spanish allow the sequence d-article+*de*, (e.g., (30)), French and Italian do not (e.g., (32)). By assuming the structure in (32b), with a null noun, Miller predicts the illformedness of this sequence in all four languages. In the present approach, where there are no null nouns, all we need to assume to explain the contrast among these languages are different constraints on the host of the d-article:

- Spanish and Catalan allow the d-article to form a compound with a following preposition *de*.
- Italian and French do not allow the d-article to attach to any preposition.[12]

Such restrictions in word formation are to be expected and are expected to vary from language to language.

The presence of the null noun in headless NPs in Miller 1992 makes it hard to explain the cross-linguistic variation noted. It would be very strange to assume that some languages allow inflection on a null stem, whereas others do not. Even if we made this assumption, we would still need to explain that there are restrictions on what can follow the null stem when it is inflected with the d-article: in Spanish and Catalan, it must be followed by some element in the same NP; if followed by a P that heads an adjunct of the d-article, it must be *de*. So, we would need to make two unnatural assumptions: that the null stem can have inflections in some languages and that the inflections impose restrictions on what follows the null stem. This supports the conclusion that both the d-article and weak prepositions should be treated as lexical sharing bound morphemes, rather than as inflections.

6 Listed compounds

D-compounds are for the most part unlisted: they are formed on-line, satisfying the constraints in the lexical entry of the d-article in (26), and are not listed in the lexicon or in a repository of words. However, since d-compounds are words, the possibility exists that certain d-compounds are listed. Certain words can only function as the second component of a d-compound. The clearest case is *qual*, which can only be used immediately following the d-article forming the "compound relative": *el qual*, *la qual*, *els quals*, *les quals*. Other combinations involving *qual* are impossible: **un qual*, **aquell qual*, **la primera qual opció* (cf. *la qual primera opció* 'which first option'), **amb qual*, etc. All we need to assume is that *qual* is a SpecA (like

[12] This constraint can probably be subsumed under a more general statement that would also exclude other combinations of the Italian and French d-article with postnominal elements, such as with a relative clause (e.g., French **le qui parle* (lit. *the who speaks*) vs. *celui qui parle* 'he who speaks').

altre ‘other’, *primer* ‘first’, etc.) that is lexically listed as forming a compound with the d-article.[13]

Another relevant case is the group of so-called strong possessives: *meu/meva* (1st pers. sg.), *teu/teva* (2nd pers. sg.), *seu/seva* (3rd pers.), etc. They are used either as SpecA or as PostN-A. (The use as postnominal adjectives, PostN-A, is found in noun phrases such as *amic meu* ‘my friend’, *molts coneguts teus* ‘many acquaintances of yours’, and (34c).) However, for many speakers, whose judgments are reported here, they can only be used as SpecA if immediately preceded by the d-article:

(34)	a.		el	meu	amic	‘my friend’
			the(M,S)	my(M,S)	friend(M,S)	
	b.	*	aquell	meu	amic	‘that friend of mine’
			that(M,S)	my(M,S)	friend(M,S)	
	c.		aquell	amic	meu	‘that friend of mine’
			that(M,S)	friend(M,S)	my(M,S)	
	d.		els	meus	dos gossos	‘my two dogs’
			the(M,P)	my(M,P)	two dogs(M,P)	
	e.	*	els	dos meus	gossos	‘my two dogs’
			the(M,P)	two my(M,P)	dogs(M,P)	

Strong possessives have a dual classification as SpecA and PostN-A. But, as SpecA, they are the second member of a d-compound (i.e., must form a compound with the d-article). The claim that the d-article forms a compound with the following word affords a simple and natural way to account for the marked distribution of certain words: relative *qual* and strong possessives when used as SpecA are listed as the second member of a d-compound.

7 Conclusions

In this paper we have seen compelling evidence for the claims: (a) that the d-article in Catalan (as in closely related languages) is a bound morpheme that attaches to a word to its right to form a word (the d-compound); (b) that the d-compound instantiates two or more terminal nodes in the c-structure corresponding to the categories contributed by its component elements: an instance of lexical sharing; (c) that weak prepositions in Catalan are also bound morphemes of the lexical sharing type; and (d) that the number of categories that a word can instantiate can be greater than two and has no upper bound.

Lexical sharing, i.e. the hypothesis that a word maps onto more than one terminal node in the c-structure, proves to be the correct way of accounting for what has sometimes been known as phrasal affixes (bound morphemes whose distribution depends on phrasal information). Alternative

[13] Notice that the cognate form in French is spelled as a single word together with the d-article: *lequel*, *laquelle*, *duquel*, *auxquelles*, *desquels*, etc.

approaches that do not allow lexical sharing account for the same (or similar) facts by complicating the theory considerably or, simply, cannot account for the relevant facts (as argued about Miller's (1992) approach). Many phenomena that seem to straddle the morphology-syntax boundary would benefit from an analysis that incorporates lexical sharing.

References

Abney, Steven. 1987. The English noun phrase in its sentential aspect. Ph.D. dissertation, MIT.

Bresnan, Joan. 2001. *Lexical-functional syntax*. Oxford: Blackwell.

Brucart, José and Lluïsa Gràcia. 1984. I sintagmi nominali senza testa. *Rivista di Grammatica Generativa* 11: 3−32.

Carlson, Greg N. 1983. Marking constituents. Frank Heny and Barry Richards (eds.), *Linguistic categories: Auxiliaries and related puzzles*, 69−98. Dordrecht: Reidel.

Forst, Martin, Tracy Holloway King, and Tibor Laczko. 2010. Particle verbs in computational LFGs: Issues from English, German, and Hungarian. *Proceedings of LFG'10*. CSLI Publications.

Gildersleeve, Basil L. and Gonzalez Lodge. 1895. *Gildersleeve's Latin Grammar*. London: Macmillan.

Lobeck, Anne. 2006. Ellipsis in DP. Martin Everaert and Henk van Riemsdijk (eds.), *The Blackwell companion to syntax*, vol. 2, 145–173. Malden, MA: Blackwell.

Martí i Girbau, Núria. 2002. El SN: els noms. Joan Solà et al. (eds.), *Gramàtica del català contemporani*, vol. 2, 1281–1339. Barcelona: Empúries.

Miller, Philip H. 1992. *Clitics and constituents in phrase structure grammar*. New York: Garland.

Miller, Philip H, Geoffrey K. Pullum, and Arnold M. Zwicky. 1997. The Principle of Phonology-Free Syntax: four apparent counterexamples in French. *Journal of Linguistics* 33:67–90.

Miller, Philip H. and Ivan A. Sag. 1997. French clitic movement without clitics or movement. *Natural Language and Linguistic Theory* 15:573–639.

Tseng, Jesse. 2003. Phrasal affixes and French morphosyntax. *Proceedings of Formal Grammar 2003*, 177–188. Technische Universität Wien.

Wescoat, Michael T. 2002. On Lexical Sharing. Ph.D. dissertation, Stanford Univ.

Wescoat, Michael T. 2005. English nonsyllabic auxiliary contractions: An analysis in LFG with lexical sharing. *Proceedings of LFG'05*, 468–486. CSLI Publications.

Wescoat, Michael T. 2007. Preposition-determiner contractions: An analysis in Optimality-Theoretic Lexical-Functional Grammar with lexical sharing. *Proceedings of LFG'07*, 439–459. CSLI Publications.

Zwicky, Arnold M. 1969. Phonological constraints in syntactic descriptions. *Papers in Linguistics* 1:411–453.

Zwicky, Arnold M. and Geoffrey K. Pullum.1986. The Principle of Phonology-Free Syntax: Introductory remarks. *Working Papers in Linguistics* 32:63–91. Columbus, OH: the Ohio State University.

OBJECT SHIFT
OR OBJECT PLACEMENT IN GENERAL?

Maia Andréasson
University of Gothenburg

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

This paper addresses the phenomenon of pronominal object shift in Danish and Swedish and to what extent it is relevant to analyse object shift as an isolated choice between two positions. A corpus investigation reveals that there are distinct information dynamical strategies for the placement of pronominal objects in these languages. These strategies have not previously been discussed in connection with object shift and involve more positions than the IN SITU position and the SHIFTED position.

On the basis of this investigation, I argue that it is necessary to explore all positions for pronominal objects in order to analyse the underlying causes of objects appearing in the position traditionally referred to as SHIFTED.

1 Pronominal object shift

Pronominal object shift is – in brief – the possibility for pronominal objects to precede the sentence adverbial in Scandinavian languages, henceforth SHIFTED, whereas full NP objects must appear following the sentence adverbial, henceforth IN SITU.

In previous studies of object shift only these two positions are investigated, see e.g.: Holmberg (1986, 1999), Hellan & Platzack (1995), Josefsson (1992, 2003), Pedersen (1993) Sells (2001), Svenonius (2002), Vikner (1994, 1997, 2005), Andréasson (2008) and several others. In this paper, I show data that lead to the conclusion that all object positions must be investigated to fully analyse why pronominal objects appear preceding sentence adverbials, i.e. as SHIFTED.

The outline of the paper is as follows. In this this section a brief presentation is made of the notion of object shift and the role of the referent's accessibility. In section 2, I outline the problems with the method of previous studies, to investigate only two of all possible object position, and I present material, methods and results from the present investigation. Section 3 is a brief summary of the findings.

1.1 The nature of the object and information dynamics

Ever since Holmberg (1986) it has been well known that object shift has structural as well as information dynamical delimitations. One of the structural delimitations concerns the nature of the object. Pronominal objects, but not full NP objects, usually precede sentence adverbials in mainland Scandinavian languages. For an example of this, see (1), where the pronoun *den*, 'it', but not the full NP *boken*, 'the book', is licensed preceding the negation *inte*, 'not'.

† I thank the audience of LFG10 at Carleton University, Ottawa, and my colleagues at the University of Gothenburg, Sweden, for helpful comments. My research on object shift is funded by the Swedish Research Council, Vetenskapsrådet.

(1) Agnes köpte {den inte /*boken inte}. [SW]
Agnes bought {it not /*the book not}.
'Agnes didn't buy the book/it.'

Another structural delimitation that Holmberg (1986) mentions is that object shift is not licensed in clauses where the lexical verb is not in V2, see (2a).

(2) a. *Agnes hade den inte köpt. [SW]
Agnes had it not bought
'Agnes had not bought it.'
b. *I know that Agnes hade den inte köpt. [SW]
I know that Agnes had it not bought
'I know that Agnes had not bought it.'

The information dynamical delimitation on object shift presented in previous studies is generally that only "non-stressed" pronominal objects shift. Pronouns with contrast interpretation and – in speech – contrastive stress, here marked with double apostrophes, must appear after sentence adverbials, see (3). In Swedish, it is not ungrammatical for an unstressed, non-contrasted pronoun to appear IN SITU, but in standard Danish unstressed pronouns following sentence adverbials is considered ungrammatical, see (4) (cf. Pedersen 1993). Unstressed pronouns are here marked with a subscribed zero. Marking for prosody will only be made when it is relevant for the analysis.

(3) Agnes$_i$ såg "David men han såg inte "henne$_i$. [SW]
Agnes$_i$ så "David men han så ikke "hende$_i$. [DA]
Agnes saw "David but he saw not "her
'Agnes saw David, but he didn't see her.'

(4) Agnes såg boken, men hon köpte {^{0}den inte /inte ^{0}den}. [SW]
Agnes så bogen, men hun købte {^{0}den ikke /*ikke ^{0}den}. [DA]
Agnes saw the book but she bought {^{0}it not/ *not ^{0}it}
'Agnes saw the book, but she didn't buy it.'

These information dynamical restrictions have given rise to the analysis of object shift where an unstressed pronoun "escapes" from a FOCUS domain. "[N]onfocused arguments have to move out of VP, the focus domain, into the presupposition domain, i.e. the space between C and VP [...]" (Holmberg 1999:23).

In this paper, I primarily discuss pronominal object shift and the data I present is from Swedish and Danish. There is, however, variation across Scandinavia, when it comes to the nature of the objects that shift. I will here only briefly mention Icelandic and the Swedish variety Övdalian. In Icelandic definite NPs too may appear preceding negation when the lexical verb is in V2 (the examples in (5) and (6) are from Vikner 2005).

(5) Af hverju las Pétur *aldrei þessa bók*? [ICELANDIC]
what read-PST Peter never this book
'Why did Peter never read this book.'

(6) Af hverju las Pétur *þessa bók aldrei*? [ICELANDIC]
why read-PST Peter this book never
'Why did Peter never read this book?'

In Övdalian (the dialect of Älvdalen, Sweden), on the other hand, no objects shift (the examples in (7) are from from Garbacz 2010).

(7) a. An såg int mig. [ÖVDALIAN]
he saw not me
'He didn't see me.'
b. *An såg mig it. [ÖVDALIAN]
he saw me not
'He didn't see me.'

The different possibilities throughout Scandinavia referred to here may be illustrated as a scale from no object shift in Övdalian to definite NP object shift in Icelandic, see (8). I will briefly return to this scale in the next section.

	Övdalian ⟶	**Swedish** ⟶	**Danish** ⟶	**Icelandic**
(8)	*no OS*	*pronominal OS* *(no OS)*	*pronominal OS*	*definite NP OS*

1.2 The type of antecedent and accessibility

In Andréasson (2008, 2009), I address the fact that the analyses presented in previous work on object shift seem to deal with only pronominal objects with NP antecedents, henceforth pro_{NP}, such as *den* in example (1) above, and overlook object pronouns with sentence antecedents, henceforth det_{S}. In these papers, I show that there is a significant difference in distribution for object pro_{NP} and det_{S},[1] and that this distributional difference may be linked to a difference in accessibility of the object referents. In the following, I will summarise the differences in distribution for pro_{NP} and det_{S} for the SHIFTED and IN SITU positions investigated in Andreasson (2008, 2009).

As mentioned above it is not ungrammatical for a non-contrasted object pronoun to appear IN SITU in Swedish. Nevertheless, non-contrasted pro_{NP} IN SITU appear to be very rare in written Swedish (Andréasson 2008). The numbers in table 1 show that Swedish and Danish are very similar when it comes to the distribution of pro_{NP} in relation to negation. All of the Danish pro_{NP} in the IN SITU position where overtly contrasted, while about half of the Swedish pro_{NP} IN SITU were not (Andréasson 2008).

[1] Preliminary investigations show that object pronouns with VP antecedents, det_{VP}, show a similar distrubutional pattern as det_{S} but here, as well as in Andréasson (2008, 2009) I deal primarily with det_{S}.

Swedish	shifted	in situ
honom/henne	115 (91%)	12 (9%)
Danish	shifted	in situ
ham/hende	325 (93%)	24 (7%)

TABLE 1: *Positions for pronominal objects with NP antecedents (*pro$_{\text{NP}}$*) 'him'/'her' in relation to negation (Andréasson 2008)*

Det$_{\text{S}}$ without contrast interpretation, as in (9), on the other hand appear IN SITU to a greater extent than the *pro*$_{\text{NP}}$ both in Swedish and – more surprisingly – in Danish, where non contrasted objects IN SITU are considered ungrammatical.

(9) – Agnes köpte boken. Förstod du inte det? [SW]
– Agnes købte bogen. Førstod du ikke det? [DA]
Agnes bought the book understood you not that
'Agnes bought the book. Didn't you understand that?'

In non-declarative sentences, Andréasson (2008) shows that as many as 76% of the object pronouns with sentence or VP antecedents in Swedish and 81% in Danish appear IN SITU. In declarative sentences, only Swedish shows a difference in distribution with relation to the *pro*$_{\text{NP}}$; 25% of the object pronouns with sentence and VP antecedents appear IN SITU. For Swedish, a difference in distribution between sentences with factive and non-factive matrix verbs was also found. In sentences with a non-factive matrix verb, as many as 72% of the *det*$_{\text{S}}$ are IN SITU in Swedish. In sentences with a factive matrix verb, only 9% appear IN SITU.

In Danish only 6% of the object pronouns with sentence and VP antecedents appear IN SITU in declarative sentences, and all the Danish examples with *det*$_{\text{S}}$ IN SITU in Andréasson (2008) are sentences where the matrix verb is non-factive. Thus, for Danish, the numbers for declarative sentences seem to match those for *pro*$_{\text{NP}}$, in table 1 above. As we will see in section 2, this is not entirely true.

In Andréasson (2008, 2009), I present an analysis where it is not the type of antecedent per se that affects the difference in distribution between *det*$_{\text{S}}$ and *pro*$_{\text{NP}}$. Rather it is the cognitive status, or accessibility, that the referents of the object pronouns are assumed to have in the mind of the listener that lies behind the distribution. The analysis that elements with different levels of accessibility are placed in different syntactic positions relates to observations for English on choices between referring expressions (Gundel, Borthen and Fretheim 1999; Gundel, Hegarty and Borthen 2003).

The accessibility hierarchy of Gundel, Hedberg and Zacharski (1993), in Figure 1, is a model of how different levels of cognitive status are linked to nominal expressions in English. The more accessible a speaker assumes a referent to be in the listener's mind, the further to the left of this scale she may go when she chooses a suitable nominal expression.

in focus	activated	familiar	identifiable	referential	type identifiable
it	*this/that/ this N*	*that N*	*the N*	*indefinite this N*	*a N*

FIGURE 1: *Givenness hierarchy, Gundel, Hedberg & Zacharski (1993)*

Gundel et al. (2003) show that when the antecedent is an NP, it is legitimate to use the pronoun *it*, the expression highest on the accessibility scale. When the antecedent instead is a non-factive clause or VP, reference must be made with the pronoun *that/this*, an expression further down on the scale (Gundel et al. 2003; Hegarty 2003). Example (10) below (from Gundel et al. 1999) illustrates that the pronoun *it* may refer to an NP, since this pronoun is not felicitous referring to the situation expressed by the entire clause *There was a snake on my desk*.

(10) a. There was [a snake]$_i$ on my desk. It$_i$ scared me.
b. [There was a snake on my desk.]$_j$ That$_j$ scared me.

In English the two highest levels of accessibility are linked to distinct word forms, the pronouns *it* and *this/that*. In Swedish,[2] this difference in accessibility is not expressed by different lexical items, but by a slight difference in pronunciation. When the antecedent is an NP, (11a), an unstressed *det*, here marked with a zero, and when the antecedent is a sentence, (11b), a slightly more stressed *det*. Please note that this slight difference in stress does not signal contrast.

(11) a. – Agnes har tydligen köpt [ett nytt dataspel]$_i$. [SW]
Agnes has obviously bought a new computer game
'Agnes obviously bought a new computer game.'
– Ja, jag har faktiskt sett ^{0}det$_i$.
yes I have actually seen it
'Yes, I actually saw it.'
b. [– Agnes har tydligen köpt ett nytt dataspel]$_i$. [SW]
Agnes has obviously bought a new computer game
'Agnes obviously bought a new computer game.'
– Ja, jag har faktiskt sett 'det.
yes I have actually seen that
'Yes, I actually noticed that.'

[2]Gundel, Borthen and Fretheim (1999) note that the choice between *it* and *that* in English correspond to a similar choice in Norwegian, between one deaccented and another, slightly more accented *det*. Discussions with native Danes give reason to believe that similar differences in stress are relevant for Danish as well.

As argued in Andréasson (2008, 2009) the distributional and phonological observations for det_S can not be explained as a difference in contrast, i.e. pronouns escaping from a focus domain. Instead the difference in stress signals two distinct levels of accessibility, and only pronouns where the antecedent has the highest level of accessibility are completely unstressed and hence licensed in the SHIFTED position. This matches the results for the pro_{NP} in table 1 and it also matches the Swedish results for det_S with factive and non-factive verbs in Andréasson (2008) summarised above.

It is well known that factive verbs trigger the presupposition that their complements have a truth value. In a sentence like (12a) below, the factive matrix verb *understand* triggers the presupposition that the speaker considers the statement *she saw me* to be true, and this presupposition remains also when the matrix verb is negated, see (12b).

(12) a. I understood that she saw me.
PRESUPPOSITION: She saw me.
b. I didn't understand that she saw me.
PRESUPPOSITION: She saw me.

Non-factive verbs on the other hand do not trigger any presupposition about the truth value of their complements. In (13) below, it is equally possible to follow up the statement *I thought that she saw me* with an affirming *...and she did* as with a negating *...but she didn't*.

(13) I thought she saw me,
a. ... and she did.
b. ... but she didn't.

When the matrix verb is factive, the truth value of the proposition represented by the subordinate clause is presupposed and assumed to be known by the listener. In this case it is felicitous to use a linguistic form that signals the highest level of accessibility – in English *it*, and in Swedish and Danish an unstressed pronoun, *det*, preceding a negation. However, if the matrix verb is non-factive, the proposition of the subordinate clause does not have a presupposed truth value, and it is not felicitous to use the linguistic expression corresponding to the highest level of accessibility or realise the pronoun in a position where only accessible elements are licensed.[3]

Andréasson (2008, 2009) assumes that the syntactic position preceding the negation, SHIFTED, is linked to the highest level of accessibility both in Swedish and in Danish.

The situation in the varieties of Scandinavian mentioned in section 1.1 may be illustrated as in Figure 2, below, showing where the negation appears in relation

[3]There are also other means of promoting a referent to the highest level of accessibility (Gundel et al. 1999). These have relevance for Swedish and Danish, and I refer to Andréasson (2008) for a presentation of them.

to different levels of cognitive status for the different languages or varieties. The different ACTVN features refer to the LFG-analysis of Andréasson (2008) and are a small development of the activation feature originally put forth by O'Connor (2006), where 0 corresponds to a referent that the speaker assumes to be cognitively fully accessible for the speaker, and with no need for further activation, i.e. *in focus* on the accessibility hierarchy.[4]

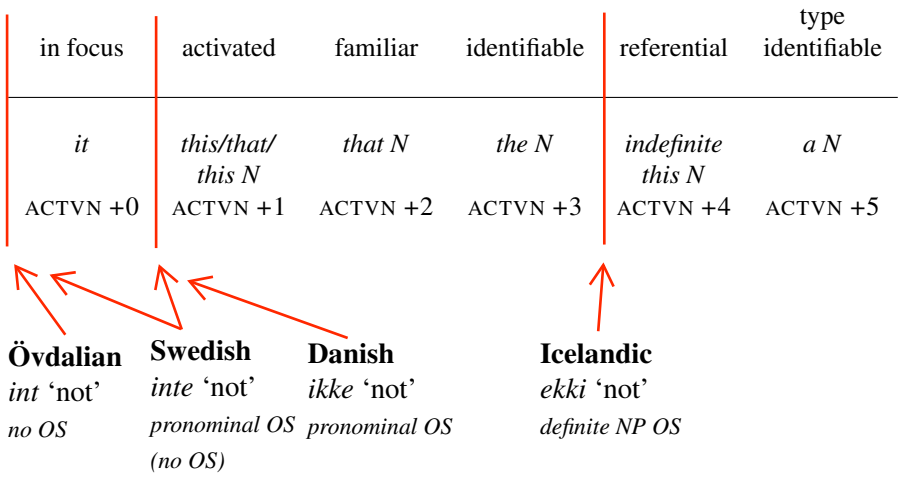


FIGURE 2: *Givenness hierarchy and object placement in relation to negation in Scandinavian varieties*

Figure 2 illustrates the fact that in Övdalian no objects shift, and the negation must precede objects regardless of their cognitive status. For Icelandic the border is between ACTVN +3 and ACTVN +4 and for Swedish and Danish the negation appears preceding object pronouns that have ACTVN +1. As mentioned earlier, Swedish allows weak objects IN SITU even if this is rare, so in some respect Swedish may be seen as intermediate between standard Danish that allows no weak objects IN SITU and Övdalian that must have all objects IN SITU.

2 Investigating all positions

One of the problems that the investigation in Andréasson (2008) fails to solve is the seemingly large number of SHIFTED det_S in declarative sentences with non-factive matrix verbs in Danish. As mentioned above, this investigation implied that as many as 94% of the det_S are shifted in these sentences. Given that the pronominal

[4] O'Connor (2006) makes use of a ±ACTVN feature, where the value "–" marks a situation where the speaker assumes that there is no need for extra lexicogrammatical marking to activate a referent in the listener's mind. This corresponds to the value 0 in this paper as well as in Andréasson (2008). Here, I use O'Connor's activation feature as a "place holder" for a future, more elaborate analysis of the architecture of the LFG i-structure than Andréasson (2008).

complements of non-factive verbs normally would represent information that is not presupposed, the numbers in the corpus investigation would point an analysis of Danish where also pronouns whose referents have a ACTVN +1 value are licensed in a SHIFTED position. As we shall see, this is nevertheless not the case.

The problem with the investigation in Andréasson (2008) and many other object shift studies lies in the method of investigation. As long as we see object shift as a binary choice between two positions, SHIFTED and IN SITU, we cannot explain the whole range of data.

In January 2008, Danish informants without any linguistic training[5] were asked to choose between the two word orders in example (14).

(14) a. Jeg tror det ikke. [DA]
I think that not
'I don't think so.'
b. Jeg tror ikke det. [DA]
I think not that
'I don't think so.'

For a speaker of Swedish the corresponding choice would be easy, all informants would choose the IN SITU word order in (14b) over the SHIFTED word order in (14a), which would correspond to the findings in the corpus study of Andréasson (2008). The Danish informants, however, did not choose the IN SITU word order. Neither did they choose the SHIFTED option. Instead, they simply refused to chose, and they stated that only a word order where the det_S was in the initial position would be acceptable, see (15).

(15) Det tror jeg ikke. [DA]
that think I not
'I don't think so.'

The word order in (15) is also possible in Swedish, and in elicitation tests Swedish informants sometimes suggest this word order as an alternative to the IN SITU word order in declarative sentences. The Danish informants, however, stated that the initial position was the only choice in the declarative sentences with non-factive matrix verbs presented to them.

Interestingly, when presented with non-declarative sentences, where the initial position is not available, the IN SITU position, see (16), was the unmarked option also for the Danish informants. This corresponds very well to the findings of the corpus investigation in Andréasson (2008, 2009) where the IN SITU position is dominant both for Swedish and Danish in non-declarative sentences, see above.

(16) a. Hvorfor tror du ikke det? [DA]
why think you not that

[5]This investigation was performed during a NORMS/ScanDiaSyn workshop in Western Jutland, January 2008. Other Danish informants also have the same intuition.

'Why don't you think so?'

b. Tror du ikke det? [DA]
think you not that
'Don't you think so?'

Both corpus data and informants' judgements suggest that object pronouns with the same cognitive status seem to be linked to different syntactic positions in Swedish and Danish, and that it is necessary to investigate all possible positions for object pronouns to be able to decide if this is the case. In Figure 3, all positions for pronominal objects are presented.[6]

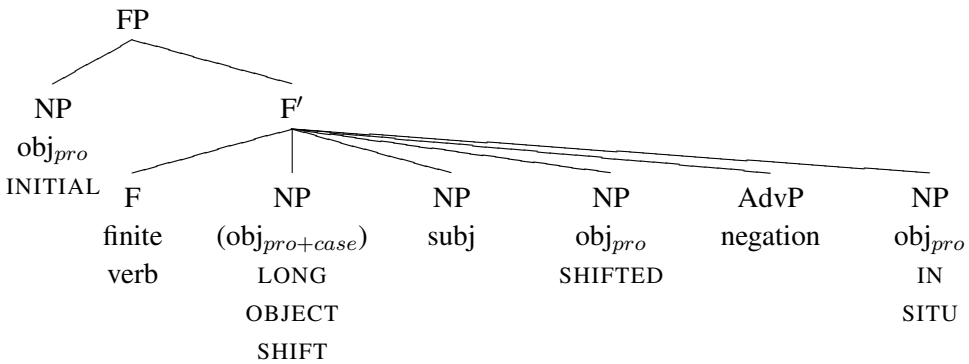


FIGURE 3: *Positions for pronominal objects in Swedish (phrase structure, see Börjars, Engdahl & Andréasson 2003; Andréasson 2007)*

In the following I will present data that strengthens the hypothesis that it may be infelicitous to investigate object shift as a choice between positions preceding and following the negation. The data instead points to the benefits of an analysis of object placement in general in the Scandinavian languages.

2.1 Methods of investigation

In Andréasson (2008), I followed the practice of previous studies and investigated object pronouns that were SHIFTED and IN SITU. I used two web corpora, for Danish the Korpus 2000[7] and for Swedish the GP04 corpus[8] and I used a search string that produces all occurrences of *det* immediately preceding or following the negation *inte/ikke*, 'not', i.e. *det inte/ikke* and *inte/ikke det*, 'it not'/'not it'.

[6]The possibility for long object shift, where a pronominal object precedes the subject in the F′ domain, is not available for object pronouns that are not case marked and will not be further discussed in this paper.

[7]Korpus 2000 (<http://korpus.dsl.dk/korpus2000/engelsk_hovedside.php3?lang=dk>) consists of about 28 million words of mixed genres, and was collected during the years 1998–2000.

[8]GP04 (<http://spraakbanken.gu.se/konk/>) consists of 19 million words from the daily newspaper Göteborgs-Posten, i.e. the total edition for 2004.

The data presented in this paper are partial results from an ongoing corpus investigation where I aim to investigate all possible positions for *det*$_S$ and – initially – get quantitative data on word orders. However, *det* is one of the most common words in Swedish and Danish; it is not case marked, so the pronoun may represent either a subject or an object function, and there is a great deal of homonymy, since *det* is also the neuter form of the definite article in both languages. Because of this I have chosen not to include the word form *det* in the search strings. Instead I perform searches for verbs that take sentential complements, with a negation in the immediate following context. The search strings, VERB [0–10 intervening words], for the corpora used are presented in (17) below.

(17) a. Korpus Dk: [word="forstår|forstod"] []{0,10} [lemma="ikke"]
b. PAROLE: [word="förstår|förstod"] []{0,10} [word="inte"]

For the current investigation I have made use of two morphologically tagged web corpora, for Danish the KorpusDK[9] and for Swedish the PAROLE corpus.[10]

So far I have investigated declarative clauses, with the factive verbs *veta/vide* 'know', and *förstå/forstå* 'understand', and the non-factive verbs *tro* and *tycka/synes*, both with the meaning 'think'/'believe'. The corpus study presented is a total investigation for these very frequent verbs in combination with a negation in these large corpora, so the quantitative data can be considered highly reliable.

2.2 Results for declaratives with non-factive and factive matrix verbs

Table 2 presents the data for the non-factive matrix verbs *tro* and *tycka/synes*, both with the meaning 'think'/'believe'. Example (18) shows the word orders relevant for the table.

(18) a. INITIAL: Det tror/tycker jag inte. [SW]
that think I not
'I don't think so.'
b. SHIFTED: Jag tror/tycker det inte.
c. IN SITU: Jag tror/tycker inte det.

[9] KorpusDK (<http://ordnet.dk/korpusdk>) consists of about 56 million words of mixed genres, and was collected during the years 1990–2000.

[10] PAROLE (<http://spraakbanken.gu.se/parole/>) consists of about 19.4 million words of mixed genres, and was collected during the years 1976–1997.

Swedish	initial	shifted	in situ
tro	67	12	27
tycka	68	1	11
Total	135 (73%)	13 (7%)	38 (20%)
Danish	initial	shifted	in situ
tro	153	56	0
synes	104	12	0
Total	257 (79%)	68 (21%)	0 (0%)

TABLE 2: *Positions for pronominal objects with non-factive matrix verbs* tro *and* tycka/synes *'believe, think'*

Table 2 shows that as many as 79% of the det_S with non-factive matrix verbs in Danish and 73% in Swedish appear in the initial position. When investigating all possible positions for object pronouns the SHIFTED position is therefore no longer dominant in Danish, as it seemed to be in Andréasson (2008). These numbers clearly reveal that an investigation of only the SHIFTED and the IN SITU position would indeed give a misleading picture.

Furthermore, the method where I searched for the verbs in combination with a negation, and not the pronoun *det*, gave an unexpected result, namely that I also got quantitative data for objects that are not realised as pronouns. It turns out that factive matrix verbs, but not non-factive, allow for object ellipsis, in table 2 called ZERO, and that Swedish and Danish differ as to what extent object pronouns are left out for these verbs.

Table 2 presents the data for the factive matrix verbs *veta/vide* 'know', and *förstå/forstå* 'understand'. Example (19) show the word orders relevant for the table.

(19) a. ZERO: Jag vet/förstår inte. [SW]
I know/understand not
'I don't know/understand.'

b. INITIAL: Det vet/förstår jag inte.
that know/understand I not
'I don't know it.'

c. SHIFTED: Jag vet/förstår det inte.

d. IN SITU: Jag vet/förstår inte det.

Swedish	zero		initial		shifted		in situ	
veta	572		243		23		20	
förstå	57		25		11		3	
Total	629	(66%)	268	(28%)	34	(4%)	23	(2%)
Danish	zero		initial		shifted		in situ	
vide	11		210		133		1	
forstå	22		130		186		1	
Total	33	(4.5%)	340	(49%)	319	(46%)	2	(0.5%)

TABLE 3: *Positions for pronominal objects with factive matrix verbs* veta/vide *'know' and* förstå/forstå *'understand'*

The numbers in Table 2 show that in Swedish there are two main options for det_S with factive matrix verbs, namely to leave out the det_S entirely (66%) or to place it in the initial position (28%). For Danish, the two main options are the initial position (49%) and the shifted position (46%). This leads to the conclusion that instead of realising a factive det_S in the shifted position, Swedish seem to prefer not to realise it at all, a fact that strengthens the assumption that its referent is highly accessible.

To conclude, if we assume that most of the det_S with non-factive matrix verbs indeed have the ACTVN value +1 (not fully accessible) and that det_S with factive matrix verbs have the ACTVN value 0 (fully accessible) we may summarise the patterns for Swedish and Danish as follows.

ACTVN 0: Both in Swedish and in Danish, only objects with the ACTVN value 0 are licensed in the shifted position. However, Swedish does not make much use of this opportunity of shifting for the verbs investigated, but prefers to leave them out.[11] Objects with ACTVN 0 are also licensed in the initial position in both languages.

ACTVN +1: Objects with the ACTVN value +1 are licensed in the IN SITU position in Swedish. In Danish they are only licensed in this position if the initial position is not available, for example in questions. Objects with ACTVN +1 are also licensed in the initial position in both languages.

2.3 Towards an OT analysis

The quantitative data presented above show some strong distributional tendencies, and qualitative analyses must also be performed. We have been able to conclude that both Swedish and Danish allow det_S with ACTVN +1 (here det_S complements of non-factive matrix verbs) in the initial position, and that for Swedish the IN SITU

[11]The zero realisation of objects is not included in Gundels et al. (1993) scale, but may be considered to be even higher on the accessibility scale than *it* or the unstressed personal pronouns in the Scandinavian languages.

position is another frequent option. I will not discuss here what factors decide when a *det*$_S$ in Swedish appears initially and when it appears IN SITU. Instead I will briefly address the fact that Danish has relatively more SHIFTED *det*$_S$ with non-factive matrix verbs, a position where only ACTVN +0 would be expected. One factor that may explain this is that Swedish and Danish seem to use different syntactic strategies for expressing contrast.

Many of the Danish sentences where an *det*$_S$ with ACTVN +1 is SHIFTED display a contrast on another element in the clause. Swedish informants report the same intuition when presented with corresponding Swedish word orders. Here we will only look at an example with subject contrast, see (20).

(20) Hvis mureren og arkitekten fastholder, at en udkradsning og
if the bricklayer and the architect maintains that a scraping and
efterfugning er tilstrækkelig, skal De tage et skriftligt forbehold.
following grouting is satisfactory shall you take a written reservation
Jeg tror det nemlig ikke. [DA]
I think it namely not
'If the bricklayer and the architect maintains that a scraping and a following grouting is satisfactory, you must put your reservations in writing. *I* am as a matter of fact not of that opinion.'

In (20), the opinions of the subject referent, the author *jeg*, diverge from those of *mureren og arkitekten* ('the bricklayer and the architect'), and *jeg* is interpreted as contrasted.

In Swedish, subject contrast is expressed syntactically and it would be preferred to use a word order with *det*$_S$ in the initial position and the subject pronoun following the negation, as in the construed sentence in example (21a). In Danish this S-ADVL < SUBJ word order is ungrammatical, see (21b).

(21) a. Det tror nämligen inte jag. [SW]
that think namely not I
'I$_{contrast}$ – as a matter of fact – don't think so.'
b. *Det tror nemlig ikke jeg. [DA]
that think namely not I
'I$_{contrast}$ – as a matter of fact – don't think so.'

In the Swedish sentence in (21a) the pronominal subject follows the sentence adverbial and it would be interpreted as contrasted without any other context. Andréasson (2007) presents an LFG-OT analysis of how information dynamical factors interact with structural patterns, and shows that pronominal subjects following sentence adverbials are always interpreted as focussed. The relevant OT constraints that reward a S-ADVL < SUBJ word order when the subject is contrasted are presented in (22).

(22) Ranking: F-OP»RESP-I

a. F-OP: Align (information dynamic operator-R, Focus domain-L), i.e. put the right edge of an information dynamic operator next to the left edge of a focus domain.

b. RESP-I: The linear order of constituents respect the information principle: i.e. GROUND < SCENE < RHEME.

F-OP is an alignment constraint that requires a focus operator to be aligned with the left edge of a focus domain, and RESP-I is a faithfulness constraint that among other things requires elements with the informational status GROUND to precede rhematic elements. In a sentence like (21a) the subject may very well be part of the GROUND but the ranking where F-OP outranks RESP-I still makes the S-ADVL < SUBJ the most felicitous.

In Danish, where the S-ADVL < SUBJ word order is ungrammatical the ranking must be the opposite: RESP-I»F-OP. The Danish means of expressing focus syntactically may instead be to place contrasted elements initially, see (20) above, a position where grammatical discourse functions, GDFs, are known to be licensed, see (23) (cf. Börjars, Engdahl & Andréasson 2003; Andréasson 2007). A det_S with higher ACTVN value may then be demoted and realised as SHIFTED to avoid a contrast interpretation.

(23) a. $\text{FP} \rightarrow \underset{(\uparrow \text{GDF})=\downarrow}{\text{XP}} \quad \underset{\uparrow=\downarrow}{\text{F}'}$

b. $\text{F}' \rightarrow \underset{\uparrow=\downarrow}{\text{F}}, \quad \underset{(\uparrow \text{SUBJ})=\downarrow}{\text{NP}}, \quad \underset{(\uparrow \text{OBJ}_{ind})=\downarrow}{\text{NP}}, \quad \underset{(\uparrow \text{OBJ}_{dir})=\downarrow}{\text{NP}}, \quad \underset{\uparrow=\downarrow}{\text{VP}}, \quad \underset{\downarrow \in (\uparrow \text{ADJ})}{\text{XP*}}$

3 Summary and outlook

In this paper, I have shown that it is not enough to look at two positions, SHIFTED and IN SITU, when exploring the underlying causes of pronominal objects appearing preceding sentence adverbials in Scandinavian languages, the phenomenon called *object shift*. Both the initial position and the possibility of object ellipsis must be considered.

We have seen that object pronouns with sentence antecedents, det_S, that do not have the highest cognitive status generally appear initially both in Swedish and in Danish. In Swedish they also frequently appear IN SITU, a position that is an option in Danish, only when the initial position is blocked by another element, i.e. in questions. Det_S that have the highest level of cognitive status generally appear initially or SHIFTED in Danish. In Swedish they are left out or appear initially. We have also seen that there are signs that contrast on another element seems to overrule some of these generalisations.

This paper presents partial results from an ongoing investigation. Detailed qualitative analyses of the different word orders will be performed as well as elicitations with Swedish and Danish informants. Other classes of factive and non-

factive verbs will be investigated, as well as sentences with temporal and modal auxiliaries, *göra* ('do') and copular verbs, both in written and in speech corpora.

4 References

Andréasson, Maia 2007. *Satsadverbial, ledföljd och informationsdynamik i svenskan* (Sentence adverbials, word order and information dynamics in Swedish). (Göteborgsstudier i nordisk språkvetenskap 7). Diss. English summary. Göteborg: Göteborgs universitet.

Andréasson, Maia. 2008. Not All Objects Are Born Alike: Accessibility as a Key to Pronominal Object Shift in Swedish and Danish. In: M. Butt & T. H. King (eds.), *The Proceedings of the LFG '08 Conference*, 26–45. Stanford: CSLI.

Andréasson, Maia 2009. Pronominal object shift – not just a matter of shifting or not. *Working Papers in Scandinavian Syntax* 84: 1–20.

Börjars, Kersti, Elisabet Engdahl & Maia Andréasson 2003. Subject and Ob- ject Positions in Swedish. In: Butt, M. & T. H. King (eds.), *Proceedings of the LFG03 Conference*, 43–58. Stanford: CSLI.

Garbacz, Piotr 2010. *Word Order in Övdalian*. Lundastudier i nordisk språkvetenskap A70. Lund: Centre for Language and Literature, Lund University.

Gundel, Jeanette, Kaja Borthen & Thorstein Fretheim 1999. The Role of Context in Pronominal Reference to Higher Order Entities in English and Norwegian. In: P. Bouquet et al. (eds.): *CONTEXT'99*, 475–478. LNAI 1688.

Gundel, Jeanette, Nancy Hedberg & Ron Zacharski 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69 (2): 274–307.

Gundel, Jeanette, Michael Hegarty & Kaja Borthen 2003. Cognitive Status, Information Structure, and Pronominal Reference to Clausally Introduced Entities. *Journal of Logic, Language and Information* 12: 281–299.

Hegarty, Michael 2003. Semantic types of abstract entities. *Lingua* 113: 891–927.

Hellan, Lars, & Christer Platzack 1995. Pronouns in Scandinavian languages: An overview. *Working Papers in Scandinavian Syntax* 56: 47–69.

Holmberg, Anders, 1986. *Word Order and Syntactic Features in the Scandinavian Languages and English*. Stockholm: Dept. of linguistics, Stockholm University.

Holmberg, Anders, 1999. Remarks on Holmberg's Generalization. *Studia Linguistica* 53: 1–39.

Josefsson, Gunlög 1992. Object shift and weak pronominals in Swedish. *Working Papers in Scandinavian Syntax* 52: 1–28.

Josefsson, Gunlög 2003. Four Myths about Object Shift in Swedish – and the Truth... In: *Grammar in Focus: Festschrift for Christer Platzack 18 November 2003*, Vol. II, 199–207. Lund: Inst. för nordiska språk, Lunds universitet.

O'Connor, Robert 2006. *Information Structure in Lexical-Functional Grammar: The Discourse-Prosody Correspondence in English and Serbo-Croatian.* Diss. Manchester: University of Manchester.

Pedersen, Karen, 1993. Letledsreglen og lighedsreglen. Novation, ekspansion og resistens. In: *Jyske studier tilegnede Magda Nyberg og Bent Jul Nielsen*, 199–218. København: C.A. Reitzels Forlag.

Sells, Peter 2001. *Structure, Alignment and Optimality in Swedish*. Stanford: CSLI Publications.

Svenonius, Peter 2002. Subject positions and the placement of adverbials. In: Peter Svenonius (ed.), *Subject, Predicates and the EPP*, 199–240. Oxford: Oxford University Press.

Vikner, Sten 1994. Scandinavian object shift and West Germanic scrambling. In: Norbert Corver & Henk van Riemsdijk (eds.), *Studies on Scrambling: Movement and non-movement approaches to free word order phenomena*, 487–517. Berlin: Mouton de Gryuter.

Vikner, Sten. 1997. The interpretation of object shift, Optimality Theory, and Minimalism. *Working Papers in Scandinavian Syntax* 60: 1–24.

Vikner, Sten 2005. Object Shift. In: Henk van Riemsdijk & Martin Everaert (eds.), *The Blackwell Companion to Syntax*, 392–436. Oxford: Blackwell.

POTTSIAN LFG

Doug Arnold and Louisa Sadler
University of Essex

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

1. Introduction

Constructions that one might broadly call 'parenthetical' (including various kinds of appositive construction) have received very little attention in the LFG literature. This is surprising because, taken broadly, parentheticals include a wide range of rather common constructions including supplemental/appositional constructions, expressives, and 'pure' parentheticals of various sorts. Apart from some brief coverage in Butt et al. (1999), the most extensive discussion of such constructions probably occurs in a series of papers in which Fortmann looks at some parenthetical constructions in German (Fortmann, 2005, 2006). Fortmann's concern is largely syntactic: in Fortmann (2005) he argues that parentheticals such as (1) are 'non-regular' constituents in that they are integrated into the c-structure, but not the f-structure, of the host clause.[1]

(1) Theo hat — ein Klempner war nicht zu erreichen — den Rohrbruch
Theo has — a plumber was not to reach — the pipe burst
selbst repariert.
self fixed

However, Fortmann (2006) shows that the structural integration facts are different for different kinds of parenthetical, specifically that at least some *verba dicendi* parentheticals are fully integrated at both c- and f-structure.

Outside LFG there has been a considerable amount of work on the syntax of parentheticals and supplemental/appositive constructions, a major focus being whether or not parenthetical elements should be analysed as syntactically integrated or syntactically orphaned (see Arnold (2007) for references). Beyond purely syntactic accounts, an important contribution is made by Potts (2005). Potts provides a very appealing account of the semantics of a range of parenthetical expressions, including supplemental/appositive (non-restrictive) relative clauses, and expressives (e.g. *damn Republican*).

In the current paper we explore whether an approach to the analysis of supplementals and appositions inspired by the work of Potts (2005) can be expressed in the framework of LFG. We concentrate on ARCs, and focus on how Potts' approach can be implemented using a resource sensitive glue approach to semantic composition (see e.g. Dalrymple, 1999; Asudeh, 2004). An LFG account of the syntax and semantics of English ARCs is a side effect of our discussion (no account seems to have been previously proposed).

The paper is structured as follows. Section 2 sets out the basic properties of English ARCs. Section 3 outlines Potts' approach. In Section 4 we describe

[†]We are grateful to the participants at LFG10 in Ottawa, Canada, for insightful and stimulating discussion, and in particular to Ash Asudeh and Mary Dalrymple. We should also thank Tracy Holloway-King for detailed comments on a earlier version of the paper.

[1]Fortmann does not provide an English translation, but it would presumably be along the line of 'No plumber being available, Theo has fixed the burst pipe himself'.

what seems to be the most direct implementation of Potts' ideas, building on the standard LFG analysis of restrictive relative clauses from Dalrymple (2001). In Section 5 we consider an alternative analysis that makes more use of the LFG projection architecture. Section 6 provides a conclusion and notes some open questions.

2 ARCs: The Facts

As is well known, supplemental or appositive relative clauses (ARCs), such as in (2), are distinguishable from restrictive relative clauses (RRCs), such as in (3), on a number of grounds. Phonologically, an ARC as in (2) is often associated with special 'comma' intonation, which is not typical of an RRC such as in (3). Semantically, in (3) the relative clause restricts the denotation of the noun *pets*, in (2) it does not, instead it simply provides additional or supplementary information about the pets (hence the terms 'non-restrictive' or 'supplementary' relative clause). As a consequence a restrictive can give rise to a 'contrast set' of entities that are in the denotation of the head noun, but do not satisfy the relative clause. Thus there is a potential antecedent for *the others* in (3), while there is no antecedent for *the others* in (2).

(2) Kim has three pets, which a neighbour looks after.
#The others fend for themselves. [ARC]

(3) Kim has three pets which a neighbour looks after.
The others fend for themselves. [RRC]

English ARCs are subject to a number of surface morphosyntactic restrictions, in particular, they must be finite and +WH, as illustrated by the contrast between (4) and (5)-(6). The corresponding RRCs are acceptable, witness (7)-(8).

(4) Kim, who Sandy spoke to, will arrive later.
(5) *Kim, for Sandy to speak to, will arrive later.
(6) *Kim, that Sandy spoke to, will arrive later.

(7) A person for Sandy to speak to will arrive later.
(8) A person that Sandy invited will arrive later.

In the remainder of this section, we will establish two key properties of ARCs in English. These are (i) that they show semantic 'wide scope' effects; and (ii) that they are syntactically integrated, at both c- and f-structure.[2]

The existence of 'wide scope' effects with ARCs is relatively well known. Consider the pair (9a)-(9b). A reading of (9a) is available in which Sandy's aunt (also) lives in Sweden, i.e. where the content of the relative clause (roughly 'she lives in Sweden') is part of the antecedent of *so*; but (9b) provides no information about where Sandy's mother lives – the ARC is

[2]See Arnold (2007) and references there for more discussion of the empirical issues.

not part of the antecedent of *so*. Similarly, in (10b) *who use the IPA* scopes outside Kim's belief set (Kim's belief is about linguists in general), while in (10a) Kim's beliefs concern those linguists who use the IPA. In (11b) *who use the IPA* similarly falls outside the scope of the question.

(9) a. Kim is visiting her aunt who lives in Sweden, and so is Sandy. [RRC]
b. Kim is visiting her mother, who lives in Sweden, and so is Sandy. [ARC]

(10) a. Kim believes that linguists who use the IPA are clever. [RRC]
b. Kim believes that linguists, who use the IPA, are clever. [ARC]

(11) a. Are linguists who use the IPA invariably clever people? [RRC]
b. Are linguists, who use the IPA, invariably clever people? [ARC]

The existence of wide scope effects with ARCs is not controversial.[3] The syntactic status of ARCs is more problematic, though we think the emerging consensus favours analyses where they are fully integrated with few, if any, structural distinctions between ARCs and RRCs.

As regards c-structure, we believe the evidence that ARCs are integrated is overwhelming. First, a range of standard constituency tests show that the ARC and the NP antecedent form a constituent. For example, it is in general impossible to separate an ARC from its antecedent in a raising context (other 'movement' phenomena show exactly the same restriction):

(12) [Kim, who Sandy invited], seemed Δ to leave early.
(13) *Kim seemed, [Δ who Sandy invited], to leave early.

The same conclusion follows from the possibilities of placement for the possessive *'s* when the possessor is modified by an ARC. If we assume that possessive *'s* attaches to the right edge of NP, the following provide convincing evidence that an ARC and its antecedent together form a NP (i.e. they are integrated at c-structure). Notice also that the phonological form of the possessive varies according to the final segment of the ARC, as one would expect if the ARC is integrated.

(14) Kim – who you hit –'s mother (wants to talk). (/s/)
(15) Kim – who you hid –'s mother (wants to talk). (/z/)
(16) Kim – who you miss –'s mother (wants to talk). (/ɪz/)
(17) *Kim's – who you miss – mother (wants to talk).

The issue of f-structure is more subtle, but what evidence there is suggests that the ARC is also integrated at f-structure.

[3]But not entirely unproblematic. It is widely assumed that ARCs invariably take wide scope, but this is not the case. Arnold (2007) discusses circumstances where ARCs appear to take narrow scope, and Harris and Potts (2010) give convincing examples, bolstered with some persuasive experimental results, which indicate that ARCs can sometimes take narrow scope.

One argument in favour of this conclusion may be provided by data concerning conditions on VP ellipsis (VPE) in English. Potsdam (1997), following Warner (1993), observes that in ellipsis of a VP with an auxiliary verb, the antecedent VP must not be finite, so (18) is grammatical but (19) is ungrammatical. This shows that the resolution process that finds the antecedent of an ellided VP needs to check some morphosyntactic properties (f-structure attributes) of potential antecedents.

(18) You may not be confused, but you really should. [be confused]
(19) *You are not confused, but you really should. [be confused]

The key point for present purposes is how this fact about VPE interacts with ARCs. VPE is possible both into and out of ARCs and RRCs. The antecedent of VPE in an ARC can be inside a restrictive as in (20a), and the antecedent of VPE in a restrictive relative clause can be inside an ARC, as in (20b). As (20c) additionally shows, both antecedent and VPE can be within ARCs. Given that VPE is subject to morphosyntactic constraints, these data provide an important argument that ARCs are contained within (accessible within) whatever representation is relevant to the statement of morphosyntactic conditions such as those noted by Potsdam: in LFG this would be f-structure. The data in (21a)-(21f) support this conclusion by showing that the nonfiniteness requirement (on VPE) determines the grammaticality. Notice in particular that the pattern of judgements is the same regardless of whether the examples involve ARCs or RRCs.

(20) a. Someone that [supports the war]$_i$ insulted Kim, who doesn't Δ_i
b. Sandy, who [supports the war]$_i$, insulted someone that doesn't Δ_i
c. Sandy, who [supports the war]$_i$, insulted Kim, who doesn't Δ_i

(21) a. Someone that [may be worried]$_i$ told Kim, who really should Δ_i
b. *Someone that [is worried]$_i$ told Kim, who really should Δ_i
c. Sandy, who [may be worried]$_i$, told someone that really should Δ_i
d. *Sandy, who [is worried]$_i$, told someone that really should Δ_i
e. Sandy, who [may be worried]$_i$, insulted Kim, who really should Δ_i
f. *Sandy, who [is worried]$_i$, insulted Kim, who really should Δ_i

At very least, these data show that the facts about VPE will be easier to describe if ARCs are part of f-structure in the same way as RRCs.

A second argument for f-structure integration comes from the fact that Right-Node Raising operates out of both ARCs and restrictives alike, as shown in the following examples:

(22) Sam met someone [that buys Δ], and Kim met someone [that sells Δ], the pretentious garbage that some call Modern Art. RRC

(23) Sam met Sandy, [who buys Δ], and Kim met Leslie, [who sells Δ], the pretentious garbage that some call Modern Art. ARC

(24) a. Sam met someone [that buys Δ], and Kim met Leslie, [who sells Δ],

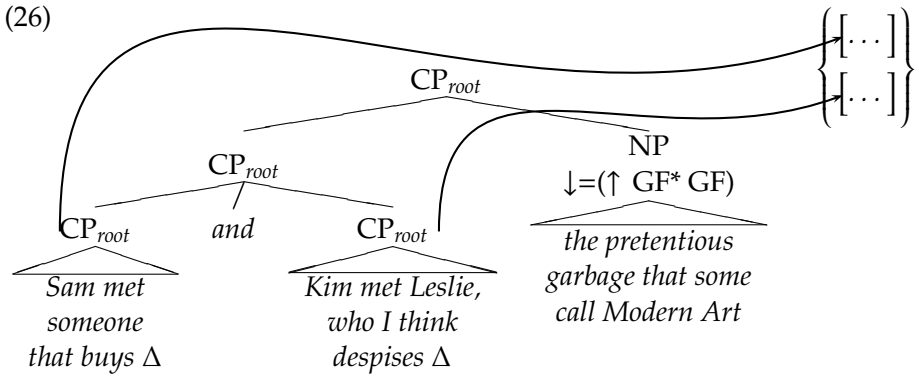
the pretentious garbage that some call Modern Art. RRC,ARC

b. Sam met Sandy, [who buys Δ], and Kim met someone [that sells Δ], the pretentious garbage that some call Modern Art. ARC,RRC

Simplifying considerably, one might have a structure for (25) such as (26). A functional uncertainty statement is associated with the RNR-ed NP (given here in maximally general form as ↓= (↑GF* GF)). Since ↑ is a set, the functional uncertainty may be solved differently in each member of the set (in each conjunct), but in each it will have to be a continuous path through the f-structure, for example, OBJ RELMOD OBJ in the first conjunct and OBJ RELMOD COMP OBJ in the second conjunct. Since it appears that these paths can reach indiscriminately into either ARCs or RRCs, the implication is that ARCs are integrated into the f-structure just as RRCs are — or at least, a description of RNR will be easier to formulate if ARCs are part of f-structure in the same way as RRCs.

(25) Sam met someone that buys Δ and Kim met Leslie, who I think despises Δ, the pretentious garbage that some call Modern Art.

(26)



3 Potts' Approach

Potts' central idea is that the interpretation of an expression involves (at least) two dimensions: an *at-issue* dimension of normal truth-conditional content, and another dimension, which Potts argues to consist of Gricean conventional implicatures — the *ci* dimension. It is to this dimension that supplemental expressions, including ARCs, make their contribution. Since we assume that one could accept the idea of a special 'supplemental' dimension of meaning without accepting Potts' Gricean characterization of it, we will not pursue the latter topic, interesting though it is. We also focus only on ARCs, leaving aside other parentheticals and supplementals, but note that Potts' own proposal has much wider scope.

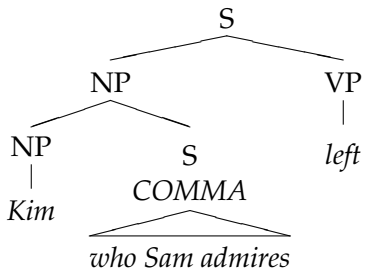
Potts' framework is type-theoretic, so the idea is that every expression is associated with a pair of meanings, and a pair of types $\langle \alpha^a, \gamma^c \rangle$, where α^a is an *at-issue* type, and γ^c is a *ci* type (note the superscripts). In what follows we

will normally suppress the superscript on *at-issue* types, to avoid notational clutter. Semantic composition involves constructing a semantic parsetree, whose nodes are subject to a number of node-admissibility conditions. These are most easily understood via an example, such as (27).

(27) Kim, who Sam admires, left.

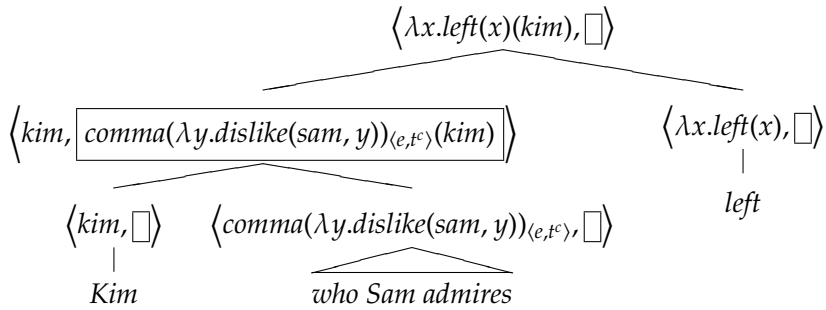
The basic syntactic structure Potts assumes is along the lines of (28) (modulo the node labels, e.g. Potts has DP where we have used NP). From the point of view of the discussion above, the only important points to note are (a) that the ARC is syntactically integrated, and (b) the presence of a COMMA feature on the ARC. This syntactic feature provides an interface between the phonology (comma intonation) and the semantics (*ci* type content).

(28)



The corresponding semantic tree is along the lines of (29), in which each node is associated with a pair denoting the *at-issue* and *ci* content, and where we have highlighted *ci* content by putting it in a box. Notice that most nodes have no *ci* content (the box is empty). We have only indicated the types of expressions where they are non-standard.

(29)



Notice that the *at-issue* content of the whole sentence is $\lambda x.left(x)(kim)$, that is $left(kim)$, which would also be the *at-issue* content of a simple sentence without a parenthetical (viz. *Kim left*). Notice likewise that the *at-issue* content of *Kim, who Sandy admires,* is the same as the *at-issue* content of *Kim* (the content of the ARC has disappeared from the *at-issue* dimension). The *ci* content of *Kim, who Sandy admires,* is produced from the *at-issue* content of *who Sandy admires* and the *at-issue* content of *Kim* in a way we will describe directly. It will turn out to be the *ci* proposition $dislike(sam, kim)_{t^c}$. Finally, notice that the *ci* content of *who Sandy admires* is empty – at the level of

the relative clause, the content is still in the *at-issue* dimension. Intuitively, what happens is that the content of the relative clause 'starts out' as *at-issue* content, but then moves to the *ci* dimension, and then plays no further part in the construction of the *at-issue* content of the main clause. This is in contrast with the content of *Kim*, which one might think of as being 'copied' from the *at-issue* dimension into the *ci* dimension, since it plays a role in the *ci* content, and also plays a role in the *at-issue* content of the clause as a whole (i.e. it is used twice).

The mechanics here involves two admissibility conditions, and the definition of *comma*, which is the semantic correlate of the COMMA feature.[4]

Potts defines *comma* as in (30) – it is simply a function that changes the type of its argument from being of type $\langle e, t\rangle$ to being of type $\langle e, t^c\rangle$; it takes a function from entities to normal propositions, and produces a function from entities to propositions in the *ci* dimension.

(30) $comma = \lambda X.\lambda x.X(x)_{\langle\langle e,t\rangle,\langle e,t^c\rangle\rangle}$

In the example above, the denotation of *who Sam admires* is of type $\langle e, t\rangle$, so *comma* will apply to it as in (31), giving (32) as the *ci* content of *Kim, who Sam admires*, as promised above.

(31) $comma(\lambda y.dislike(sam, y))_{\langle e,t^c\rangle} =$
$\lambda X.\lambda x.X(x)(\lambda y.dislike(sam, y))_{\langle e,t^c\rangle} =$
$\lambda x.dislike(sam, x)_{\langle e,t^c\rangle}$

(32) $\lambda x.dislike(sam, x)_{\langle e,t^c\rangle}(kim) =$
$dislike(sam, kim)_{t^c}$

The first of Potts' node admissibility conditions is given schematically in (33a), and exemplified in (33b). This is used for 'normal' (*at-issue*) content. (The order of daughters in these trees is not relevant; we have used α for the functor daughter, and β for the argument; more generally, functor and argument are identified by their types, which we have not bothered to indicate here).

(33) a. $\langle \alpha(\beta), \square\rangle$
$\langle \beta, \square\rangle \quad \langle \alpha, \square\rangle$

b. $\langle \lambda x.left(x)(kim), \square\rangle$
$\langle kim, \square\rangle \quad \langle \lambda x.left(x), \square\rangle$
Kim *left*

In words: the *at-issue* content of the mother is produced by applying the *at-issue* content of the functor argument to the *at-issue* content of the argument daughter. The *ci* content is not touched or affected in any way (note, in particular, that it is not percolated). Essentially, this is just normal

[4]Potts assumes a third admissibility condition, which licenses *ci* content which does not take any piece of *at-issue* content as an argument. This is used for the parenthetical part of examples like *Kim (and you won't believe this) ate fifty eggs*. It is not relevant here.

semantic composition.

The second is given schematically in (34a) and exemplified in (34b).

(34) a. $\langle \beta, \boxed{\alpha(\beta)} \rangle$ b. $\langle kim, \boxed{comma(\lambda y.dislike(sam, y))_{\langle e,t^c \rangle}(kim)} \rangle$

$\langle \beta, \square \rangle$ $\langle \alpha, \square \rangle$ $\langle kim, \square \rangle$ $\langle comma(\lambda y.dislike(sam, y))_{\langle e,t^c \rangle}, \square \rangle$

Kim *who Sandy admires*

Here the *at-issue* content of the mother is just that of the argument daughter, but as well as being 'passed up' to the mother, the content of this daughter is also used as an argument in the *ci* content of the mother. The functor in the *ci* content comes from the *at-issue* content of the functor daughter. This admissibility condition requires α to be a function from *at-issue* content to *ci* content: $\langle \sigma^a, \tau^c \rangle$ for some types σ, τ — in the example in (34b) the function is of type $\langle e, t^c \rangle$. Producing this sort of type from a 'normal' *at-issue* type is of course the semantic effect of the COMMA feature. One way of thinking of this node admissibility condition is as moving the functor into the *ci* dimension, and 'copying' the argument into the *ci* dimension (the argument also remains in the *at-issue* dimension; in resource logic terms, this means the argument is 'consumed twice').

Semantic parsetrees are interpreted according to the following principle (ignoring intensionality). Let T be a semantic parsetree with the *at-issue* term α_{σ^a} on its root node (that is, a semantic expression α of type σ^a), and distinct terms $\beta^1_{\tau^c}, \ldots, \beta^n_{\tau^c}$ on its nodes, then the interpretation of T is the tuple in (35).

(35) $\left\langle [\![\alpha_{\sigma^a}]\!]^{M,g}, [\![\beta^1_{\tau^c}]\!]^{M,g}, \ldots, [\![\beta^n_{\tau^c}]\!]^{M,g} \right\rangle$

That is, a tuple consisting of the interpretation of the *at-issue* content, and all the *ci* content from anywhere in the tree, all interpreted relative to the same model and variable assignment.

Several features of this approach are worth noting. First, we have seen there are expressions of type $\langle \sigma^a, \tau^c \rangle$, that is, functors which take *at-issue* content to *ci* content (ARCs are of this type, and *comma* is designed to produce expressions of this type). However, Potts sets up the type theory so that there are no expressions which work the other way: there are no expressions of type $\langle \tau^c, \sigma^a \rangle$. Intuitively, this means there are no expressions which take *ci* content and move or copy it into the *at-issue* dimension. There is only one way traffic from the *at-issue* dimension into the *ci* dimension.

The fact that there are no functions from *ci* content means that *ci* content escapes the scope of all 'normal' (*at-issue*) operators, and makes *ci* content 'scopeless'. The effect of (35) is to combine *ci* content at the highest level – to give it 'wide scope'. Potts' approach thus captures what we take to be two key features of ARCs: syntactic integration and semantic wide scope.

Second, notice that the *ci* dimension is not percolated around the semantic parsetree. Instead, once an appositive has been formed, it is left where it is until the interpretation of the whole parsetree assimilates it to the interpretation of the whole main clause. This further underlines the limited possibilities for interaction across the dimension: once an appositive has been formed it is entirely inaccessible to the *at-issue* content — as though it was not part of the higher semantic tree. Potts considers and rejects introducing *ci* projection and *ci* storage mechanisms, on the grounds that the approach adopted makes the conceptual separation of the dimensions clearer.

The third, and for current purposes the most interesting, feature of the approach is the potential problem it raises for resources sensitive approaches to semantic interpretation — as we have seen, some content (e.g. the content of *Kim* above) is used twice, once in the *at-issue* dimension, once in the *ci* dimension.

Potts is aware of this issue, and sketches out a possible solution (Potts, 2005, p85ff). He exemplifies it with 'expressive' adjectives. The basic idea is that the semantics of an expressive adjective like *damn* is such that the semantics of *damn Republican* is associated with an *at-issue/ci* pair like (36), where the *at-issue* content is just the normal content of *Republican*, which can be used in the normal way (as in, e.g. *The Republicans will win*), and the *ci* content predicates *damn* of *Republican*, conveying a negative attitude to Republicans.

(36) $\left\langle Republican_{\langle e,t\rangle}, \boxed{damn(Republican)_{t^c}} \right\rangle$

Producing this is mainly a matter of giving the right semantics to *damn* in the lexicon — specifically, giving it a type that yields a *ci* proposition when applied to its argument. We need not pursue this here. What matters here is that issues of resource sensitivity arise with *damn* as they do with ARCs above, because the semantics of *Republican* is used in both *at-issue* and *ci* dimensions.

Potts suggests this issue can be addressed by associating *damn* with a resource like (37). He gives a derivation of *damn Republican* as in (38).[5]

(37) $f \multimap [f \otimes p^c]$

The idea is that *damn* should consume the resource associated with the content of a noun like *Republican* to produce a composite ('tensor') resource, consisting of another normal resource (f), and a *ci* resource (p^c), thus duplicating the resource associated with *Republican* and avoiding the apparent

[5]Here we use the notation Potts uses, with 'meaning' and 'glue' expressions linked by '$\rightsquigarrow$'. Below we will use a different notation, this will typically involve a name (abbreviation) for a meaning constructor, followed by a meaning expression (a piece of lambda calculus) and a glue expression separated by a comma, so we would have something like [**Republican**] $\lambda x.Republican(x) : f$ in place of (38a).

problem of 'double consumption' of resources.[6]

(38) a. [**Republican**] $\rightsquigarrow f$
b. [**damn**] $\rightsquigarrow f \multimap [f \otimes p^c]$
c. [**damn-Republican**] $\rightsquigarrow f \otimes p^c$

Potts does not develop this suggestion beyond this description of a resource for expressive adjectives, in particular, he does not consider whether the approach can be generalized to deal with *ci* content that is not lexically based. In what follows, we attempt to remedy this, filling out details, and applying the approach to the analysis of ARCs.

4 An LFG-Glue Implementation

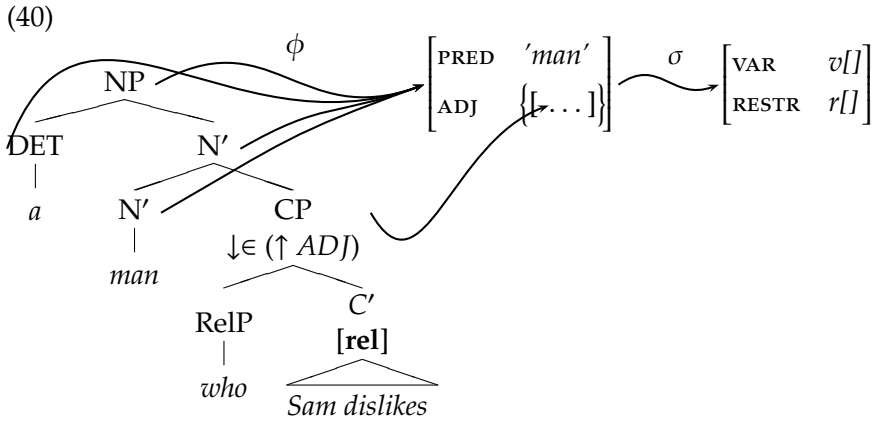
Assuming ARCs are syntactically integrated like RRCs, then a reasonable starting point for our analysis is the approach to English RRCs in Dalrymple (2001, 416ff), which provides an account of the syntax and the semantics of English RRCs. Since our account makes crucial use of the meaning constructors that she proposes, we begin by outlining her analysis.

4.1 Restrictive Relatives (Dalrymple, 2001)

Dalrymple's c-structure is as in (40), where the RRC is adjoined to N′, and the [**rel**] meaning constructor is associated with C′.

(39) A man who Sam dislikes left.

(40)



The core of Dalrymple (2001)'s glue approach to RRCs is the meaning constructor [**rel**] which is associated with the C′ node in RRCs. The meaning constructor [**rel**] is defined as in (41), where we use $v_{\langle e \rangle}$ and $r_{\langle t \rangle}$ as abbrevi-

[6]Strictly speaking, the issue is not 'double consumption', but 'multiple consumption'. In an example like *The damn Republicans, who I despise, will win*, the resource associated with *Republicans* is involved in three ways: in the main clause, as an argument of *damn*, and in the ARC. More complicated examples involving still more resource consumption can easily be imagined.

ations for ((ADJ $\in \uparrow)_\sigma$ VAR) and ((ADJ $\in \uparrow)_\sigma$ RESTR) , and $h_{<e>}$and $g_{<t>}$are ($\uparrow$RELPRO)$_\sigma$ and $\uparrow_\sigma$, respectively.[7] Intuitively, $[h_{<e>} \multimap g_{<t>}]$ corresponds to the resource contributed by the relative clause, which is a function from an individual to a proposition. Since $[v_{<e>} \multimap r_{<t>}] \multimap [v_{<e>} \multimap r_{<t>}]$ is the sort of thing one associates with a nominal modifier, [**rel**] is a meaning constructor which combines with a particular kind of (semantically incomplete) clause to produce a nominal modifier. In (42) we spell out the glue side of the definition of [**rel**] with abbreviations expanded.

(41) [**rel**] $\lambda P.\lambda Q.\lambda X.P(X) \wedge Q(X) : [h_{<e>} \multimap g_{<t>}] \multimap [[v_{<e>} \multimap r_{<t>}] \multimap [v_{<e>} \multimap r_{<t>}]]$

(42) [($\uparrow$ RELPRO)$_\sigma$ $\multimap \uparrow_\sigma$] $\multimap$
[[((ADJ $\in \uparrow)_\sigma$ VAR) $\multimap$ ((ADJ $\in \uparrow)_\sigma$ RESTR)] $\multimap$
[((ADJ $\in \uparrow)_\sigma$ VAR) $\multimap$ ((ADJ $\in \uparrow)_\sigma$ RESTR)]]

Inside the relative clause itself the relevant meaning constructors are as shown in (43). The resource (43a) is basically a one-place predicate (i.e. it will consume an entity to provide a proposition), and the resource associated with *who*, (43b), will consume (43a) to produce (43c): *who* is taken to add the restriction that the argument in question is human.

(43) a. [**Sam-dislikes**] $\lambda Y.dislike(Sam, Y) : h_{<e>} \multimap g_{<t>}$
b. [**who**] $\lambda Q.\lambda X.person(X) \wedge Q(X) : [h_{<e>} \multimap g_{<t>}] \multimap [h_{<e>} \multimap g_{<t>}]$
c. [**who-Sam-dislikes**]
$\lambda Y.person(Y) \wedge dislike(Sam, Y) : h_{<e>} \multimap g_{<t>}$

[**rel**] will combine with [**who-Sam-dislikes**] to produce a relative clause meaning, i.e. a nominal modifier, as noted above.

(44) a. [**who-Sam-dislikes**] $\lambda Y.person(Y) \wedge dislike(Sam, Y) : h_{<e>} \multimap g_{<t>}$
b. [**rel**] $\lambda P.\lambda Q.\lambda X.P(X) \wedge Q(X) : [h_{<e>} \multimap g_{<t>}] \multimap [[v_{<e>} \multimap r_{<t>}] \multimap [v_{<e>} \multimap r_{<t>}]]$
c. [**rel-who-Sam-dislikes**]
$\lambda P.\lambda Q.\lambda X.P(X) \wedge Q(X)(\lambda Y.person(Y) \wedge dislike(Sam, Y))$
$\lambda Q.\lambda X.[\lambda Y.person(Y) \wedge dislike(Sam, Y)](X) \wedge Q(X)$
$\lambda Q.\lambda X.person(X) \wedge dislike(Sam, X) \wedge Q(X) : [v_{<e>} \multimap r_{<t>}] \multimap [v_{<e>} \multimap r_{<t>}]$

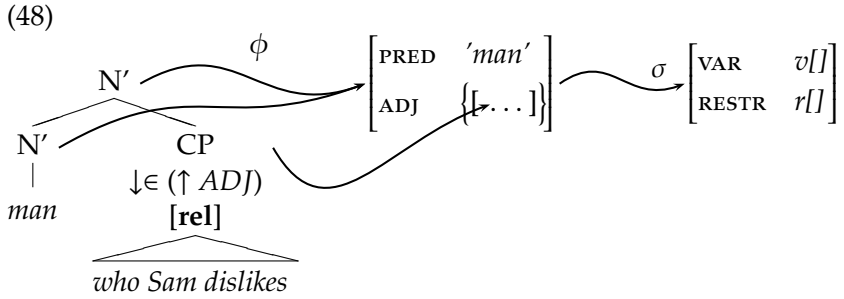
Finally, the relative clause as a whole consumes the nominal meaning and produces a (suitably restricted) nominal meaning, as in (47).

(45) [**man**] $\lambda Z.man(Z) : v_{<e>} \multimap r_{<t>}$
(46) [**rel-who-Sam-dislikes**]
$\lambda Q.\lambda X.person(X) \wedge dislike(Sam, X) \wedge Q(X) : [v_{<e>} \multimap r_{<t>}] \multimap [v_{<e>} \multimap r_{<t>}]$
(47) [**man-rel-who-Sam-dislikes**]
$\lambda Q.\lambda X.person(X) \wedge dislike(Sam, X) \wedge Q(X)(\lambda Z.man(Z))$
$\lambda X.person(X) \wedge dislike(Sam, X) \wedge \lambda Z.man(Z)(X)$
$\lambda X.person(X) \wedge dislike(Sam, X) \wedge man(X) : [v_{<e>} \multimap r_{<t>}]$

[7] In (41) we give both the complete paths and the abbreviations naming the pieces of semantic structure, but generally below we give only the abbreviations.

Note that Dalrymple (2001) introduces **[rel]** inside CP as an annotation to C′. This is not crucial, however, and we can recast the analysis of the N′ *man who Sam dislikes* as in (48) — where **[rel]** is associated with CP — without any changes being required to either the f-structure or the meaning side of the meaning constructors. The only changes required are in the associated glue expression, where small changes to the paths to the σ structures are necessary. (The reason this is possible is that semantic derivations using glue are not compositional in the c-structure, of course). The revised definition of **[rel]** is show in abbreviated form in (49); (50) presents the glue expression with abbreviations expanded. This is not crucial to our approach, but it will open up the possibility of a simplification of the treatment we present in the next section.

(48)



(49) **[rel]** $\lambda P.\lambda Q.\lambda X.P(X) \wedge Q(X) : [h_{<e>} \multimap g_{<t>}] \multimap [[v_{<e>} \multimap r_{<t>}] \multimap [v_{<e>} \multimap r_{<t>}]]$

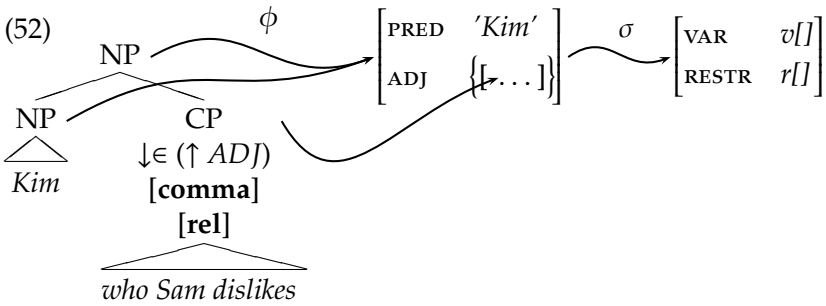
(50) $[\ (\downarrow \text{RELPRO})_\sigma \multimap \downarrow_\sigma\ ] \multimap [[\ (\uparrow_\sigma \text{VAR}) \multimap (\uparrow_\sigma \text{RESTR})\ ] \multimap [\ (\uparrow_\sigma \text{VAR}) \multimap (\uparrow_\sigma \text{RESTR})\ ]]$

4.2 ARCs

We now turn to the analysis of ARCs. Alongside **[rel]** we introduce a meaning constructor **[comma]** in the semantics. We will assume the modification of Dalrymple's approach to English relative clauses made in (48) and (49), so that these meaning constructors are associated with the CP node corresponding to the relative clause itself, which is syntactically integrated as sister of an NP head, as in (52). Notice that, apart from adjunction to N′ vs NP, the only difference between this, and structure given for the restrictive in (48) is the presence of **[comma]** in (52).

(51) Kim, who Sam dislikes, left.

(52)



Recall that Potts uses *comma* to change the type of the relative *who Sam dislikes* so that it becomes a function to *ci* content (type $\langle e, t^c \rangle$), and uses the schema in (34) to ensure that the content of *Kim* is available in both *at-issue* and *ci* domains. We therefore want **[comma]** to consume **[rel-who-Sam-dislikes]** (the meaning of the relative clause) and produce a resource which will consume the meaning of *Kim* (which is of type *e*), and produce a 'tensor' meaning constructor which combines the meanings of *Kim* and the relative clause (a *ci* meaning of type t^c, with roughly the meaning of *Sam dislikes Kim*, formed by applying the relative clause meaning to the meaning of *Kim*). To bring this about we define **[comma]** as in (53).[8]

(53) **[comma]** =
$\lambda P.\lambda Y.[Y, (P(\lambda Z.true))(Y)]$:
$[[v_{\langle e \rangle} \multimap r_{\langle t \rangle}] \multimap [v_{\langle e \rangle} \multimap r_{\langle t \rangle}]] \multimap [l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}]]$

Here $[v_{\langle e \rangle} \multimap r_{\langle t \rangle}] \multimap [v_{\langle e \rangle} \multimap r_{\langle t \rangle}]$ is the glue expression associated with a nominal modifier such as a relative clause (a function from noun meanings to noun meanings), using abbreviations introduced above; and l abbreviates $\uparrow_\sigma$ — the σ-projection of the mother NP. On the glue side, then, **[comma]** will consume an RRC-like resource and produce a resource whose glue is of the form $l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}]$ — the kind of resource Potts suggested in (37) above; viz a resource which will itself consume an NP meaning ($l_{\langle e \rangle}$) to produce a 'tensor resource' consisting of (a) an NP meaning ($l_{\langle e \rangle}$), and (b) a propositional resource in the *ci* domain ($l_{\langle t^c \rangle}$). On the meaning expression side, **[comma]** is a function that applies to an RRC meaning expression, does some type lowering (this is the purpose of the $\lambda Z.true$ expression), and yields an expression $\lambda Y.[Y, T]$, a function from individuals to a pair of meaning expressions.

The end result will be that we will get a meaning constructor like (54) for *Kim, who Sam dislikes*. In what follows, we will spell out the process by

[8]Our version of **[comma]** is less general than Potts'. This is because we want to build directly on the existing treatment of RRCs. Since semantically RRCs are of common noun modifiers of type $\langle\langle e, t\rangle, \langle e, t\rangle\rangle$, and ARCs are NP modifiers (type $\langle e, t\rangle$), this means we have to do some type lowering, and we have chosen to build this in to the definition of **[comma]**. Potts is not concerned with the relation between RRCs and ARCs, and so ignores this. Of course, it is not obvious that combining type lowering with the *ci* to *at-issue* type manipulation in this way is the right thing to do. However, we will not pursue the issue here.

which this result is produced, step by step.

(54) $[Kim, (person(Kim) \wedge dislikes(Sam, Kim) \wedge true)] : l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}$

The RRC-like resource that [**comma**] consumes is shown in (55) — this is just the nominal modifier resource standardly associated with relative clauses (and produced in turn by [**rel**]) — it is in fact exactly what we had for an RRC above, cf. (44c).

(55) [**rel-who-Sam-dislikes**]
$\lambda Q.\lambda X.person(X) \wedge dislike(Sam, X) \wedge Q(X) :$
$[v_{<e>} \multimap r_{<t>}] \multimap [v_{<e>} \multimap r_{<t>}]$

The effect of combining [**comma**] with [**rel-who-Sam-dislikes**] is shown in (56). On the glue side, [**comma**] consumes the resource [**rel-who-Sam-dislikes**] to produce a resource which is a function that will take an entity resource and produce the required tensor resource. The effect of applying the function [**comma**] to the (relative clause meaning) argument is spelled out in the reduction steps in (56), resulting in a lambda function corresponding to a one-place predicate. This in turn will apply to [**Kim**] to produce the tensor resource in (58).

(56) [**comma-rel-who-Sam-dislikes**]
$\lambda P.\lambda Y.[Y, (P(\lambda Z.true))(Y)]([\textbf{rel-who-Sam-dislikes}]) : l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}]$
$\lambda Y.[Y, ([\textbf{rel-who-Sam-dislikes}](\lambda Z.true))(Y)] : l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}]$
$\lambda Y.[Y, (\lambda Q.\lambda X.person(X) \wedge dislikes(Sam, X) \wedge Q(X)(\lambda Z.true))(Y)] : l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}]$
$\lambda Y.[Y, (\lambda X.person(X) \wedge dislikes(Sam, X) \wedge \lambda Z.true(X))(Y)] : l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}]$
$\lambda Y.[Y, (\lambda X.person(X) \wedge dislikes(Sam, X) \wedge true)(Y)] : l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}]$

(57) [**Kim**] $Kim : l_{\langle e \rangle}$

(58) [**Kim-comma-rel-who-Sam-dislikes**]
$\lambda Y.[Y, (\lambda X.person(X) \wedge dislikes(Sam, X) \wedge true)(Y)](Kim) : l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}$
$[Kim, (\lambda X.person(X) \wedge dislikes(Sam, X) \wedge true)(Kim)] : l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}$
$[Kim, (person(Kim) \wedge dislikes(Sam, Kim) \wedge true)] : l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}$

Thus, corresponding to *Kim, who Sam dislikes,* we have, on the meaning side, a pair of meanings corresponding to *Kim* and the proposition that Kim is a person and Sam dislikes Kim. On the glue side, we have a 'tensor' resource consisting of two resources, one in the *at-issue* dimension, and one in the *ci* dimension.

A further step is now required. In order to deal with these resources separately, we will need a new inference rule, which we will call *at-issue-ci-split* (*ACiS*), to separate the two parts of the tensor resource so that they may be used separately in subsequent proof steps. The rule is formulated as in (59); it is inspired by the Context Split rule of Dalrymple (2001, 297). In (60) we show a proof using it.

(59) $$\frac{[M, M'] : R_e \otimes R_{t^c}}{M : R_e \quad M' : R_{t^c}} \text{ ACiS (at-issue-ci-split)}$$

(60)

$$\dfrac{\lambda Y.left(Y) : l_{\langle e\rangle} \multimap s_{\langle t\rangle} \quad \dfrac{[Kim, (person(Kim) \wedge dislikes(Sam, Kim) \wedge true)] : l_{\langle e\rangle} \otimes l_{\langle t^c\rangle}}{Kim : l_{\langle e\rangle} \quad (person(Kim) \wedge dislikes(Sam, Kim) \wedge true) : l_{\langle t^c\rangle}}\ \text{ACiS}}{left(Kim) : s_{\langle t\rangle} \quad (person(Kim) \wedge dislikes(Sam, Kim) \wedge true) : l_{\langle t^c\rangle}}$$

If, following Potts, we assume there are no functions from *ci* types, then this will ensure that the content of an ARCs will not be in the scope of any normal operator. However, we must still find a way of integrating the *at-issue* and *ci* content, since both are relevant for truth conditions. Recall that Potts (2005) dealt with this by having a principle which collects all *ci* content from the semantic parsetree (cf (35)). Our goal now is to find a resource sensitive way of dealing with this issue.

The simplest approach is to introduce an '!' ('of course') meaning constructor on the root S, which will (iteratively, and non-deterministically) select *ci* propositions and combine them with the main clause content. Suppose, for example, that the 'start' rule for the grammar is associated with a [**root-cp**] meaning constructor, as in (61), where [**root-cp**] is defined as in (62).

(61) $Root \rightarrow \quad CP$
[**root-cp**]

(62) [**root-cp**] $\lambda q.\lambda p.(p \wedge q) : !\,[Some_{t^c} \multimap [\uparrow_\sigma \multimap \uparrow_\sigma]]$

The idea is that this meaning constructor can be applied as many times as required to consume resources of type t^c, 'and'-ing them together and then combining them with the main clause content.

The main question is how the relevant *ci* resources are to be located , i.e. how $Some_{t^c}$ should be defined in (62). This is straightforward. The following functional uncertainty expression, which will pick out meaning resources of type t^c associated with any piece of f-structure, will do the job, giving (64) as the full definition of [**root-cp**].

(63) $(\downarrow GF*)\sigma_{t^c}$

(64) [**root-cp**] $\lambda q.\lambda p.(p \wedge q) : !\,[(\downarrow GF*)\sigma_{t^c} \multimap [\uparrow_\sigma \multimap \uparrow_\sigma]]$

With respect to the example in hand, one solution for the functional uncertainty expression will be the resource of type t^c associated with *Kim, who Sam dislikes*, namely [**Kim-comma-rel-who-Sam-dislikes**], as in (58). We can procede as follows. First [**root-cp**] applies to this, as in (65), the result can then be applied to the *at-issue* content to produce the interpretation of the whole main clause, as in (66).

(65) [**root-cp**] ([**Kim-rel-who-Sam-dislikes**])
$\lambda q.\lambda p.(p \wedge q)(person(Kim) \wedge dislikes(Sam, Kim) \wedge true)$
$\lambda p.(p \wedge person(Kim) \wedge dislikes(Sam, Kim) \wedge true) : [\uparrow_\sigma \multimap \uparrow_\sigma]$

(66) [**root-cp-Kim-rel-who-Sam-dislikes**] ([**left-Kim**])
$\lambda p.(p \wedge person(Kim) \wedge person(Kim) \wedge dislikes(Sam, Kim) \wedge true)(left(Kim))$

$$left(Kim) \wedge person(Kim) \wedge person(Kim) \wedge dislikes(Sam, Kim) \wedge true :\uparrow_\sigma$$

This gives us the desired interpretation for the whole clause, but it deviates from the strict Pottsian approach, because [**root-cp**] must clearly be a function from expressions of type t^c, which are otherwise forbidden. However, since this should be the only such deviation, it is perhaps not problematic.

To summarise: we have shown how a Potts' style approach to ARCs can be implemented in a resource sensitive way, based on the standard LFG analysis of RRCs. Apart from a straightforward, and empirically harmless, reformulation of Dalrymple's treatment of RRC's, we have taken over Potts' type theory, and introduced a new meaning constructor [**comma**], and an inference rule, *ACiS*, to extract *ci* content, and a further new meaning constructor, [**root-cp**] to integrate it. We have the following two rules for RRCs and ARCs respectively:

(67) a. $N' \rightarrow N' \quad CP$
$\downarrow\in (\uparrow ADJ)$
[rel]

b. $NP \rightarrow NP \quad CP$
$\downarrow\in (\uparrow ADJ)$
[rel]
[comma]

The special restrictions on ARCs (viz that they are finite and +WH) can easily be expressed by adding the annotations (↑TENSE), and (↑WH=$_c$+) to (67b). Notice that we can use exactly the same rules to describe the internal structure of ARCs and RRCs.

We can, however, simplify things further. Given that [**comma**] and [**rel**] are associated with the same node in (67b), we can fold them together, to produce something simpler. If we introduce the meaning constructor [**rel**$_{arc}$], as in (68), we can dispense with the trivial (though harmless) *true* conjunct used above, and procede in fewer steps.[9]

[**rel**$_{arc}$] consumes the one-place predicate [**who-Sam-dislikes**], in (69), and produces the tensor resource (70). This in turn consumes [**Kim**] to produce the tensor resource corresponding to [$Kim, person(Kim) \wedge dislike(Sam, Kim)$] exactly as before, see (72).

(68) [**rel**$_{arc}$] $\lambda P.\lambda Z.[Z, P(Z)] : [h_{<e>} \multimap g_{<t>}] \multimap [l_{\langle e\rangle} \multimap [l_{\langle e\rangle} \otimes l_{\langle t^c\rangle}]]$
(69) [**who-Sam-dislikes**]
$\lambda Y.person(Y) \wedge dislike(Sam, Y) : h_{<e>} \multimap g_{<t>}$
(70) [**rel**$_{arc}$**-who-Sam-dislikes**]
$\lambda P.\lambda Z.[Z, P(Z)]\ (\lambda Y.person(Y) \wedge dislike(Sam, Y))$
$\lambda Z.[Z, \lambda Y.person(Y) \wedge dislike(Sam, Y)(Z)]$
$\lambda Z.[Z, person(Z) \wedge dislike(Sam, Z)] : l_{\langle e\rangle} \multimap [l_{\langle e\rangle} \otimes l_{\langle t^c\rangle}]$
(71) [**Kim**] $Kim : l_{\langle e\rangle}$
(72) [**Kim-rel**$_{arc}$**-who-Sam-dislikes**]

[9]Our main reason for introducing this simplification is to simplify the discussion in the following section. It is in no way essential. In fact, on may wonder if it is not retrograde from a theoretical point of view, because it takes us further from Potts' idea of a very general statement of a single operation for type manipulation.

$\lambda Z.[Z, person(Z) \wedge dislike(Sam, Z)](Kim)$
$[Kim, person(Kim) \wedge dislike(Sam, Kim)] : l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}$

This can be split, and the parts combined with the *at-issue* content exactly as before.

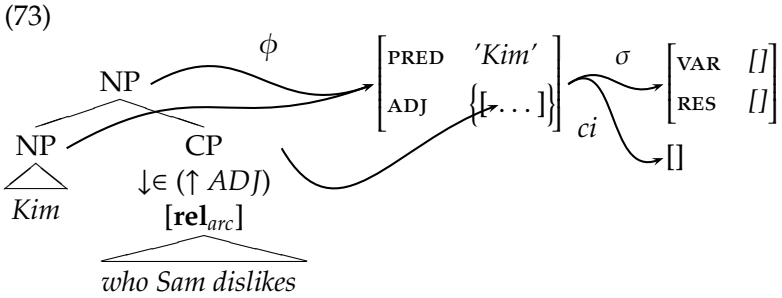
With or without this simplification, this is a promising result — it suggests that we can directly incorporate Potts's approach, and his analyses of particular phenomena, in a straightforward and simple manner into the LFG framework. In the following section we consider an alternative which avoids the multiplication of types characteristic of the Potts' approach.

5 An Alternative Implementation

The previous section built directly on Potts' idea of handling supplementary/appositive meanings by introducing non-standard types, and apparatus for manipulating them.

A potential alternative in the LFG architecture would be to use the projection architecture directly, introducing an additional projection for *ci* content, the idea being that the *ci* content would be distinct only in terms of which projection it is associated with. This will permit us to dispense with Potts' non-standard *ci* types. The architecture which this approach involves is exemplified in (73) — in addition to the σ projection from f-structure to semantic structure (we might now say '*at-issue*' semantic structure), we introduce a parallel projection, the *ci*-projection, from f-structure to '*ci*-structure'. In this section we outline a formalization of this alternative.

(73)



We start from the $[\mathbf{rel}_{arc}]$ resource introduced above and repeated here as (74) for convenience. Recall that this consumes the resource corresponding to a relative clause and produces a function from the meaning of the head NP to the tensor resource:

(74) $[\mathbf{rel}_{arc}]\ \lambda P.\lambda Z.[Z, P(Z)] : [h_{<e>} \multimap g_{<t>}] \multimap [l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}]]$

The new idea is that the *ci* meaning should be associated with a separate projection, hence we replace (74) by (75): (76) spells out the paths to the f-structure in full.

(75) $[\mathbf{rel}_{arc}]\ \lambda P.\lambda Z.[Z, P(Z)] : [h_{<e>} \multimap g_{<t>}] \multimap [l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes \uparrow_{ci}]]$

(76) $[\mathbf{rel}_{arc}]\ \lambda P.\lambda Z.[Z, P(Z)] : [\ (\downarrow \text{RELPRO})_\sigma \multimap \downarrow_\sigma\ ] \multimap [\ \uparrow_\sigma \multimap [\uparrow_\sigma\ \otimes\ \uparrow_{ci}]\ ]$

Careful comparison of (75) with (74) will reveal exactly one difference — where in the former, the second part of the tensor resource is an resource of type t^c, in the latter it is a resource associated with the *ci* projection, $\uparrow_{ci}$. We have not indicated its type, but it is the standard type t. This, with some obvious and straightforwardly related modifications, is almost the only thing we need to change to implement the approach in the projection architecture.

The derivation proceeds as before, only the final result is different. The newly defined meaning constructor $[\mathbf{rel}_{arc}]$ consumes the meaning of *who Sam dislikes* to produce the meaning of an ARC as shown in (79). The resulting meaning constructor is a function from an NP meaning (that of the host NP of the relative clause) to a tensor resource corresponding to this NP meaning and a (propositional) meaning in the *ci* projection.

(77) $[\mathbf{rel}_{arc}]\ \lambda P.\lambda Z.[Z, P(Z)] : [h_{<e>} \multimap\ g_{<t>}] \multimap\ [l_{\langle e\rangle}\ \multimap [l_{\langle e\rangle} \otimes\ \uparrow_{ci}]]$

(78) **[who-Sam-dislikes]**
$\lambda Y.person(Y) \wedge dislike(Sam, Y) : h_{<e>} \multimap\ g_{<t>}$

(79) **[rel$_{arc}$-who-Sam-dislikes]**
$\lambda P.\lambda Z.[Z, P(Z)]\ (\lambda Y.person(Y) \wedge dislike(Sam, Y))$
$\lambda Z.[Z, \lambda Y.person(Y) \wedge dislike(Sam, Y)(Z)]$
$\lambda Z.[Z, person(Z) \wedge dislike(Sam, Z)] : l_{\langle e\rangle}\ \multimap [l_{\langle e\rangle} \otimes\ \uparrow_{ci}]$

(80) **[Kim]** $Kim : l_{\langle e\rangle}$

(81) **[Kim-rel$_{arc}$-who-Sam-dislikes]**
$\lambda Z.[Z, person(Z) \wedge dislike(Sam, Z)](Kim)$
$[Kim, person(Kim) \wedge dislike(Sam, Kim)] : l_{\langle e\rangle} \otimes\ \uparrow_{ci}$

As before, we need to split the tensor resource, so that the resource associated with *Kim* can be consumed separately by composition steps accessing resources in σ structure. The reformulation of the splitting rule is straightforward. The resources we want to split are now characterised by the projection they are associated with, rather than their type (we have changed the name of the rule accordingly):

(82) $$\frac{[M, M'] : R_\sigma \otimes\ R_{ci}}{M : R_\sigma \qquad M' : R_{ci}}\ \text{SCiS (sigma-ci-split)}$$

We also need to be able to re-integrate content from σ- and *ci*-projections. The most straightforward approach involves a trivial reformulation of the **[root-cp]** meaning constructor: what we now want is that **[root-cp]** picks up resources of type t associated with the *ci*-projection of various pieces of f-structure, and 'and's them together with the content of the σ projection:

(83) $[\mathbf{root\text{-}cp}]\ \lambda q.\lambda p.(p \wedge q) : ![(\downarrow GF*)ci_t \multimap\ [\uparrow_\sigma \multimap \uparrow_\sigma]]$

This is a straightforward re-formulation of the definition we gave in the previous section, and we will not repeat the discussion.[10]

[10] To ensure that *ci* content does not appear in the scope of other operators, we must also

However, once we have dispensed with the need to associate *ci* content with Potts' non-standard types, there are other possibilities. In particular, we can avoid the use of the '!' 'of course' meaning constructor. We will briefly outline how.

To begin with, we will introduce the abbreviation $\uparrow_{R\sigma}$ to designate the σ-projection of the root f-structure. It can be defined by the inside-out functional uncertainty expression (84), in which the off-path constraint ensures that the solution to the functional uncertainty statement is not the value of any attribute.[11]

(84) $\left(\neg \overleftarrow{GF}\ GF{*}\ \uparrow \right)_{\sigma}$

Now suppose that instead of being of type t, as we have assumed, *ci* resources are of type $\langle t, t \rangle$: specifically, functions from the σ projection of the root f-structure to the σ projection of the root f-structure (i.e. from $\uparrow_{R\sigma}$ to $\uparrow_{R\sigma}$).

The general form of a *ci* meaning constructor would be (85), where **CI** is the content specific to the particular construction. So, for *Kim, who Sam dislikes* it might be as in (86).

(85) $\lambda p.(p \wedge \mathbf{CI}) : \uparrow_{R\sigma} \multimap \uparrow_{R\sigma}$

(86) **[Kim-rel$_{arc}$-who-Sam-dislikes]**
$\lambda p.(p \wedge person(Kim) \wedge dislike(Sam, Kim)) : \uparrow_{R\sigma} \multimap \uparrow_{R\sigma}$

Since the ordinary (*at-issue*) meaning associated with the root f-structure of *Kim, who Sam dislikes, left* is:

(87) **[left-Kim]** $left(Kim) : \uparrow_{R\sigma}$

The 'full-interpretation', involving reintegration of the *ci* content, will be obtained by applying each *ci* meaning constructor to the normal *at-issue* meaning constructor associated with the root f-structure:

(88) **[Kim-rel$_{arc}$-who-Sam-dislikes]**
$\lambda p.(p \wedge person(Kim) \wedge dislike(Sam, Kim)) : \uparrow_{R\sigma} \multimap \uparrow_{R\sigma}$

(89) **[Kim-rel$_{arc}$-who-Sam-dislikes]** (**[left-Kim]**)
$\lambda p.(p \wedge person(Kim) \wedge dislike(Sam, Kim))(left(Kim))$
$left(Kim) \wedge person(Kim) \wedge dislike(Sam, Kim) : \uparrow_{R\sigma}$

---

stipulate that no ordinary meaning constructor (i.e. other than **[root-cp]**) is permitted to access the *ci*-projection.

[11] We could have used this idea of directly accessing the root f-structure with the Pottsian approach we considered in Section 4.2. If we have b., rather than a., as the glue expression associated with **[comma]**, then the glue expression for **[Kim-comma-rel-who-Sam-dislikes]** will be $[l_{\langle e \rangle} \otimes \uparrow_{R\sigma}]$, and the *ci* content will be directly associated with the root f-structure. Nothing else would need to change.

a. $[[v_{\langle e \rangle} \multimap r_{\langle t \rangle}] \multimap [v_{\langle e \rangle} \multimap r_{\langle t \rangle}]] \multimap [l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes l_{\langle t^c \rangle}]]$

b. $[[v_{\langle e \rangle} \multimap r_{\langle t \rangle}] \multimap [v_{\langle e \rangle} \multimap r_{\langle t \rangle}]] \multimap [l_{\langle e \rangle} \multimap [l_{\langle e \rangle} \otimes \uparrow_{R\sigma}]]$

We will not pursue this here, because associating all *ci* content with the syntactic root is a departure from Potts' approach. It is not clear if it would have empirical consequences.

This resource is available both as the final interpretation of the whole clause, or as input to be consumed by any other *ci* resources that may exist.

6 Conclusion and Further Work

Potts' approach provides an interesting account of the semantics of *inter alia* supplementals (including ARCs) and expressives by introducing additional semantic types alongside the normal types used to capture *at-issue* content. We have explored how the central insights of this approach can be expressed within the LFG formalism. We have suggested two ways: one more or less directly encodes Potts idea, including his non-standard types. Our contribution here has been to fill out Potts' sketch of how his approach could be made compatible with a resource sensitive approach, and show that the idea is workable. We have also suggested an alternative which uses only standard types, and which exploits the projection architecture of LFG. We leave open the question whether there is any empirical basis for favouring one approach over the other.

References

Doug Arnold. Non-Restrictive relatives are not orphans. *Journal of Linguistics*, 43(2):272–309, 2007.

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

Miriam Butt, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. *A Grammar Writer's Cookbook*. CSLI Pubs., Stanford, CA, 1999.

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

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

Christian Fortmann. On Parentheticals (in German). In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG '05 Conference*, Stanford, CA, 2005. CSLI Pubs.

Christian Fortmann. The complement of *verba dicendi* parentheticals. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG '06 Conference*, University of Konstanz, Germany, 2006.

Jesse A. Harris and Christopher Potts. Perspective-shifting with appositives and expressives. *Linguistics and Philosophy*, 32(6):523–552, 2010.

Eric Potsdam. English verbal morphology and VP ellipsis. In *Proceedings of the 27th Meeting of the North East Linguistic Society*, pages 353–368, Amherst, Mass, 1997. GLSA, University of Massachusetts at Amherst.

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

Anthony R. Warner. *English Auxiliaries: Structure and History*. Cambridge University Press, Cambridge, 1993.

PUTTING IT ALL TOGETHER: AGREEMENT, INCORPORATION, COORDINATION AND EXTERNAL POSSESSION IN WUBUY (AUSTRALIA)

Brett Baker*, Kate Horrack**, Rachel Nordlinger*
and Louisa Sadler***

*University of Melbourne, ** University of New England
and *** University of Essex

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010
CSLI Publications
http://www-csli.stanford.edu/

Abstract

In this paper we examine the interaction of a number of grammatical phenomena in Wubuy, a polysynthetic language from northern Australia, and show how they can be given a comprehensive analysis within the framework of LFG. While each of these phenomena – noun incorporation, verbal agreement, coordination and external possession – has received various treatments within the LFG literature, no one study has addressed the compatibility of these analyses under interaction, despite the fact that they frequently co-occur in the world's languages. We use data from Wubuy to showcase the effects of this interaction, and investigate the implications for LFG and for LFG analyses of polysynthetic languages more generally.

1 Introduction

In this paper we examine the interaction of a number of grammatical phenomena in Wubuy, a polysynthetic language from northern Australia, and show how they can be given a comprehensive analysis within the framework of LFG. While each of these phenomena – noun incorporation, verbal agreement, coordination and external possession – has received various treatments within the LFG literature, no one study has addressed the compatibility of these analyses under interaction, despite the fact that they frequently co-occur in the world's languages. We use data from Wubuy to showcase the effects of this interaction, and investigate the implications for LFG. We show how standard LFG treatments of agreement and coordination combine effortlessly with the analysis of incorporation presented in Nordlinger and Sadler (2008) (henceforth NS08) to account for the complex Wubuy data. We also provide an analysis of the external possession construction (building on earlier work in LFG, e.g. Laczkó 1995, 1997, Schrock 2007, Lødrup 2009) that can likewise interact appropriately with the rest of the grammar, providing a single unified account of a range of empirical facts. As well as accounting for the Wubuy data, this work has implications for LFG analyses of polysynthetic languages more generally.

2 Wubuy

Wubuy (previously known as Nunggubuyu (Heath 1980, 1981, 1984)) is an Australian language of the Gunwinyguan family (Alpher, Evans, & Harvey 2003) which also includes Bininj Gunwok, Ngalakgan, Jawoyn, and others. It is spoken as a primary means of communication by adults over the age of 50 in the small remote community of Numbulwar, NT (c. 60 L1 speakers). It has not been fully acquired by children since the 50s, though many children and young adults understand and use it to varying degrees.[1]

[†]We are grateful to the participants at LFG10 in Ottawa, Canada, for insightful and stimulating discussion.

[1]The data reported here come from Brett Baker and Kate Horrack's fieldnotes from fieldwork carried out with speakers in Darwin and Numbulwar between July-Sept 2009 and during Baker's previous field trips. Examples given here should not be further cited without seeking additional permission from Baker, and giving explicit ac-

3 Incorporation in Wubuy: overview

Wubuy, like many polysynthetic languages, allows for productive incorporation of body parts, as shown in the following examples[2] in which we see *-ḻanarr* 'nail' (1), *-yarrga* 'flipper' (2), (3) and *-yirr* 'foliage' (4) incorporated into the verbal word:[3]

(1) *na-ḻanarr ngayawinyinyung nga-ni-ḻanarr-wawayuwaa*
MASC.TOP-nail 1SG.GEN 1SG-3MASC-nail-cut.PC
'I was cutting off my nails (MASC).' IPC

(2) *nga-wu-yarrga-nagiina yii-ngarrugaḷij-*(inyung)*
1SG-NEUT-flipper-cook.PR FEM.OBL-dugong-GEN
'I'm cooking the dugong's (FEM) flipper (NEUT).' IPC

(3) *nga-ngu-yarrga-gambana (ngarra-ngarrugaḷij)*
1SG-3FEM-flipper-roast.PR FEM.TOP-dugong
'I'm roasting the dugong's (FEM) flipper (NEUT).' EPC

(4) *niini-ma-yirr-mangi mana-wuluru mana-ma-manjarr-gadhuwa*
1DUMASC-VEG-foliage-get.PC VEG.TOP-*acacia.sp* VEG.TOP-VEG.REL-leaves-new
'We two (excl.) got new leaves (NEUT) of the acacia species (VEG).' EPC

In fact, incorporated body parts participate in two different construction types, as these examples demonstrate.[4] The relevant distinction is between the **Internal Part-Whole or Possession Construction** (IPC) in which the possessum or part is in construction with the whole (and so the possessor is coded solely as an argument of the possessum), and the **External Whole/Possessor Construction** (EPC) (or Possessor Raising) in which both the whole/possessor and the part/possessum are arguments of the verb. Thus the possessum-possessor relationship can be expressed in two different syntaxes and in both of these it is possible to incorporate the possessum or body part.

knowledgement of the source. Many thanks to Gaḷiḷiwa Nunggarrgalu, Didamain Uibo, Leonie Murrungun, and especially Ginyibuwa Murrungun for sharing their insights into the language.

[2]Unless otherwise specified, all of the examples cited here come from (a subset of) the authors' fieldnotes.

[3]Abbreviations: FEM, MASC, NEUT, RESID, VEG, COLL: noun classes feminine, masculine, neuter, residual, vegetable, collective, TOP: topic form of noun class prefix, OBL: oblique form of noun class prefix, REFL: reflexive, REL: relational ('part') form of noun class prefix (a type of oblique marking), LOC: locative, DAT: dative, GEN: genitive, PC: past continuous, PP: past punctual/present perfective, PR: present, PROX: proximate. In the Wubuy orthography, retroflex consonants are represented with an underscore.

[4]Note that, in common with other Gunwinyguan languages, Wubuy allows noun incorporation into both verbs and adjectives. For ease of presentation, in this paper we will focus on incorporation into verbs. We have shown elsewhere that the analysis of incorporation in NS08 extends naturally to an account of incorporation into adjectives also (see Baker and Nordlinger, 2008).

In the Internal Possession Construction in (1) and (2), the incorporated body part is itself the direct object argument: the verb agrees with it directly (showing MASC or NEUT, respectively, object agreement in this case), and a doubled external NP appears in direct (unmarked) case, as is appropriate for (subjects and) direct objects in Wubuy. In the IPC, the possessor must be marked with the genitive case, as (1) and (2) also demonstrate. Example (5) shows that the IPC construction need not involve noun incorporation:

(5) *anaani* *ana-wanja* *wu-warra-gayiyn* *(ngayawinyinyung)*
NEUT.PROX NEUT.TOP-arm 3NEUT-DUMMY-ache.PP 1SG.GEN
'This arm(NEUT) of mine is aching/sore.' IPC

Examples (3) and (4), on the other hand, exhibit the External Possession Construction, in which the whole (or possessor) is encoded as direct object. This is evidenced by (i) the fact that the object verb agreement (in (3), *-ngu-*) shows noun class agreement with 'dugong' (i.e. FEM) and not 'flipper' (NEUT); and (ii) the lack of genitive/oblique case marking on the external possessor NP, which shows it to be a core argument of the verb. The incorporated body part may be doubled by an external NP, which now must appear in oblique case (as in (4) above, and (6), (7) below) showing it *not* to be an object argument of the verb.[5] Example (8) shows that incorporation of the body part is not obligatory in EPC constructions – but that the external NP expressing the part remains in oblique case irrespective of whether or not it is doubled by an incorporated nominal.

(6) *ngaya nga-laan-barrlhiyn* *yii-laan-duj*
1SG 1SG-knee-sore.REFL.PP MASC.OBL-knee-LOC
'I have sore knee(s)/I am sore in the knee(s)/my knee(s) is/are sore.' EPC

(7) *ngu-warraga-wagiwayn,* *ama-rulbu-rruj*
3FEMSG/3FEMSG-upper.back-hit.PP VEG.OBL-back-LOC
'She hit her in the upper back, in the back.' (Heath (1980, 49 ex. 6.3)) EPC

(8) *ana-ngarrgu* *nga-rang* *a-lhuganda-rruj*
RESID.TOP-'roo 1SG/RESID-spear.PP NEUT.OBL-shin-LOC
'I speared the kangaroo in the lower leg.' EPC

These two incorporation constructions are schematized in (9), in which the bolded elements are those which refer to the part:[6]

[5]The relational noun class marking exhibited on the external nominal *mana-ma-manjarr-gadhuwa* in (4) is a type of oblique marking. We discuss relational noun class marking further in section 6.

[6]Note that, in the interests of clarity, we are focussing only on the incorporation of non-subject arguments in this paper. In fact, Wubuy, like many incorporating languages, allows incorporation of intransitive subjects also, in both IPC and EPC constructions, as illustrated in the following examples.

(9) IPC + NI: SUBJ-AGR - **obj-agr** - **ni-of-part**- VERBSTEM
EPC + NI: SUBJ-AGR - OBJ-AGR - **ni-of-part**- VERBSTEM

Despite the difference in predicate-argument relations, and the morphosyntactic reflexes of this, in both types of incorporation there is no reduction in the valency of the verb (see (1) and (3)). And in both constructions, the incorporated body part can also be doubled by an external noun (see (1) and (6)). We also find external modifiers referring to the incorporated nominal, as in (10) and (11). Thus, both incorporation constructions are clearly of the classifier type (Rosen 1989).

(10) *nga-ni-l̲anarr-wawayuwaa (na-)wulawaa*
1SG-3MASC-toenail-cut.PC MASC.TOP-two
'I cut two toenails (MASC).' IPC

(11) *ngaya anaani nga-l̲anarr-wawayuwiini, ngayajbaj anaani*
1SG PROX 1SG-toenail-cut.REFL.PC, me.myself PROX
'I cut this/these toenail(s).' (lit. 'I cut myself, toenails.') EPC

Table 11 provides a summary of the empirical facts laid out above, showing that both IPC and EPC constructions can be syntactic (i.e. non-incorporated) or contain classifier incorporation (as shown by the possibility of doubling and modification of the incorporated noun).

Part-Whole Expression Type	Syntactic	Incorporate Part	NI+ Doubling	NI + Modify
IPC	(5)	(2)	(1)	(10)
EPC	(8)	(3)	(6)	(11)

Table 1: Part Incorporation construction types

In the remainder of the paper we will show how these two different incorporation constructions and their morphosyntactic properties follow straightforwardly from analyses of classifier noun incorporation (NS08) and external possession (e.g. Schrock 2007) in the LFG literature. In section 6 we also provide an analysis of the complex interaction with coordination. Then in section 7 we provide some initial remarks on the semantics.

(i) *naagi, ni-yarra-wuldhiyn na-yarra*
MASC.PROX 3MASC-nail-cut.PP MASC.TOP-nail
'The nail(s) (MASC) is/are cut.' IPC

(ii) *an'-agalgi nga-ra-yilgiini*
NEUT.TOP-yesterday 1SG-tooth-poke.REFL.PC
'Yesterday I poked my tooth/I poked myself in the tooth (MASC).' EPC

The analysis we present will ultimately need to be extended to include the incorporation of intransitive subjects (e.g. through disjunctions of grammatical functions in the appropriate places). We put this aside for future work, but don't expect it to have any conceptual impact on the basic analysis presented here.

4 Incorporation and Internal Possession Construction

An analysis of incorporation in an IPC construction, as in (12) below, follows straightforwardly from standard LFG treatments of verbal agreement, and the treatment of noun incorporation presented in NS08.[7] The crucial characteristics of this construction are:

- OBJ agreement with the part/possessum
- part/possessum may undergo NI
- possessor/whole appears in an oblique case (genitive)

Consider an example such as (12). The verb involves four morphs: the first element is a subject (agreement) marker indicating that the SUBJ is 1SG. The second element is an object (agreement) marker, indicating that the OBJ is of NEUT gender. The third element is the incorporated nominal stem *yarrga* 'flipper', and the final element is the verbal stem itself. The nominal corresponding to the possessor of the (incorporated) body-part carries a feminine gender prefix (which also marks the noun as having an oblique (i.e. non-direct) case) and is obligatorily marked with GEN case, as a dependent of the (incorporated) body-part.

(12) *nga-wu-yarrga-nagiina* *yii-ngarrugalij-*(inyung)*
1SG-NEUT-flipper-cook.PR FEM.OBL-dugong-GEN
'I'm cooking the dugong's flipper.' IPC

Building on the analysis of classifier incorporation provided by NS08, we assume that the lexical entry associated with the (fully derived and inflected) verb in (12) is that provided in (13), and that the (simplified) f-structure corresponding to the clause is that in (14). Because this is a case of classifier incorporation, the verb maintains its valency (hence the PRED value in the first line of the lexical entry). The incorporate has the grammatical function status of an OBJ but we allow for the OBJ itself to be a set, which allows for doubling (and for coordination). The PRED value of the incorporated nominal (IN) indicates that the incorporated nominal subcategorises for a POSS function (the 'whole').[8]

(13) *nga-wu-yarrga-nagiina*
(↑ PRED) = 'cook<(SUBJ)(OBJ)>' verb maintains its valency

[7] Previous work in LFG on both valence reducing (compounding) and valence preserving (classifier) incorporation includes Ball (2004); Asudeh (2007); Duncan (2007); Nordlinger and Sadler (2008); Baker and Nordlinger (2008) as well as Mohanan (1995); Wescoat (2002) on Hindi and Manning (1996); Bresnan (2001) on West Greenlandic. Discussion of the relationship between our analysis and these alternative approaches is provided in NS08. We build especially on Asudeh (2007) in section 7.

[8] Of course the lexical description in (13) could equally well describe a (monomorphemic) verb with the specific lexical meaning 'cook a flipper': we provide here the full form lexical entry for simplicity but assume that this is the result of some lexical process operating in the morphology to combine the nominal and verbal stems in the case of NI.

$(\uparrow \text{OBJ } (\in)) = \downarrow$	NI
$(\downarrow \text{PRED}) = \text{‘flipper}\langle(\text{POSS})\rangle\text{’}$	a PRED value for the IN
$(\downarrow \text{INDEX PERS}) = 3$	
$(\downarrow \text{INDEX GEND}) = \text{NEUT}$	
$(\uparrow \text{OBJ INDEX GEND}) = \text{NEUT}$	from the OBJ agr marker
$(\uparrow \text{OBJ INDEX PERS}) = 3$	
$(\uparrow \text{SUBJ INDEX PERS}) = 1$	
$(\uparrow \text{SUBJ INDEX NUM}) = \text{SG}$	
$(\uparrow \text{SUBJ PRED}) = \text{‘PRO’}$	

(14)
$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 1 \end{bmatrix} \\ \text{PRED} & \text{‘PRO’} \end{bmatrix} \\ \text{PRED} & \text{‘COOK}\langle(\text{SUBJ})(\text{OBJ})\rangle\text{’} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{‘FLIPPER}\langle(\text{POSS})\rangle\text{’} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{NEUT} \end{bmatrix} \\ \text{POSS} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{FEM} \\ \text{CASE} & \text{GEN} \\ \text{PRED} & \text{‘DUGONG’} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The fact that the lexical entry for the verb optionally constructs a set-valued OBJ allows for the doubling of the incorporated noun in examples like (1) (as per the NS08 apposition analysis), and will also allow for coordination of the incorporated noun with an external noun (see section 6 below).[9]

Thus, using existing analyses of classifier incorporation, we can account for the IPC incorporation constructions without further modification required.

5 External Possession Construction

Recall that in the external possession construction, it is the whole/possessor which is coded as a direct argument of the verb, the incorporated (part) noun can be doubled with an external noun in oblique case (as in (6)), but the whole is doubled by a noun in direct (unmarked) case, as in (15).

[9] As Nordlinger and Sadler (2008, fn 14) note, this allows for two alternative analyses when the OBJ consists of only a single element: one where there is no set, and another in which there is a set with a singleton member. We assume that the latter analysis can be ruled out in the syntax by additional constraints, if required.

(15) *ngarra-ngarrugalij, nga-ngu-yarrga-gambana*
FEM.TOP-dugong, 1SG-3FEM-flipper-roast.PR
'I'm roasting the dugong's flipper.' EPC

The crucial characteristics of this construction are of course a combination of the characteristics of the EPC and NI of body-parts, namely:

- OBJ agreement is with the whole
- the whole appears in direct (unmarked) case
- OBJ agreement with the whole/possessor suggests that *dugong* raises to occupy the grammatical function otherwise associated with the whole phrase *dugong's flipper*
- the part/possessum optionally undergoes NI
- the part/possessum is optionally doubled by a noun with oblique case or in an oblique noun class/case form

5.1 Analysis of EPC in Wubuy

The syntactic part of our analysis of the EPC builds on earlier syntactic analyses in LFG (Laczkó, 1995, 1997; Schrock, 2007; Lødrup, 2009). These accounts treat EPCs as arising from alternative semantic forms for verbs (which satisfy the appropriate semantic restrictions) in which the verb in question is taken to subcategorise for an additional syntactic argument, with structure sharing between the OBJ and the POSS function in the f-structure of the part/possessum. Lødrup (2009) makes explicit the fact that this analyis extends the LFG treatment of control and raising to the nominal domain (using functional control). A key issue which arises for syntactic accounts is whether the OBJ in an EPC construction is thematic or not (neither Schrock (2007) not Lødrup (2009) provide any semantic analysis of the construction). For present purposes we follow Schrock (2007) in assuming that the OBJ in EPC constructions from transitive verbs is non-thematic, so that it is not a semantic argument of the verb itself. But nothing in the analysis hinges on this particular assumption.[10] Lodrup is less explicit about the thematicity of the OBJ in the standard possessor raising construction from transitive verbs, but he is also concerned with the productive possessor raising construction from agentive (unergative) intransitive verbs in Norwegian, and here he notes that the OBJ will be non-thematic.[11]

[10] Although he takes the possessor as non-thematic with respect to the verb, Schrock (2007) also considers that the semantic restrictions on the EPC (namely, that it is usually restricted to inalienable or part/whole possession) may suggest that it may be better to treat both elements as semantic arguments of the verb so that these restrictions can be captured. However, it seems to us that these restrictions on the distribution of the construction should in fact be captured at a more abstract, semantic level.

[11] Under the assumptions of Lexical Mapping Theory, it then follows that for intransitives, possessor raising is restricted to unergatives. The single argument of an unaccusative would be marked [-r]. However (by the assumptions of LMT) a non-thematic argument can only be [-r]. Since languages are generally assumed to have only one intrinsically assigned [-r] argument, intransitive possessor raising will be found only with unergative verbs.

The data that we are primarily concerned with here involves an alternation between an OBJ part-whole construction (IPC) and an EPC in which the whole (possessor) raises to OBJ function while the part or possessum is assigned to a less central function (we are not concerned here with possessor raising from/to SUBJ). The oblique case marking on the unincorporated part noun in Wubuy suggests that OBL is the appropriate grammatical function for the part noun in these Wubuy EPC constructions. Thus, we can capture the syntax of the (OBJ) EPC in Wubuy in terms of a lexical rule that allows the alternatives shown in (16), subject of course to further conditioning restrictions on the class of two place predicates which undergo the alternation.[12]

(16) a. (↑ PRED) = '<(SUBJ)(OBJ)>'

b. (↑ PRED) = '<(SUBJ)(OBL)>(OBJ)'
(↑ OBL POSS) = (↑ OBJ)

The f-structure associated with the EPC construction in (17) is therefore that provided in (18):

(17) *ana-ngarrgu* *nga-rang* *a-lhuganda-rruj*
RESID.TOP-'roo 1SG/RESID-spear.PP NEUT.OBL-shin-LOC
'I speared the kangaroo (RESID.) in the shin (NEUT).'

(18)
$$\begin{bmatrix} \text{OBJ} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{RESID} \\ \text{PRED} & \text{'KANGAROO'} \end{bmatrix} \\ \text{SUBJ} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 1 \\ \text{PRED} & \text{'PRO'} \end{bmatrix} \\ \text{PRED} & \text{'SPEAR<(SUBJ)(OBL)>OBJ'} \\ \text{OBL} & \begin{bmatrix} \text{PRED} & \text{'SHIN<(POSS)>'} \\ \text{CASE} & \text{LOC} \\ \text{POSS} & \end{bmatrix} \end{bmatrix}$$

5.2 Incorporation and External Possession Construction

With this analysis of the EPC in place, the analysis of an EPC combined with nominal incorporation of the part then follows straightforwardly. An example of this construction is given in (19). In this example, the OBJ agreement marker on the verb codes a VEG argument, indicating that the OBJ is *-aalburrunggu* 'turkey' and not the incorporated nominal *-laga-* 'leg'. The information

[12]Ultimately, we assume that the lexical rule itself can be dispensed with given a more articulated view of the syntax-lexicon interface in which lexically governed argument structure alternations are captured using some version of linking theory, but we leave that matter to one side in this paper.

associated with the verb in (19) is provided in (20). The effects of the EPC lexical rule are shown in the first two lines. The effects of NI are the same as those for the IPC construction in (13).

(19) *man'-aa<u>l</u>burrunggu, nga-ma-<u>l</u>aga-wagiwaa*
VEG.TOP-turkey, 1SG-VEG-leg-break.PC
'I broke the turkey's (VEG) legs (NEUT).'

(20) (↑ PRED) = 'break<(SUBJ)(OBL)>(OBJ)'
(↑ OBJ) = (↑ OBL POSS) by EPC
(↑ SUBJ INDEX PERS) = 1
(↑ SUBJ INDEX NUM) = SG
(↑ SUBJ PRED) = 'PRO'
(↑ OBJ INDEX GEND) = VEG
(↑ OBJ INDEX PERS) = 3
(↑ OBL (∈)) = ↓ by NI
(↓PRED) = 'LEG<(POSS)>'
(↓INDEX PERS) = 3
(↓INDEX GEND) = NEUT

(21)
$$\begin{bmatrix} \text{OBJ} & \begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{VEG} \end{bmatrix} \\ \text{PRED} & \text{'TURKEY'} \end{bmatrix} \\ \text{SUBJ} & \begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 1 \end{bmatrix} \\ \text{PRED} & \text{'PRO'} \end{bmatrix} \\ \text{PRED} & \text{'BREAK<(SUBJ)(OBL)>OBJ'} \\ \text{OBL} & \begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{NEUT} \end{bmatrix} \\ \text{PRED} & \text{'LEG<(POSS)>'} \\ \text{POSS} & \text{———} \end{bmatrix} \end{bmatrix}$$

Thus, by way of summary, the EPC construction maps the second argument onto OBL, while the NI construction incorporates the second argument into the verb (independent of whether its function is OBJ or OBL):

EPC maps 2nd argument to OBL
(↑ PRED) = 'break<(SUBJ)(OBL)>(OBJ)' EPC
(↑ OBJ) = (↑ OBL POSS) EPC
NI incorporates 2nd argument into verb (here denoted by GF):
(↑ GF (∈)) = ↓ by NI
(↓PRED) = 'LEG<(POSS)>''
(↓INDEX PERS) = 3
(↓INDEX GEND) = NEUT

As noted above, interesting issues arise concerning how such sub-generalizations over sets of lexical elements should be captured, both in the case of 'possessor raising' and in cases of nominal incorporation. In the latter case, we may assume (as in NS08), that a morphological operation attaches a nominal stem into a verbal stem and adds some f-structure information. There are clear restrictions on the set of nominals which can be incorporated, but we see from the interaction of bodypart NI with both the EPC and the IPC constructions that it is not restricted to a particular GF. Using sublexical trees to schematize the relationship between verbal and nominal stem, we see that at least the following possibilities are attested in Wubuy.

V	⟶	N	V
		(↑ OBL|OBJ(∈)) = ↓	↑ = ↓

6 Interactions with Coordination

NS08 (and also Baker and Nordlinger 2008 (BN08)) note that their analysis of NI allows in principle for an interaction of noun incorporation with coordination — the theory would allow an external (bodypart) NP to be coordinated with the incorporated noun, which heads a GF at f-structure. The two different incorporation constructions (IPC and EPC) predict that, if such coordination is possible, the form of the coordinated external noun should reflect the different functions of the incorporated nominal: in an IPC construction, where the incorporated noun is an OBJ, the external noun should be unmarked. In an EPC construction, in which the incorporated noun is an OBL, the external coordinand should be marked with an oblique case form. In fact, this is exactly what we find in the data.

In (22) (the IPC), the part is the OBJ argument and so coordinates with other direct (unmarked) NPs, despite being incorporated:

(22) *wirri-wudu-miyn, marri andhiri, marri bagalang wirri-ma-ngarrgiwayn*
3PL/3NEUT-liver-get.PP and heart and eye 3PL-3VEG-cut.out.PP
'They got the liver (NEUT), and heart (NEUT), and the eye (VEG) they cut out.' IPC

In the EPC construction in (23), on the other hand, the part is an OBL and so coordination must be with other oblique NPs for the construction to be grammatical. This is shown by the fact that the external part nouns in the following example cannot be in direct (unmarked) form, but must be in relational noun class form, in which part nouns take double noun class prefixation to

agree with the noun class of the possessor. Although these nouns do not take an overt oblique case suffix, we regard these forms as obliques since part nouns in relational noun class cannot control verb agreement (in contrast to part nouns in the IPC construction). There is persuasive evidence that forms in relational noun class form are not OBJ in this construction, as nouns prefixed with relational noun class appear never to occupy direct argument positions. As the following constrasting pair shows, an unmarked NP conjunct would be ungrammatical here.

(23) *man'-aa<u>l</u>burrunggu, nga-m'-a<u>n</u>ja-wagiwaa marri mana-ma-<u>l</u> aga*
VEG.TOP-turkey, 1SG-VEG-arm-break.PC and VEG.TOP-VEG.REL-leg
'I broke the wings (lit. 'arms' NEUT) and the legs (NEUT) of the turkey (VEG).' EPC

(24) **man'-aa<u>l</u>burrunggu, nga-m'-a<u>n</u>ja-wagiwaa marri ana-<u>l</u>aga*
VEG.TOP-turkey, 1SG-VEG-arm-break.PC and NEUT-leg
'I broke the wings (lit. 'arms' NEUT) and the legs (NEUT) of the turkey (VEG).' EPC

The coordination of incorporated body parts with external NPs has received almost no mention in the literature (although Van Geenhoven (1998, p792) provides without further discussion the Greenlandic example in (25) from Sadock (1991, p20)), and would seem to violate many standard accounts of coordination based on constituent structure.

(25) *Marlu-raar-p-u-q affar-mik-lu*
two-catch-IND-[-TR]-3SG half-INS-and
'He caught two and a half.'

However, it follows directly from the interaction of NS08's (appositional) analysis of nominal incorporation, and standard LFG analyses of coordination (e.g. Dalrymple 2001) as developed to accommodate various types of discontinuous coordination in Sadler and Nordlinger (2010).

Example (26) provides the lexical entry for the (first) IPC verb in (22), showing the analysis of the incorporated body part as projecting either the OBJ or a member of the OBJ (NS08). Following assumptions laid out in NS08, external NPs are also annotated with $(\uparrow \text{OBJ } (\in)) = \downarrow$ in the c-structure, resulting in the (partial) f-structure in (27) for the first clause in (22). Note that case agreement amongst the coordinands is enforced by the fact that case is a distributive feature.[13]

(26) *wirri-wudu-miyn*
$(\uparrow \text{PRED}) = \text{'get}\langle(\text{SUBJ})(\text{OBJ})\rangle\text{'}$
$(\uparrow \text{OBJ } (\in)) = \downarrow$
$(\downarrow\text{PRED}) = \text{'liver'}$
$(\downarrow\text{INDEX PERS}) = 3$
$(\downarrow\text{INDEX GEND}) = \text{NEUT}$
$(\uparrow \text{OBJ INDEX GEND}) = \text{NEUT}$

[13]We use NOM here to refer to the direct (unmarked) case that is found on subjects and objects in Wubuy.

(27)
$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{‘PRO’} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{PL} \end{bmatrix} \end{bmatrix} \\ \text{PRED} & \text{‘GET<(SUBJ)(OBJ)>’} \\ \text{OBJ} & \begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{NEUT} \end{bmatrix} \\ \text{CONJ} & \text{AND} \\ & \left\{ \begin{array}{l} \begin{bmatrix} \text{PRED} & \text{‘LIVER’} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{NEUT} \end{bmatrix} \\ \text{CASE} & \text{NOM} \end{bmatrix} \\ \begin{bmatrix} \text{PRED} & \text{‘HEART’} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{NEUT} \end{bmatrix} \\ \text{CASE} & \text{NOM} \end{bmatrix} \end{array} \right\} \end{bmatrix} \end{bmatrix}$$

In the EPC construction, the ‘raised’ possessor is a non-thematic object of the verb, and is identified with the possessor selected by the incorporated nominal (which is itself an OBL) (28). As shown in the associated (partial) f-structure (29), the POSS will distribute appropriately across all members of the coordinated set.

(28) *nga-ma-laga-wagiwaa*
(↑ PRED) = ‘break<(SUBJ)(OBL)>(OBJ)’
(↑ OBJ) = (↑ OBL POSS)
(↑ OBL (∈)) = ↓
(↓PRED) = ‘LEG<(POSS)> ’
(↓INDEX PERS) = 3
(↓INDEX GEND) = NEUT
(↑ OBJ INDEX GEND) = VEG

(29)

$$\begin{bmatrix}
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'TURKEY'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{VEG} \end{bmatrix} \end{bmatrix} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 1 \end{bmatrix} \end{bmatrix} \\
\text{PRED} & \text{'BREAK<(SUBJ) (OBL)>(OBJ)'} \\
\text{OBL} & \begin{bmatrix} \text{INDEX} & [\text{GEND} \quad \text{NEUT}] \\ \left\{ \begin{array}{l} \begin{bmatrix} \text{PRED} & \text{'LEG<(POSS)>'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{NEUT} \end{bmatrix} \\ \text{CASE} & \text{OBL} \\ \text{POSS} & \end{bmatrix} \\ \begin{bmatrix} \text{PRED} & \text{'ARM<(POSS)>'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{GEND} & \text{NEUT} \end{bmatrix} \\ \text{CASE} & \text{OBL} \\ \text{POSS} & \end{bmatrix} \end{array} \right\} \end{bmatrix}
\end{bmatrix}$$

7 Semantics

In this section we provide a preliminary account of how the semantics of Wubuy NI might be handled in LFG, building on both NS08 and in particular on Asudeh (2007), which provides an account of the semantics of non-valency preserving Niuean pseudo-incorporation (see also Van Geenhoven (1998), Farkas and de Swart (2003), Chung and Ladusaw (2003) and Asudeh and Ball (2005)).

The fundamental distinction between compounding (non-valency preserving) incorporation (as found in Niuean) and classifier incorporation (as in Wubuy), is that the latter involves a subcategorised argument of the verb - that is, there is evidence that the IN continues to bear a syntactic grammatical function subcategorised by the verbal stem, rather than being syntactically inert (as is the case in so-called compounding incorporation). The term pseudo-incorporation refers to the fact that Niuean incorporation appears to be syntactic rather than morphological (but this is orthogonal to the semantic treatment). We begin therefore by summarising the approach taken in Asudeh (2007), which uses the resource sensitive glue approach to semantic composition which is standard in LFG (see e.g. Dalrymple, 1999; Asudeh, 2004).

In terms of the syntax, Asudeh (2007) introduces a non-valent GF labelled INCORPORATE. Modifiers of the incorporate will occur freely in the syntax (Niuean does not exhibit doubling). The following illustrates the approach: (31) is the phrase structure rule introducing the the incorporation structure and (32) is the f-structure for the example in (30). Note that the non-thematic INCORPORATE corresponds to a semantic argument of the verb ($(\uparrow_\sigma \text{ARG}) = \downarrow_\sigma$).

(30) *Ne inu kofe a Sione.*
PAST drink coffee ABS Sione
'Sione drank coffee.' (Niuean)

(31) $$\begin{array}{lccc} \text{V}^0 \longrightarrow & \text{V}^0 & \hat{\text{N}} \\ & \uparrow = \downarrow & (\uparrow \text{INCORPORATE}) = \downarrow \\ & & (\uparrow_\sigma \text{ARG}) = \downarrow_\sigma \end{array}$$

(32) $$\begin{bmatrix} \text{INCORPORATE} & \begin{bmatrix} \text{PRED} & \text{'COFFEE'} \end{bmatrix} \\ \text{PRED} & \text{'DRINK}\langle(\text{SUBJ})\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'SIONE'} \\ \text{CASE} & \text{ABS} \end{bmatrix} \end{bmatrix}$$

In this approach, the incorporate is syntactically a non-projecting nominal (as in (33)), and corresponds semantically to a property: the $\hat{\text{N}}$ is derived by lexical rule from a N with no semantic change.

(33) *kofe: $\hat{N}$* $\lambda X.\ \textit{coffee}(X)$: $(\uparrow_\sigma \text{VAR}) \multimap (\uparrow_\sigma \text{RESTR})$

A lexical process converts the unincorporating verb to an incorporating verb, that is, it relates (34) to (35).

(34) *-inu: V* $(\uparrow \text{PRED}) = \text{drink}\langle(\text{SUBJ})\ (\text{OBJ})\rangle$
$\lambda X.\lambda Y.drink(X,Y)$: $(\uparrow \text{SUBJ}_\sigma) \multimap (\uparrow \text{OBJ}_\sigma) \multimap \uparrow_\sigma$

(35) *-inu: V* $(\uparrow \text{PRED}) = \text{drink}\langle(\text{SUBJ})\rangle$
$\lambda P.\lambda X.\exists Y[drink(X,Y) \wedge P(Y)]$:
$[(\uparrow_\sigma \text{ARG VAR}) \multimap (\uparrow_\sigma \text{ARG RESTR})] \multimap$
$[\ (\uparrow \text{SUBJ}_\sigma) \multimap \uparrow_\sigma\ ]$

On the syntactic side, since Niuean NI is non-valency preserving, the syntactic subcategorisation properties of the verb differ in the input and output. In terms of the semantics, the input verbal stem in (34) is associated with the standard meaning constructor for a transitive verb, defining a function from the semantics of one nominal argument (here the SUBJ) to a function from the semantics of the second nominal argument (the OBJ) to the semantics of the sentence

as a whole. On the other hand, the *output* meaning constructor consumes a nominal meaning (that is, a property, rather than an NPmeaning) to create a function from the SUBJ meaning to the meaning of the sentence. It uses existential closure over the properties corresponding to the incorporate and its dependents.

Our preliminary sketch of the interaction of the IPC/EPC with noun incorporation and coordination is based on this (property modification) approach. We proceed step by step, initially abstracting away from the interaction with coordination. Consider an example involving NI (of the bodypart) and the IPC, as in (36).

(36) *nga-wu-yarrga-nagiina* *yii-ngarrugalij-*(inyung)*
1SG-NEUT-flipper-cook.PR FEM.OBL-dugong-DAT
'I'm cooking the dugong's flipper.' IPC

The (relevant sub-part) of the entry for a (non-incorporating) verbal stem in an IPC construction is as in (37), while (38) shows the related verbal stem with an IN: the incorporated nominal is the head of the OBJ in the IPC.[14]

(37) *-nagiina* IPC
(↑ PRED) = 'cook<(SUBJ)(OBJ)>'
$\lambda X.\lambda Y.\ cook(X,Y)$: (↑ SUBJ)$_\sigma$ ⊸ (↑ OBJ)$_\sigma$ ⊸ ↑ $_\sigma$

(38) *-wu-yarrga-nagiina* IPC+NI
(↑ PRED) = 'cook<(SUBJ)(OBJ)>'
$\lambda P.\lambda X.\exists Y[cook(X,Y) \wedge\ P(Y)]$:
[(↑ $_\sigma$ ARG VAR) ⊸ (↑ $_\sigma$ ARG RESTR)] ⊸ (↑ SUBJ)$_\sigma$ ⊸ ↑ $_\sigma$
(↑ OBJ) = ↓
(↓PRED) = 'flipper'
$\lambda X.flipper(X)$: (↓$_\sigma$ VAR) ⊸ (↓$_\sigma$ RESTR)
(↑ $_\sigma$ ARG) = ↓$_\sigma$

The verbal meaning constructor specifies a function from a nominal property (or collection of properties) to a one-place predicate, that is a function from a nominal argument (corresponding to the SUBJ) to the semantics of the sentence as a whole. The NI introduces a nominal property. In the simplest case (that is, where there are no nominal modifiers), the verbal meaning constructor applies directly to the NI meaning constructor, resulting in the meaning constructor in (39).

(39) $\lambda X.\exists Y[cook(X,Y)\ \wedge\ flipper(Y)]$: (↑ SUBJ)$_\sigma$ ⊸ ↑ $_\sigma$

The EPC differs from the IPC in terms of syntax, but shares the same semantic (argument) structure, and therefore the meaning constructor is the same as above (modulo the GF labels

[14]Here we follow Asudeh (2007) in existentially quantifying over the variable associated with the nominal property, but see further below.

associated with the semantic arguments the predicate consumes), as shown in (41). The reentrant f-structure (ie the non-thematic OBJ) is consumed once in producing the semantics of the OBL. Although we do not formulate it here, it is clear that the EPC construction itself is subject to a number of semantic restrictions: a familiar restriction in a number of languages is to restrict the applicability of the EPC to cases of inalienable possession and indeed a version of this restriction essentially limits it to bodyparts in Wubuy. We do not formulate this further restriction here, but assume that in a more complete account the lexical process capturing the IPC-EPC alternation would capture this semantic relation between the POSS (the whole) and its governing PRED (the part).[15]

(40) *nga-ngu-yarrga-gambana (ngarra-ngarrugalij)*
1SG-3FEM-flipper-roast.PR FEM.TOP-dugong
'I'm roasting the dugong's (FEM) flipper (NEUT).' EPC

(41) (↑ PRED) = 'cook<(SUBJ)(OBL)>(OBJ)' EPC
(↑ OBJ) = (↑ OBL POSS)
$\lambda X.\lambda Y.\ cook(X,Y)$: (↑ SUBJ)$_\sigma$ ⊸ (↑ OBL)$_\sigma$ ⊸ ↑ $_\sigma$

The EPC can combine with NI, which incorporates the OBL bodypart. The semantics of the NI is just as described above for the IPC case: the incorporated stem consumes a property (rather than an entity):

(42) *-ngu-yarrga-nagiina* EPC+NI
(↑ PRED) = 'cook<(SUBJ)(OBL)>(OBJ)'
(↑ OBJ) = (↑ OBL POSS)
$\lambda P.\lambda X.\exists Y[cook(X,Y) \wedge P(Y)]$:
[(↑ $_\sigma$ ARG VAR) ⊸ (↑ $_\sigma$ ARG RESTR)] ⊸ (↑ SUBJ)$_\sigma$ ⊸ ↑ $_\sigma$
(↑ OBL) = ↓
(↓PRED) = 'flipper'
$\lambda X.flipper(X)$: (↓$_\sigma$ VAR) ⊸ (↓$_\sigma$ RESTR)
(↑ $_\sigma$ ARG) = ↓$_\sigma$

Again, the result of having the verbal constructor consume the NI meaning directly would be $\lambda X.\exists Y cook(X,Y) \ \wedge \ flipper(Y)$: (↑ SUBJ)$_\sigma$ ⊸ ↑ $_\sigma$. The following table summarises:

[15]Similarly, incorporation itself is subject to significant semantic restrictions of a similar sort, which would be added as additional constraints on the morphological process of NI in a more complete account.

IPC ($\uparrow$ PRED) = 'cook<(SUBJ)(OBJ)>' $\lambda X.\lambda Y.\ cook(X,Y)$: ($\uparrow$ SUBJ)$_\sigma$ $\multimap$ ($\uparrow$ OBJ)$_\sigma$ $\multimap$ $\uparrow_\sigma$
EPC ($\uparrow$ PRED) = 'cook<(SUBJ)(OBL)>(OBJ)' $\lambda X.\lambda Y.\ cook(X,Y)$: ($\uparrow$ SUBJ)$_\sigma$ $\multimap$ ($\uparrow$ OBL)$_\sigma$ $\multimap$ $\uparrow_\sigma$
NI $\lambda P.\lambda X.\exists Y[cook(X,Y) \wedge P(Y)]$:[($\uparrow_\sigma$ ARG VAR) $\multimap$ ($\uparrow_\sigma$ ARG RESTR)] $\multimap$ ($\uparrow$ SUBJ)$_\sigma$ $\multimap$ $\uparrow_\sigma$ ($\uparrow$ OBJ|OBL) = $\downarrow$ ($\uparrow_\sigma$ ARG) = $\downarrow_\sigma$ $\lambda X.flipper(X)$: ($\downarrow_\sigma$ VAR) $\multimap$ ($\downarrow_\sigma$ RESTR)

The account sketched above needs further modification in order to accommodate the fact that, as we have seen in the examples above, the incorporate may potentially be a member of a coordinate structure. Thus the incorporate potentially contributes a member to the set corresponding to the grammatical function in question in the syntax, while the semantics comes from that grammatical function as a whole. The modification required is therefore rather straightforward and is along the following lines:

(43)

Change this:	($\uparrow$ OBL) = $\downarrow$	($\uparrow_\sigma$ ARG) = $\downarrow_\sigma$
Into this:	($\uparrow$ OBL ($\in$)) = $\downarrow$	($\uparrow_\sigma$ ARG) = ($\uparrow$ OBL)$_\sigma$

With this further modification, the EPC-NI would now look as follows :

(44) $\lambda P.\lambda X.\exists Y[cook(X,Y) \wedge P(Y)]$:
[($\uparrow_\sigma$ ARG VAR) $\multimap$ ($\uparrow_\sigma$ ARG RESTR)] $\multimap$ ($\uparrow$ SUBJ)$_\sigma$ $\multimap$ $\uparrow_\sigma$
($\uparrow$ OBL ($\in$)) = $\downarrow$ ($\uparrow_\sigma$ ARG) = ($\uparrow$ OBL)$_\sigma$
$\lambda X.flipper(X)$: ($\downarrow_\sigma$ VAR) $\multimap$ ($\downarrow_\sigma$ RESTR)

This sketch of the lines along which an account of the semantics of the relevant constructions may be developed raises a number of issues for future work. Foremost amongst these are the following:

- The semantics of the verb in cases of NI is such that it combines with a nominal (property) meaning rather than an entity meaning. This will allow for stranded modifiers and appositions, doubling and coordination provided all of these are at the property rather than the entity level. It is an open question whether this property modification view is correct (for example, it will require the nominal coordination constructor to operate below the entity level).

- Related to the above, we have followed Asudeh (2007) here in introducing existential closure at the level of the (lexical) meaning constructor, which ultimately commits us to the existence of a particular entity. This may be incorrect, and also predicts that doubling

by something with entity semantics is impossible (because the slot is already saturated). In their approach, Chung and Ladusaw (2003) assume that binding off is done at the level of the lexical meaning, and bind off the variable much later in the derivation. This approach does offer the advantage of considerably more flexibility, and could be adapted to approach to meaning composition if the data suggests that such a move is justified.

- An alternative (outlined in Asudeh (2007)) involves type-shifting the IN to produce a nominalisation of the property (of type e). The verbal constructor would then be effectively the same as in non-incorporating cases.

(45) $\lambda X.\lambda Y.cook(X,Y)$: $(\uparrow_\sigma \text{ ARG}) \multimap (\uparrow \text{SUBJ})_\sigma \multimap \uparrow_\sigma$
$(\uparrow \text{OBL } (\in)) = \downarrow \quad (\uparrow_\sigma \text{ ARG}) = (\uparrow \text{OBL})_\sigma$
$\lambda X.flipper(X)$: $(\downarrow_\sigma \text{ VAR}) \multimap (\downarrow_\sigma \text{ RESTR})$
$\lambda P.^\cap P$: $[(\downarrow_\sigma \text{ VAR}) \multimap (\downarrow_\sigma \text{ RESTR})] \multimap \downarrow_\sigma$

8 Conclusion

In this paper we have shown how it is possible to provide a comprehensive analysis of bodypart incorporation in Wubuy and in particular how existing analyses of different aspects of the grammar — external possession, incorporation, agreement and coordination — interact to provide a single analysis of the complex empirical facts — including an analysis of the coordination of external NPs with incorporated nominals. Furthermore, we have shown how a semantic analysis can be straightforwardly integrated with our morphosyntactic approach to provide a comprehensive analysis of these different construction types, using the standard glue approach to semantic composition and building on existing proposals for the semantics of incorporation in the literature. Overall, the approach we outline here to the phenomenon of bodypart incorporation in Wubuy highlights the strength of LFG in accommodating typologically diverse languages, and will have important implications for the analysis of polysynthetic languages cross-linguistically.

References

Ackerman, Farrell. 1990. The Morphological Blocking Principle and Oblique Pronominal Incorporation in Hungarian. In P. F. Katarzyna Dziwirek and E. Mejías-Bikandi, eds., *Grammatical Relations: A Cross-Theoretical Perspective*, pages 1–20. Stanford, CA: CSLI Publications.

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

Asudeh, Ash. 2007. Some notes on pseudo noun incorporation in Niuean. Unpublished, Carlton University.

Asudeh, Ash and Douglas Ball. 2005. Niuean incorporated nominals as non-projecting nouns. Presented at LSA, Oakland.

Baker, Brett and Rachel Nordlinger. 2008. Noun-adjective compounds in Gunwinyguan languages. In M. Butt and T. H. King, eds., *Proceedings of the LFG08 Conference*. Stanford, CA: CSLI Publications: http://www-csli.stanford.edu/publications.

Ball, Douglas. 2004. Pseudo noun incorporation and argument structure in Niuean. In M. Butt and T. H. King, eds., *Proceedings of the LFG04 Conference*, pages 1–20. Stanford, CA: CSLI Publications: http://www-csli.stanford.edu/publications.

Barry Alpher, Nicholas Evans and Mark Harvey. 2003. Proto Gunwinyguan verb suffixes. In N. Evans, ed., *The Non-Pama-Nyungan languages of Northern Australia*, pages 305–352. Canberra: Pacific Linguistics.

Bresnan, Joan. 2001. *Lexical Functional Syntax*. Oxford: Blackwell.

Chung, Sandra and William A. Ladusaw. 2003. *Restriction and Saturation*. Boston, MA: MIT Press.

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*, vol. 34 of *Syntax and Semantics*. New York, NY: Academic Press.

Duncan, Lachlan. 2007. Analytic Noun Incorporation in Chuj and K'ichee' Mayan. In M. Butt and T. H. King, eds., *Proceedings of the LFG07 Conference*, pages 163–183. Stanford, CA: CSLI Publications: http://www-csli.stanford.edu/publications.

Farkas, Donka and Henriette de Swart. 2003. *The Semantics of Incorporation*. Stanford, CA: CSLI Publications.

Heath, Jeffrey. 1980. *Nunggubuyu myths and ethnographic texts (NMET)*. Canberra: AIAS.

Heath, Jeffrey. 1981. Aspectual "skewing" in two Australian languages (Mara, Nunggubuyu). In P. Tedeschi and A. Zaenan, eds., *Tense and Aspect [Syntax and Semantics Vol. 14]*, pages 91–102. New York: Academic.

Heath, Jeffrey. 1984. *Functional Grammar of Nunggubuyu*. Canberra: AIAS.

Laczkó, Tibor. 1995. *The Syntax of Hungarian Noun Phrases – A Lexical-Functional Approach*. Frankfurt am Main: Peter Lang GmbH.

Laczkó, Tibor. 1997. Action nominalization and the possessor function within Hungarian and English noun phrases. In *Acta Linguistica Hungarica*, vol. 44, pages 413–75.

Lødrup, Helge. 2009. Looking possessor raising in the mouth: Norwegian possessor raising with unergatives. In *Proceedings of the LFG09 Conference*. Stanford, CA: CSLI Publications: http://www-csli.stanford.edu/publications.

Manning, Christopher D. 1996. *Ergativity: Argument Structure and Grammatical Relations*. Dissertations in Linguistics. Stanford, CA: CSLI Publications. Revised and corrected version of 1994 Stanford University dissertation.

Mohanan, Tara. 1995. Wordhood and lexicality: Noun incorporation in Hindi. *Natural Language & Linguistic Theory* 13(1):75–134.

Nordlinger, Rachel and Louisa Sadler. 2008. From juxtaposition to incorporation: an approach to Generic-Specific constructions. In M. Butt and T. H. King, eds., *Proceedings of the LFG08 Conference*. Stanford, CA: CSLI Publications: http://www-csli.stanford.edu/publications.

Rosen, Sara. 1989. Two types of noun incorporation: a lexical analysis. *Language* 64:294–317.

Sadler, Louisa and Rachel Nordlinger. 2010. Nominal Juxtaposition in Australian languages: An LFG analysis. *Journal of Linguistics* 46:415–452.

Sadock, Jerrold M. 1991. *Autolexical syntax: A theory of parallel grammatical representations*. Chicago: Chicago University Press.

Schrock, Terrill. 2007. Syntactic Exorcism: Toward an LFG analysis of Swahili external possession. Tech. rep.

Van Geenhoven, Veerle. 1998. *Semantic incorporation and indefinite descriptions: semantics and syntactic aspects of noun incorporation in West Greenlandic*. Stanford, CA: CSLI Publications.

Wescoat, Michael. 2002. *On Lexical Sharing*. Ph.D. thesis, Stanford University.

PASHTO (ENDO-)CLITICS IN A PARALLEL ARCHITECTURE

Tina Bögel
Universität Konstanz

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

This paper examines the phenomenon of Pashto (endo-)clitics, which are subject to both prosodic and syntactic contraints. These clitics challenge the view of prosody as being derivative from the syntax (e.g. Selkirk 1984) and the Principle of Lexical Integrity (Bresnan and Mchombo 1995) in that Pashto allows clitics to be inserted into the morphological word. However, these challenges can be resolved by assuming an architecture that views syntax and prosody as independent but interacting dimensions of grammar trying to align with each other as much as possible (see Bögel et al. (2009) for an approach within LFG). This paper presents data showing that it is the prosodic component that must account for the placement of the clitics *within* words, which leads to the conclusion that in cases of misalignment, the prosodic component takes precedence over the syntactic component, although this causes a violation of the Principle of Lexical Integrity.

1 Introduction

This paper examines the phenomenon of Pashto second position (endo)clitics as described by Tegey (1977).[1] Pashto is an Eastern Iranian language spoken in parts of Afghanistan and Pakistan by an estimated 40 million speakers. Clitics are quite common in this language; this paper mainly focuses on one group of second position (2P) clitics that have special properties which challenge the common understanding of the interaction of morphology, syntax and phonology. These clitics are subject to both syntactic and prosodic constraints and different approaches have been developed describing their placement in a clause. In general, Pashto clitics are placed following the first item of a sentence (the verb in (1a)). However, in the context of a stress alternation that accompanies a difference in aspect, this group of 2P clitics can appear as *endo*clitics — clitics that are placed *within* a word as in (1b):

(1)	a.	ṭakwαhə́	me		b.	ṭák	me	wαhə
		shake.IMPERF	I			shake$_1$-	I	-shake$_2$.PERF
		'I was shaking it.'				'I shook it.'		

(Tegey 1977, 92)

Assuming that clitics are syntactic items in their own right, this phenomenon clearly poses a problem to the Principle of Lexical Integrity as stated in Bresnan (2001, 92), in that no syntactic item may intervene in a morphological word.

[†]I would like to thank Miriam Butt, Bariş Kabak, Tracy Holloway King, Astrid Krähenmann, Ghulam Raza and the audiences of the LFG conference and the MFG workshop for their help and their useful comments on the topic.

[1]For a general debate of 2P clitics see Halpern and Zwicky (1996) and references therein.

(2) **Lexical Integrity:**
Morphologically complete words are leaves of the c-structure tree and each leaf corresponds to one and only one c-structure node.

There have been different approaches to solving this problem. However, most of them are insufficient and do not fully account for the data. Thus, in this paper, I first describe the general 2P clitics and the prosodic and syntactic constraints that have to be considered (section 2). I then describe the phenomenon of endoclitics in Pashto and show that these are 2P clitics as well, following yet another prosodic constraint (section 3). In the case of Pashto endoclitics, prosody seems to be able to postlexically place a clitic after an accent-bearing element. The consequence is that prosody appears to be capable of overruling syntax and actually interacts with the morphological word. This phenomenon falls in with the assumption of a parallel architecture as introduced for LFG by Bögel et al. (2009), where prosody and syntax are assumed to be decoupled, but interacting modules of a grammar (following e.g. Zec and Inkelas (1990) and Lahiri and Plank (2009)).[2]

By assuming the syntactic and prosodic components to being parallel, clitics can be viewed as subject to prosodic and syntactic constraints simultaneously. This property is shared by clitics in other languages (e.g., Serbian/Croatian/Bosnian; see Franks and King (2000) for an overview), which seems to be natural, given that clitics have syntactic functions on the one hand and are prosodically deficient items on the other hand. By granting prosody an independent and strong position, I will show in section 5 how a morphological word can be interrupted by a clitic via the satisfaction of prosodic constraints, thus avoiding a violation of the Principle of Lexical Integrity.

2 Pashto Clitics

The group of 2P clitics discussed in this paper involves personal pronouns, modals and adverbials all of which are listed in Table 1:

Weak Pronoun	Num.&Pers.	**Modal**	Translation	**Adverbial**	Translation
me	1. Sg	ba	will, should	xo	really
de	2. Sg	de	should, let	no	then
ye	3. Sg				
am / mo	1. Pl				
am / mo	2. Pl				
ye	3. Pl				

Table 1: Pashto second position clitics and potential endoclitics

All of these clitics behave in the same way with regard to their position. If more than two clitics cooccur, they are placed in a fixed template, shown in (3).

[2]This architecture questions the architectural assumptions that view prosody as derivative of syntax, following the tradition of proposals made by Selkirk (1986) and Nespor and Vogel (1986); see also the criticism of general “syntactocentrism” by Jackendoff (2010).

(3)	1	2	3	4	5	6	7	8	
	xo	ba	am	am/mo	me	de	ye	no	(Roberts 1997, 5)

Note that in position 6, the clitic *de* can be only represented once even though this phonological shape can refer to two different clitics (see Table 1).

2.1 Syntactic Constraints

At first glance, these clitics seem to be common 2P clitics: they are placed after the first word of a sentence. Thus, in (4), the weak pronoun *ye* 'he' follows the noun *angur* 'grapes' while the modal clitic *ba* 'maybe' is placed after the adjective *nαroğa* 'sick' in (5).[3]

(4) angur <u>ye</u> rαwṛə
grapes he brought
'He brought grapes.' (Tegey 1977, 138)

(5) nαroğa <u>ba</u> wi
sick maybe is
'Maybe he is sick.' (Tegey 1977, 84)

However, the host of the clitic is not necessarily the first word of the sentence. Example (6a) shows that the element serving as a host for the clitic can be a syntactic constituent, in this case a coordinated noun phrase. This coordination may not be interrupted ((6b)):

(6) a. [xušαl aw patang]$_{NP}$ <u>ba</u> <u>ye</u> dər ta rαwṛi
Koshal and Patang will it you to bring
'Koshal and Patang will bring it to you.' (Tegey 1977, 84)

b. *[xušαl <u>ba</u> <u>ye</u> aw patang]$_{NP}$ dər ta rαwṛi

The same is true for postpositional phrases, where the clitic is not allowed in between the postposition and its argument:

(7) laylα na <u>de</u> αxistə (*laylα <u>de</u> na αxistə)
Layla from you buy
'You were buying it from Layla.' (Tegey, 1977, 114)

In constructions involving more than one sentence, the clitic may not appear outside of the clause in which it functionally originates, but instead is inserted after the first element therein. Thus, in (8), the two clitics are part of their individual clauses, occupying the second position respectively.

[3]Throughout the text, clitics are underlined.

(8) [tor <u>me</u> wəlidə] [magar [spin <u>me</u> wə nə lidə]]
Tor I saw but Spin I PERF not saw
'I saw Tor, but I didn't see Spin.' (Tegey 1977, 127)

Pashto provides numerous examples for the relatively regular syntactic placement of 2P clitics. However, since this paper is mainly concerned with the prosodic contraints and the interaction between prosody and syntax, the following sections will primarily focus on the influence of prosody on the placement of 2P clitics.

2.2 Prosodic Constraints

Up to this point, the constraints responsible for the positioning of the clitics in the above examples can be more or less explained syntactically. However, prosody plays a crucial part as well, as can be seen in (9) and (10), where the clitic is placed after the first item bearing lexical stress.

(9) rα ta te rα ṭolawə́l <u>de</u>
me for from_it here collect.IMPERF you
'You were collecting them for me from it (and bringing them) here.'
(Tegey 1977, 119)

(10) rα ta pe gαnḍə́ <u>de</u>
me for by_him sew.IMPERF you
'You were having him sew it for me.' (Tegey 1977, 119)

The elements preceding the verb belong to another group of Pashto clitics (Tegey's "Type II clitics"), which are usually placed in front of the verb. These clitics are all prosodically unstressed material, which forces the 2P clitic to appear after the first stressed element at the very right edge of the phrase even though Pashto is a fairly rigid verb-final language. The verb in the above constructions is the first element of the sentence bearing stress and hence the only proper host for the prosodically deficient clitics. Note furthermore that if the verb has constrastive or focus accent, the Type II clitics follow the verb. In these cases, the clitic in question is placed in between the verb and the Type II clitics ((11), cf. (10)) — all other positions are ungrammatical:

(11) gαnḍə́ <u>de</u> rα ta pe (*gαnḍə́ rα ta pe <u>de</u>)
'You were having him *sew* it for me.' (Hock 1996, 235)

Based on these examples one could argue that it is simply the head of the clause that the clitic attaches to. The following examples contradict this hypothesis in that the clitic is clearly attached to the stressed element, even though this element is *not* the head of the clause in (12a) (in contrast to (12b)):

(12) a. rα sará <u>de</u> wi b. rα sara wí <u>de</u>
me with let be me with be let
'Let it be with me.' 'Let it *be* with me.' (Tegey 1977, 121)

The hypothesis that the clitic attaches to the first accent-bearing element of the sentence is also confirmed when it comes to endoclitics. Although endoclitics are most common in simple verb-clitic constructions, they also appear in expressions where every element of the sentence apart from the verb is unstressed. This leads to an alternate version of example (9):

(13) rα ta te rα [ṭól de kṛəl]$_V$
me for from_it here collect$_1$ - you -collect$_2$.PERF
'You collected them for me from it (and brought them) here.'
(Tegey 1977, 119)

Here, the clitic is inserted *into* the verb following the part of the verb that bears the main accent and thus reacting to an verb-internal stress shift that comes along with a change in aspect, the main environment for the phenomenon of endoclisis as described in the following section.

3 Pashto Endoclitics

Like South-Asian languages in general, Pashto is an argument-dropping language (e.g. Butt (2007) and references therein). Sentences can therefore consist of only a verb and a clitic. The endoclitics mainly appear in these short sentences in the context of a stress alternation that accompanies a difference in aspect as in example (1), repeated here for convenience:

(14) ṭák me wαhe
shook$_1$ - I -shook$_2$.PERF
'I shook it.' (Tegey 1977, 92)

In Pashto, the *perfective* aspect of the verb is accompanied by a verb-internal stress shift placing the main stress on the first foot of the verb, while the verb in the *imperfective* aspect carries the main stress on the last foot of the verb. With regard to the stress shift, Pashto verbs fall roughly into three classes, depending on their word-internal structure. Since these structures are essential to the correct placement of the clitics, it is necessary to analyse them more closely in order to find the appropriate (prosodic or syntactic) unit on which the clitics depend. Thus, the different verb classes, their internal characteristics and their behavior concerning the placement of clitics will be introduced below.

3.1 Monomorphemic Class I Verbs

Class I verbs are monomorphemic. In the imperfective, these verbs bear stress on the last foot; the clitic is placed after the verb ((15a)). In the perfective aspect however, class I verbs take on a perfective prefix *wə*- which receives the main stress. In this case, the clitic occurs after the prefix and in front of the stem ((15b)):

(15) a. **imperfective**
təxnawə́la me
tickle I
'I was tickling (her).'

b. **perfective**
wə́ me təxnawəla (*wə́təxnawəla me)
PERF I tickle
'I tickled (her).' (Tegey 1977, 86)

3.2 Bimorphemic Verbs

In contrast to class I verbs, class II and III form the perfective by means of a stress shift from the last to the first foot of the verb without adding a perfective prefix. The verbs of both classes are bimorphemic. Class II verbs consist of a derivational prefix and a root. In the imperfective aspect, the stress is on the second foot of the verb — the clitic is placed after this ((16a)). The perfective is formed via a stress shift from the last to the first foot of the verb. The clitic is then placed after this first foot as in example (16b), i.e. after the derivational prefix.

(16) a. **imperfective**
ṭelwɑhə́ me
push I
'I pushed (it).'

b. **perfective**
ṭél me wɑhə (*télwɑhə me)
PREF I push
'I was pushing (it).' (Tegey 1977, 92)

Class III verbs are complex predicates consisting of a adjective, adverb or noun and a light verb and form the largest group of verbs in Pashto. Their behavior with respect to clitics is the same as with the class II verbs in that there is a verb-internal stress shift that goes along with a change in aspect, and that the clitic will be positioned after the first foot in the perfective ((17)) and after the whole verb in the imperfective.

(17) **perfective**
póx me kə
cook I do
'I cooked (it).' (Tegey 1977, 98)

With class III verbs, one can easily identify the single elements of the word because they are complex predicates. Thus, an analysis in favor of treating all three elements as postlexically independent items seems likely.

With class II verbs on the other hand, the separation of the elements is not as clear-cut, but one could argue that the derivational prefix might itself be a 'lexical word' (Anderson 2005), e.g. a clitic. Assuming that clitics are postlexical elements that occupy separate syntactic nodes, the class II verb in (16b) would thus lead to a c-structure representation similar to Figure 1:

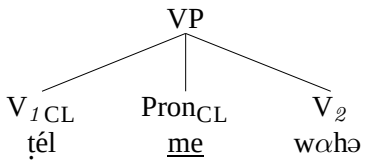


Figure 1: Clitics as postlexical elements

However, there is a group of verbs within class II which do not contain any identifiable derivational prefix.

(18) a. **imperfective**
bαylodə́ me
lose I
'I was losing (it).'

b. **perfective**
bə́αy me lodə
lose1- I -lose2
'I lost (it).' (Tegey 1977, 93)

That is, the element after which the clitic is placed (in the above example *'bαy'*) does not constitute a morpheme with a separate meaning. It is therefore rather difficult to argue in favor of a clitic status of *'bαy'* as in Figure 1, if the morpheme is not identifiable as such and furthermore holds a unique position within the language, i.e. it cannot be found in any other word.

3.3 The Special Class of A-initial Verbs

Apart from the three classes introduced above, there is small group of verbs that can have *alternating* stress in the imperfective, but form the perfective with the perfective prefix of class I (*wə-*), thus adopting properties of all three classes. Within this group, there are verbs that begin with consonants, which do not show any special behavior in the imperfective: even if the stress is on the front vowel, the clitic is placed after the verb.

However, there is a small number of verbs in this group with an initial vowel *a-* which show a very distinct behavior with regard to the alternating stress shift in the imperfective. If the stress falls on the second foot, the clitic is placed after the verb ((19a)). If it falls on the initial vowel *a-* however, the clitic is placed directly after the vowel as in (19b), thus acting like the class of bimorphemic verbs.

(19) a. **imperfective — stress on the *second* foot**
ağustə́ me
wear I
'I was wearing it.' (Tegey 1977, 89)

b. **imperfective — stress on the *first* foot**
á me ğustə
wear1- I -wear2
'I was wearing it.' (Tegey 1977, 89)

Apart from the group of verbs discussed in example (18), these *a*-initial verbs are of special interest, because they cannot be clearly identified as bimorphemic verbs and thus display "real" endoclisis. It has been argued that the *a*- was a prefix/clitic (Kaisse 1981, Anderson 2005) from a diachronic perspective, but this cannot be confirmed for all *a*-verbs[4] — furthermore, syncronically, the initial *a*- does not have a recognizable prefix/morpheme-function, as Tegey explicitly states in his thesis (Tegey 1977, 89). The same can be said of the remainder of each form — *ǧust* and all other "remaining" roots are not identifiable as separate morphemes. Hence, additionally to the group of class II verbs where the clitic is inserted after a morphologically unidentifiable item (as in (18)), we have another group of verbs that poses a problem[5] to a postlexical analysis as in Figure 1 and thus seems to violate the Principle of Lexical Integrity.

4 The Postlexical Status of the Clitics

Instead of assuming a postlexical analysis as in Figure 1, another option would be to consider the clitic as being generated in the lexicon, as a part of the morphological word itself, thus preventing the violation. However, there is evidence supporting the fact that the clitic is inserted into the verb *postlexically*. As has been mentioned before, the *a*-initial verbs take the perfective prefix *wə*- like class I verbs. In contrast to the consonant-initial verbs, however, perfective *a*-verbs display *vowel coalescence*, a process that is part of Lexical Phonology (see the overview in Spencer (1996)). In example (20a), the adjacency of the perfective prefix *wə*- and the initial *a*- results in a fusion: *wα*-. In the event of clitic insertion after the perfective prefix, the fused vowel is still present ((20b)), providing evidence that the clitic has been inserted into the word postlexically, that is: after the lexical process of vowel coalescence.

(20) a. **Vowel coalescence — without clitic**
tə ye wαxla (*wə axla)
you it PERF.buy
'You buy it.' (Tegey 1977, 149)

b. **Vowel coalescence — with clitic**
wα ye xla
PERF.buy$_1$ it buy$_2$
'Buy it.' (Tegey 1977, 163)

Given the above examples, Tegey argues that clitic placement takes place after the process of vowel coalescence:

[4]The verbs, where the *a*-vowel cannot be identified as a prefix diachronically are claimed to have been reanalysed (Kaisse 1981).

[5]In that one cannot assign separate syntactic nodes to the two parts of the verb and the clitic.

data				**process**
wə́	axla		... ye	
wáα	xla		... ye	vowel coalescence
wáα	ye	xla		clitic placement

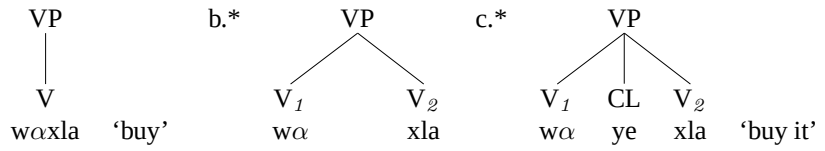
Table 2: Tegey's approach: vowel coalescence before clitic placement

Another argument supporting the analysis of clitics as postlexical elements is that these clitics do not only occur in the context of endoclisis, but act as normal 2P clitics as described in section 2. These clitics exhibit a low degree of selection with respect to their hosts, thus fullfilling a major criterion for a postlexical status (Zwicky and Pullum 1983). The only positional requirement these clitics have to fulfill is the second position. The host of this second position, however, can vary between syntactic constituents and prosodic units as described above. Thus, nobody would question the postlexical status of the 2P *en*clitics — and it seems to be peculiar to describe the same set of clitics as postlexical elements in one context and as lexical in another context, especially since the lexical *endo*clitics can be described as 2P clitics as well: they follow the first accent-bearing foot.

5 Analysis

The lexical process of vowel coalescence described above prohibits the prefix and the verb stem from occupying separate syntactic terminal nodes. That is, they must be viewed as a morphological unit. The monomorphemic verbs thus definitely need to be treated as lexical units and thus can not be represented as in (21b), but are restricted to a representation as in (21a):

(21) a. b.* c.*



The data on vowel coalescence presented in section 4 leads to the assumption that endoclitics have a postlexical status. Assuming that clitics in general are independent lexical items that should occupy a separate syntactic node leads to a conflict in the case of endoclitics: The integration of the clitic occupying a separate syntactic node would force the verb to split illicitly as demonstrated in (21c). Furthermore, the integration of a postlexical element violates the Principle of Lexical Integrity, which states that a syntactic rule must not interfere with a morphological word.

Numerous approaches to solve this problem have been suggested, but most of them do not provide a satisfactory account of the data. In the following section, the major approaches, their advantages and disadvantages will be discussed to see if some of the findings can be taken as a basis for further development.

5.1 Previous Approaches

5.1.1 Prosodic Inversion

Halpern (1995) proposes Prosodic Inversion (PI) for Pashto (and 2P clitics in general). Halpern assumes a basic underlying (syntactic) structure where the enclitic is swapped with the next available host to its right if no host to the left is provided. This analysis is convenient for a theory like LFG, because it allows functional information to be gathered before the clitic is moved into its prosodically determined position (see e.g. Austin and Bresnan 1996, Nordlinger 1998, Bögel et al. 2010). However, the question remains how the clitic ended up in its syntactic position in the first place. Furthermore, in the specific case of Pashto endoclitics, the violation of the Principle of Lexical Integrity is still given in that the clitic still moves into the word. Even though Halpern assumes PI for Pashto endoclitics, describing them as subcategorizing for a metrical foot, he does not comment upon the problem of Lexical Integrity. Thus, PI might be a possible approach if the above mentioned issues can be resolved, but it does not resolve the architectural issues by itself.

5.1.2 A Different View of Architecture

Kaisse (1981), working within generative grammar, attempts to solve the architectural problem by stating that no phonological rule should precede a syntactic rule, thus assuming that the phonological component is placed after the syntactic component. Kaisse views stress assignment as part of the morphological component, marking a category with phonological information in a first step. In a second step, the clitic moves into its position via syntactic movement rules. It is after the clitic placement that the phonological process of *vowel coalescence* takes place. Kaisse thus assumes an architecture like the following:

$$\begin{array}{cl} \text{morphological component} & \rightarrow \text{stress assignment} \\ \downarrow & \\ \text{syntactic component} & \rightarrow \text{clitic placement} \\ \downarrow & \\ \text{phonological component} & \rightarrow \text{vowel coalescence} \end{array}$$

Figure 2: Kaisse's architectural assumption

This architectural view stands in contrast to the assumption that syntax and prosody form two parallel and interacting, but independent modules (e.g. Inkelas and Zec 1990). Furthermore, it does not provide a satisfactory account of the endoclitic phenomenon. In order to avoid the problem of endoclitics, Kaisse claims that Pashto displays no real endoclisis by arguing that all Pashto verbs allowing endoclisis can be described as bimorphemic i.e. as containing a prefix after which the clitic attaches. However, even if this could be verified (but see the argumentation in section 3.2. and 3.3.), the prefixes would still be part of the morphological word — thus there would still be a violation in this approach.

5.1.3 Optimality Theory

van der Leeuw (1997), Roberts (1997) and Anderson (2005) analyse Pashto endoclitics within Optimality Theory (OT) (see e.g. McCarthy 2001). Roberts states that Pashto endoclitics should be viewed in phonological terms only, although he also assumes that phonological phrases are derived from maximal syntactic projections (along the lines of Selkirk (1986)). In his approach principles of OT select the output form. Anderson also assumes the OT constraint ranking ((22)):

(22) Integrity(DP), Integrity(PP), Integrity(PPhrase), NonInitial(cl_i, IP) $>>$ LeftMost(cl_i, IP)

(Anderson 2005, 154)

which reads as: "The clitic is oriented towards the left edge of the IP; however, it must not appear in the initial position and the integrity of the DP, the PP and the Phonological Phrase[6] appearing in this initial position must be preserved". Thus, Anderson assumes the phonological phrase to be the prosodic host for the clitics. He follows Kaisse in that he dismisses the existence of endoclitics, but views all *a*-initial verbs as complex verbs as well.[7] In his approach the status of a phonological word is assigned to any lexical element that bears stress. Phonological phrases are then constructed on the basis of phonological words — allowing phonological phrases to consist of only one phonological word as well. The question remains open as to how this initial verb sequence is analysed in prosodic terms — to consider this first element a "phonological word" or even a "phonological phrase" (Roberts 1997) seems odd, particularly if not even the prefixal status is confirmed (*a*- initial verbs and subgroup of class II verbs). However, even assigning an independent phonological word status to prefixes seems disproportionate, especially since it looses this status so easily if the stress is on the second part of the verb. As an alternative, I suggest that this initial element should be described as "the foot bearing main stress" (along the lines with Kopris and Davis (2005) and Halpern (1995)).

	Roberts (1997)	**Anderson (2005)**	**New suggestion**
prosodic unit	Phonological Phrase	Phonological Word/Phrase	Foot
example	(P-Phr)-cl (P-phr) á <u>me</u> xistələ	(P-Wrd)-cl (P-Wrd) á <u>me</u> xistələ	(<u>x</u>)-cl (x <u>x</u>) á <u>me</u> xistəl(ə)

Table 3: Determining the prosodic host

However, not all approaches to clitics share the belief in prosody as the driving factor. In a later account of Pashto clitics, Roberts (2000) discards his former prosodic approach and claims that the clitics are actually agreement morphemes

[6]Thus accounting for examples like (9), where the unstressed material is ignored by the clitic.

[7]Note that Anderson does not comment on the group of class II verbs that cannot be analysed as complex verbs, but allow endoclitics ((18)). Kaisse mentions them briefly, but in a separate context.

merged into a high position in the clause, reducing the phonological operations to a minimum. The only operation Roberts considers as a "last resort" is Prosodic Inversion to explain the endoclitics.

5.1.4 Word Order Domains/ Topological Fields

Dost (2005) argues strongly against Roberts and proposes an interaction of syntax and prosody to be involved in clitic placement. His approach is based on word order domains/topological fields in combination with a Head-driven Phrase Structure Grammar (HPSG) architecture (Pollard and Sag 1994). His distinction between vertical hierarchy and linear precedence is along the lines of Bögel et al. (2010), but he does not support the idea of Prosodic Inversion and suggests instead that even though the verbs consist of *one* syntactic atom, they remain separable in terms of the word order domain; that is, these verbs contain more than one domain and are thus internally complex. However, the architecture neither makes reference to word-internal feet or stress nor (as Dost himself points out) does it give a full account of what the interaction between the prosodic and the syntactic component should look like. Further research needs to be done in that direction. Dost also emphasizes the fact that the Principle of Lexical Integrity is not violated in a strict sense in that he does not apply *syntactic* processes to interfere with the internal structure of the word, but views clitic placement as resulting from the interaction of syntax, prosody and word order domains — an idea fully supported by this paper.

5.1.5 Lexical Sharing

Another approach that has been lately debated within the LFG community is Lexical Sharing (Wescoat 2002). Lexical Sharing is an application allowing two terminal nodes in the c-structure to share one lexical item. To achieve this, Wescoat assumes that each word is separated from its terminal node and put into a linearly ordered set, the *l-structure*. Wescoat then introduces a structural correspondence between c- and l-structure in the form of a *lexical exponent mapping* λ that generally is a one-to-one mapping between the terminal node and the word, but also allows for two or more terminal nodes to refer to one word. The following formula refers to the f-structure of the lexical component of the current node:

(23) $\varphi(\lambda(*))$, in short: $\Downarrow$

where λ represents the mapping from *l-* to *c-structure* and φ the mapping from *c-* to *f-structure*. A short description of the basic components of the theory is given in Table 4:

l-structure	→ lexical-exponence rules, contribute an independent set of functional descriptions
λ	→ maps the words to the terminal nodes of c-strucure
c-structure	→ two syntactically aligned terminal nodes may share one lexical item
φ	→ transports functional information to f-structure

Table 4: The basic features of Lexical Sharing

Udi Person Markers

Wescoat (2009) applies this theory to Udi person markers (Harris 2002), another form of endoclitics. He views the Udi person markers as "instantiation-altering morphemes" — a word containing a person marker instantiates two terminal nodes. Such an application would, of course, solve the problem posed by the Pashto endoclitics discussed here. Thus, this section takes a close look at Wescoat's application of Lexical Sharing to Udi person markers.

According to his theory, a word can be the lexical exponent of two terminal nodes. Wescoat gives several examples of what these lexical exponence rules should look like; in principle they are ordered as in example (24), where the Udi verb *beɣ-al* 'watch' is combined with an enclitic *le*:

(24) beɣ-al-**<u>le</u>** ←

V	PM
(↓ PRED) = 'WATCH<(↓ SUBJ),(↓ OBJ)>'	(↑ SUBJ) = ↓
(↓ TNS) = FUT	(↓ PERS) = 3
⇓ = ↓	(↓ NMB) = SG

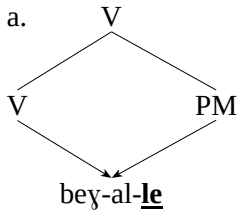
(Wescoat 2009)

Generally clitics are viewed as independent syntactic items that are phonologically dependent on a host; however, Wescoat uses the term "morpheme" to describe the person markers and refers to Harris' alignment constraints based on Optimality Theory to explain the exact position of these clitics. Even though his argumentation seems to be quite straightforward, the question remains whether he views these person markers as being generated in the lexicon as part of the word or as being attached to the word in a later process. If the former is true, then this process would be highly inefficient, since these "morphemes" must be allowed to attach to a large variety of hosts. On the other hand, following the latter assumption, if the clitic is attached to the word later on, it is unclear how it acquired this position, let alone the position *within* the word, without violating the Principle of Lexical Integrity. Still, as mentioned before, Wescoat's analysis is interesting in that it allows two nodes to correspond to one lexical item. Thus I will pursue his appoach a bit further and take a close look on what happens within the syntax.

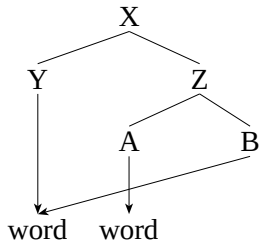
The terminal node PM corresponding to the person marker, which is aligned with the verb by morphological alignment constraints and thus associated with the terminal node V in the lexical exponence rule ((24)), is positioned by syntactic

constraints. It shares a mother node with the terminal node of the verb, leading to a representation as in (25a). The relative order of the terminal nodes must be preserved, avoiding representations as in (25b), where two nodes correspond to one lexical item crossing another lexical item:

(25) a.



*b.



According to his analysis, even though Udi person markers can be found in all kinds of positions in the sentence, their constituent can be adjoined to either S, VP or V, thus allowing for the functional information of the marker to be picked up by the f-structure. Even though it is not deeply relevant to the analysis, it remains unclear how this syntactic placement of the PM terminal node would be treated within c-structure in the case of endoclisis: is it the position preceding or following the verb that is reserved for the terminal node of the endoclitic? And on what grounds is this decision made?

Pashto clitics and Lexical Sharing

After this brief survey of the implementation of Udi person markers, I now turn to Pashto to see if Lexical Sharing can be applied to this phenomenon as well, solving the main problem of allowing two syntactic nodes corresponding to one lexical item.

A lexical exponence rule as used for Udi person markers in (24) could be postulated for Pashto endoclitics as in (26).

(26) verb$_1$ - clitic -verb$_2$ ← V CL
... ...

Assuming that the words including the clitics have to be generated in the lexicon, this approach will lead to an immense number of forms within the lexicon:

1. *Each verb will have to be listed with each clitic as* endo- *and* en*clitic*: As we can see in section 2, the group of clitics discussed in this paper involves 10 different clitics, which would all have to be listed.

2. *Each non-verbal host in the language has to be listed with these clitics as well*: Pashto clitics and second position clitics in general show a high promiscuity with respect to their host; thus, approximately half of the words in the language can act as potential hosts and must be encoded as such within the lexicon.

3. *Each of the word-clitic combinations under 1. and 2. can be combined with any other clitic*: All of the clitics in question can cooccur, thus all possible cooccurences have to be listed in the lexicon as well

The required listing of clitics and their hosts in the lexicon would be a very unsatisfying and inelegant solution. And even if we ignore this and continue with a syntactic representation, we encounter further problems. Since these clitics can possibly cooccur with each other, each one of them can instantiate an independent syntactic node. Given this, one has to consider the template in (3), which forces a certain order on the clitics in question. Using the corresponding terminal nodes, this template would look like (27):

(27)

ADV	MOD	PRON	PRON	PRON	PRON/MOD	PRON	ADV
1	2	3	4	5	6	7	8
xo	ba	am	am/mo	me	de	ye	no

If several clitics are involved in an expression, the order in (27) would have to be maintained in the instantiated terminal nodes. However, it is hard to imagine how the order of terminal nodes in the first line of (27) could be justified within syntax. If syntactic constraints were involved we would not encounter an adverb at the beginning and at the end interrupted by numerous pronouns and modals as sketched in Figure 3:[8]

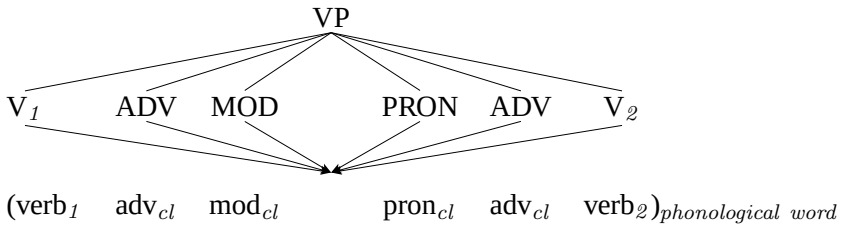


Figure 3: Multiple endoclitics in a verb represented by Lexical Sharing (sketch)

I conclude that the terminal nodes "sharing" one lexical item are obviously not internally ordered by syntactic constraints[9] — their position in the syntax is difficult to justify. This goes along with the fact that the positioning of Pashto (endo)clitics is regulated by prosody, a component that has not been dealt with (yet) within Lexical Sharing.

To sum up, although Lexical Sharing is one possible way of representing one lexical item with two terminal nodes, it cannot be applied to Pashto (endo)clitics for the reasons listed above.

[8] A reordering of the nodes is excluded as well, since this might lead towards the direction of representations as in (25b).

[9] However, if the order of the clitics within the template is on the basis of prosodic/phonological constraints or if the order is 'accidental' is difficult to say. One also could assume that the syntax has a special category for these clitics, e.g. CL_1, CL_2... In this case, the ordering would not be an issue for the c-structure placement.

5.2 Proposed Approach

The evaluation presented above thus excludes certain possibilities of dealing with endoclisis:

1. The clitics and all possible hosts have to be listed in the lexicon — an unsatisfying and inelegant solution.

2. The combination of the word and the clitic cannot instantiate two or more terminal nodes, because the internal ordering of the clitics' terminal nodes is not syntactically justified.

Thus, other approaches have to be taken into consideration. One possible solution is the separation of the linear order of prosodic and syntactic elements. This has been suggested before by proposals like Prosodic Inversion (see section 5.1.1.) and specifically within LFG recently by Bögel et al. (2010). In this approach, the syntactic representation includes the clitics in the first position of the sentence, thus gathering the information for the f-structure from this location. The prosodic representation then determines the position of the clitics as pronounced in an utterance via the application of Prosodic Inversion, placing the clitic in the correct second position as determined by syntactic or prosodic constraints.

However, this approach faces the same problem concerning lexical integrity as this paper in that the movement of a clitic *into* a word causes a violation of the Principle of Lexical Integrity. Thus, a solution to this problem could also be seen as an extension to this approach. In general, it would be desirable to represent the actual prosodic succession of elements in the syntax as well as securing functional information. However, such an approach does not (yet) exist — further research needs to be done. Generally though, the basic concept of viewing prosody and syntax as two independent but interacting components would seem to be just right for Pashto endoclitics.

Hock (1996), also viewing prosody as decoupled from syntax, states that second position clitics should be placed after the first "accented element" of an inital verbal clitic group and after the first accent-bearing constituent elsewhere. With these constraints, one can account for all of the above Pashto examples. Thus, the requirement of a clitic to attach to a host is a strong prosodic requirement. In an architecture that assumes syntax and prosody as interacting, but decoupled dimensions, the placement of Pashto endoclitics can be explained due to prosodic (and not syntactic) constraints. This assumption is represented in the following (rough) architecture, showing the prosodic effects in a parallel architecture with a class II verb ((28)):

(28) ṭél me wɑhə
PREF I push
'I pushed (it).' (Tegey 1977, 92)

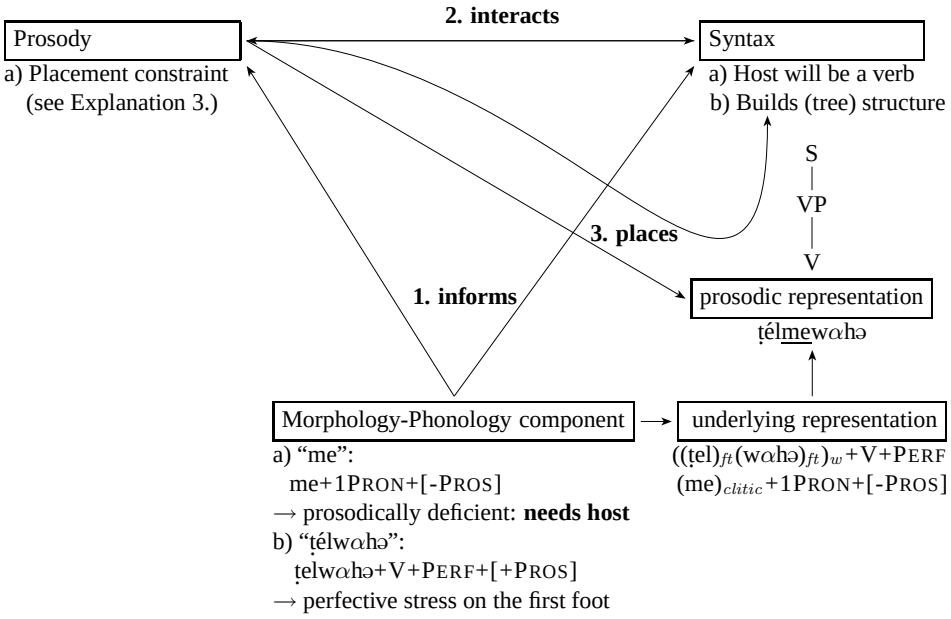


Figure 4: A representation of the parallel architecture

Both lexical items, the verb *ṭelwαhə* 'push' and the clitic *me* 'I' are analysed within the MORPHOLOGY-PHONOLOGY component. Here, the verb receives perfective stress on the first foot because of the perfective aspect ([+PERF]) and is indicated as forming a prosodic word ([+PROS]); the clitic on the other hand is marked as prosodically deficient ([−PROS]). The component then (1.) **informs** syntax and prosody of the properties of the lexical items, i.e. the analysis of the words. The clitic is recognized as a prosodically deficient item ([−PROS]). It needs to be attached to a host. The functional and phonological information of each lexical item is stored in the *underlying representation*. PROSODY and SYNTAX (2.) **interact** in that they share information on structure and intonation. Syntax provides the information that it will be the verb that has to be the host of the clitic. PROSODY is responsible for the placement of the prosodically deficient item. Depending on the host, the clitic is palced in a certain position. Since the host is a verb, the placement is subject to a prosodic constraint: the clitic has to be placed after the first stressed foot of the verb. Prosody thus (3.) **places** the clitic after the first stressed foot of the host, which leads to the *prosodic surface representation*: *ṭélmewαhə*.

Thus, the above architecture tries to represent the "parallel" approach in which a morphological component interfaces with the LFG syntax and an independent prosodic representation. These three components are each independent and governed by independent rules and principles. However, they must also interact and the complex nature of their interaction is brought out nicely by phenomenon such as the Pashto endoclitics, allowing prosody to interact with the lexical word and thus not violate the Principle of Lexical Integrity.

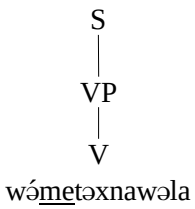
(29) PROSODIC INTERACTION PRINCIPLE:

While syntax may not intervene in the word-internal structure after the morphological word is formed, prosody still has access to the internal structure of the prosodic word (e.g. the footing).

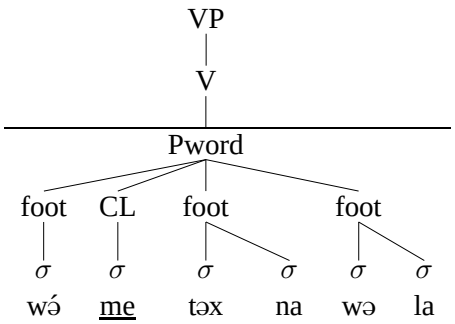
If it is the case that there is no host where a prosodically deficient item can attach to at the right edge, the prosody has the power to overrule the Principle of Lexical Integrity and place the clitic appropriately according to the prosodic structure. In Pashto, this would be after the first accent-bearing foot in a verb-initial clause, leading to a syntactic representation as in (30b) and a prosodic representation as in (30c), where the clitic is inserted after the first foot bearing main accent.

(30) a) wə́ me təxnawəla
PREF I tickle
'I tickled (her).'

b)



c) Proposed tree with syntactic and prosodic structure of (30a).



6 Conclusion

This paper presents the problem of Pashto endoclitics, which challenge established views of the prosody-syntax interface and notions of lexical integrity. Different approaches are evaluated and conclusions are drawn from these evaluations. The result is the proposal of a solution involving an architecture of grammar in which morphology, syntax and prosody are taken to be independent, but interacting modules of grammar. As much as possible, the three components align with one another; however misalignments are also allowed. In particular, prosody is allowed to misalign with syntax when a prosodically deficient item like a clitic needs to be placed in a prosodically appropriate position. In the case of Pashto endoclitics, this is after the first accent-bearing foot. Thus, postlexical prosodic requirements are taken to allow for the placement of material into a morphologically well-formed

and complete word, thus evading a violation of the Principle of Lexical Integrity. However, this paper is only a first proposal in this direction — further research is necessary, especially on the exact nature of the prosody-syntax interface.

References

Anderson, Stephen. 2005. *Aspects of the Theory of Clitics*. Oxford University Press.

Austin, Peter and Bresnan, Joan. 1996. Non-Configurationality in Australian Aboriginal Languages. *Natural Language and Linguistic Theory* 14(2).

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.

Bögel, Tina, Butt, Miriam, Kaplan, Ronald M., King, Tracy Holloway and Maxwell III, John T. 2010. Second Position and the Prosody-Syntax Interface. In *Proceedings of LFG10*, CSLI Publications.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Blackwell.

Bresnan, Joan and Mchombo, Sam. 1995. The Lexical Integrity Principle: Evidence from Bantu. *Natural language and Linguistic Theory* 13(2), 181–254.

Butt, Miriam. 2007. The Role of Pronominal Suffixes in Punjabi. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling and Chris Manning (eds.), *Architecture, Rules, and Preferences*, pages 341–368, CSLI Publications.

Dost, Ascander. 2005. A Domain-Based Approach to 2P Clitics in Pashto. In *Proceedings of the Texas Linguistics Society IX Conference*, CSLI Publications.

Franks, Steven and King, Tracy Holloway. 2000. *A Handbook of Slavic Clitics*. Oxford University Press.

Halpern, Aaron. 1995. *On the Placement and Morphology of Clitics*. CSLI Publication.

Halpern, Aaron and Zwicky, Arnold (eds.). 1996. *Approaching Second: Second Position Clitics and Related Phenomena*. CSLI Publications.

Harris, Alice C. 2002. *Endoclitics and the Origins of Udi Morphosyntax*. Oxford University Press.

Hock, Hans. 1996. Who's on First? Towards a Prosodic Account of P2 Clitics. In Aaron Halpern and Arnold Zwicky (eds.), *Approaching Second: Second Position Clitics and Related Phenomena*, pages 199–270, CSLI Publications.

Inkelas, Sharon and Zec, Draga (eds.). 1990. *The Phonology-Syntax Connection*. CSLI Publications.

Jackendoff, Ray. 2010. *Meaning and the Lexicon*. Oxford University Press.

Kaisse, Ellen M. 1981. Separating Phonology from Syntax: A Reanalysis of Pashto Cliticization. *Journal of Linguistics* 17, 197–208.

Kopris, Craig A. and Davis, Anthony R. 2005. *Endoclitics in Pashto: Implications for Lexical Integrity*. Presented at the Fifth Mediterranean Morphology Meeting, Sept 15-18, Fréjus, France.

Lahiri, Aditi and Plank, Frans. 2009. What Linguistics Universals can be true of. In Sergio Scalise, Elisabetta Magni and Antonietta Bisetto (eds.), *Universals of Language Today*, pages 31–58, Springer.

McCarthy, John. 2001. *A Thematic Guide To Optimality Theory*. Cambridge University Press.

Nespor, Marina and Vogel, Irene. 1986. *Prosodic Phonology*. Foris.

Nordlinger, Rachel. 1998. *Constructive Case: Evidence from Australian Languages*. CSLI Publications.

Pollard, Carl and Sag, Ivan. 1994. *Head-driven Phrase Structure Grammar*. University of Chicago.

Roberts, Taylor. 1997. The Optimal Second Position in Pashto. In Geert Booij and Jeroen van de Weijer (eds.), *Phonology in Progress - Progress in Phonology: HIL Phonology Papers III*, pages 367–401, Holland Academic Press.

Roberts, Taylor. 2000. *Clitics and Agreement*. Ph. D. thesis, Massachusetts Institute of Technology.

Selkirk, Elisabeth. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press.

Selkirk, Elisabeth. 1986. On Derived Domains in Sentence Phonology. *Phonology Yearbook* 3, 371–405.

Spencer, Andrew. 1996. *Phonology*. Blackwell.

Tegey, Habibullah. 1977. *The Grammar of Clitics: Evidence from Pashto and Other Languages*. Ph. D. thesis, University of Illinois.

van der Leeuw, Frank. 1997. *Clitics: Prosodic Studies*. Holland Academic Graphics.

Wescoat, Michael T. 2002. *On Lexical Sharing*. Ph. D. thesis, Stanford University.

Wescoat, Michael T. 2009. Udi Person Markers and Lexical Integrity. In *Proceedings of LFG09*, CSLI Publications.

Zec, Draga and Inkelas, Sharon. 1990. Prosodically Constrained Syntax. In Sharon Inkelas and Draga Zec (eds.), *The Phonology-Syntax Connection*, pages 365–378, CSLI Publications.

Zwicky, Arnold and Pullum, Geoffrey K. 1983. Clitization vs. Inflection: English *n't*. *Language* 59, 502–513.

SECOND POSITION AND THE PROSODY-SYNTAX INTERFACE

Tina Bögel†, Miriam Butt†, Ronald M. Kaplan‡,
Tracy Holloway King‡, and John T. Maxwell III*

Universität Konstanz†, Microsoft Corp.‡, PARC*

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

Bögel et al. (2009) outlined a new architecture for modeling the interaction between prosody and syntax: we proposed an arrangement of interacting components in which prosodic information is developed in a module that operates independently of the syntax while still allowing for syntactic rules and preferences to be conditioned on prosodic boundaries and other features. This architecture allows for misalignments between prosodic units and syntactic constituency, but it incorporates a Principle of Prosodic Preference that causes syntactic structures that do not coincide with prosodic boundaries to be dispreferred. In this paper, we extend the proposal to account for so-called second position clitics. These are clitics that are interpreted syntactically as if they are immediate constituents of a clause, but their appearance after the first prosodic word may embed them in lower constituents and thus insulate them from normal clausal interpretation. We meet this theoretical challenge by adding to the architecture a mathematically restricted "interface mapping" in the form of a regular relation that mediates between the divergent syntactic and prosodic requirements that clitics must jointly satisfy.

1 Introduction

In Bögel et al. (2009), we outlined a new architecture for modeling the interaction between prosody and syntax.[1] As an alternative to the co-description approach first suggested by Butt and King (1998), we proposed an arrangement of interacting components in which prosodic information is developed in a module that operates independently of the syntax while still allowing for syntactic rules and preferences to be conditioned on prosodic boundaries and other features. Under this architecture the terminal string of the syntactic tree (the LFG c-structure) is a sequence of lexical formatives intermixed with additional features that also satisfy the constraints of the prosodic component. This allows for misalignments between prosodic units and syntactic constituency because syntactic rules can ignore prosodic information that would otherwise disrupt syntactic patterns. However, the architecture also incorporates a Principle of Prosodic Preference that causes syntactic structures that do not coincide with prosodic boundaries to be dispreferred.

The present paper explores how this modular prosody-syntax architecture can be extended to apply to an additional class of syntax-prosody misalignments, the second position clitics that appear in many languages. Second position clitics have presented a challenge to many if not all theoretical frameworks — as a consequence, there is a substantial literature on the subject (i.a., Halpern and Zwicky (1996), Franks and King (2000), Anderson (2005) and references therein). The crucial aspects of the problem, from an architectural point of view, are demonstrated by the following example from Serbian/Croatian/Bosnian (SCB):

[1]We would like to thank Damir Ćavar, Steve Franks and Olya Gurevich for discussion of the data and implications for theoretical issues. We would also like to thank Mary Dalrymple for detailed discussion of architectural issues and the LFG10 audience for intensive discussion of our proposal, especially Alex Alsina, Ash Asudeh, Joan Bresnan, Aaron Broadwell, and Rachel Nordlinger.

(1) [Taj *joj ga je* čovek]$_{\text{NP}}$ poklonio.
that her it AUX man presented
'That man presented her with it.' (Halpern 1995, 26)

The clitic sequence *joj ga je* appears as an interruption between the demonstrative *taj* and the head noun *čovek* of what would otherwise be an ordinary initial NP, and those clitics contribute feature and argument information not to the interpretation of the NP that they are contained within but to the clause containing that NP.

Layered on top of these syntactic issues is the interaction with prosody: these clitics can appear in the second position of a prosodic unit without regard to syntactic alignments. This is illustrated by the prosodic bracketing in (2). In this paper, we adopt the usual convention of using square brackets to surround syntactic constituents and smooth parentheses to mark prosodic constituents. The clitics are shown in italics. As per assumptions in the literature, we show the clitics attaching to the prosodic word to their left and thus forming a new prosodic word (e.g., Selkirk (1995), Inkelas (1990)). That is, the clitics form one, larger prosodic word with their host.[2]

(2) (((Taj)$_\omega$ *joj ga je*)$_\omega$ (čovek)$_\omega$)$_p$ (poklonio)$_p$
that her it AUX man presented
'That man presented her with it.'

On the traditional LFG view that the terminal string of the c-structure is a direct representation of the prosodically-ordered sequence of lexical formatives, the problem is first to account for the fact that the string underlying the initial NP (or any other phrasal category) contains otherwise unexpected clitics, and then to arrange for the functional contributions of those NP-internal clitics to affect the clausal f-structure even though that f-structure is not accessible through the usual $\uparrow$ metavariable. The prosody-syntax architecture as originally proposed deals only with misalignments of constituent boundaries and so does not offer an immediate solution to this problem.

The paper is organized as follows. In section 2 we provide an overview of the basic second position patterns that have been discussed in the literature, presenting data from Russian and SCB. In section 3 we recap the modular architecture proposed in Bögel et al. (2009). In the remaining sections we add to this architecture a mathematically restricted "interface mapping" that reconciles the positional requirements that the prosodic component imposes on these clitics with independent syntactic constraints on phrasal organization and functional interpretation.

[2]For our purposes it is irrelevant whether the clitics are represented at the phonological/prosodic level as individual elements or as a clitic group (e.g., see Nespor and Vogel (1986) and Hayes (1989)).

2 Second Position Clitics: Data

In this section we consider data from Russian and SCB to illustrate possible patterns of second position clitic placement.[3] Russian is the simpler case, since the position of the relevant clitics is defined solely in prosodic terms. SCB clitics can appear in either prosodically defined second position (after the first prosodic word) or in positions that can be characterized in syntactic terms (after the first clause-level constituent). There is much controversy in the SCB theoretical linguistics literature as to whether SCB clitic placement is purely prosodic (Radanović-Kocić 1996), purely syntactic (Franks and Progovac 1994, Progovac 1996), or a combination of both (Halpern 1995); for discussion of this controversy and the major arguments for the different approaches, see chapter 10 of Franks and King (2000).[4] In this paper, we show how purely prosodic clitic placement and placement defined by a combination of prosodic and syntactic factors can both be captured within our extended LFG framework; purely syntactic clitic placement can be straightforwardly captured in traditional LFG theory. We leave the exact nature of the SCB data as a matter of continued research.

Prosodic conditioning bears most directly on the architecture for prosody/syntax interaction and is the main focus of this paper. We present a formal analysis of this in section 4 and, for the sake of completeness, we also provide an analysis of clitics that come after the first syntactic constituent in section 5.

2.1 Second Prosodic Word Languages: Russian

The Russian interrogative marker *li* is an example of a clitic that always appears after the first prosodic word in the clause over which it has semantic scope, i.e. the clause whose interrogativity is marked by *li*'s presence.[5] This is shown in (3) where the presence of *li* indicates that the subordinate clause is interrogative. The fact that *li* occurs after the main verb of the subordinate clause results in a neutral reading (King 1994).

[3]Russian and SCB also have simple clitics (in the sense of Zwicky (1977)), e.g., many of the prepositions are simple clitics. Their surface realization poses no issues for LFG theory.

[4]More nuanced differences within SCB apparently exist between language and dialectal varieties. For example, Diesing et al. (2009) suggest that while Serbian allows for both syntactic and prosodic placement, Croatian only allows clitics after the first prosodic word. Damir Ćavar (p.c., September 2010) notes that differences in preference have been observed between Bosnian and Serbian and that differences may exist between standard Croatian and other Croatian dialects. We merely note that according to the current state-of-the-art in the literature, both syntactic and prosodic constraints seem to play a role. Diesing et al. (2009) further report differing preferences on clitic placement depending on whether the host is an argument or a predicate.

[5]The interrogative *li* can be used in matrix clauses as well as in subordinate clauses. However, in Contemporary Standard Russian and in spoken Russian, matrix uses of *li* are rare and often sound stilted. For this reason we use subordinate clauses in our examples. The pattern of clitic placement is identical in matrix and subordinate clauses for those speakers that still use *li* in matrix clauses.

(3) a. Oni ne znajut, rabotaet *li* Maša na ètom zavode.
they not know work Q Masha at this factory
'They don't know whether Masha works at this factory.'

b. Oni ne znajut (((rabotaet)$_\omega$ *li*)$_\omega$ Maša na ètom zavode).

Other items can be the host of *li*. In (4) the subject hosts *li* and the result is the focusing of the subject, as reflected by the clefting in the translation.

(4) a. Oni ne znajut, Maša *li* rabotaet na ètom zavode.
they not know Masha Q work at this factory
'They don't know whether it is Masha who works at this factory.'

b. Oni ne znajut, (((Maša)$_\omega$ *li*)$_\omega$ rabotaet na ètom zavode).

These examples tell us little about the exact positioning of *li* because the clitic appears after items that are both a single prosodic word and a single syntactic constituent. However, if a more complex syntactic constituent appears initially in the subordinate clause, it is the first prosodic word within that constituent that hosts *li*. This is shown in (5) in which the fronted phrase is a complex PP. The preposition and demonstrative form a single prosodic word[6] while the head noun of the NP within the PP forms a separate prosodic word.

(5) a. Oni ne znajut, na ètom *li* zavode rabotaet Maša.
they not know at this Q factory work Masha
'They don't know whether it is at this factory that Masha works.'

b. Oni ne znajut, (((na ètom)$_\omega$ *li*)$_\omega$ zavode rabotaet Maša).

c. Oni ne znajut, [na ètom *li* zavode]$_{\mathrm{PP}}$ rabotaet Maša.

In contrast, example (6) shows that *li* cannot appear after the entire PP.

(6) a. *Oni ne znajut, na ètom zavode *li* rabotaet Maša.
they not know at this factory Q work Masha
'They don't know whether it is at this factory that Masha works.'

b. *Oni ne znajut, ((na ètom)$_\omega$ ((zavode)$_\omega$ *li*)$_\omega$ rabotaet Maša).

c. *Oni ne znajut, [na ètom zavode]$_{\mathrm{PP}}$ *li* rabotaet Maša.

Thus the examples in (5) and (6) form a minimal pair demonstrating that it is prosodic constituency and not syntactic constituency that is important for the placement of *li*. We also note that the interrogativity indicated by the presence of *li* in (5) applies to the entire subordinate clause, even though the linear order places the clitic within the PP.[7]

[6]Most Russian prepositions are proclitics that form a prosodic word with the word to their right.

[7]There is an additional focus on the PP, or a subconstituent of it (King 1994), but the main interrogative scope is the entire clause.

A final constraint on the placement of *li* is that its prosodic host must lie within the subordinate clause. That is, the last word of the matrix clause cannot serve as its host, even though that word may satisfy the prosodic requirements of the enclitic. The different versions in (7) demonstrate that the order of the words in the subordinate clause other than *li* are immaterial for the ungrammaticality. Variations in the order of words in the matrix clause are also immaterial for the ungrammaticality. In (8a) the matrix verb cannot host *li*, while in (8b) the matrix subject cannot host *li*, even though both are full prosodic words.

(7) a. *Oni ne znajut, *li* rabotaet Maša na ètom zavode.
they not know Q work Masha at this factory
'They don't know whether Masha works at this factory.'

b. *Oni ne znajut, *li* Maša rabotaet na ètom zavode.

c. *Oni ne znajut, *li* na ètom zavode rabotaet Maša.

(8) a. *Deti ne znajut, *li* rabotaet Maša na ètom zavode.
children not know Q work Masha at this factory
'The children don't know whether Masha works at this factory.'

b. *Ne znajut deti, *li* rabotaet Maša na ètom zavode.

The ungrammaticality of (8) is explained if the subordinate clause forms a prosodic domain separate from that of the matrix clause, as generally assumed for the prosody-syntax interface (Selkirk 1986, 2001). That is, major syntactic categories such as IP, CP or VP are generally assumed to align with an intonational phrase and constitute an independent prosodic domain.

With respect to Russian, the observation is that *li* must be placed after a prosodic host within a larger prosodic domain such as the intonational phrase corresponding to the CP. This placement constraint is again exemplified in (9).

(9) a. Oni ne znajut, (((rabotaet)$_\omega$ *li*)$_\omega$ Maša na ètom zavode).
they not know work Q Masha at this factory
'They don't know whether Masha works at this factory.'

b. *Oni ne znajut, (*li* rabotaet Maša na ètom zavode).

To summarize, the Russian interrogative clitic *li* must appear after the first prosodic word in its larger prosodic domain. This prosodic word may or may not correspond to a syntactic constituent. This is shown schematically in (10).

(10) a. $\surd$ [main clause] [((xxx)$_\omega$ *li*)$_\omega$ . . .]

b. * [main clause] [(xxx)$_\omega$ ((xxx)$_\omega$ *li*)$_\omega$ . . .]

c. * [main clause] [*li* (xxx)$_\omega$ . . .]

2.2 Second Prosodic or Syntactic Constituent Languages: SCB

Serbian/Croatian/Bosnian (SCB) has been argued to allow clitic placement after either the first prosodic word or the first syntactic constituent in the clause (Halpern 1995).[8] Consider the minimal pair in (11) and (12). In (11) the clitic cluster occurs after the complex subject noun phrase, while in (12) the clitic cluster occurs after the first prosodic word in the subject noun phrase, namely the demonstrative modifier. The scopal interpretation of the clitics is identical in both sentences; that is, they are arguments and auxiliaries of the main verb.

(11) a. Taj čovek *joj ga je* poklonio.
that man her it AUX presented
'That man presented her with it.' (Halpern 1995, 26)

b. [Taj čovek]$_{\mathrm{NP}}$ *joj ga je* poklonio.

(12) a. Taj *joj ga je* čovek poklonio.
That her it AUX man presented
'That man presented her with it.' (Halpern 1995, 26)

b. [Taj *joj ga je* čovek]$_{\mathrm{NP}}$ poklonio.

c. ((Taj)$_{\omega}$ *joj ga je*)$_{\omega}$ čovek poklonio.

The clitic host can also be the verb, as in (13). As verbs are both prosodic words and syntactic constituents ((13b,c)), these constructions do not provide much evidence for conditioning factors. The verb is always the clitic host when it is the only non-clitic in the clause as in (13d), the pro-dropped version of (13a).

(13) a. Poklonio *joj ga je* taj čovek.
presented her it AUX that man
'That man presented her with it.'

b. [Poklonio]$_{\mathrm{V}}$ *joj ga je* taj čovek.

c. ((Poklonio)$_{\omega}$ *joj ga je*)$_{\omega}$ taj čovek.

d. Poklonio *joj ga je*.
presented her it AUX
'(He) presented her with it.'

[8] There are situations in which the clitics appear to be in third position (Ćavar and Wilder 1994a,b). These have largely been analyzed as sentences in which the first constituent is prosodically and syntactically separated from the remainder of the clause (e.g., via topicalization or contrastive focusing via left extraposition). We will not address these examples here, although we believe that our analysis can account for the data if the previous claims are correct about the unusual prosodic and syntactic structure of these sentences. That is, we would expect the topicalized or focused material to make up a prosodic domain that is independent from the main clause. Clitic placement would then proceed independently in each prosodic domain, which Radanović-Kocić (1988) has argued to be within an intonational phrase.

The clitics cannot appear sentence initially, as in (14a). That is, they are enclitics and must have a prosodic host to their left. (14b) provides a version of the sentence in which *joj* has been replaced with the full pronoun *njoj*. In this case there is a prosodic word that the clitics can attach to, and so the example is grammatical.

(14) a. **Joj ga je* taj čovek poklonio.
her it AUX that man presented
'That man presented her with it.'

b. Njoj *ga je* taj čovek poklonio.
her it AUX that man presented
'That man presented her with it.'

The relevant domain for SCB clitic placement is the clause, not the sentence, as was also seen for the Russian interrogative marker *li*. This is shown in (15), where the clitics appear in the subordinate clause, prosodically hosted by the complementizer. Effectively, the complementizer occupies the first position of the clause and the clitics are in second position.

(15) a. Ona tvrdi [da *joj ga je* taj čovek poklonio].
she claims that her it AUX that man presented
'She claims that that man presented her with it.'

b. Ona tvrdi ((da)$_\omega$ *joj ga je*)$_\omega$ taj čovek poklonio.

As seen in (15), in a subordinate clause the complementizer must host the clitics. They cannot be hosted by another prosodic (16a) or syntactic (16b) constituent.

(16) a. *Ona tvrdi [(da)$_\omega$ (taj)$_\omega$ *joj ga je* čovek poklonio].
she claims that that her it AUX man presented
'She claims that that man presented her with it.'

b. *Ona tvrdi [da [taj čovek]$_{\text{NP}}$ *joj ga je* poklonio].

As also illustrated by (15), in embedded clauses the clitics are clause initial in that they immediately follow the complementizer which serves as the prosodic host. Thus the ungrammaticality in (14a) is the result of violating a prosodic, not a syntactic, requirement.

Although clitics may appear in third or even later position in examples involving topicalization or focusing (see fn. 8), clitics generally cannot appear after the second prosodic word in a clause. This is shown in (17).

(17) a. Marija *ga* neće doneti.
Maria it NEG.will bring
Maria won't bring it.

b. *(Marija)$_\omega$ (neće)$_\omega$ *ga* doneti.
Maria NEG.will it bring
Maria won't bring it. (Halpern 1995, 67)

Finally, the SCB clitics generally form a cluster and cannot occur in different positions in the clause, even if those positions are otherwise possible second position sites. This is shown in (18) where some of the clitics are hosted by the first prosodic word while others are hosted by the first syntactic constituent. This generally results in ungrammaticality.[9]

(18) a. *Taj *joj ga* čovek *je* poklonio.
that her it man AUX presented
'That man presented her with it.'

b. *[(Taj)$_\omega$ *joj ga* čovek]$_{\text{NP}}$ *je* poklonio.

c. *[(Taj)$_\omega$ *joj* čovek]$_{\text{NP}}$ *ga je* poklonio.

In summary, SCB clitics can appear after either the first prosodic word or the first syntactic constituent in their clause, and they generally cluster together. These possibilities are outlined in (19).[10]

(19) a. $\surd$ XP *clitics* . . .

b. $\surd$ ((xxx)$_\omega$ *clitics*)$_\omega$ . . .

c. $\surd$ [main clause] [C *clitics* . . .]

d. * *clitics* . . .

e. * XP XP+ *clitics* . . .

f. * [main clause] [C XP *clitics* . . .]

g. * [main clause] [C ((xxx)$_\omega$ *clitics*)$_\omega$ . . .]

h. * . . . *clitics* . . . *clitics* . . .

[9]However, Damir Ćavar (p.c. September 2010) notes that instances of split clitic clusters do exist. Again, these examples involve topicalization, in particular VP-topicalization in which a pronominal clitic remains in the topicalized VP and the auxiliary clitics appear in the main clause.

[10](19e) may be possible with topicalization or contrastive focus fronting; (19h) may be possible when the clitic clusters are in separate clauses or with instances of VP topicalization.

3 Architecture for Syntax-Prosody Interactions

Our analysis of prosodically-conditioned second position clitics is framed within the architecture proposed by Bögel et al. (2009). We review some of its key properties as context for the present discussion.

Bögel et al. (2009) proposed that prosodic information is represented in a component that operates independently of the syntax, thus allowing easy description of misalignment phenomena. We also proposed a simple way of making prosodic information accessible to syntax, so that syntactic rules and preferences can be conditioned on prosodic boundaries. We place the prosodic and syntactic components of the grammar in a configuration such that the terminal string of the syntactic tree is a sequence of lexical formatives intermixed with features determined by the prosodic component. Depending on how they are distributed with respect to syntactic groupings, those features may or may not have an impact on the syntax.

Bögel et al. (2009)'s proposed architecture assigns an extended interpretation to the ordinary rules of a conventional LFG grammar. The effect of this extended interpretation for particular rules is equivalent to including in the grammar some additional rules that are systematically related to the originals. We argued that this is generally the case: the behavior of every syntactic rule according to the proposed architecture can be modeled by a finite expansion to a set of rules that could have been written in standard, pre-existing notations. In other words, the architectural principles in (20) can be implemented as metagrammatical operations that systematically transform the rules of a conventional grammar.

(20) a. An independent prosodic component interprets various prosodic properties to determine the boundaries of prosodic phrases.

b. Prosodic boundaries are visible to the syntax as distinct symbols in the terminal string of the syntactic constituent structure.

c. Prosodic boundary symbols augment but do not disrupt syntactic patterns.

d. The syntactic component obeys a Principle of Prosodic Preference: syntactic structures with constituent boundaries that do not coincide with prosodic boundaries are dispreferred.

As a consequence, this architecture implies no changes to the mathematical and computational properties of the syntactic component.

A conventional LFG grammar contains a set of c-structure rules of the form:

(21) CAT → RHS

where CAT is a nonterminal category and the right-hand side RHS denotes a regular language over categories annotated with functional (or other co-describing) constraints. To implement the architectural specifications, we replace each such rule with another rule of the form:

(22) CAT → (LB) RHS / [LB|RB] (RB)
Disprefer

The prosodic brackets (L(eft)B(racket) and R(ight)B(racket)) belong to the terminal and nonterminal vocabularies of the enlarged grammar, in accordance with (20b). The right-side of the original rule is replaced by a rule expansion which allows for the parsing of prosodic brackets. The categories of the original right-hand side can be optionally preceded by a left prosodic bracket (as indicated by the parentheses) and optionally followed by a right prosodic bracket. In addition, the expansion will match a daughter sequence that would match the RHS regular expression if all occurrences of either LB or RB in that sequence are ignored. The | indicates a disjunction and the / is a notation for the "Ignore" operator first introduced by Kaplan and Kay (1994).[11] The effect of this use of the Ignore operator is to implement property (20c) of the architecture: it ensures that occurrences of prosodic brackets cannot disrupt otherwise valid phrase-structure expansions.

The "Disprefer" annotation implements the Principle of Prosodic Preference (20d). Whenever a prosodic bracket is ignored in the middle of the RHS, the structure is assigned a dispreference optimality mark. The effect of this is to determine a ranking over possible syntactic analyses, as described by Frank et al. (1998). The only brackets that are not dispreferred are those that match the optional LB and RB categories, the ones that appear on the edges of constituents. Replacement rules produced in this way by metagrammatical expansion thus provide dispreferences only for misaligned prosodic brackets, as required.

Bögel et al. (2009) remark that this architectural conception does not depend on the internal details of the prosodic component, but they speculate that the mapping between its inputs and outputs may be regular in nature. If this is the case, then the combination of prosody with an LFG syntax has no more generative capacity than the syntactic module alone, and prosodic mappings can be characterized by well-known and convenient notations for specifying regular relations. Our account of second position clitics is consistent with this speculation but also not dependent on it: we introduce a separate regular relation to define with formal precision an interface mapping that stands between the syntactic and prosodic components.

4 Clitics in Prosodic Second Position

Two issues must be addressed in order to account for clitics in second prosodic position. First, we have to ensure that the functional information carried by the clitics is projected to a clausal f-structure. This is despite the fact that the clitics appear in the prosodic string between units corresponding to lexical items that may not be immediate daughters of the syntactic clause. Second, we have to allow for the clitics to be properly placed in the prosodic string, after the clause-initial prosodic

[11] It is included in the Xerox finite-state machine calculus (Beesley and Karttunen 2003) and in the c-structure notation of the XLE system (Crouch et al. 2010).

word. In our analysis the syntactic and prosodic components have a shared responsibility: the syntactic component deals with the clausal scope of functional information while the prosodic component provides for proper placement.

Clitics would naturally have clausal functional scope if they appeared as immediate daughters of the clause node in the syntactic c-structure. This can be achieved by a simple extension of the c-structure rule that derives the normal patterns of clausal daughter sequences, as schematized in (23).

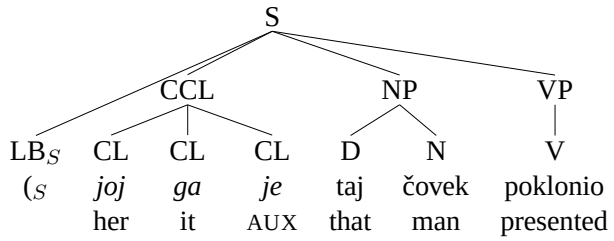
(23) $\mathrm{S} \rightarrow \mathrm{RHS}_S$

RHS_S denotes the possible expansions of the clausal S node, with left and right prosodic brackets possibly ignored or dispreferred according to the metarule convention in (22). We replace (23) by a rule that allows for clausally-scoped clitic clusters (CCL) to appear optionally as prefixes of normal S expansions, as in (24).

(24) $\mathrm{S} \rightarrow \mathrm{LB}_S \quad \underset{\uparrow=\downarrow}{(\ \mathrm{CCL}\ )} \quad \mathrm{RHS}_S$

LB_S is a distinguished pre-terminal that marks the left edge of clauses and enables syntactic and prosodic constraints to be aligned with respect to clause boundaries. CCL expands to the set of clitic sequences that can appear together in second position, for example the singleton *li* in Russian or the sequence *joj ga je* for SCB.[12] The ↑=↓ annotation provides for the clause-level functional scope of the clitics, as required. The tree in (25) is the c-structure that this rule assigns to our core SCB example *Taj joj ga je čovek poklonio*.

(25)

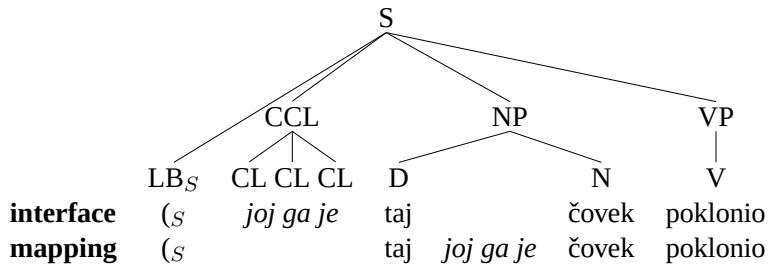


This syntactic configuration does not reflect the position of the clitics as attested in the prosodic representation. Due to prosodically determined constraints, the clitics are realized in second prosodic position. They cannot be realized at the beginning of a matrix sentence (because there is nothing prosodically heavy enough in front), and the clausal boundary LB_S is a barrier that prevents the initial clitics of a subordinate clause from attaching to a preceding matrix word. We resolve the conflict between these inconsistent requirements by distinguishing more carefully between the string of prosodic elements and the formatives of the c-structure terminal string. Rather than assuming the exact identity of these two

[12]We assume that the particular clitic sequences are licensed by clitic-specific rules or templates that express appropriate generalizations. These details are not relevant to the present discussion.

representations, as we earlier proposed, we now add to the Bögel et al. (2009) architecture an explicit mapping as an interface between the syntactic and prosodic components. This interface mapping defines a correspondence between representations that are typically identical, as before, but it also allows for a limited amount of misalignment. Specifically, it provides a correlation between the first-position clitics in the c-structure terminal string with their attested realization after the first prosodic word.[13] Diagram (26) illustrates this mapping by showing the c-structure terminal string and the corresponding prosodic string as separate representations.

(26)



Since the clitics are drawn from a given set of lexical/prosodic formatives and since they cluster according to a fixed set of patterns, we know that there are only a finite number of clitic sequences that are subject to the interface mapping. This fact enables us to provide a characterization of the mapping within the formal space of regular relations. Let

(27) $\mathrm{CS} = \{\mathrm{CS}_1, \mathrm{CS}_2, \ldots, \mathrm{CS}_n\}$

denote the finite set of clitic sequences, the lexical/prosodic sequences that can be realizations of the CCL category. For SCB the sequence CS_1 might be the string *joj ga je*. Also let *W* stand for any prosodic word, presumably marked by distinc-

[13]Our proposal that the clitics appear in clause-initial syntactic position but second prosodic position is consistent with the intuition behind *Prosodic Inversion* (Halpern 1995). According to Prosodic Inversion, syntactically clause-initial clitics are realized in second prosodic position because they must attach to a preceding prosodic word. Halpern does not as clearly separate the two levels of representation or characterize the formal properties of the inversion mechanism. He also does not embed his clitic proposal in a general architectural framework for syntax/prosody interactions.

Our proposal contrasts starkly with assumptions in the transformational literature that 2P clitics are moved from the particular positions where full form equivalents of the clitics may be base-generated (see Klavans (1982) for an overview of early strategies). Movement is needed by transformational theories to account for the fact that clitics do not co-occur with full forms. This is not an issue for our analysis because the Uniqueness Condition of LFG does not allow PRED semantic forms coming from different phrasal positions to fill the same grammatical function.

Our approach also contrasts with the LFG-based lexical-sharing analysis of Wescoat (2002). He relies on lexical rules to attach sequences of clitics as suffixes to all words in the lexicon, and then depends on a correlated family of modified c-structure rules that anticipate the categories of those clitics in fronted syntactic position. His solution thus treats second position clitics as a purely syntactic/lexical phenomenon that operates without reference to independent prosodic generalizations.

tive prosodic-word brackets.[14] Then the interface mapping is the regular relation denoted by the following expression:

(28) $[\ \Sigma^*\ (\ \bigcup_{CS_i}\ [\ (_S\ \ CS_i{:}0\ \ W\ \ 0{:}CS_i\ ]\ )\ \Sigma^*\ ]^*$

In this traditional notation, the term Σ^* stands for any number of prosodic items. According to this relation, a clitic cluster appearing at the beginning of any c-structure clause (as indicated by the $(_S$ symbol visible to both prosody and syntax) may be treated by one of the expressions inside the optional union. The term CS_i:0 indicates that there is nothing (denoted by 0) on the prosodic side of the map corresponding to a particular cluster on the syntactic side (say CS_1=*joj ga je* of our SCB example). The following prosodic word *W* is unchanged in the mapping. After that word the term 0:CS_i indicates that the same i^{th} cluster appears on the prosodic side corresponding to nothing on the syntactic side. The effect is that strings with syntactically clause-initial clitic sequences are related to strings where those particular clusters appear on the other side of an adjacent word. The optionality of the union and the final asterisk allow for any number of such correlations to occur in a given sentence.

The finiteness of the set of possible clitic sequences is crucial for restricting the mathematical complexity of this analysis: correspondences between the elements of an unbounded set would require formal power lying beyond the capacity of regular relations.

Thus on our account the placement and interpretation of clitic clusters follows from the interaction of syntax and prosody: rule (24) restricts clitics so that they can appear and be functionally interpreted only at the beginnings of syntactic clauses, relation (28) provides for optional misalignment around the first prosodic word, and the misalignment becomes mandatory because clause-initial enclitics would otherwise lack a prosodic host. It is important to note that it is not accidental that our rule (24) generates the clitics in clause-initial position as opposed to some other syntactic positions. This is a necessary consequence of the fact that the clitics are realized in second prosodic position but have clausal functional scope together with our hypothesis that the mapping between syntactic and prosodic representations is characterized as a regular relation. Regular relations are mathematically restricted devices that cannot be sensitive to recursive structure, are therefore unable to identify the right boundary of initial syntactic constituents with possibly deep embeddings, and therefore cannot reliably define correspondences between clitics in prosodic second position and any clause-level positions further to the right. For the same reason a second position clitic cannot directly constrain the f-structures corresponding to any rightward NPs or other sub-clausal constituents.

[14] If a language had a different prosodic constraint, e.g., it placed clitics after the first prosodic phrase, then one could similarly define a *P*, which would stand for any prosodic phrase, etc.

5 Clitics in Syntactic Second Position

The relation in (28) defines a correspondence only for clause-initial clitic sequences. Clusters not immediately preceded by $(_S$ will match against the Σ^*s, and their syntactic and prosodic positions will be the same. In particular, clitics that come after the first syntactic constituent (as in (11a), repeated as (29)) are not displaced by this relation.

(29) [Taj človek]$_{\text{NP}}$ *joj ga je* poklonio.
that man her it AUX presented
'That man presented her with it.'

We require an extension to the c-structure rule for S to ensure that clitics not realized after the first prosodic word can appear in the second syntactic position and only in that position. We modify the S rule to allow also for the optional appearance of clitic clusters in the second position of every acceptable sequence of clausal daughters. This is specified schematically in (30).

(30) $\text{S} \rightarrow \text{LB}_S \;\; [\; \underset{\uparrow=\downarrow}{(\text{CCL})\ \text{RHS}_S} \mid \underset{\uparrow=\downarrow}{\text{Second}(\text{RHS}_S, \text{CCL})} \;]$

The expression Second(x, y) indicates the insertion of y in the second position of every string in a regular language x.[15] The schematic rule (30) thus allows c-structure clusters in either clause-initial or clause-second position, but not both. This means that clusters in a single clause cannot be split across the two positions, so that ungrammatical strings such as (18) are not possible. Since the category CCL is not found in any other phrase-structure rule, these are the only environments in which clitics can appear. If RHS$_S$ includes a NP–VP sequence as one of its expansions, the daughter paths in (31) will be included as instances of rule (30).

(31) a. $\text{LB}_S \quad \underset{(\uparrow \text{SUBJ})=\downarrow}{\text{NP}} \quad \underset{\uparrow=\downarrow}{\text{VP}}$

b. $\text{LB}_S \quad \underset{\uparrow=\downarrow}{\text{CCL}} \quad \underset{(\uparrow \text{SUBJ})=\downarrow}{\text{NP}} \quad \underset{\uparrow=\downarrow}{\text{VP}}$

c. $\text{LB}_S \quad \underset{(\uparrow \text{SUBJ})=\downarrow}{\text{NP}} \quad \underset{\uparrow=\downarrow}{\text{CCL}} \quad \underset{\uparrow=\downarrow}{\text{VP}}$

[15] Second(x, y) can be easily implemented by means of the ignore and intersection regular expression operators:

$$\text{Second}(x, y) \equiv x/y \cap [[\Sigma - y]\ y\ [\Sigma - y]^*]$$

The first term introduces the possibility of y appearing anywhere in the middle of the strings in the language x while the other term imposes the restriction that only occurrences after the first element of each of those strings are permitted.

Thus, languages like Russian which only have the prosodic placement option will have the rule in (24), while languages like SCB with a combination of prosodic and syntactic clitic placement will have the rule in (30). The rules we have used to account for second position clitic placement in LFG are repeated in (32).

(32) a. Prosodic Second Position:

$$\text{S} \rightarrow \text{LB}_S \quad \underset{\uparrow=\downarrow}{(\text{CCL})} \quad \text{RHS}_S$$

b. Prosodic or Syntactic Second Position:

$$\text{S} \rightarrow \text{LB}_S \quad [\ \underset{\uparrow=\downarrow}{(\text{CCL})}\ \text{RHS}_S \mid \text{Second}(\text{RHS}_S, \underset{\uparrow=\downarrow}{\text{CCL}})\ ]$$

c. Interface Mapping:

$$[\ \Sigma^*\ (\ \bigcup_{\text{CS}_i}\ [\ (_S\ \ \text{CS}_i{:}0\ \ W\ \ 0{:}\text{CS}_i\ ]\ )\ \Sigma^*\ ]^*$$

6 Examples

In this section we illustrate how our proposed architecture and the augmented rule set apply to some of the SCB data. In particular, we consider the three situations which any analysis needs to account for: no clitic cluster is present; the clitic cluster is prosodically placed; the clitic cluster is syntactically placed.

6.1 No Clitic Cluster

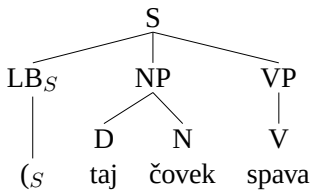
First consider the case where there are no clausal clitics, as in (33).

(33) a. Taj čovek spava.
that man sleeps
'That man sleeps.'

b. $\text{LB}_S \quad \underset{(\uparrow \text{SUBJ})=\downarrow}{\text{NP}} \quad \underset{\uparrow=\downarrow}{\text{VP}}$

This receives a straightforward LFG analysis, with the only unusual element the initial LB_S. The c- and f-structures for (33a) are shown in (34).[16]

(34) a.



b.

$$\begin{bmatrix} \text{PRED} & \text{'sleep}\langle\text{SUBJ}\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'man'} \\ \text{SPEC} & \text{that} \end{bmatrix} \\ \text{TENSE} & \text{present} \end{bmatrix}$$

[16] We represent only the left prosodic bracket (LB_S) because this is the prosodic boundary that can be reliably identified via a %L boundary tone (cf. Godjevac 2000, Radanović-Kocić 1988).

6.2 Clitic Cluster in Prosodic 2nd Position

Next consider when the clitic cluster follows the first prosodic word, as in (35a), which uses the realization in (35b).

(35) a. Taj *joj ga je* čovek poklonio.
that her it AUX man presented
'That man presented her with it.'

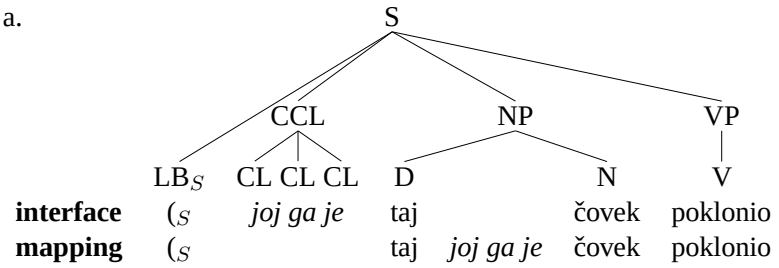
b.

LB_S	CCL	NP	VP
	↑=↓	(↑ SUBJ)=↓	↑=↓

Given lexical entries as in (36) for the clitics, this will result in the c- and f-structures in (37).

(36) *joj* (↑ OBJ2 PRED)='pro'
(↑ OBJ2 PERS)=3
(↑ OBJ2 NUM)=sg
(↑ OBJ2 GEND)=fem
ga (↑ OBJ PRED)='pro'
(↑ OBJ PERS)=3
(↑ OBJ NUM)=sg
(↑ OBJ GEND)=masc
je (↑ SUBJ PERS)=3
(↑ SUBJ NUM)=sg
(↑ TENSE)=past

(37) a.



b. $$\begin{bmatrix} \text{PRED} & \text{'present<SUBJ,OBJ,OBJ2>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'man'} \\ \text{PERS} & 3_{je} \\ \text{NUM} & \text{sg}_{je} \\ \text{SPEC} & \text{that} \end{bmatrix} \\ \text{OBJ}_{ga} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{masc} \end{bmatrix}_{ga} \\ \text{OBJ2}_{joj} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{fem} \end{bmatrix}_{joj} \\ \text{TENSE} & \text{past}_{je} \end{bmatrix}$$

In the c-structure in (37a) it is the interface mapping that allows for the clitics to be hosted by the first prosodic word, here the demonstrative *taj*. (The clitic forms are shown as italicized indices on the f-structure for expository purposes.)

6.3 Clitic Cluster in Syntactic 2nd Position

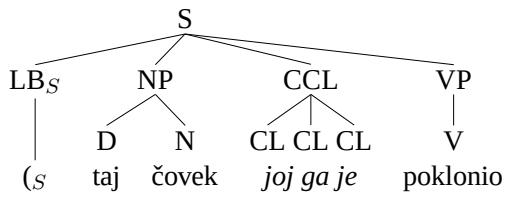
Next consider when the clitic cluster follows the first syntactic constituent, as in (38a), which uses the realization in (38b) (licensed by the expression *Second(x,y)*).

(38) a. Taj čovek *joj ga je* poklonio.
that man her it AUX presented
'That man presented her with it.'

b. LB_S NP CCL VP
$(\uparrow \text{SUBJ})=\downarrow$ $\uparrow=\downarrow$ $\uparrow=\downarrow$

Using the same lexical entries for the clitics as in (36), the c- and f-structures are shown in (39). Note that the resulting f-structure is identical whether the clitics follow the first prosodic word or the first syntactic constituent; it is only the c-structure and interface correspondences that differ.

(39) a.



b. F-structure: same as (37b)

Thus, we account for the standard instances of second position clitics, either syntactic or prosodic, with the small set of rules in (32) along with a conception of the prosody-syntax interface in which prosody and syntax are taken to be separate and independent modules of grammar that interact through a restricted interface mapping. We have not discussed instances of clitic placement in clauses involving topicalization or contrastive focus. These depend on another set of syntactic and prosodic correlations; however, we are confident that our approach will extend to the more complex data.[17]

7 Conclusion

In Bögel et al. (2009), we outlined a new architecture for modeling the interaction between prosody and syntax: we proposed an arrangement of interacting components in which prosodic information is developed in a module that operates independently of the syntax while still allowing for syntactic rules and preferences to be conditioned on prosodic boundaries and other features. This was made possible because the terminal string of the syntactic tree was taken to be a sequence of lexical formatives intermixed with additional features that also satisfy the constraints and reflect the generalizations of a separate prosodic component. This architecture allows for misalignments between prosodic units and syntactic constituency, but it also incorporates a Principle of Prosodic Preference that disprefers syntactic structures that do not coincide with prosodic boundaries.

In this paper, we extended the 2009 proposal to account for prosodically second position clitics by allowing for clausally-scoped clitics to appear optionally as prefixes of the normal expansions of the S c-structure rules. We introduced a separate interface mapping to mediate between the possibly conflicting requirements of the syntactic and prosodic components. The interface mapping defines a correspondence between separate syntactic and prosodic strings that are typically identical, as in the original proposal, but now are allowed to differ in a limited number of ways. We showed how such a mapping can correlate the first-position clitics in the c-structure terminal string with their attested realization after the first prosodic word. This mapping is a regular relation that satisfies the clitics' need for a prosodic host.

As we have noted, it is a consequence of our analysis that clitics in second prosodic position can only have clausal functional scope and cannot directly modify the f-structures corresponding to other constituents. This is because a regular relation is mathematically restricted in its ability to recognize and operate on the recursive structures of the syntactic component. For the same reason our account

[17] Radanović-Kocić (1988) also notes that appositions and relative clauses show interesting clitic placement patterns. We assume that these have to do with the prosodic phrasing of such clauses and that the basic generalization holds that clitics appear either after the first syntactic constituent or after the first prosodic word in the relevant prosodic and syntactic domain.

also predicts that second position clitics will always be drawn from a closed, finite set that a regular relation is able to correlate around arbitrary but locally specified prosodic units.

The approach to second position clitics described in this paper thus provides an elegant and new account of this phenomenon that fits comfortably within the collection of formal mechanisms that already exist within the LFG theoretical framework. This account offers further support for the architecture of the prosody-syntax interface that we previously proposed: prosody and syntax operate as separate modules that interact by virtue of a limited amount of shared information that can mutually constrain the behavior of both components.

References

Anderson, Stephen. 2005. *Aspects of the Theory of Clitics*. Oxford University Press.

Beesley, Kenneth and Karttunen, Lauri. 2003. *Finite State Morphology*. 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 Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG09 Conference*, CSLI On-line Publications.

Butt, Miriam and King, Tracy Holloway. 1998. Interfacing Phonology with LFG. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG98*, CSLI Publications.

Ćavar, Damir and Wilder, Chris. 1994a. "Clitic Third" in Croatian. In *Linguistics in Postdam*, pages 25–63.

Ćavar, Damir and Wilder, Chris. 1994b. Long Head Movement? Verb-Movement and Cliticization in Croatian. *Lingua* 93, 1–58.

Crouch, Dick, Dalrymple, Mary, Kaplan, Ronald M., King, Tracy, Maxwell, John and Newman, Paula. 2010. *XLE Documentation*. On-line documentation, Palo Alto Research Center (PARC).

Diesing, Molly, Filipović-Đurđević, Dušica and Zec, Draga. 2009. Clitic Placement in Serbian: Corpus and Experimental Evidence. In Sam Featherston and Susanne Winkler (eds.), *The Fruits of Empirical Linguistics*, pages 61–75, Mouton de Gruyter.

Frank, Anette, King, Tracy Holloway, Kuhn, Jonas and Maxwell III, John T. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG98 Conference*, CSLI On-line Publications.

Franks, Steven and King, Tracy Holloway. 2000. *A Handbook of Slavic Clitics*. Oxford University Press.

Franks, Steven and Progovac, Ljiljana. 1994. On the Placement of Serbo-Croatian Clitics. In *Proceedings of the 9th Biennial Conference on Balkan and South Slavic Linguistics, Literature, and Folklore*, pages 69–78.

Godjevac, Svetlana. 2000. An Autosegmental/Metrical Analysis of Serbo-Croatian Intonation. In Jennifer S. Müller, Tsan Huang and Craige Roberts (eds.), *Varia*, pages 79–142, Working Papers In Linguistics No. 54, Ohio State University.

Halpern, Aaron. 1995. *On the Placement and Morphology of Clitics*. CSLI Publications.

Halpern, Aaron and Zwicky, Arnold. 1996. *Approaching Second — Second Position Clitics and Related Phenomena*. CSLI Publications.

Hayes, Bruce. 1989. The Prosodic Hierarchy in Meter. In Paul Kiparsky and Gilbert Youmans (eds.), *Rhythm and Meter*, pages 201–260, Academic Press.

Inkelas, Sharon. 1990. *Prosodic Constituency in the Lexicon*. Garland Publishing.

Kaplan, Ronald M. and Kay, Martin. 1994. Regular Models of Phonological Rule Systems. *Computational Linguistics* 20, 331–478.

King, Tracy Holloway. 1994. Focus in Russian Yes-no Questions. *Journal of Slavic Linguistics* 2, 92—120.

Klavans, Judith. 1982. *Some Problems in a Theory of Clitics*. Ph. D. thesis, University College London, revised version published by Indiana Linguistics Club.

Nespor, Marina and Vogel, Irene. 1986. *Prosodic Phonology*. Foris.

Progovac, Ljiljana. 1996. Clitics in Serbian/Croatian: Comp as the Second Position. In Aaron Halpern and Arnold Zwicky (eds.), *Approaching Second: Second Position Clitics and Related Phenomena*, pages 411–428, CSLI Publications.

Radanović-Kocić, Vesna. 1988. *The Grammar of Serbo-Croatian Clitics: A Synchronic and Diachronic Perspective*. Ph. D. thesis, University of Illinois.

Radanović-Kocić, Vesna. 1996. The Placement of Serbo-Croatian Clitics: A Prosodic Approach. In Aaron L. Halpern and Arnold M. Zwicky (eds.), *Approaching Second: Second Position Clitics and Related Phenomena*, pages 429–445, CSLI Publications.

Selkirk, Elisabeth. 1986. On Derived Domains in Sentence Phonology. *Phonology Yearbook* 3, 371–405.

Selkirk, Elisabeth. 1995. The Prosodic Structure of Function Words. In Jill Beckman, Laura Walsh Dickey and Suzanne Urbanczyk (eds.), *Papers in Optimality Theory*, pages 439–70, GLSA Publications, also in Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition, eds. James L. Morgan and Katherine Demuth, 187-214: Lawrence Erlbaum Associates.

Selkirk, Elisabeth. 2001. The Syntax-Phonology Interface. In N.J. Smelser and Paul B. Baltes (eds.), *International Encyclopedia of the Social and Behavioral Sciences*, pages 15407–15412, Pergamon.

Wescoat, Michael. 2002. *On Lexical Sharing*. Ph. D. thesis, Stanford University.

Zwicky, Arnold. 1977. On Clitics. Technical Report, Indiana University Linguistics Club.

COMPLEMENTS OF ADJECTIVES: A DIACHRONIC APPROACH

Nigel Vincent & Kersti Börjars

The University of Manchester

Proceedings of the LFG2010 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

In this paper we challenge the traditional view that adjectives do not subcategorise for the grammatical function OBJ. We argue instead that the more general cross-linguistic restriction is against the assignment of accusative case outside the domain of a governing verb or preposition. This may however be violated in particular languages as we show by comparing Old Swedish, in which adjectives may unusually take accusative complements as first noted by Platzack (1982a, b) and Maling (1983), with Latin, where an accusative complement of an adjective is not possible. We then explore the diachronic developments into modern Swedish and more generally the modern Germanic languages and contrast them with the changes that have taken place in the modern Romance languages. We show that there are significant differences between the two language families in the way prepositions compensate for the loss of morphological case. We also suggest an alternative to Maling's account of the history of English *near*.

1. Introduction[1]

Can an adjective have an object? Traditional grammar says no (Huddleston & Pullum 2001: 527). In a similar vein, Principles & Parameters Case Theory relies on the inability of nouns and adjectives to assign objective case to explain the distribution of English *of* (Chomsky 1981: 50-1). Compare too the theory of categories proposed by Jackendoff (1977), according to which adjectives are [-obj, –subj], thus contrasting with verbs: [+subj, +obj], nouns: [+subj, –obj] and prepositions:[–subj, +obj].

[1] We are grateful to those who have provided examples and/or contributed to discussions of this paper at LFG 2010 and at the annual conference of the Linguistics Association of Great Britain held at the University of Leeds in September 2010. These include: Wiebke Brockhaus-Grand, Martin Forst, Joan Maling, Christer Platzack, Louisa Sadler, Eva Skafte Jensen, Merethe Damsgaard Sørensen, and Bo A. Wendt. We thank too the anonymous referees of the abstract when it was submitted for LFG 2010, and Miriam Butt for her comments on the pre-final draft. Errors and omissions remain our own responsibility.
Correspondence address: {kersti.borjars, nigel.vincent}@manchester.ac.uk

Maling (1983) and Platzack (1982a,b) by contrast noted in the earlier stages of the Germanic languages the existence of a category of so-called 'transitive adjectives', by which they mean adjectives taking NP complements (Platzack 1982b: 274). In the Old Swedish examples in (1) (cited after Platzack 1982a, b) the adjectives *lyþoghe* 'obedient' and *vis* 'sure' take complements respectively in the dative and the genitive:

(1) a. *at i ärin guþi lyþoghe*
COMP 3PL be.PRS.PL God.DAT obedient
'that you are obedient to God'

b. *þäs är iak vis*
that.GEN be.PRS.SG 1SG sure
'of that I am sure'

Platzack points out that the complements of the same adjectives in modern Swedish are marked by prepositions, which is what one might expect given that the modern language has lost the earlier system of nominal case inflection. Yet the issue is not simply one of replacing cases by prepositions since, as he also notes, already in Old Swedish many adjectives took PP complements and some adjectives occur with either an NP or a PP complement.

More recently, within LFG, Mittendorf & Sadler (2008) consider the Welsh construction exemplified in (2), and propose an analysis according to which the adjective *byr* 'short' takes *Siân* as SUBJ and *thymer* as OBJ:

(2) *Mae Siân yn fyr ei thymer.*
is S PRED short her temper
'Siân is short-tempered.'

Al Sharifi & Sadler (2009) argue along similar lines in respect of the Arabic adjectival construct in (3) suggesting that *-l-wağh-i* 'the face' bears the relation OBJ to *ğamī-l-at-u* 'beautiful' and that this relation is realised as GEN(itive) in the context of this construction:

(3) *imraʔ-at-un ğamī-l-at-u -l-wağh-i*
woman-F-NOM beautiful-F-NOM the-face-GEN
'a woman with a beautiful face'

Data of this kind raise three questions: a) what grammatical functions (GFs) should be assigned to adjectival complements? b) how do morphological case

and prepositions interact in the realization of these GFs? c) how do the relations between the underlying GFs and their overt realizations change over time? The broader issue of what properties a function should have in order to be classified as OBJ will not be dealt with directly here; we refer the reader to Börjars & Vincent (2008) for discussion.

In this paper we address these questions through a comparative analysis of the history of cases and prepositions which depend on adjectives in Romance and Germanic. After further discussion of some of the basic theoretical issues (section 2), we compare the distribution of adjectives and case in Old Swedish (OSw) (section3.1) and Latin (section 3.2). We then review the possible pathways for change in general terms (section 4) before charting the particular history of these constructions in the two language families (section 5). Finally, in section 6 we draw out the theoretical consequences that we suggest follow from our account.

2. Theoretical preliminaries

The architecture of LFG assigns, via f-structure, a key role to grammatical functions or relations (GFs). We need to ask therefore what links there are on the one hand between the f-structure and the syntactic categories of c-structure, and on the other between f-structure and the m(orphological)-structure system of case for those languages which have one. A central issue concerns the GFs that can occur as arguments of adjectives. The set of sub-categorizable GFs includes: SUBJ, OBJ, OBJ_{Θ}, OBL, XCOMP and COMP. Whereas a verb, depending on its semantics, may in principle subcategorize for any of these, it has standardly been assumed that adjectives differ from verbs in allowing for all except OBJ and OBJ_{Θ}. In this respect, LFG is no different from traditional grammar and the various versions of Chomskyan syntax mentioned in our opening paragraph (see Platzack 1982a,b; van Riemsdijk 1983 and Ohkado 1990 for discussion of the issues that the construction gives rise to within a Chomskyan framework and how these can be dealt with).

If we are to pursue the question of whether this received wisdom is correct, it is important to distinguish between a GF and its morpho-syntactic realisation. Thus, if a language like Latin does not allow accusative arguments of adjectives, this could in principle be explained at the level of f-structure by a constraint that forbids OBJ as the argument of A, or at the level of m-structure by a constraint that OBJ cannot be realised as ACC within an AP. Put another way, Old Swedish, which does allow accusative complements of adjectives, could be exceptional in allowing its adjectives to sub-categorise for OBJ or it could be unusual in permitting ACC to be assigned inside the AP. One of our main concerns in the

present paper is to try to find grounds for deciding between these two alternatives. Our discussion focuses in the main on simple adjectives, although in section 3.2 we touch briefly on the properties of some Latin participial constructions which lie at the border between adjectives and verbs.

A further issue concerns the relations between adjectives and the categorial — as opposed to functional — status of their complements. Thus, Maling (1983: 254) articulates a widely held assumption when she writes that: '… there is something essentially correct about the idea that it is less natural for A and N to take NP complements than for V and P to do so …'. And if we are dealing with a language without nominal case such as English she is surely right; hence her discussion of apparently exceptional examples such as *worth further consideration* or *like your sister*, which we pick up below. For a language like Latin, on the other hand, Maling's remark is less obviously true, since examples such as *plenus rimarum* 'full of chink.GEN.PL' and *similis matri* 'like his mother.DAT.SG' abound.[2] The heart of the problem therefore seems to reside in the nature of a language's system for marking functional dependency rather than in the theory of grammatical categories.

3. Adjectives and case in the older languages

In this section we compare the patterns of case distribution in Old Swedish (OSw) (section 3.1) and Latin (section 3.2). Both languages have nominal case systems, but exhibit significant differences in the co-occurrence of the cases with adjectives. Most relevantly, as we shall see, OSw allows complements of adjectives in the accusative whereas Latin does not. To set the scene, the following table shows the correspondences between the cases reconstructed for Proto-Indo-European (PIE) and Germanic and Latin, and the different ways the original system has been reduced in Latin compared to OSw, which inherits unchanged the Proto-Germanic four case system.[3]

---

[2] We assume here that bare case forms such as *rimarum* 'chink.GEN.PL' and *matri* 'mother.DAT.SG' are indeed NPs. In systems which exploit extensive inventories of functional heads, these forms might be KPs or something similar, but within such a system of course not even verbs would co-occur with NPs. Either way, it remains true that in languages like English and French bare NPs do not usually co-occur with adjectives whereas they do in languages like Latin or Russian.

[3] We exclude the vocative and nominative from consideration since they would not have been available to mark internal arguments, whether verbal or adjectival. In addition, Latin has a small number of residual locatives but they too are not relevant here. See Meiser (1992) for further discussion and references.

Proto-Indo-European	Proto-Germanic	Latin
Accusative	Accusative	Accusative
Genitive	Genitive	Genitive
Dative	Dative	Dative
Ablative	Genitive OR Dative	Ablative
Instrumental	Dative	Ablative
Locative	Dative	Ablative
Table I: Case correspondences between Latin and Germanic		

Noteworthy here is that Latin has retained the Indo-European ablative in a range of functions, one of which is to mark the complements of some adjectives, whereas the corresponding adjectives in those Germanic languages which retain case-marking take a dative or a genitive. Fuller investigation of these more detailed case-marking differences between the two families will however have to be put off to a future occasion.

3.1 Old Swedish and Germanic

As we have already seen in (1), OSw has a range of different cases which can depend on adjectives, among which the following are remarkable for their ability to combine with an accusative (data once more derived from Platzack 1982 a,b):

(4) *lönlikin* 'clandestine', *rätter* 'suitable', *godher* 'kind',
mögheliker 'possible', *þækkeliker* 'delightful'

With some adjectives, on the other hand, a prepositional construction is already attested in OSw as an alternative to an NP complement:

(5) *fri* (*af*) 'free (of)'
milder (*ivir*) 'lenient (to)'
rädder (*for*) 'afraid (of)'
vis (*a*/*op*/*af*) 'certain (of)'

Similar patterns to Old Swedish are also found in Old Danish, though in that language we have not yet come across any unambiguous accusative complements of adjectives. Thus, in (6a) the adjective *skuldich* 'owing, indebted' combines

with the dative pronoun *hannum* 'to him', and in (6b) *oss* 'us' co-occurs with the co-ordinated adjectives *høriige eller ludiige* 'attentive or obedient':[4]

(6) a. *och* *noger* *borger* *ær* *hannum* *noget* *skuldich*
and some citizens be.3PL.PRS he.DAT some owing
'and some citizens owe something to him' (1452, Rsv V.306)

b. *at* *the ey* *skulle* *være* *oss* *høriige* *eller* *ludiige*
that he not should be us.DAT attentive or obedient
'that he should not listen to and obey us'
(1502, Rosenv. GL D I,27)

For Old Norse, Faarlund (2004) notes only adjectival complements with the dative and the genitive and interprets their distribution in semantic terms. He writes (p.99), for example, that 'Adjectives taking dative complements are first of all those that denote a state of mind or an attitude … This is the benefactive or recipient role, which is the basic meaning of the dative case', while for genitives he comments (p.101) that 'typical genitive complements of adjectives have a partitive meaning'. Though the tendencies identified by Faarlund are likely to be correct, Platzack (1982a,b) is more cautious about the connection and further work is required on historical data to get a clearer picture of the correspondences between case on the one hand and the semantic relation between the adjective and its complement on the other. As with Old Swedish, so in Old Norse there are alternations between case marked NPs and prepositional constructions; compare the examples in (7) [= Faarlund's (24b) and (29a)]: in (7a) the complement of *fusír* 'eager' is in the genitive case whereas in (7b) *búinn* 'ready' takes a PP introduced by *til* 'to'.

(7) a. *er* *fúsir* *váru* *fararinnar*
who eager be.3.PST journey.GEN.DEF
'who were eager to leave' (Kkr II.308.9)

b. *nú* *em* *ek* *búinn* *til* *ferðar*
now be.1SG.PRS 1SG prepared to journey.GEN.DEF
'Now I am ready to go.'

[4] The form *oss* here could in principle be either accusative or dative but, given the available evidence of other forms, there is no reason to treat it as anything but dative in this context.

For the history of Dutch, van der Horst (2008) provides examples of adjectives with genitive and dative complements down to the point in time when the case system was lost, but makes no mention of adjectives taking NPs in the accusative case (see also Broekhuis To appear).

3.2 Latin

A semantic account of the distribution of cases with adjectives is also characteristic of traditional Latin grammars, which talk of the genitive of quality, dative of similarity and so on. The relevant cases here are three: genitive, dative and ablative. In the words of Serbat (1996: 371): 'Après adjectif, tous les cas ou tours prépositionnels sont possibles, à l'exception remarquable de l'Ac[cusatif], qui est comme réservé au verbe.' [After an adjective all cases and prepositional phrases are possible, with the notable exception of the accusative, which is as it were reserved for the verb.] We set out in (8) a selection of adjectives grouped according to the cases they typically govern:

(8) DATIVE *similis* 'like', *aequus* 'equal', *iunctus* 'joined', *aptus* 'suited', *gratus* 'pleasing', *carus* 'dear'
ABLATIVE *natus* 'born', *dignus* 'worthy', *vacuus* 'free', *oriundus* 'descended from'
GENITIVE *plenus* 'full', *fecundus* 'fertile', *cupidus* 'greedy', *memor* 'mindful', *ignarus* 'ignorant', *peritus* 'skilled'

A semantic account seems to fit naturally for ablatives which express the origin or source, since this is a function independently associated with the ablative case. In other instances the historically appropriate semantics is no longer transparent. Thus Latin has ablatives in expressions like *crine ruber* 'redhaired, *lit*. red in the hair.ABL' and *mente captus* 'insane, *lit*. caught in the mind.ABL', where the meaning is appropriate for an earlier locative which was subsequently conflated with the ablative. It is notable too that many of the adjectives which take dative or genitive in Latin take the corresponding case in Germanic, which reinforces the argument for an account along the lines indicated by Faarlund in his comments, quoted above, on the Old Norse data.[5]

Pinkster (1990: 58ff) contrasts the semantic approach, in which arguments of both verbs and adjective bear the same case according to the meaning of the

[5] For a thorough and richly documented exploration of the semantic bases of adjectival cases in Latin see the relevant sections of Serbat (1996).

predicate expressed, with a structural one whereby objects of verbs are assigned accusative and the same arguments inside noun or adjective phrases are assigned genitive. This principle of genitive as the structural case inside NP or AP can be traced back to the attempt by Benveniste to provide a unified account of the diverse uses of the genitive catalogued by de Groot (1956). Benveniste (1962: 18) writes: '… dans la conception esquissée ici, la fonction du génitif se définit comme résultant d'une transposition d'un syntagme verbal en syntagme nominal.' [… on the view sketched here, the function of the genitive is defined as the result of transposing a verb phrase into a noun phrase.]

An instructive pattern from this point of view is the three-way alternation exemplified in (9):

(9) a. *laborem fugit* 'he shuns work.ACC'
b. *laborem fugiens* 'shunning work.ACC'
c. *laboris fugiens* 'shunning work.GEN'

The verb here is *fugire* 'to shun' whose finite forms take a direct object in the accusative case as in (9a). The present participle *fugiens* by contrast is attested with both accusative objects as in (9b) and genitive objects as in (9c). Not surprisingly when the participle retains its verbal force it takes the accusative case as in (10):

(10) *quibus* *pacem* *atque* *amicitiam* *petentibus*
who.DAT peace.ACC and friendship.ACC seek.PPRT.DAT
'to those seeking peace and friendship' (Caes. *BG*. 4,18,3)

On the other hand the meaning of such a participle can often be close to adjectival in which case the genitive is possible, as in (11), where the participle is co-ordinated with the adjective *fortis* 'strong' and both bear the typically adjectival superlative suffix *-issim-*:

(11) *vir* *fortissimus* *et*
man.NOM.M.SG strong.SUP.NOM.M.SG and
amantissimus *rei publicae*
love.PPRT.SUP.NOM.M.SG state.GEN.SG
'a very strong man and most loving of the state' (Cic *Cat* 4.17)

Such a participial genitive falls neatly under Benveniste's generalization about what he calls, in the passage already quoted, the genitive of transposition.[6]

We have said that there are no Latin accusatives depending on adjectives of the kind that Platzack has described for OSw. Apparent exceptions are the expressions in (12), where adjectives of dimension such as *altus* 'high, deep', *longus* 'long', *latus* 'wide' and *crassus* 'thick' are found with an NP in the accusative denoting extent:

(12) a. *longus* *binos* *pedes*
long two feet.ACC
'two feet long'
b. *latus* *digitos* *tres*
wide fingers.ACC three
'three fingers wide'

The reason for the accusative here is not objecthood but rather the fact that expressions of extent, spatial and temporal, in Latin require the accusative:

(13) a. *trabes* *distantes* *inter* *se* *binos* *pedes*
beams.NOM distant.NOM between REFL two.ACC feet.ACC
'beams distant from each other two feet' (Caes *BG* 7, 23,1)
b. *Gorgias* *centum* *et* *novem* *vixit* *annos*
G 100 and 9 live.PRF.3SG year.ACC.PL
'Gorgias lived for 109 years.' (Quint, 3, 1, 9)

We have here then a good instance of Pinkster's (1990: 59) principle that case forms inside APs commonly reflect the functions of similar items at sentence level.[7] In Latin then the strong constraint against an accusative inside the domain of AP and VP can be overridden only by adverbial uses of the accusative whereas in OSw the constraint is weaker and some OBJ accusatives are also admitted.

Before concluding this section we should note that from the earliest stages of Latin we find adjectives that are complemented by PPs instead of case forms, and

---

[6] For fuller discussion and exemplification of the genitive with participles tending in the same direction as Benveniste, see Serbat 1996: 395-399.

[7] It is perhaps worth noting that corresponding expressions are unmarked by prepositions in English: *two metres tall, five inches thick,* etc. In other words, in both languages, measure phrases, although they are not objects, bear the marking appropriate to a verbal object.

instances where case forms and PPs are alternatives after the same adjective, just as we have observed for OSw. Thus for example we have both *oriundi ex Etruscis* 'sprung from the Etruscans' with the preposition *ex* and *caelesti semine oriundi* 'sprung from divine.ABL seed.ABL' with a dependent ablative. This is part of the wider phenomenon of alternation between cases and prepositions in both semantic and grammatical uses (see Molinelli 1996 and section 5.1 below).

4. Some routes for change

Given the systems of morphological case as markers of adjective dependency for Latin and the earlier stages of Germanic that we have sketched in section 3, we may then ask: what are the possible routes for change if a language loses that system of morphological case? The options are various and include:

i) Prepositions may take over the function of marking grammatical relations. In particular a default preposition may be used to license arguments outside the verbal domain, as happens with *of* in English and *de* in French.

ii) The head item may change status from adjective to preposition. This is commonly argued to be what has happened in the history of English with words like *worth* and *like* (Maling 1983).

iii) 'Transitive' adjectives — in the sense of adjectives with bare NP-complements — may survive as marked options or historical relics, as Maling (1983) suggests is true of English *near*.

We shall see in section 5 that these scenarios play out in interestingly different ways within our chosen families of Romance and Germanic.

5. The historical developments

5.1 From Latin to Romance

As is well known, the Latin case system disappears in the course of the evolution to the Modern Romance languages.[8] While nominative and accusative as the markers of SUBJ and OBJ give way to fixed pre- and post-verbal positioning, the other cases are in different ways replaced by PPs. Particularly striking in this regard is the development of the Latin preposition *de*, originally meaning 'away,

[8] Old French retains a morphological distinction between nominative and accusative but uses *de* in a range of constructions corresponding to the Latin genitive. Romanian retains a morphological genitive which co-exists with constructions using *de* in ways that are too complex to document here and which do not in any case affect the overall thrust of our argument.

down', to become the marker of nominal dependence across the whole family. Two properties are worth noting in this connection: first that *de* + NP replaces GEN in all its functions, and second that there is evidence for the functional equivalence of GEN and *de* + NP from a very early date. In her otherwise exhaustive survey of the uses of the genitive and its replacement by *de* it is interesting that Molinelli (1996) does not include any examples of adjectival complements. However, the following examples —where *de* + NP in (14a) and (15a) alternate with the genitive NPs in (14b) and (15b) — suffice to complete her account; further alternations of this kind are well documented across the full historical span of the language.[9]

(14) a. *his* *de* *rebus* *conscium* *esse* *Pisonem*
this.ABL.PL DE thing ABL.PL complicit.ACC.SG be.INF Piso.ACC
'that P was complicit in these deeds' (Cic *Att* 2.24.3)

b. *si* *conscius* *Dymno* *tanti*
if complicit.M.NOM.SG Dymnus.DAT.SG so great.GEN.SG
sceleris *fuissem*
crime. GEN.SG be.PRF.SBJ.1SG (Curt 6.10.20)
'if I had been an accomplice of Dymnus in so great a crime'

(15) a. *de* *agricultura* *peritissimus*
DE agriculture.ABL.SG skilled.SUPERL.NOM.SG.M
'very skilled in agriculture' (Varro *RR* 1.2.10)

b. *multarum* *rerum* *peritus*
many.GEN.PL.F thing. GEN.PL.F skilled.NOM.SG.M
'skilled in many things' (Cic *Font* 7,15)

The generalization that emerges therefore is that the Romance languages continue the basic Latin pattern but realise it by different means. The historical evidence thus provides further confirmation that Benveniste (1962) was right to discern a unity in the apparently diverse and multifarious uses of the Latin genitive. In more recent terminology, the genitive is the structural case assigned within the noun phrase. The fact that in the Romance languages this genitival relation is realised via a grammatical preposition rather than a case inflection does not alter the continuity of the underlying structural pattern. In consequence it is possible to identify constructions with *de* corresponding to almost all the

[9] NB Latin *de* governs the ablative. Ex (14) is cited by Pinkster (1990: 66).

different functions traditionally assigned to the genitive. Table II illustrates this for a representative sample of constructions, where the Latin genitives are in bold. (The language here is French but analogous examples could be constructed in all the modern Romance languages.)

Construction	**Latin**	**French**	**Gloss**
Comp of N	*rex* ***regum***	*le roi des rois*	'the king of kings'
Comp of A	*avidus* ***gloriae***	*avide de gloire*	'eager for glory'
Comp of V	*memini* ***vivorum***	*je me souviens des vivants*	'I remember the living'
Partitive	*multi* ***civium***	*beaucoup des citoyens*	'many of the citizens'
Quality	*vir* ***magnae eloquentiae***	*un homme d'une grande éloquence*	'a man of great eloquence'
Possessive	*domus* ***regis***	*le palais du roi*	'the king's palace'
Table II: Latin genitive and French *de* compared			

The one exception to this generalisation is the genitive dependent on a participle as in example (11) above. Arguably in this instance, however, the reason lies in the altered status of the participle, which has either dropped out of use or has been retained only in lexicalised adjectives and nouns such as *intéressant* 'interesting', *puissant* 'powerful', *aimant* 'lover' and the the like.[10]

5.2 From Old to Modern Swedish

Platzack (1982a, 1982b) considers 62 adjectives that took an NP in the dative, genitive or accusative in OSw. Of these, 23 no longer exist in the modern

[10] It is worth noting in this context a further pattern, which traditional etymological accounts record but do not explain, whereby a number of prepositions in Romance contain an apparently pleonastic *de*. Thus, in Italian we find *dopo* 'after' < DE POST, *da* 'from' < DE AB, *davanti* 'in front of' < DE AB ANTE, etc. The lexical semantic content of the modern prepositions here is found in the corresponding Latin simple prepositions *post* 'after', *ab* 'from by', *ante* 'in front of, before'. It is tempting to see these as instances in which the *de* serves to mark the dependent role within the clause not only for NPs but also for some PPs (cf Vincent 1997).

language, 17 take only a PP complement, 13 take either a PP or an NP, and 9 take only an NP.[11] A first option to consider is the possibility that the elements which take an NP complement have in fact developed into prepositions, as is persuasively argued for the English *worth* and *like* by Maling (1983). The data in (16), however, show that an item like *trogen* 'faithful' has all the characteristics of an adjective. Thus, it distributes like an adjective in that it can occur attributively as in (16a), predicatively as in (16b) or as an adjunct as in (16c). Not all of these adjectives can occur attributively, but they still have adjectival properties relating to agreement. As the examples in (17) demonstrate, *övermäktig* takes an NP complement, but like any other adjective it agrees with the gender and number of its controlling noun.

(16) a. *en sin husse trogen hund*
a POSS master faithful dog
'a dog faithful to its owner'

b. *Hunden är trogen sin husse*
dog.DEF be.PRS faithful POSS master
'The dog is faithful to his master'

c. *Sin husse trogen vägrade hunden att gå.*
POSS master faithful refuse.PST dog.DEF COMP go.INF
'Faithful to his master, the dog refused to go.'

(17) a. *Verkligheten blev oss övermäktig.*
reality.COM.DEF become.PST 1PL.OBJ overpowering.COM.SG
'Reality overpowered us.'

b. *Livet blev oss övermäktigt.*
life.NT.DEF become.PST 1PL.OBJ overpowering.NT.SG

---

[11] There are even some adjectives which take two nominal complements:

i) *Jag var skyldig honom mitt stöd.*
I was liable.COM 3SG.M.OBJ my support
'I owed him my support.'

ii) *Jag känner mig värd resten av chokladkakan.*
I feel 1SG.OBJ worth.COM.SG rest of chocolate bar
'I feel I have earned the rest of the chocolate bar.

In this respect the situation in Swedish parallels that found in modern German.

'Life overpowered us'

c. *Utmaningarna* *blev* *oss* *övermäktiga.*
challenges.PL.DEF become.PST 1PL.OBJ overpowering.PL
'The challenges overpowered us.'

These elements then have all the hallmarks of adjectives and they are not just occasional or exceptional examples. Indeed some adjectives which took NP or PP in OSw, such as *liker* 'like' and *værþugher* 'worthy', take only NP in modern Swedish. Furthermore, some of the modern Swedish adjectives which take NP complements are not attested at all in OSw. Although non-attestation does not always mean non-existence, it is likely that some of these adjectives entered the language after case was lost. In short, the sequence A+NP is a genuine pattern of Swedish grammar, and cannot simply be dismissed as a "historical residue".

The complement can precede the adjective as in (18), a pattern which is uncharacteristic of Swedish, and which has been taken to betray German influence, although more research is required on the nature of the contact situation that could have engendered this change.

(18) a. *Regeringen* *är* *inte* *uppgiften* *vuxen.*
government be.PRS NEG task.DEF adult
'The government is not up to the task.'

b. *Hunden* *är* *sin husse* *trogen.*
dog.COM.DEF be.PRS POSS master faithful.COM
'The dog was faithful to his master.'

For most adjectives which can occur with an NP complement, this can either precede or follow, while for some the complement can only follow. Adjectives taking NP complements have entered the language over a long period of time and, although contact with German may have influenced the historical development of the construction, there can be no doubt that it is now an intrinsic property of the language, and hence any analysis of Swedish syntax needs to account for this distribution.

5.3 Other Germanic languages

As mentioned in section 3.1, descriptions of early forms of Germanic do mention adjectives with NP complements, but they do not mention the possibility of these complements occurring with accusative case. Platzack (1982a,b) show that they

do occur in OSw and it may be that further investigation will reveal that they also do in other early Germanic varieties. What is clear is that the Germanic languages that have lost case do indeed have adjectives with NP complements, as the following examples from Danish (19a), Norwegian (19b) and Dutch (19c) show:

(19) a. *Nu* *er* *vi* *kvit* *den lykke* *vi havde.*
now be.PRS we rid the happiness we have.PST
'Now we have lost the happiness we had.'

b. *Ho* *var* *rädd* *björnen.*
she be.PST afraid bear.DEF
'She was afraid of the bear.'

c. *KNVB directeur Kesler* *is* *Rutten en Jol* *zat.*
KNVB director Kesler is Rutten and Jol fed up
'The KNVB director Kesler is fed up with Rutten and van Jol.'

The pattern of an adjective with an NP complement in non-case languages, which as we have said has been claimed by many to be rare, appears then to be common across the relevant Germanic languages.

It is also worth noting that there are adjectives in modern German, a language which has preserved case, which take an NP complement in the accusative:[12]

(20) a. … *waren* *die Bürger* *den* *arroganten Aktivismus* *leid.*
were the citizens the.ACC arrogant activism fed up
'the citizens were fed up with the arrogant activism.'

b. *Dann wird* *auch den* *starken* *Regen* *gewohnten*
then becomes also the strong.ACC[13] rain accustomed
Landwirten *der Boden* *zu nass.*
farmers the ground too wet

[12] We are grateful to Martin Forst for pointing this out to us. (20a) is a simplified version of a corpus example supplied to us by Martin. Thanks also to Wiebke Brockhaus-Grand for discussion of the German examples. We hope to undertake a more detailed comparison of modern German and Old Swedish in future work.
[13] *Starken* as a form is not unambiguously accusative, but in this environment it can be shown to be so.

'Then the ground became too wet even for the farmers who were used to heavy rain.'

5.4 The special case of near

We mentioned in section 4 that Maling (1983) treats English *near* as a historical relic. The reason she does so is that, while at first sight this item might seem to fall into the same class as *worth* and *like* with the distribution of a preposition — compare *She lives near/beside/opposite the church* — it retains the classically adjectival property of allowing comparison (examples culled from a Google search):

(21) a. *Which public control economy is* ***nearest a mixed economic system****?*
b. *Mascherano edges* ***nearer the exit***

The evidence of Latin suggests however that this may not be an arbitrary property of the English *near,* as the historical residue scenario might imply, but that there are semantic factors at work. Thus, in Latin the word meaning 'near' appears to be prepositional in its ability to co-occur with an NP in the accusative, and indeed is usually so treated in grammars and dictionaries, as in (22):

(22) *prope amnem*
near river.ACC
'near the river'

However, here too we can find the same pattern of a comparative (*propius* 'nearer') or a superlative (*proxime* 'nearest') co-occurring with the accusative in apparent defiance of what we have said above about the absence of accusative complements of adjectives in Latin:

(23) a. *propius* *urbem*
near.COMP city.ACC
'nearer the city'
b. *proxime* *hostem*
near.SUPERL enemy.ACC
'nearest the enemy'

It seems therefore that English *near* is not a historical relic, in the sense of an arbitrary survival of an earlier pattern, but rather has special properties, arguably

related to its semantics, which are independently attested in other languages.[14] It should be pointed out that this cannot simply be due to the gradability of the meaning of NEAR, since there are prepositions in the same languages which are conceptually gradable but which do not inflect like an adjective.

6. Consequences and conclusions

Let us now try to sum up the overall theoretical consequences of our analysis. First, it should be clear that we believe there are good grounds for allowing OBJ (and OBJ_{Θ}) to be sub-categorised by adjectives as well as verbs. In other words, we argue that predicates of all kinds may take the full range of sub-categorisable functions. At the sentential level, accusative is the case which marks the GF OBJ just as nominative is the case which marks SUBJ. In other words the object of a verb is a structural relation marked by the accusative case. Oblique complements of verbs are by contrast semantically motivated and realised either by an oblique case or by a PP. Neither accusative nor nominative, however, are in the unmarked circumstance appropriate in the nominal domain, where instead GEN holds sway. It is only when the ACC has a function other than that of marking OBJ, as with the so-called 'accusative of extent', that it can and does occur inside the NP as well as at sentence or clause level.

More generally, the realisations of grammatical relations depend on the relations between case and prepositions in particular (stages of) languages, and in this respect Romance and Germanic provide an instructive minimal contrast. We have seen that the Latin genitive is genuinely structural and that in the daughter languages it is replaced across the board with the equally structural preposition descended from *de*. We would then argue that in these languages, OBJ is marked by the preposition *de* in the nominal domain. In the Nordic languages, by contrast, there is no single preposition that fulfils this function, with the consequence that there is a much wider range of items that realise the relation OBJ. Thus, compare the range of prepositions used in English and Danish to mark nominal dependence (examples from Allan *et al* 1995: para 732; the Danish prepositions are highlighted in bold):

[14] It is interesting to note the way Romance descendants of these items have developed. Thus the superlative *proximus* yields the Italian adjective *prossimo* 'next', complements of which are marked by the preposition *a* 'to': e.g. *prossimo alla stazione* 'next to the station'. French *proche* 'near' is a derivative of *prochain* 'next' (< Lat *proximanus*) and takes complements marked by *de*: e.g. *un café proche de la gare* 'a café near the station'.

(24)	the discovery of America	*opdagelsen **af** Amerika*
	the manager of the company	*chefen **for** firmaet*
	the time of departure	*tidspunktet **for** afrejsen*
	a professor of physics	*en professor **i** fysik*
	the crew of the ship	*mandskabet **på** skibet*
	the murder of Duncan	*mordet **på** Duncan*
	the cause of the fire	*årsagen **til** branden*

Given these two different diachronic scenarios, the surprising language is in fact English, which genetically belongs to the Germanic family but has a realisation system based on the single preposition *of* which is analogous to that of Romance *de/di*. The most plausible explanation for this state of affairs is the contact between Old English and Old French in the Norman period, which has led to the Romance, and ultimately Latin, system of structural genitive marking being applied in the alien context of a Germanic language. Dutch would appear to hold a middle ground, in that *van* distributes much like *of* in (24), but the language still has adjectives with NP complements.

Two outstanding diachronic questions are: a) why did Swedish and the other North Germanic languages lose their case system and yet still preserve a significant number of bare NP complements with adjectives? and b) what licensed the bare NPs in the first place? The answer to the first of these questions seems to lie in part at least in the fact there never was the across-the-board equivalence between a prepositional construction and the genitive case that we have evidenced in Latin. No single preposition emerged to inherit the role of the genitive. Here too contact may be a factor, this time between the North Germanic languages and German. It is also possible that the Germanic genitive, despite being cognate with the Latin genitive, had come to assume a rather different function in those languages and therefore did not have the same unified structural role as its Latin congener. This in turn would have meant a different distribution of functions across genitive and accusative cases, and in particular would have licensed bare accusatives along with other cases inside the AP, thus answering our second question. These however are issues which we will have to address in detail on a future occasion.

References

Allan, Robin, Philip Holmes & Tom Lundskær-Nielsen. 1995. *Danish. A comprehensive grammar*. London: Routledge.

Al Sharifi, Budour & Louisa Sadler. 2009. The adjectival construct in Arabic. In Miriam Butt & Tracy Holloway King (eds) *Proceedings of the LFG09 conference*. Stanford, Ca: CSLI Publications, pp. 26-43. http://csli-publications.stanford.edu/

Benveniste, Emile. 1962. Pour l'analyse des fonctions casuelles: le génitif latin. *Lingua* 11: 10-18. Reprinted in E. Benveniste (1974) *Problèmes de linguistique générale 1*, Paris: Gallimard, 140-148.

Börjars, Kersti & Nigel Vincent. 2008. Objects and OBJ. In Miriam Butt & Tracy Holloway King (eds) *Proceedings of the LFG08 Conference,* Stanford, Ca: CSLI Publications, pp. 150-168. http://csli-publications.stanford.edu/

Broekhuis, Hans. To appear. Adjectives and adjective phrases. *Modern grammar of Dutch*, ed. by H. van Riemsdijk. Tilburg: University of Tilburg.

Faarlund, Jan Terje. 2004. *The syntax of Old Norse*. Oxford: Oxford University Press.

de Groot, A.W. 1956. Classification of the uses of a case illustrated on the genitive in Latin. *Lingua* 6: 8-65.

van der Horst, J.M. 2008. *Geschiedenis van de Nederlandse syntaxis*. Leuven: Universitaire Pers Leuven.

Maling, Joan. 1983. Transitive adjectives: a case of categorial reanalysis. In *Linguistic categories: auxiliaries and related puzzles*. *Vol 1*, eds. Frank Heny and Barry Richard, 253-289. Dordrecht: D Reidel.

Meiser, Gerhard. 1992. Syncretism in Indo-European languages - motives, process and results. *Transactions of the Philological Society* 90: 187-218.

Mittendorf, Ingo, and Sadler, Louisa. 2008. NP would like to meet GF. In Miriam Butt & Tracy Holloway King (eds) *Proceedings of the LFG08 conference*. Stanford, Ca: CSLI Publications, pp. 373-393. http://csli-publications.stanford.edu/

Molinelli, Piera. 1996. Casi e preposizioni in latino: lo sviluppo del genitivo e del dativo. Una prospettiva (anche) sociolinguistica. *Linguistica e Filologia* 3: 73-125.

Ohkado, Masayuki. 1990. Transitive adjectives and the theory of Case. *Lingua* 81: 241-64.

Pinkster, Harm. 1990. *Latin syntax and semantics*. London: Routledge.

Platzack, Christer. 1982a. Transitive adjectives in Swedish: a phenomenon with implications for the theory of abstract Case. *Linguistic Review* 2:39-56.

Platzack, Christer. 1982b. Transitive adjectives in Old and Modern Swedish. In *Papers from the 5th International Conference on Historical Linguistics*, ed. Anders Ahlqvist, 273-282. Amsterdam: John Benjamins.

van Riemsdijk, Henk. 1983. The Case of German adjectives. *Linguistic categories: auxiliaries and related puzzles. Vol 1*, ed. by F. Heny & B. Richard, 223–52. Dordrecht: D Reidel.

Serbat, Guy. 1996. *Grammaire fondamentale du latin VI: L'emploi des cas en latin*. Louvain: Peeters.

Vincent, Nigel. 1997. Prepositions. In Martin Maiden & Mair Parry (eds) *The dialects of Italy*, London: Routledge, pp. 208-213.

TWO MOVEMENT PARADOXES IN ZAPOTEC

George Aaron Broadwell
University at Albany, State University at New York

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract: Bresnan (2001) has drawn attention to *movement paradoxes* in syntactic theory. These are cases where the category of a constituent in a derived position differs from the category of the same constituent when it is *in situ*. San Dionisio Ocotepec Zapotec (SDZ), an Otomanguean language of Mexico, also shows a movement paradox. For a number of verbs with semantics like 'cover/fill/be spread' which subcategorize for a Theme and a Location, the usual argument realization has two NPs after the V. However, when the Theme occurs in an "extraposed" position, it may optionally appear as a PP. This phenomenon finds a natural explanation in LFG, which allows the same grammatical function to have a different categorial realization in different parts of the c-structure.

1 Movement paradoxes and syntactic theory

Bresnan (2001) has drawn attention to *movement paradoxes:*

1) a. *This theory accounts for $_{CP}$[that languages are learnable]
   b. $_{CP}$[That languages are learnable] is accounted for by this theory.

2) a. *We talked about $_{CP}$[that he was sick] for days.
   b. $_{CP}$[That he was sick], we talked about for days.

Theories which use movement to derive passive and topicalization must posit some category-change or other mechanism to account for the facts. In contrast, a theory with base-generated passive subjects and topics faces no such difficulties.

2 Background

San Dionisio Ocotepec Zapotec (SDZ) is an Otomanguean language of Mexico.[1] The most neutral word order is VSO:

---

[1] I thank Ash Asudeh, Joan Bresnan, Miriam Butt, Brook Lillehaugen, Pamela Munro, and Mike Galant for helpful comments on this paper, along with other members of the audience at the 2009 SSILA session on Zapotec languages and the LFG 2010 conference. Special thanks are due to Luisa Martínez, who supplied all the data for this paper.

The orthography for San Dionicio Ocotepec Zapotec is adapted from the practical orthographies for other Zapotec languages spoken in the Valley of Oaxaca. In the SDZ orthography, <x> = /ʒ/ before a vowel and /ʃ/ before a consonant, <xh> = /ʃ/, <dx> = /ʤ/, <ch> = /ʧ/, <c> = /k/ before back vowels, <qu> = /k/ before front vowels, <e> = /ɛ/ and <ey> = /e/. Doubled vowels are long. SDZ is a language with four contrastive phonation types: breathy <Vj>, creaky <V'V>, checked <V'>, and plain <V>. High tone is marked with an acute accent, low with a grave. Nominal tones are affected by position within the intonational phrase, and so nouns may show slightly varying tones from example to example.

A SDZ verb root must always be preceded by some aspect prefix. The citation form of a SDZ verb contains the habitual aspect prefix /r- ~ rr-/. Sample lexical entries show the verb root, while mentions in the text show the citation form.

Ordinary affixes are separated from the stem by the hyphen; clitics are separated by =. Glosses use the following abbreviations: aff = affirmative, com = completive aspect, def = definite future aspect, hab = habitual aspect, neg = negative, neu = neutral aspect, p = possessed, pot = potential aspect, pred = predicative, 1s =1st person singular, 3 = 3rd person human (ordinary respect level), 3i = 3rd person inanimate.

3) Ù-zìì' Juáàny tòyby xhùmbréèjl. VSO
com-buy Juan a hat

'Juan bought a hat.'

In addition to this word order, SDZ also has several word orders in which one or more constituents with a special discourse function precede the verb. Of these variants, one in which the subject appears in the internal topic/focus position ([Spec, IP]) is particularly frequent, yielding SVO order:

4) Juáàny ù-zìì' tòyby xhùmbrèjl. SVO
Juan com-buy a hat

'Juan (TOP/FOC) bought a hat.'

We also see wh-elements in [Spec, CP]:

5) ¿Túú ù-dííny bè'cw?
who com-hit dog

'Who hit the dog'

6) ¿Xhíí cùn ù-dííny Juààny bè'cw?
what with com-hit Juan dog

'What did Juan hit the dog with?'

The unusual order in example (6) also shows pied-piping with inversion (Broadwell 2001).

3 Movement paradox I: Cover themes in San Dionisio Ocotepec Zapotec

3.1 Normal vs. 'displaced' orders

For a number of verbs with semantics like 'cover/fill/be spread over' which subcategorize for a Theme and a Location, the usual argument realization has an NP and a PP after the V.

7) Rr-sè'w nìjs lòò yùù.
hab-cover water on floor

'Water covers the floor.'

If the Theme argument of a verb like 'cover/fill/be spread over' is fronted, it may occur with the preposition *cùn* 'with':[2]

[2] *Cùn* is an old borrowing from Spanish *còn* 'with'.

8) a. Cùn nìjs rr-sè'w lòò yùù
with water hab-cover on floor

'Water (TOP/FOC) covers the floor.'

b.) ¿ Xhíí cùn rr-sè'w lòò yùù?
what with hab-cover on floor

'What covers the floor?'

A fronted NP in these environment is also possible:

9) a.) Nìjs rr-sè'w lòò yùù.
water hab-cover on floor

'Water (TOP/FOC) covers the floor.

b.) ¿Xhíí rr-sè'w lòò yùù?
water hab-cover on floor
'What covers the floor?'

3.2 The impossibility of PP *in situ*

Cùn 'with' may not appear if the Theme is *in situ,* either immediately after the verb or after the other PP:

10) a. *Rr-sè'w cùn nìjs lòò yùù.
hab-cover with water on floor

b. *Rr-sè'w lòò yùù cùn nìjs.
hab-cover on floor with water

(*intended:* Water covers the floor.)

Thus SDZ shows a movement paradox – we have fronted PPs, but an *in-situ* PP is ungrammatical.

The Zapotec alternation is not found with all Theme arguments, but only with those in the 'cover/fill/be spread' class where the Theme is in complete contact with a Location. I will refer to this subclass of Themes as Cover-Themes. Using a slighly modified version of the formalism of Jackendoff (1990:160ff), the semantic representation of such verbs contains the following:

11) $[_{Event}$ INCH $[_{State}$ BE $([_{Thing}\ ]_i, [_{Place}\ \ IN_d/ON_d\ [_{Thing}\ ]_j\ ])\ ]\ ]$

IN_d and ON_d are distributive versions of the IN and ON locational predicates.

3.3 Analogous alternations in the causative counterparts

There is also a homophonous causative version of this verb which subcategorizes for Agent,

Cover-Theme, and Location.

12) Rr-sè'w Juààny dà' lòò yùù
hab-cover Juan mat on floor

'Juan puts the mat on the floor/
Juan covers the floor with the mat.'

These sentences also show an alternation in the realization of the Cover-Theme: NP when *in-situ*, but PP when fronted, as seem in the following:

13) a.) Cùn dà' rr-sè'w Juààny lòò yùù.
with mat hab-cover Juan on floor

'Juan puts the mat (TOP/FOC) on the floor/
Juan covers the floor with the mat.'

b.) ¿Xhíí cùn rr-sè'w Juààny lòò yùù?.
what with hab-cover Juan on floor

'What does Juan put on the floor/
What does Juan cover the floor with?'

The Cover-Theme may also be fronted as an NP:

14) a.) Dà' rr-sè'w Juààny lòò yùù.
mat hab-cover Juan on floor

'Juan puts the mat (TOP/FOC) on the floor./
'Juan covers the floor with the mat (TOP/FOC).'

b.) ¿Xhíí rr-sè'w Juààny lòò yùù.
what hab-cover Juan on floor

'What does Juan put on the floor?'
'What does Juan covers the floor with?'

As seen in the non-causative version of 'cover', it is impossible for the Theme argument to appear as a PP when *in situ:*

15) *Rr-sè'w Juààny cùn dà' lòò yùù
hab-cover Juan with mat on floor

(*intended* 'Juan puts the mat on the floor/Juan covers the floor with the mat.')

3.4 Other verbs that participate in this alternation

We find a similar alternation with a few other verbs like *rr-dá'* 'fill' and *r-yè'* 'spread' (causative *rr-zè'*). Consider the following examples:

16) a.) Rr-dá' nìjs lè'èn rì'.
hab-fill water in jug

'Water fills the jug/The jug is filled with water.'

b.) (Cùn) nìjs rr-dá' lè'èn rì'.
(with) water hab-fill in jug

'Water (TOP/FOC) fills the jug.'

c.) *Rr-dá' cùn nìjs lè'èn rì'.
hab-fill with water in jug.

(*intended*: 'Water fills the jug.')

17) a.) N-yé' màntèquíì lòò gèèt_xtíìly
neu-spread butter on bread

'The butter is spread on the bread.'

b.) (Cùn) màntèquíì n-yé' lòò gèèt_xtíìly
(with) butter neu-spread on bread

'The butter (TOP/FOC) is spread on the bread.'

c.) *N-yé' cùn màntèquíì lòò gèèt_xtíìly
neu-spread with butter on bread

(*Intended:* 'The butter is spread on the bread.')

3.5 Similar verbs that do not alternate

Other verbs that have a Theme and Location do not show this alternation, though they have apparently similar semantics.

Rrdòòyby 'be splashed/splattered with' (causative *rr-tóóyby)* and *rr-dxàj* (causative *rr-cháj*) 'be splashed/spattered' fail to show this alternation:

18) a.) Ù-dòòyby bààny x-càrr=à.
com-splatter mud p-car=1s

'Mud splattered my car.'

b.) *Cùn bààny ù-dòòyby x-càrr=à.
with mud com-splatter p-car=1s

(*Intended:* 'Mud (TOP/FOC) splattered my car.')

19) a.) Rr-dxàj nìjs lòò làády
hab-splashed water on clothes

'Water is splashed on the clothes.'

b.) *Cùn nìjs rr-dxàj lòò làády
with water hab-splashed on clothes

'Water is splashed on the clothes.'

We also see an apparently similar verb *rrgàè'by* 'smear' which does not alternate:

20) a.) Ù-dàè'by Màríí zàj lòò gèèt.
com-smear Maria lard on tortilla

'Maria smeared lard on the tortilla.'

b.) *¿Xhíí cùn ù-dàè'by Màríí lòò gèèt?
what with com-smear Maria on tortilla

(*Intended* 'What did Maria smear on the tortilla?')

The difference between these verbs and those that show the alternation is that spraying, splashing, and smearing do not involve covering a location distributively. Covering and filling do involve complete contact of the Theme with all of the Location.

4 Cross-linguistic comparisons

4.1 Figure-ground alternations

Pinker (1989) examines verbs in English which express a locative relationship between Figure and Ground and finds that they fall into four syntactic classes, depending on whether the Theme or the Location is presented as Ground (NP) and whether both Theme and Location are obligatory.

a.) Non-alternating Figure verbs (Theme NP | Location PP)

21) *John poured water into the glass.*
**John poured the glass with water.*

b.) Non-alternating Ground verbs (Location NP | Theme PP)

22) *John filled the glass with water.*
**John filled the water into the glass.*

c.) Figure-alternating verbs (*pile, spray, load*) (expression of Location optional)

23) *John loaded books (on the table).*
John loaded the table with books.

d.) Ground-alternating verbs (*stuff, paint, wrap)* (expression of Theme optional)

24) *John stuffed feathers into the pillow.*
John stuffed the pillow (with feathers).

For alternating verbs like *load* or *stuff* in English, the simplest treatment is to say that they have two subcategorizations, related by a lexical rule. In one subcategorization, the Theme has an NP realization (*feathers; books*) and in the other it has a PP realization (*with feathers; with books*). However, the NP/PP realization in English is not related to the in-situ/displaced status of the argument:

25) a.) They loaded hay on the wagon. (*NP realization of Theme*)
b.) What did they load on the wagon?
c.) *With what did they load on the wagon?

26) a.) They loaded the wagon with hay. (*PP realization of Theme*)
b.) With what did they load the wagon?
c.) *What did they load the wagon?

The SDZ alternation instead always has the Location realized as PP. The Theme is variably realized as NP or PP in extraposed positions, but as NP only *in-situ*. Schematically, the pattern is as follows:

27) [NP/ PP], [$_S$ V (NP) PP]
Theme *Causer* *Location*

[$_S$ V (NP) NP/*PP PP]
Causer *Theme* *Location*

There is a family resemblance between the English spray/load alternations and the SDZ alternation, but they are not exactly the same.

4.2 Alternations in other languages

Similar alternations are found in many languages, including Korean (Kim, Landau, and Phillips 1999), Hungarian[3] (Ackerman 1992), and Modern Greek (Kordoni 2003):

[3] Ackerman (1992) shows that the Hungarian alternation is accompanied by a change in the prefix which indicates the telicity of the action.

28) **Hungarian**

a.) A paraszt (rá=)rakta **a szénát** a szekerre
the peasant (onto)=load the hay:acc the wagon:sublative

'The peasant loaded hay onto the wagon.'

b.) A paraszt meg=rakta a szekeret **szénával.**
the peasant perf=load the wagon hay:instr

'The peasant loaded the wagon with hay.'

29) **Korean**

a.) Yumi-ka **kirul-ul** pyek-ey chilha-ess-ta.
Yumi-nom oil-acc wall-loc paint-past-decl

'Yumi painted the oil onto the wall.'

b.) Yumi-ka pyek-ul **kirul-elo** chilha-ess-ta.
Yumi-nom wall-acc oil-applic paint-past-Decl

'Yumi painted the wall with the oil.'

30) **Greek**

a.) O georgos fortose **to ahiro** sto karo.
the farmer:nom load:past:3s the hay:acc onto:the wagon

'The farmer loaded the hay onto the wagon.'

b.) O georgos fortose to karo **me** **ahiro**
the farmer:nom load:past:3s the wagon with hay:acc

'The farmer loaded the wagon with hay.'

The generalization seems to be that Cover-Themes frequently show whatever morphosyntax is found with Instruments, whether this is an adposition (Zapotec, English, Greek) or a case marker (Korean, Hungarian).

5 Toward a solution

In order to capture the similarities between SDZ and the other languages cited, I argue that we need the combination of a lexical rule and Lexical Mapping Theory to produce two entries for verbs like *rr-sè'w* 'cover'. (31a) shows the Cover-Theme encoded as Theme; (31b) shows the Cover-Theme encoded as Instrument:

31) *sè'w,* V,

a. $[_{Event}$ INCH $[_{State}$ BE $([_{Thing}\ ]_A, [_{Place}\ \ IN_d/ON_d\ [_{Thing}\ ]_A\ ])\ ]\ ]$
SUBJ OBL$_{Loc}$

b. $[_{Event}$ INCH $[_{State}$ BE $([$WITH $[_{Thing}\ ]_A\ ], [_{Place}\ \ IN_d/ON_d\ [_{Thing}\ ]_A\ ])\ ]\ ]$
SUBJ OBL$_{Loc}$

In these lexical entries, subscript A marks the elements of syntactic structure which must be linked to some syntactic role.

Note that the entry does not say that the first argument must be a NP or that OBL$_{Loc}$ must be a PP. This is determined by the phrase-structure rules of the language, where the mapping between categories and functions is constrained by the annotations.

The SDZ lexical rule that relates (31a) to (31b) is approximately as follows:

32) *Zapotec Cover-Theme Rule*

In a Lexical Conceptual Structure which contains

$[_{Event}$ INCH $[_{State}$ BE $([_{Thing}\ ]_A, [_{Place}\ \ IN_d/ON_d\ [_{Thing}\ ]_A\ ])\ ]\ ],$

$[_{State}$ BE $[_{Thing}]_A$ ...] is optionally replaced by $[_{State}$ BE [WITH $[_{Thing}]\ ]_A$ ...]

The Zapotec Cover-Theme Rule is a language-particular variant of a lexical rule found in many languages which allows a Cover-Theme to be expressed as an instrumental.[4]

In many other languages just cited, a rule like the Cover-Theme rule acts in concert with a rule which we might call the Distributive Location Rule. This rule allows a Location with IN$_d$/ON$_d$ to appear as an unrestricted GF (SUBJ or OBJ). [5]

The combination of the two rules gives us familiar alternations like the English *load* alternation cited above:

33) a.) They loaded hay on the wagon.
b.) They loaded the wagon with hay.

The Zapotec alternation is important, however, since it shows that the two rules are not necessarily linked. Zapotec allows Cover-Themes to alternate between NP and 'with NP' but does not allow Locations to alternate between PP and NP.

Note that sentences like the following, which are parallel to (33b), are ungrammatical, since

---

[4] Jackendoff (1990:161ff) gives a comparable English rule, though his WITH-Theme Adjunct Rule covers a much wider range of verb classes where *with* appears.

[5] The English rule allowing a Location with IN$_d$/ON$_d$ to have the feature [-r], and thus link to SUBJ/OBJ is a language-particular instantiation of a more general process allowing Locations to be [-r]. Locative inversion in Chicheŵa (Bresnan and Kanerva 1989) is probably another member of this family of rules.

Zapotec does not have any lexical rule that allows a Location to surface as SUBJ or OBJ.

34) *Rr-sè'w Juààny yùù cùn dà'
hab-cover Juan floor with mat

(*Intended*: Juan covered the floor with mat.)

6 The movement paradox again

I will assume that SDZ contain phrase structure rules of the following sort:

35) CP –> (COMP) (XP) IP
↑=↓ (↑INTERROG)=↓ ↑=↓
(↑GF)=↓

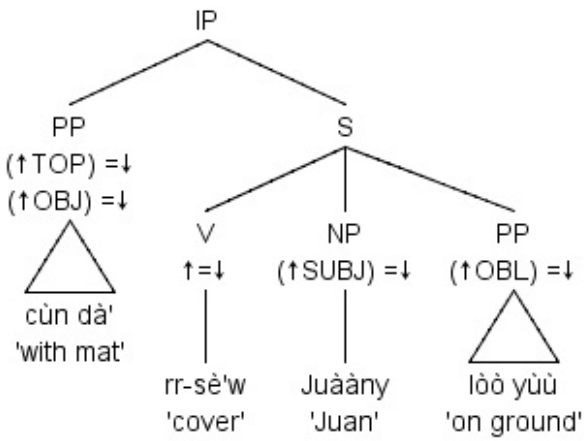
IP –> (Infl) (XP) S
↑=↓ (↑TOPIC)=↓ ↑=↓
(↑GF)=↓

S –> V (NP) (NP) (NP) PP*
↑=↓ (↑SUBJ)=↓ (↑OBJ_θ)=↓ (↑OBJ)=↓ (↑OBL_θ)=↓

The TOPIC and INTERROG functions may be assigned to any XP, but the SUBJ, OBJ, and OBJ_θ positions are restricted to NPs. Crucially, a postverbal PP cannot be a SUBJ or OBJ.

With these considerations in mind, the proposed c-structure representation of the movement paradox case is as follows:

36)



From the LFG perspective, the movement 'paradox' is not a paradox at all. A lexical rule which allows Cover-Themes to optionally appear as *with*-phrases has applied to make the Theme a SUBJ or OBJ. However, the phrase structure (PS) rules of Zapotec do not allow *in-situ* PPs as SUBJ or OBJ. Thus we only see the effect of the lexical rule when the argument is in a position outside S.

7 Movement paradox II – Comitative applicatives

7.1 Basic facts

SDZ has an optional comitative applicative /-néé/. Thus the following two sentences express the same content:

37) Ù-dàw Màrííí gèèt cùn Juáány
com-eat Maria tortilla with Juan

'Maria ate tortillas with John.'

38) Ù-dàw-**néé** Màrííí Juáány gèèt.
com-eat-applic Maria Juan tortilla

'Maria ate tortillas with John.'

When the comitative applicative appears, the OBJ_{Comit} appears between the SUBJ and the OBJ, as shown in the phrase structure rules in (35) above.

It is not possible to have both the applicative and the preposition *cùn* when a comitative object is *in-situ*:

39) *Ù-dàw-**néé** Màrííí cùn Juáány gèèt.
com-eat-applic Maria with Juan tortilla

'Maria ate tortillas with John.'

As we shall see in the following section, however, when comitative objects are in fronted or extraposed positions, the *cùn* may reappear, giving rise to a second movement paradox.

7.2 Extraposition possibilities in sentences with comitatives

Comitative objects are largely like other objects in their syntax. In sentences with regular and comitative objects, both can be questioned.

40) ¿Xhíí ù-dàw-**néé** Màrííí Juáány?
what com-eat-applic Maria Juan

'What did Maria eat with Juan?'

41) ¿Túú ù-dàw-**néé** Màrííí gèèt?
who com-eat-applic Maria tortilla

'Who did Maria eat tortillas with?'

It is also possible to question both the regular object and the object of the preposition *cùn* 'with' in the sentences without an applicative:

42) ¿Xhííí ù-dàw Màrííí cùn Juáány?
what com-eat Maria with Juan

'What did Maria eat with Juan?'

43) ¿Túú cùn ù-dàw Màrííí gèèt?
who with com-eat Maria tortilla

'Who did Maria eat tortillas with?'

It is surprising, therefore that a questioned comitative object may appear with the preposition *cùn.* Recall that this is ungrammatical *in-situ*:

44) ¿Túú cùn ù-dàw-**néé** Màrííí gèèt?
who with com-eat-applic Maria tortilla

'Who did Maria eat tortillas with?'

45) *Ù-dàw-**néé** Màrííí cùn Juáány gèèt.
com-eat-applic Maria with Juan tortilla

'Maria ate tortillas with John.'

7.3 A lexical rule for Comitatives

The lexical rule for adding a comitative might be as follows:

46) Zapotec Comitative Applicative[6]

/X/, V, $[_{Event}$ CAUSE $[_{Thing}$ $]_{A}$ ...] ==>

/X-néé/, V, $[_{Event}$ CAUSE $[_{Thing}$ $]_{A}$ [WITH $[_{Thing}$ HUMAN $]]_{A\ [+r]}$ ...]

[6] I have tentatively included CAUSE in such lexical rules, since it seems that comitatives do not appear with all lexical classes, but are confined to those with causative (or agentive) semantics.

Likewise, I have tentatively made the Comitative [+Human], since my language consultant is very reluctant to accept non-human co-subjects (e.g. ?* *Ù-dàw-néé Màrííí bè'cw gèèt* 'Maria ate the tortilla with a dog.'). It is not absolutely clear to me whether this is a grammatical fact or a fact about Zapotec ideas of human/non-human co-agency (or both).

Rákosy (2003) suggests the label Partner for the Thematic role associated with Comitatives, and notes that this role is associated with the feature [+r] in Lexical Mapping Theory. Combining this idea with Jackendoff's notation, the A[+r] subscript indicates that the comitative is marked as a [+r] argument.

This rule combined with regular rules of linking will produce a lexical entry like the following for *r-àw-néé* 'eat with':

47) /àwnéé/, V,

$$[_{Event}\ \text{CAUSE}\ [_{Thing}\ \alpha\ ]_{A}\ [\text{WITH}\ [_{Thing}\ \text{HUMAN}\ ]]_{A\ [+r]},\ [\text{GO}\ [\text{FOOD}]_{A}\ [\text{TO}\ [\text{IN}\ [\text{MOUTH-OF}\ [_{Thing}\ \alpha]\ ]]]]$$

SUBJ OBJ$_{Comit}$ OBJ

Because the phrase structure rules for Zapotec require an *in-situ* OBJ$_{\theta}$ to be an NP, we account for the ungrammaticality of the following example (repeated from above):

48) *Ù-dàw-**néé** Màrií cùn Juáàny gèèt.
com-eat-applic Maria with Juan tortilla

'Maria ate tortillas with John.'

However, just as with the Cover-Themes seen above, the initial positions [Spec, CP] and [Spec, IP] allow any XP to appear, with any GF. Thus we correctly predict the grammaticality of the following:

49) ¿Túú cùn ù-dàw-**néé** Màrií gèèt?
who with com-eat-applic Maria tortilla

'Who did Maria eat tortillas with?'

In an example like this, the initial PP will bear the OBJ$_{Comit}$ function, even though a PP cannot bear this GF when *in-situ*.

8 Conclusions

8.1 The role of LCS in linking theory

Standard Lexical Mapping Theory starts with Argument Structure, usually represented as a predicate plus a list of the Thematic roles that it assigns. Thus the Argument structures for the verbs *rr-sè'w* 'cover' and *rr-gàè'by* 'smear' would be as follows:

50) COVER <Agent Theme Location>
SMEAR<Agent Theme Location>

It is clear that this is not sufficient detail to distinguish these verbs so that we know that *rr-sè'w* 'cover' shows a NP/PP alternation for its Theme, but *rr-gàè'by* 'smear' does not.

Following Jackendoff (1983, 1990), I will assume that the lexical entry of a verb contains its Lexical-Conceptual Structure, which gives an explicit representation of its lexical semantics. Thematic

roles are conventional labels for particular arguments of underlying predicates like CAUSE, GO, AFFECT, etc.

The conventional Argument Structure can be thought of as a projection of Lexical-Conceptual Structure, where we look only at the parts of LCS that link to arguments. Lexical Mapping Theory regulates the way in which these arguments are realized as grammatical functions. But the Thematic Roles assigned to arguments by Lexical Mapping Theory are not themselves primitives – they are simply recurrent portions of Lexical-Conceptual Structure.

Lexical Mapping Theory need not be abandoned *in toto*, but the conventional starting point of LMT – a list of Thematic Roles – only works well for the most frequently-discussed and clear-cut verb classes. When we look at verbs with more complex lexical semantics, Argument Structure is too impoverished a representation to be able to account for the full range of effects. We need instead a more detailed representation like Lexical Conceptual Structure.

8.2 LFG architecture and movement paradoxes

Lexical-Functional Grammar gives a simple account of the two movement paradoxes discussed here. A few features of the LFG architecture are key to the solution. First is a theory of phrase-structure rules that allows a specified categorial realization (NP or PP) in some positions and a free categorial realization (XP) in others. Second is a concept of lexical entries that specifies the semantics and grammatical function, but not the categorial realization of arguments.

LFG is different in these respects from other syntactic theories such as Minimalism. Minimalism eschews the sort of detailed phrase structure rules that allow us to distinguish in-situ realizations of Cover-Themes and Comitative objects from their extraposed varieties (using simple versions of X-bar theory). Minimalism also has a theory of lexical entries which specifies the categorial realization of subcategorized arguments.

Thus movement paradoxes show that these aspects of the LFG architecture correctly model natural language in a way that is difficult for theories with different approaches to phrase-structure rules and lexical entries.

9 Appendix – Right extraposition

We also find effects similar to those mentioned above when Cover-Themes and Comitative Objects appear in right-extraposed positions.

The condition for comitative objects is that they optionally appear with *cùn* if they are right-extraposed via Heavy NP Shift:

51) Ù-dàw-**néé** Chéé gèèt [(cùn) dáád ní=ù-dè'd mèèl lòò=èby.]
com-eat-applic José tortilla (with) man rel=com-give money to=3

'José ate tortillas with the man who gave him money.'

Heavy NP-shift is near obligatory for sentences of this sort; NP and PP are both very marginal *in-situ:*

52) *?Ù-dàw-**néé** Chéé [(cùn) dáád ní=ù-dè'd mèèl lòò=èby] gèèt
com-eat-applic José (with) man rel=com-give money to=3 tortilla

'José ate tortillas with the man who gave him money.'

Another environment for a PP realization of the Cover-Theme is found with verbs of this sort. That is Rightward Extraposition:

53) Rr-sè'w Juààny lòò yùù, cùn dá'.
hab-cover Juan on floor with mat

'Juan put on the floor the mat/
Juan covered the floor with the mat.'

However, the Cover-Theme apparently cannot be Right Extraposed as an NP:

54) *Rr-sè'w Juààny lòò yùù dá'.
hab-cover Juan on floor mat

(*Intended:* Juan put on the floor the mat.)

This may be due to the fact that PPs can be rather freely extraposed, while NPs are only extraposed when sufficiently heavy.

I have not attempted a formalization of these facts here, since they appear to require a characterization of heaviness which would go beyond this scope of this paper.

10 References

Ackerman, Farrell. 1992. Complex predicates and morphosyntactic relatedness: The locative alternation in Hungarian, in Ivan Sag and Anna Szabolcsi (eds.) *Lexical matters*. Stanford: CSLI.

Bresnan, Joan. 2001. *Lexical Functional Syntax*. Oxford: Blackwell.

Bresnan, Joan and Jonni Kanerva. 1989. Locative inversion in Chichewa: A case study for factorization in grammar. *Linguistic Inquiry* 20:1-50.

Broadwell, George Aaron. 2001. Optimal order and pied-piping in San Dionicio Zapotec. in Peter Sells, (ed.) *Formal and Empirical Issues in Optimality Theoretic Syntax*. Stanford: CSLI Publications.

Jackendoff, Ray. 1983. *Semantics and cognition.* Cambridge, MA: MIT Press.

Jackendoff, Ray. 1990. *Semantic structures*. Cambridge, MA: MIT Press.

Kim, Meesook, Barbara Landau, and Colin Phillips. 1999. Cross-linguistic differences in children's syntax for locative verbs. in *Proceedings of the 23rd Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press. http://www.ling.umd.edu/colin/research/papers/bucld23.pdf (Accessed July 29, 2008).

Kordoni, Valia. 2003. "Valence alternations in Modern Greek: an MRS analysis." in *Proceedings of the 9th International Conference on HPSG*. Stanford: CSLI http://csli-publications.stanford.edu/HPSG/3/kordoni.pdf (Accessed December 26, 2008).

Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Rákosy, György. 2003. Comitative arguments in Hungarian. Uil-OTS Yearbook. (http://tigris.unideb.hu/~aai/admin/file_1519.pdf)

TWO APPROACHES TO AUTOMATIC MATCHING OF ATOMIC GRAMMATICAL FEATURES IN LFG

Anton Bryl and Josef van Genabith
CNGL, School of Computing,
Dublin City University

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

The alignment of a bilingual corpus is an important step in data preparation for data-driven machine translation. LFG f-structures provide bilexical labelled dependencies in the form of lemmas and core *grammatical functions* linking those lemmas, but also important *grammatical features* (TENSE, NUMBER, CASE, etc.) representing morphological and semantic information. These grammatical features can often be translated independently from the lemmas or words. It is therefore of practical interest to develop methods that align grammatical features which can be considered translations of each other (e.g. the number features of the corresponding words in the source and target parts of the corpus) in data-driven LFG-based MT. In a parallel grammar development scenario, such as ParGram, this is to a large extent captured through manually hardcoding the correspondences in the hand-crafted grammars, using similar or identical feature names for similar phenomena across languages. However, for a completely automatic learning method it is desirable to establish these correspondences without human assistance. In this paper we present and evaluate two approaches to the automatic identification of correspondences between atomic features of LFG (and similar) grammars for different languages. The methods can be used to evaluate the correspondence between feature names in hand-crafted parallel grammars or find correspondences between features in grammars for different languages where feature alignments are not known.

1 Introduction

Recent attention to deep linguistic representations, such as LFG f-structures, in syntax-based statistical machine translation (SMT) (Avramidis and Kuhn, 2009; Graham and van Genabith, 2009; Riezler and Maxwell III, 2006) poses new problems for processing richly annotated data. In our work we focus on one such problem, namely on deepening the automatic cross-language structure alignment by adding the possibility to align not only words, but also atomic f-structure features.

Each lexical node (essentially a node with a PRED feature) in an f-structure is characterized by a number of atomic-valued features, which contain information about case, tense, gender, etc. Though the sets of features differ for different languages, they are far from being disjoint. A number of features, such as number or case are shared between many languages. It is possible to hardcode the knowledge about these similarities in the grammars; that is, to give the same names to the same (or similar) linguistic properties for grammars for different languages (Butt et al., 1999). However, when one wants to make use of correspondences between such features in a language-agnostic syntax-based SMT system, such "feature name alignment" between source and target grammars cannot simply be taken for granted. Moreover, the degree of correspondence may differ from

[†]This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) at Dublin City University. The authors thank Ms. Yvette Graham, whose preprocessed experimental data was reused for evaluation in this paper.

feature to feature across grammars. This motivates the need for an automatic way to judge correspondences between atomic features in f-structure representations for arbitrary language pairs.

In this paper we show that, provided that we have parsers for two languages and a word-aligned parallel corpus for these languages, it is possible to automatically identify some of the correspondences between atomic-valued grammatical features. Once identified, these feature pairs can further be used to improve the coverage of transfer-based statistical machine translation. For example, if the algorithm identifies that NUM in English and NUM in German generally *co-vary* (we presume that we do not have any prior knowledge about this correspondence), then we can safely induce transfer rules (from aligned parsed bitext corpora) which abstract over the number feature, effectively providing a back-off to more specific transfer rules (which include NUM information), safe in the knowledge that number features can be transferred independently in the majority of cases. At the same time, if an English feature and a German feature, even though they were intended to capture the same (or similar) phenomena by the grammar writers, do not change simultaneously with enough systematicity, then these features are not safe for use in feature-by-feature translation, and one should keep using more specific "building blocks", e.g. complete sub-f-structures, to create transfer rules.

We present two algorithmic solutions to the problem in question, one assessing the frequency of covariation of feature pairs and the other calculating the mutual predictability of source language (SL) and target language (TL) features. We evaluate both methods on German-English Europarl data. We show that the first method identifies a number of correspondences correctly, without false positives, though some theoretically expected correspondences are not identified. The second method proved to be less precise; it is able to detect many correct correspondences, but produces false positives as well.

The paper is organized as follows: Section 2 explains the motivation, Sections 3 and 4 describe the two proposed methods together with the experimental evaluation, and Section 5 presents conclusions.

2 Motivation

Let us first briefly describe the way in which transfer rules for LFG-based SMT may be induced (Graham and van Genabith, 2009). Consider a simplified example where transfer rules are extracted from a word-aligned English-German string pair "big cat" ⇒ "große Katze". After parsing, the English and German f-structures are as shown in Figure 1.

The word alignment determines which German word corresponds to which English word and allows us to abstract over the adjuncts in our example. As the German sub-f-structure with PRED='groß' corresponds to the English sub-f-structure with PRED='big', we can replace both with x and obtain the generalized transfer rule in Figure 2.

$$\begin{bmatrix} \text{PRED} & \text{‘cat’} \\ \text{NUM} & \text{sg} \\ \text{CASE} & \text{nom} \\ \text{ADJUNCT} & \left\{ \begin{bmatrix} \text{PRED} & \text{‘big’} \end{bmatrix} \right\} \end{bmatrix} \qquad \begin{bmatrix} \text{PRED} & \text{‘Katze’} \\ \text{NUM} & \text{sg} \\ \text{CASE} & \text{nom} \\ \text{ADJUNCT} & \left\{ \begin{bmatrix} \text{PRED} & \text{‘groß’} \end{bmatrix} \right\} \end{bmatrix}$$

Figure 1: Parsed structures

$$\begin{bmatrix} \text{PRED} & \text{‘cat’} \\ \text{NUM} & \text{sg} \\ \text{CASE} & \text{nom} \\ \text{ADJUNCT} & x \end{bmatrix} \Rightarrow \begin{bmatrix} \text{PRED} & \text{‘Katze’} \\ \text{NUM} & \text{sg} \\ \text{CASE} & \text{nom} \\ \text{ADJUNCT} & x \end{bmatrix}$$

Figure 2: Generalized Transfer Rule

To perform this abstraction, the method does not need to know that ADJUNCT in English and ADJUNCT in German are in correpondence. Word alignment alone is sufficient. In fact, this approach will work equally well in the case when e.g. an active construction is translated as a passive construction involving an object and subject argument switch; the word alignment still tells us what corresponds to what. If the method relied on the correspondences of the grammatical function names, its automaticity would be compromised to a certain extent: before being able to learn transfer rules, it would require a team of linguists to manually synchronize the names of grammatical functions for the two languages. Such an alignment cannot be taken for granted if our goal is a fully automatic approach.

Consider the resulting transfer rule (Figure 2) and how it can be improved. It is obvious that the rule could abstract not only over the adjunct *grammatical function*, but also over the number, and maybe also over the case *grammatical feature*. For this, however, we do not have word alignment as our guide. Grammatical functions are bilexical labelled dependencies, and two pairs of aligned words allow us to easily identify a translational correspondence between the grammatical functions which relate them. Grammatical features, such as number and case, are, on the other hand, connected with only one head word. Unless we know which particular German feature corresponds to English NUM, we have no way to align the features and use this alignment for producing more abstract and general transfer rules.

There are two ways to establish correspondences between the features of two languages. One is to turn to human judgement (e.g. use grammars with synchronized terminology and rely on the same-name correspondence), thus making the method less automatic and more dependent on human preprocessing of the data. The other way is to try to extract this correspondence automatically from the data.

An attempt to explore this second way is what the present paper is focused on.

As the methods presented here are intended to compute what is often the judgement of grammar designers, they can also help them in their work. That is, the approaches can be used in grammar design to match a new against an existing grammar, collecting empirical evidence for the correspondence between the features of the two languages.

3 Method 1: Searching for Covariation

3.1 Definition

The idea behind the first method is that if a feature **A** in one language corresponds to a feature **B** in another language, then a change in the value of **A** in a certain phrase frequently corresponds to a change in the value of **B** in the aligned translation of this phrase.

To apply the method to a given pair of languages, we need a parser for each of the languages, and a parallel corpus. The basic procedure (without optimization) is given in Figure 3. It can be seen that for each atomic feature occurring in the first language we will get either a corresponding feature from the second language, or a result "NONE" which means that there is no corresponding feature. Once the features are matched, it is easy to establish correspondences between their values; indeed, if we know that feature **Lang1.A** is translated into feature **Lang2.B**, we just have to calculate which value of **Lang2.B** co-occurs most frequently with each particular value of **Lang1.A**.

It is easy to see that the method is asymmetric with respect to the languages involved. This is meaningful in the context of SMT: even if a change in feature **Lang1.A** can be safely translated with a corresponding change in feature **Lang2.B**, it does not mean that the correspondence will work equally well in the opposite direction.

The method as described above assumes that, apart from the focus feature under consideration, all other local features remain unchanged. However, it is easy to make the method iterative: that is, to remove those features which were already matched from the corpus and to re-run the same algorithm again. This allows the method to make use of the node-pairs which differ in more than one atomic feature.

3.2 Evaluation

In this section we present an experimental evaluation of the algorithm, and discuss the results.

For our experiments we used 219,667 sentences from the German-English part of the Europarl corpus (Koehn, 2005) parsed into f-structures with the XLE parser (Kaplan et al., 2002) using English (Riezler et al., 2002) and German (Dipper, 2003; Rohrer and Forst, 2006) LFGs. For word alignment Giza++ (Och et al., 1999) was used.

1. Word-align the bitext corpus using a standard method based on co-occurrence of lexemes.
2. Parse both language sections of the corpus with the corresponding parsers. *We will say that a node x of an f-structure in the first language corresponds to a node y in the second language if the heads of these two nodes match in the word alignment. The notation is $y = a(x)$.*
3. For each atomic feature **A** occurring in the first language the following actions are performed:
   (a) For the first language all pairs (x_1, x_2) of sub-structures are identified which differ in the value of feature **A** but agree in the values of all other atomic features. For example, for **A**=`NUM`, if in one local f-structure the word "dog" occurs in plural and in another sentence in singular, while the rest of the features, including e.g. case, carry the same values, these two occurrences form a pair (Figure 4).
   (b) For each such pair:
      i. Compare the aligned structures $a(x_1)$ and $a(x_2)$ in the other language of the bitext.
      ii. If they differ only in the value of a single atomic feature **B**, increase the counter $C_{A,B}$ and the counter C_A
      iii. If they are identical, increase the counter $C_{A,NONE}$ and the counter C_A
4. *Search for correspondences:*
   while at least one feature **A** of the first language with $C_A > 0$ has no correspondence selected, repeat:
   (a) Find the correspondence with the highest normalized score: $(A, B) = \arg\max_{A,B} \frac{C_{A,B}}{C_A}$
   (b) Record **B** as the correspondence to **A**: add the pair (`A` $\Rightarrow$ `B`) to the output.
   (c) For all possible values of B_i set $C_{A,B_i} := 0$ *(only one correspondence is chosen for each feature of the first language; once matched, feature **A** is excluded from further consideration).*
   (d) If $B \neq NONE$, for all possible values of A_i set $C_{A_i,B} := 0$ *(each answer except "NONE" may be chosen only once; once chosen, feature **B** is excluded from further consideration).*

Figure 3: The algorithm.

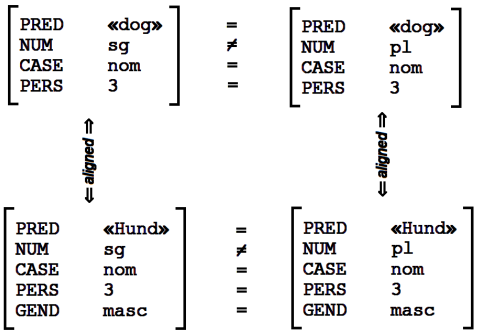


Figure 4: Simultaneous change of the values of `NUM` in parallel data. Finding such a situation, the algorithm increases the probability counter for the (`Eng.NUM` ⇒ `Ger.NUM`) correspondence.

Grammatical features whose values are non-atomic but are sub-structures without a `PRED`, e.g. `TNS-ASP`, were also treated as atomic features. Their values were considered equal if and only if the sub-structures were completely identical. Two iterations of the algorithm were run. As the algorithm is asymmetric with respect to the order of languages, the procedure was performed in both directions: correspondences for German atomic features were found in English, and vice versa. Additionally, we ran the first iteration of the method excluding the step 4.c from the algorithm; that is, permitting one feature of the second language to be chosen as a correspondence for several features of the first language.

The results are presented in Table 1. As the XLE LFG grammars for different languages are very consistent (Butt et al., 1999), all the pairs identified consist of same-name features. There are no false positives, but some expected correspondences were not identified, e.g. (`ATYPE` ⇒ `ATYPE`) for English-to-German.

The exclusion of the step 4.c has no effect on German-to-English results. For English-to-German, two more pairs emerge: (`PFORM` ⇒ `CASE`) and (`DEG-DIM` ⇒ `DEGREE`). We observe that these pairs, though not corresponding to exact matches, are far from being random: the second pairs up two closely related attributes, while the first reflects the similarity of semantic functions expressed by prepositions and noun cases.

4 Method 2: Measuring the Predictability

4.1 Definition

The second method makes use of the predictability of target features by the source features. For each possible pair (**Lang1.A**, **Lang2.B**) we calculate the best possible accuracy of the deterministic prediction of the value of **Lang2.B** in the target language structure by the value of **Lang1.A** in the aligned source language f-structure (the absence of a certain feature in a structure is considered here a special *feature*

a. German-to-English

Pair	N_{it}	$\frac{C_{A,B}}{C_A}$
(NUM ⇒ NUM)	1	0.88
(TNS-ASP ⇒ TNS-ASP)	1	0.70
(CLAUSE-TYPE ⇒ CLAUSE-TYPE)	1	0.62
(CASE ⇒ CASE)	1	0.51
(ATYPE ⇒ ATYPE)	1	0.79
(COMP-FORM ⇒ COMP-FORM)	2	0.92
(PASSIVE ⇒ PASSIVE)	2	0.64

b. English-to-German

Pair	N_{it}	$\frac{C_{A,B}}{C_A}$
(NUM ⇒ NUM)	1	0.67
(TNS-ASP ⇒ TNS-ASP)	1	0.81
(CASE ⇒ CASE)	1	0.86
(DEGREE ⇒ DEGREE)	1	0.98
(PASSIVE ⇒ PASSIVE)	2	0.55

Table 1: Experimental results. N_{it} is the number of iteration on which the pair emerged. $\frac{C_{A,B}}{C_A}$ is the normalized score (see algorithm in Figure 3).

absent value of this feature; so in the end each feature is considered to be occurring in each structure, sometimes with this special value). This accuracy is calculated as follows:

$$Acc_{best}(Lang1.A, Lang2.B) = \frac{1}{N} \times \sum_{a \in \mathfrak{V}(Lang1.A)} \max_{b \in \mathfrak{V}(Lang2.B)} \mathfrak{C}(a, b), \quad (1)$$

where $\mathfrak{V}(.)$ is the set of values of the given feature, $\mathfrak{C}(.,.)$ is the counter of co-occurrences of values, and N is the total number of aligned pairs of substructures.

The idea behind the above formula is the following: for each input (a value of Lang1.A) a deterministic predicting algorithm gives only one output (a value of Lang2.B); then, the best possible answer in each case is the value of Lang2.B which collocates with the given value of Lang2.A most frequently.

The accuracy of the pick-most-frequent baseline is then subtracted from the value of Acc_{best}. The resulting value, that is the increase in prediction accuracy over the baseline, is used as the matching scores for the pair of features. Once the matching scores for all feature pairs are calculated, a greedy algorithm is used to establish a one-to-one correspondence.

Pair	Score
(NUM ⇒ NUM)	0.3481754277389169
(PERS ⇒ PERS)	0.34524304388880844
(NTYPE ⇒ NTYPE)	0.3207387801123815
(CASE ⇒ CASE)	0.20730678870766603
(PASSIVE ⇒ PASSIVE)	0.16556148908789317
(CLAUSE-TYPE ⇒ CLAUSE-TYPE)	0.14868700122581083
(PRON-TYPE ⇒ PRON-FORM)	0.1256236959989163
(VTYPE ⇒ VTYPE)	0.11999399711277062
(TNS-ASP ⇒ STMT-TYPE)	0.11956901394608822
(PRON-FORM ⇒ PRON-TYPE)	0.09944473293128155
(ATYPE ⇒ DEG-DIM)	0.08733271268083366
(DEGREE ⇒ DEGREE)	0.0526062756732995
(PTYPE ⇒ PTYPE)	0.047626004188740335
(DET-TYPE ⇒ DET-TYPE)	0.031052988760523315
(ADJUNCT-TYPE ⇒ ADJUNCT-TYPE)	0.021342123401830882
(NUMBER-TYPE ⇒ NUMBER-TYPE)	0.010531614099347783
(ADV-TYPE ⇒ ADV-TYPE)	0.004314907214222062
(COMP-FORM ⇒ COMP-FORM)	0.0032471370079325762
(DEG-DIM ⇒ ATYPE)	0.0027398133527054827
(PRT-FORM ⇒ TNS-ASP)	3.2670580938708133E-4

Table 2: Experimental results for Method 2: German-to-English.

4.2 Evaluation

For this experiment we used the same data and the same alignment tool as in the previous case (see Section 3.2).

In Tables 2 and 3 we present the results. Though many correspondences are established correctly, there are also clear false positives, e.g. TNS-ASP is matched incorrectly in all cases. One possible reason is the great number of possible values of TNS-ASP, which results in a noisy estimate of prediction accuracy (see the formula in Section 4.1). It may be better to treat sub-features of TNS-ASP as separate features when using this method.

5 Conclusion

In this paper we presented two methods for the automatic identification of functionally similar atomic features in Lexical-Functional Grammars for a given pair of languages. The experimental evaluation shows that the first method, based on the covariation of feature values, is capable of finding strong correspondences for the German-English language pair. In our experiments it produced no false positives and we consider it suitable for inclusion in an MT system. The second method,

Pair	Score
(NUM ⇒ NUM)	0.36233400755141965
(NTYPE ⇒ NTYPE)	0.35840689747679527
(PERS ⇒ PERS)	0.35432041871466496
(CASE ⇒ CASE)	0.24313951002097026
(PASSIVE ⇒ PASSIVE)	0.1679174895181886
(CLAUSE-TYPE ⇒ CLAUSE-TYPE)	0.1540895997322606
(TNS-ASP ⇒ TNS-ASP)	0.1346346672049787
(PRON-TYPE ⇒ PRON-FORM)	0.13356424085389743
(VTYPE ⇒ VTYPE)	0.12779907858337175
(DEGREE ⇒ DEG-DIM)	0.08904991028870966
(HUMAN ⇒ PRON-TYPE)	0.08029658512744847
(ATYPE ⇒ ATYPE)	0.062197614516362514
(PTYPE ⇒ PTYPE)	0.058877433526656406
(DET-TYPE ⇒ DET-TYPE)	0.02897588353336317
(ADJUNCT-TYPE ⇒ ADJUNCT-TYPE)	0.02128900050599558
(NUMBER-TYPE ⇒ NUMBER-TYPE)	0.014094832337500382
(ADV-TYPE ⇒ ADV-TYPE)	0.006284438577315726
(COMP-FORM ⇒ COMP-FORM)	0.003548609441797891
(DEG-DIM ⇒ DEGREE)	6.374747500235733E-4

Table 3: Experimental results for Method 2: English-to-German.

based on mutual predictability of features, proved to be less precise, producing a considerable number of false positives; some improvement is needed before the method becomes accurate enough for practical use.

ParGram XLE LFG grammars are developed manually in parallel with close alignment of the feature space used in the grammars across languages (Butt et al., 1999). It would be interesting to test these methods also on non-parallel grammars developed independently, such as the HPSG and LFG grammars for English and Norwegian, respectively, used in the LOGON project (Lønning et al., 2004). Running the algorithms on two different grammars for one language (e.g. English LFG and English HPSG) might also be of interest.

References

Avramidis, E. and Kuhn, J. 2009. Exploiting XLE's finite state interface in LFG-based statistical machine translation. In *Proceedings of LFG'09*.

Butt, M., King, T. H., Niño, M.-E. and Segond, F. 1999. *A Grammar Writer's Cookbook*. CSLI Publications.

Dipper, S. 2003. *Implementing and Documenting Large-scale Grammars - German LFG*. Ph. D.thesis, IMS, University of Stuttgart.

Graham, Y. and van Genabith, J. 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, 91, 37–46.

Kaplan, R. M., King, T. H. and Maxwell III, J. T. 2002. Adapting existing grammars: the XLE experience. In *COLING'02, Workshop on Grammar Engineering and Evaluation*, pages 29–35.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit X*, pages 79–86.

Lønning, J. T., Oepen, S., Beermann, D., Hellan, L., Carroll, J., Dyvik, H., Flickinger, D., Johannessen, J. B., Meurer, P., Nordgård, T., Rosén, V. and Velldal, E. 2004. LOGON. A Norwegian MT Effort. Uppsala, Sweden.

Och, F. J., Tillmann, C. and Ney, H. 1999. Improved Alignment Models for Statistical Machine Translation. In *Proceedings of EMNLP99*, pages 20–28.

Riezler, S., King, T. H., Kaplan, R. M., Crouch, R., Maxwell III, J. T. and Johnson, M. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *ACL'02*, pages 271–278.

Riezler, S. and Maxwell III, J. T. 2006. Grammatical Machine Translation. In *HLT-ACL*, pages 248–255.

Rohrer, C. and Forst, M. 2006. Improving coverage and parsing quality of a large-scale LFG for German. In *In Proceedings of the Language Resources and Evaluation Conference (LREC-2006)*.

USING TEMPLATES TO ACCOUNT FOR ENGLISH RESULTATIVES

Elizabeth Christie
Carleton University

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

This paper attempts to account for the English Resultative through the use of templates in Lexical Functional Grammar. This account will involve the use of LFG templates as an interface between the syntax and semantics and employ Glue Semantics as a logical means to combine the components. It will introduce two new templates for the data, and attempt to show how the notions introduced by Asudeh et al. (2008) can be extended to fit the resultative without having to introduce any new formal mechanisms.

1 Introduction

The English Resultative is a secondary predication relationship in which the main verbal predicate is affected by the existence of a resultant state. This resultant state describes the outcome of the action which was performed on the patient of the main predicate. There are five subtly different resultant structures, shown below in (1) through (5).

(1) **The transitive resultative**

He hammered the metal flat.

(2) **The transitive resultative with a non-subcategorized NP** [1]

Bryan drank Shannon under the table.

(3) **The fake reflexive**

John danced himself breathless.

(4) **The intransitive resultative with a non-subcategorized NP**

Kim ran the pavement thin.

(5) **The unaccusative resultative**

The river froze solid.

The base case of the resultative structure is usually the transitive resultative, as it is the simplest and most straight-forward case to begin from. In this case, the main verbal predicate is transitive, and takes its usual object. However, there is a second predicate which acts on the object, resulting in a change of its state. Thus,

[†]I am very thankful to Louisa Sadler, for pointing me in this direction in the first place, and Ash Asudeh for his patient comments on the many drafts of my poster and presentation.

[1]My thanks go to the editors for pointing out the similarity between the second and fourth types of the resultative, as the transitive verb in the example can also have an intransitive reading. At this point I do not have a solution for this issue, and will leave it to the reader to decide if they are two formations or only one, as the number of resultative formations does not affect my application of templates to the structure.

the meaning in (1), above, is not ‘a substance was hit’ but ‘a substance was hit *until it changed into another shape’*.

It is this change of state that semantically differentiates the resultative from another secondary predication relationship; the depictive (shown in (6)). In the depictive the object is not changed by the second predicate, but is described by it. In the example, the fish does not become raw during the serving action, but is raw when the action occurs. The two structures are also differentiated syntactically, as the depictive can be acceptably incorporated into a fronting structure (as in (7)), while the resultative cannot (as in (8)).

(6) John served the fish raw.

(7) Raw, John served the fish.

(8) ? Pink, Kim dyed her hair.

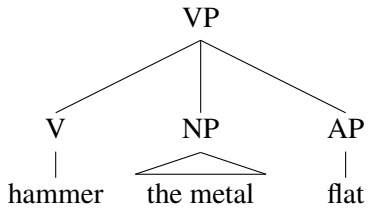
For the sake of expedience in this examination, I will focus only on the two most straightforward cases of the resultative, the transitive resultative and the unaccusative resultative. The difference between the two structures is that unlike the transitive resultative (described above), the unaccusative verb does not take an object, so the action of both the main and secondary predicates are carried out upon the subject of the sentence. Thus, in (5) above, the meaning is not ‘there was a freezing event involving a river’ but ‘during a freezing event, a river became solid’.

2 Previous Treatments

2.1 Carrier and Randall (1992) (Government and Binding)

Due to the requirements of the Government and Binding approach, the resultative must be licensed by the nodes in the tree structure, and thus Carrier and Randall propose a ternary branching tree structure (shown in (9)) in order to account for the resultative. According to Carrier and Randall (1992) the ternary branching tree describes the resultative structure best as it allows for both the s-selection and M-Command relationships which are required for grammaticality. These relationships, according to Haegeman (1994), allow for the correct assignment of semantic (theta) roles in the structure, thus allowing the structure to admit only those syntactic trees which would result in a grammatical sentence.

(9)



However, the account favoured by Carrier and Randall (1992) is not entirely intuitive. Government and Binding Theory requires that ”lexical information is syntactically represented” (Haegeman, 1994, pg. 55), and since there are two syntactic possibilities for the primary predicates used in the resultative (one regular use and one resultative use), there would have to be two lexical entries for resultative verbs, one for each use of the verbs. This duplication, however, misses the generalization that it is not the main predicate that changes in the resultative, but the state of the object of the main predicate.

2.2 Boas (2003) (Construction Grammar and Event Frame Semantics)

Boas asserts that all of the conditions for the formation of the resultative can be stated in the semantics of the lexical entry for the verb, within one of the multiple event-frames available, and that the syntactic rules need only store the form that the resultative takes, thus yielding two syntactic rules, shown in (10) and (11).

(10) [NP V NP XP] where X = A, N or P[2]

(11) [NP V XP] where X = A, N or P

In this distinctly semantic approach, Boas (2003) is able to account for the formation of the resultative through each individual lexical entry. Boas's C-Structure rules grossly overgenerate, but, according to his account, this overgeneration is deemed unacceptable by the individual semantics of the primary predicates themselves. Since each predicate contains a finite number of event frames which can be utilized in a given utterance, it is the semantics of the predicate which determines the grammaticality or acceptability of the sentence. Thus, in (12) the sentence is deemed unacceptable due to the fact that hammering a substance would not change its colour (it would change its shape). Similarly, in (13), the sentence is deemed unacceptable because of the fact that dying something does not change its shape (but would change its shape).

(12) *He hammered the metal pink.

(13) *She dyed the metal flat.

Although his account does manage to account for the resultative completely, it seems to lack an ability to generalize repetitive information, and relies heavily on semantics to disqualify unacceptable utterances. In the end, the account given by Boas (2003) can at times seem weighed down by the need to specify each utterance individually, and misses the observable fact that the resultative, although varied in its formation, can be described as a repetitive feature of the English language.

[2] Boas (2003) page 2, example (1.2)

2.3 Simpson (2006) (LFG)

In the account given by Simpson (2006), the description of the resultative is taken into the realms of LFG, and leaves behind the empirical requirements of the other formalisms. Simpson (2006) states that the resultative uses a regular set of restrictions, which can be logically stated within the lexicon. She posits the use of a powerful generative tool called a general lexical rule. A general lexical rule is a sort of interface between the syntax and the lexicon, as it is not stored in either place, but can be called for by a lexical entry before it is applied to a syntactic form.

Since there is a correlation between the base meaning of the verb and its resultative counterpart, namely that the object of the main verb will act as the subject of the XCOMP, Simpson (2006) posits the general lexical rule, shown in (14). This rule extends the semantic frame of the verb with an XCOMP and adds a control equation to the lexical definition which states that the subject of the XCOMP will be the object of the main verb.

(14) XCOMP Addition Rule (Simpson, 2006):

- Add a resultative attribute XCOMP
- Add the control equation XCOMP SUBJ = verb's OBJ

This lexical rule must then be specified for in the lexical entries of the resultative predicates. This allows for the fact that the resultative can only be predicated of some verbs (for example those which require a change of state) and cannot be predicated of many others (like many unergative verbs). In this way, the verb itself specifies the restrictions on the type of resultative predicate it can take, but the form of the resultative is specified outside of the lexicon in the XCOMP Addition Rule and the syntactic form of the utterance.

However, by requiring the rule to be specifically called for in each lexical entry that can potentially use the resultative, Simpson (2006) does not formally account for the generalization that the verb which calls the rule must permit a change of state. In her account there is an ability for the general lexical rule to potentially overgenerate, as it is a very powerful means of stating regularities, and must be tightly constrained. Finally, in her account, Simpson (2006) states that the general lexical rule must be applied in a sequential order for it to function correctly, as it must always be applied before any passivization. This seems slightly unintuitive in an LFG approach, as the acceptability of a sentence is usually evaluated in the interaction between the different structures and not in the sequential ordering of constraints.

3 Previous Use of Templates

Templates are a means of summarizing repetitive information, by building structures from smaller pieces of information rather than stating each of those pieces

each time they are used. For instance, by using the template in (15), below, we can summarize the fact that all transitive verbs have an f-structure which involves a subject and an object. This then allows for the lexical definition in (16) which takes the predicate KISS and inserts it into the transitive template to give us the more familiar form of (↑PRED) = 'KISS< (↑SUBJ)(↑OBJ) >'. Templates were introduced for use in computational settings so that programs were more straightforward to write and edit, and were first cited academically by Dalrymple et al. (2004).

(15) TRANSITIVE(FN) = (↑PRED) = 'FN< (↑SUBJ)(↑OBJ) >'

(16) kiss V $\lambda e.kiss(e) : (\uparrow_\sigma \text{REL})$
@TRANSITIVE(kiss)
$\left(\begin{array}{l} \lambda P \lambda x \lambda y \lambda e.P(e) \wedge agent(e) = x \wedge patient(e) = y: \\ (\uparrow_\sigma \text{REL}) \multimap (\uparrow\text{SUBJ})_\sigma \multimap (\uparrow\text{OBJ})_\sigma \multimap \uparrow_\sigma \end{array} \right)$

However, as part of their approach to the English-*way* construction and the Swedish Directed Motion Construction (DMC), Asudeh et al. (2008) use templates as an interface between the lexicon and the syntax, giving the templates a 'crucial theoretical role' (Asudeh et al., 2008). In their approach Asudeh et al. (2008) use a template hierarchy to allow for templates to either incorporate all parts of a previous template or select between two previous templates and incorporate the information it needs. For example, using a simple disjunction, a template for a predicate can specify information either from its base case (say a transitive interpretation) or its exceptional case (say a resultative interpretation). In this way, the resultative is licensed in the lexical entry, but that entry is able to supply only the predicate name, and the templates build the meaning and structure of the sentence.

By pairing template use with Glue Semantics, Asudeh et al. (2008) are able to create an interface between the syntactic structure, the lexicon, and the semantic structure. This is because each stage in the template absorption process, the template equation is paired with a Glue equation so that the semantics of the sentence is built up in the same pieces of information as the structure of the sentence. Thus each level adds something to the semantics of the sentence, and passes that on to the next template to be absorbed into the overall meaning.

4 Proposed Extension

4.1 Semantic Representations

Before beginning on the new application of templates, we must examine the semantic representations we wish to use for the resultative formation. In both (17) and (18) we can see that the semantic form reflects the fact that the object or patient of the main predicate (the event) is the subject or experiencer of the secondary predicate (the state).

(17) John hammered the metal flat.

$$\exists e \exists s.hammer(e) \ \wedge \ flat(s) \ \wedge \ result(e,s) \ \wedge \ agent(e) = John$$

$$\wedge \, patient(e) = metal \ \wedge \ experiencer(s) = metal \ \wedge \ cause(s) = John$$

(18) The river froze solid.

$$\exists e \exists s.freeze(e) \ \wedge \ result(e,s) \ \wedge \ solid(s)$$

$$\wedge \, patient(e) = river \ \wedge \ experiencer(s) = river$$

The main difference between the two semantic forms is that the transitive form needs an agent for the primary predicate and the unaccusative form does not. This reflects the usual uses of these verbs, as a transitive verb normally requires an agent and a patient, and an unaccusative verb normally only requires a patient.

4.2 Proposed Templates

In (19) and (20), below, we can see how the resultative templates could be built. In (19) we can see the generalized semantic equation with two predicate variables and two entity variables, thus leading to the use of both a subject and an object. The semantic equation is tied to a Glue equation which takes the meaning of the main predicate, combines it with the meaning of the secondary predicate and then adds in the meanings of the SUBJ and OBJ. It then takes in those meanings and combines them with the event's semantics and the state's semantics to yield the semantics of the overall equation, thus yielding the final semantics of the utterance.

(19) RESULT-T(FN) = @TRANSITIVE-X(FN)

$$\lambda P \lambda R \lambda x \lambda y.\exists e.\exists s.P(e) \wedge R(s) \wedge \ result(e,s)$$
$$\wedge \, agent(e) = x \ \wedge \ patient(e) = y$$
$$\wedge \, experiencer(s) = y \ \wedge \ cause(s) = x:$$
$$(\uparrow_\sigma \text{REL}) \multimap ((\uparrow \text{XCOMP})_\sigma \text{REL}) \multimap$$
$$(\uparrow \text{SUBJ})_\sigma \multimap (\uparrow \text{OBJ})_\sigma \multimap (\uparrow_\sigma \text{EVENT}) \multimap$$
$$((\uparrow \text{XCOMP})_\sigma \text{STATE}) \multimap \uparrow_\sigma$$

(20) RESULT-U(FN) = @INTRANSITIVE-X(FN)

$$\lambda P \lambda R \lambda x.\exists e.\exists s.P(e) \wedge R(s) \wedge \ result(e,s)$$
$$\wedge \, patient(e) = x \ \wedge \ experiencer(s) = x:$$
$$(\uparrow_\sigma \text{REL}) \multimap ((\uparrow \text{XCOMP})_\sigma \text{REL}) \multimap$$
$$(\uparrow \text{SUBJ})_\sigma \multimap (\uparrow_\sigma \text{EVENT}) \multimap$$
$$((\uparrow \text{XCOMP})_\sigma \text{STATE}) \multimap \uparrow_\sigma$$

Similarly, in (20) the generalized semantic equation contains two predicate variables, for the event and state. However, instead of two entity variables, there is only one, which reflects the lack of agentive role in unaccusative verbs, and

allows for the object role to be omitted. The associated Glue equation takes the meaning of the relation for the main predicate and adds to it the secondary relation before absorbing the information in the SUBJ. It then takes the meanings of the two predicates and yields the overall meaning of the sentence.

4.3 Proposed C-Structure Rules

With the semantics of the resultative coming from the template, and the restrictions on the resultative coming from the lexicon, we can add in resultative C-structure rules to allow for the formation of the resultative in both the C- and F-structures. This is done by adding in the rules in (21) and (22).

(21) V' → V NP $\{NP|AP|PP\}$
↑=↓ (↑OBJ) =↓ (↑XCOMP) =↓
(↓SUBJ) = (↑OBJ)
@RESULT-T((↑PRED FN))

(22) V' → V $\{NP|AP|PP\}$
↑=↓ (↑XCOMP) =↓
(↓SUBJ) = (↑SUBJ)
@RESULT-U((↑PRED FN))

Each of the rules accounts for a different kind of resultative. (21) accounts for the transitive resultative, as it contains an object position, and links that OBJ to the subject of the resultative XCOMP with the control equation (↓SUBJ) = (↑OBJ). It also calls for the resultative template using the '@' symbol to direct the grammar to the appropriate template, and takes the main predicate to be used in the template.

(22) creates the resultative in a similar manner, taking the main predicate to be used in the associated template. However, it differs in that it does not contain an object position and thus cannot use the same control equation. Instead, (22) links the SUBJ of the main predicate to the SUBJ of the XCOMP through the control equation (↓SUBJ) = (↑SUBJ).

4.4 Example lexical entries

In order for the template call to override the normal use of the resultative verbs, we must make that normal use something that can be optional, and rely upon the principles of completeness and coherence to allow or disallow utterances based on whether or not the verbal element contains a predicate. Since all sentences must have at least one verbal predicate in order to be complete, making the normal use of the verb optional does not overgenerate as any f-structure without a verbal predicate would not be deemed grammatical. Thus, in (23) and (24) the main predicates are optional and can be overridden by a template call.

(23) hammer V $\lambda e.hammer(e) : (\uparrow_\sigma \text{REL})$
@TRANSITIVE(hammer)
$\left(\begin{array}{l} \lambda P \lambda x \lambda y \lambda e.P(e) \wedge agent(e) = x \wedge patient(e) = y: \\ (\uparrow_\sigma \text{REL}) \multimap (\uparrow \text{SUBJ})_\sigma \multimap (\uparrow \text{OBJ})_\sigma \multimap \uparrow_\sigma \end{array} \right)$

(24) freeze V $\lambda e.freeze(e) : (\uparrow_\sigma \text{REL})$
@INTRANSITIVE(freeze)
$\left(\begin{array}{l} \lambda P \lambda x \lambda y \lambda e.P(e) \wedge agent(e) = x \wedge patient(e) = y: \\ (\uparrow_\sigma \text{REL}) \multimap (\uparrow \text{SUBJ})_\sigma \multimap (\uparrow \text{OBJ})_\sigma \multimap \uparrow_\sigma \end{array} \right)$

5 Conclusions and Future Work

In conclusion, the use of templates introduced by Asudeh et al. (2008) can be nicely extended for use with the English Resultative. It allows for a space between the syntactic structure and the semantic form which can be used to store information and build meanings outside of the normal working space. This allows for an interface between the syntax and semantics which acts as more than just a correlation and becomes an integral part of the theory. A resultative template for English would have to be broken down into sub-templates in order to account for the different forms the resultative can take, but it does allow for an elegant description of the phenomenon.

In future research we can begin to flesh out the formations needed for the other forms of the resultative and rigorously test the application of the templates. By fully studying this construction, we can begin to see more of the power that this use of templates can bring to our generative capabilities. We can also begin to look into resultatives cross-linguistically, to see if they behave similarly in other languages and if templates can be used in those occasions as well.

References

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 the LFG08 Conference*, pages 68–88, CSLI Publications.

Boas, Hans C. 2003. *A Constructional Approach to Resultatives*. CSLI Publications.

Carrier, Jill and Randall, Janet H. 1992. The Argument Structure and Syntactic Structure of Resultatives. *Linguistic Inquiry* 23(2), 173–234.

Dalrymple, Mary, Kaplan, Ronald M and Holloway King, Tracy. 2004. Linguistc Generalizations Over Descriptions. In *On-line Proceedings of the LFG2004 Conference*, pages 199–208, CSLI Publications.

Haegeman, Liliane. 1994. *Introduction to Government and Binding Theory Second Edition*. Blackwell Publishers Inc.

MacGregor, Elizabeth. 2009. *Using Templates to Account for English Resultatives*. Masters Thesis, University of Essex.

Simpson, Jane. 2006. Resultatives. In Miriam Butt and Tracy Holloway King (eds.), *Lexical Semantics in LFG*, pages 149–161, CSLI Publications, reprint of Papers in Lexical Functional Grammar (1983).

LESS-TRAVELLED PATHS FROM PRONOUN TO AGREEMENT: THE CASE OF THE URALIC OBJECTIVE CONJUGATIONS

Elizabeth Coppock and Stephen Wechsler
Lund University University of Texas at Austin

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

Building on Bresnan and Mchombo's (1987) theory that the transition from pronoun to agreement marker constitutes the loss of a PRED 'pro' specification on an affix, we explore the idea that the historical path from pronoun to agreement marker can involve the loss of person and number feature specifications as well. We apply this idea to object agreement in the Uralic languages, with particular attention to Ostyak and Hungarian, and propose that person and number specifications on object agreement affixes, historically derived from bound pronouns, were lost independently at different stages. We then consider the more general hypothesis that the special distribution of person agreement can be explained as a consequence of its historical origin in incorporated pronouns, with loss of the person feature as a complicating factor. Preliminary typological evidence supports this view over Baker's (2008) theory of person agreement.

1 Introduction

Agreement is sometimes in person and sometimes not. One factor is the lexical category of the target: while verbs tend to agree in person, adjectives tend not to. For example, in French, verbs agree in person with the subject but adjectives do not:

(1) a. Je suis belle.
I be.PRES.1SG beautiful.FEM.SG
'I [female] am beautiful.'

b. Tu es belle.
You be.PRES.2SG beautiful.FEM.SG
'You [female, singular] are beautiful.'

c. Elle est belle/*beau.
She be.PRES.3SG beautiful.FEM.SG/beautiful.MASC.SG
'She is beautiful.'

The finite verb registers the person and number values of the subject, while the adjective shows the subject's number and gender, but not its person. As revealed by broad typological studies, this pattern is common among the world's languages: verbs generally take prominence over other categories for the marking of person agreement (Stassen 1997; p. 38).

The distribution of person agreement can be traced to its historical origin: agreement inflections that encode person derive from incorporated personal pronouns, which therefore can include the pronominal features of person, number and gender (Lehmann 1988, Wechsler to appear). In the Index/Concord theory of agreement (Wechsler and Zlatić 2003), these pronoun-derived morphemes are

†We thank the audience at the LFG 2010 conference for helpful feedback, especially Farrell Ackerman, Miriam Butt, György Rákosi, Mary Dalrymple, Ida Toivonen and Ash Asudeh.

Index agreement inflections. In contrast, Concord agreement comprises number, gender and case, and is derived from other sources: for example, gender marking is thought to be derived from classifiers. Since verbal agreement tends to derive from pronoun incorporation, it tends to involve person; adjectival agreement derives from other sources and therefore does not.

Nonetheless, sometimes verbal agreement does not include person among its features. One example of this is object agreement in Northern Ostyak. In Northern Ostyak (a.k.a. 'Khanty'; Uralic), verbs agree with their subjects in person and number, but with their objects only in number, and not person. For example, a plural object triggers the same plural suffix *-l* on the verb regardless of whether the object is in third person (2) or first person (3) (Nikolaeva 1999):

(2) Ma täm kälang wel-sə-l-am
I this reindeer kill-PAST-PLOBJ-1SGSUBJ
'I killed these reindeer.'

(3) Xŭnśi näng mŭng-iluw xälśa want-lə-l-an?
when you we-ACC where see-PRES-PLOBJ-2SGSUBJ
'When did you see us where?'

Baker (to appear) takes aim at the distinction between Index and Concord, arguing that it cannot explain the situation in Northern Ostyak:

> Wechsler and Zlatić (2003) divide the phi-features into two groups, the Concord features and the Index features, and stipulate that adjectives access the first set and verbs the second set. This account has no obvious way of saying why verbs can access the person feature when agreeing with their subject, but not when agreeing with their theme object... [T]here is no reason to expect verbs in a language to show Index agreement with their subjects and Concord agreement with their objects.

Our answer to Baker's objection, which we put forth in this paper, is that the verb-object agreement in Northern Ostyak should not be analyzed as Concord agreement. Instead, both subjects and objects trigger Index agreement, and both subject and object verbal inflections descend from incorporated pronouns. But in the case of objects, only third person pronouns were incorporated, and the third person specifications on those affixes were subsequently lost, giving rise to the Northern Ostyak system. Different features were lost in different Uralic languages, leading to variation in the systems of object agreement in Eastern Ostyak, Samoyedic languages, and Hungarian.

Baker (2008, to appear) proposes a rather different explanation for the distribution of person agreement: he posits that the person feature is subject to a special universal structural condition. In Section 6 we broaden our scope to look at typological predictions of the two theories. We argue that the cross-linguistic facts support our view, namely that the distribution of person agreement can be explained as

Stage 1: Pronoun	**Stage 2: Loss of reference**	**Stage 3: Loss of person**
(↓PRED) = 'pro'	~~(↓PRED) = 'pro'~~	~~(↓PRED) = 'pro'~~
(↓INDEX PERS) = p	(↓INDEX PERS) = p	~~(↓INDEX PERS) = p~~
(↓INDEX NUM) = n	(↓INDEX NUM) = n	(↓INDEX NUM) = n
(↓INDEX GEND) = g	(↓INDEX GEND) = g	(↓INDEX GEND) = g

Table 1: Stages of lexical specification in the transition from pronoun to agreement

a consequence of its historical origin in incorporated pronouns: pronoun-derived agreement includes person, except when the person feature is lost over time.

2 Feature loss

Feature loss can be modeled in LFG as loss of f-description equations, as Bresnan and Mchombo (1987) demonstrate (see also Butt 2007). They analyze the object markers in Chicheŵa as morphologically bound pronouns, having a lexical entry as in the first column of Table 1. The PRED 'pro' specification of such a morphologically bound pronoun may then become optional; this is how Bresnan and Mchombo analyze Chicheŵa's subject markers. If the PRED 'pro' is eliminated entirely, as in the second column of Table 1, the result is a grammatical agreement marker.

In this paper, we explore the idea that this process may target other features as well, such as person or number features. For example, the person feature could be eliminated, as in the third column of Table 1. Toivonen (2001) uses this idea to explain dialectal variation in Finnish possessive suffixes, showing that both number and person feature specifications can be lost. The suffix *-nsa*, which is third person in Standard Finnish, is unrestricted with respect to person in one dialect of Finnish. The suffix *-ni,* which is restricted to first person singular in the standard dialect, is unrestricted with respect to person *and* number in some South-West dialects. In other South-West dialects, *-ni* has retained its first person specification, while losing its number specification. In the same dialects, the second person singular suffix *-si* has lost its number specification, retaining its person specification. These facts show that feature loss can target either person or number specifications, and that the loss of one feature may take place independently of the loss of another. We will propose a similar account of variation in object agreement across the Uralic languages.

3 Object agreement in Uralic

Among the Uralic languages (comprising the Finno-Ugric and Samoyedic families), object agreement is found in Samoyedic (Nenets, Enets, Nganasan, Selkup), Mordvinian, and Ugric (including Hungarian, and the Ob-Ugric languages Ostyak

	SUBJECTIVE	SG.OBJ	DU.OBJ	PL.OBJ
1SG	tulêm	tulim	tulêylam	tullam
2SG	tuwên	tulin	tulêylan	tullantul
3SG	tuwêl	tultê	tulyêl	tulêl(lê)
1DU	tulmên	tulmên	tulêylamên	tullamên
2DU	tuletên	tulêtên	tulêylin	tullin
3DU	tulyên	tulêten	tulêylin	tullin
1PL	tuloy	tuluy	tulêyloy	tulloy
2PL	tultêy	tulêtên	tulêylin	tullin
3PL	tuwêlt	tulil	tulêylal	tullal

Table 2: Northern Ostyak conjugations for *tu-, tuy* 'bring' (Honti 1984; 107, taken from Kortvély 2005)

and Vogul). These languages vary in the feature specifications of their object agreement morphology. In Hungarian (É. Kiss 1987) and Samoyedic (Hajdú 1968), there are two conjugations, subjective and objective, and the objective conjugation is used when the object is definite and third person. Ob-Ugric languages have a subjective conjugation and several objective conjugation paradigms, one for each of the possible number values (singular, dual, or plural) of the object (Honti 1984, Kálman 1965). In Northern Ostyak, the use of one of these objective conjugations is further conditioned by whether the object is third person (Gulya 1966). In Mordvinian, the verb genuinely agrees in both person *and* number with the object (Keresztes 1989, see also Béjar 2008). We will address Northern and Eastern Ostyak, Hungarian, and Samoyedic in detail here. For reasons that we do not have space to go into in this paper, we believe that the Mordvinian system arose independently, and we will leave it out of the discussion for now. We begin with Northern Ostyak.

3.1 Northern Ostyak

3.1.1 Object agreement in number but not person

As shown above in (2) and (3), verbs agree in number but not person with the object in Northern Ostyak: a third person plural object and a first person plural object trigger the same plural marking on the verb. The full paradigm is given in Table 2, which shows that there is a subjective conjugation along with three objective conjugations, one for each number value (singular, dual, and plural).

As Nikolaeva (1999) documents, use of an objective conjugation requires that the object be topical. Focussed objects, as in (4), appear with the subjective conjugation:

(4) tam kalang we:l-əs / *we:l-s-əlli
this reindeer kill-PAST.3SGSUBJ kill-PAST-SGOBJ.3SGSUBJ
'It was this reindeer that he killed.'

Given that the objects that trigger the objective conjugation are topical, one might suspect that they are really floating topics, anaphorically linked to a bound pronominals realized by the objective conjugation affixes. Nikolaeva (1999) shows that this is not the case, however; the objective conjugation suffixes have, in effect, lost their PRED 'pro' specification. Her evidence is as follows: (i) external NPs receive case, (ii) when the external NP is a pronoun, it does not have a contrastive discourse function, (iii) the external NP may not appear in the periphery of the clause, and (iv) the verb does not specify all of the features of the pronoun, only number.

In summary, the basic facts about Northern Ostyak object agreement we propose to account for are: that it is sensitive to number, without being sensitive to person; that objects that trigger the objective conjugation must be topical; and that the objective conjugation suffixes are not pronominal.

3.1.2 Analysis

We propose to analyze Northern Ostyak object agreement as Index agreement without person; the object agreement affixes in Northern Ostyak derive from incorporated third person pronouns whose person specifications were subsequently lost.

Our proposed development begins with an incorporated pronoun stage. We suggest that there were three third person pronouns that were incorporated, one for each number value (singular, plural, and dual). This is schematized in (5). We crucially assume that only third person pronouns were incorporated in this ancestor of Northern Ostyak. Evidence for that assumption is that in its neighbor, Eastern Ostyak, the objective conjugation is restricted to third person objects (see §3.2 below).

(5) **Stage 1: Incorporated pronouns (only in third person)**

V_{aff} $(\uparrow \text{OBJ}) = \downarrow$
$(\downarrow \text{PRED}) =$ 'pro'
$(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$
$(\downarrow \text{INDEX PERS}) = 3$
$(\downarrow \text{INDEX NUM}) = n$ where $n \in \{\text{SG}, \text{DU}, \text{PL}\}$

We assume that the bound pronominal also comes with a topicality condition, encoded by the third equation. Indeed, it has often been observed that it is highly topical pronouns that can be reduced to bound forms. The notation for the topicality condition follows Dalrymple and Nikolaeva (to appear); it specifies that the value of the DF ('discourse function') attribute in the σ (semantic) projection of the node corresponding to the affix is TOPIC. This will cause its semantics to appear as the value of the TOPIC attribute at the level of information structure, by a principle relating semantic structure to information structure.

Then, we propose, the PRED 'pro' specification was lost, leaving the topicality condition in place.

(6) **Stage 2: Loss of PRED 'pro', retention of topicality condition**

V_{aff} $(\uparrow \text{OBJ}) = \downarrow$
~~$(\downarrow \text{PRED}) = \text{'pro'}$~~
$(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$
$(\downarrow \text{INDEX PERS}) = 3$
$(\downarrow \text{INDEX NUM}) = n$ where $n \in \{\text{SG}, \text{DU}, \text{PL}\}$

This explains the modern-day topicality condition on the use of the objective conjugation, and the fact that the objective conjugation suffixes are not pronominal.

Next, the person specification was lost.[1]

(7) **Stage 3: Loss of person specification:**

V_{aff} $(\uparrow \text{OBJ}) = \downarrow$
~~$(\downarrow \text{PRED}) = \text{'pro'}$~~
$(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$
~~$(\downarrow \text{INDEX PERS}) = 3$~~
$(\downarrow \text{INDEX NUM}) = n$ where $n \in \{\text{SG}, \text{DU}, \text{PL}\}$

At this stage, the objective conjugation suffixes start being used with first and second person objects as well.

To summarize, we propose the following events in the development of Northern Ostyak: (i) incorporation of third person pronouns; (ii) loss of PRED 'pro', retention of topicality; (iii) loss of the person specification. This accounts for the main facts about object agreement in Northern Ostyak.

3.2 Eastern Ostyak

As noted above, one reason to believe that a person specification was present at an earlier stage of Northern Ostyak is that it is reportedly still present in its neighbor, Eastern Ostyak. Gulya's (1966) grammar of Eastern Ostyak says (p. 115), "the definite [i.e. objective] conjugation ... expresses not only a definite object *of the third person*, but its number as well" (emphasis added). This implies that the objective conjugation appears with third person objects (as in the Northern Ostyak example (2)), but not first or second person objects, unlike Northern Ostyak (recall example (3)).

We hypothesize that Eastern Ostyak instantiates Stage 2 (given in (6)), and that Stage 1, where third person object pronouns were incorporated into the verb, is a common ancestor of Northern and Eastern Ostyak. When these incorporated pronouns lost their PRED 'pro' specification, they retained their third person and number specifications. The result was Eastern Ostyak: object agreement in number, and only in third person.

[1] Assuming that there are distinct affixes for each number value, this means that person was lost independently in each affix. Support for this idea could come from the existence of dialects where the person specification is lost in some but not all numbers. However, we know of no such evidence.

	SUBJECTIVE	OBJECTIVE
1SG	-ok/-ek/-ök	-om/-em/-öm
2SG	-(a)sz/-(e)sz or -ol/-el/-öl	-od/-ed/-öd
3SG	∅	-ja/-i
1PL	-unk/-ünk	-juk/-jük
2PL	-(o)tok/-(e)tek/-(ö)tök	-játok/-itek
3PL	-(a)nak/-(e)nek	-ják/-ik

Table 3: Verb conjugations in Hungarian

3.3 Hungarian

We posited above that the Northern Ostyak system arose from the loss of the person feature specification from those (formerly) third person pronouns, with the retention of number. Now we suggest that the Hungarian system is also derived from an Eastern Ostyak type system through feature loss, like Northern Ostyak. However, in Hungarian, the *person* specification was retained, and the *number* specification was lost.

Modern Hungarian has two subject-verb agreement paradigms, subjective and objective, shown in Table 3. The choice of conjugation is conditioned by the definiteness of the object. If the object is definite, as in (8), the objective conjugation is used; if the object is indefinite, as in (9), then the subjective conjugation is used.

(8) Lát-om a madar-at
see-1.SG.DEF the bird-ACC
'I see the bird'

(9) Lát-ok egy madar-at
see-1.SG.INDEF a bird-ACC
'I see a bird'

Somewhat mysteriously, the use of the objective conjugation is also sensitive to the person of the object. While a third person pronoun triggers the objective conjugation, as in (10), first and second person pronouns do not, as shown in (11).

(10) Lát-ják őt/őket
see-3.PL.DEF it/them
'They see it/them'

(11) Lát-nak engem/téged/minket
see-3PL.INDEF me/you/us
'They see me/you/us'

We will refer to this as the *third person restriction* in Hungarian. Note that the Hungarian objective conjugation is not sensitive to the number value of the object.

We suggest that the third person restriction can be viewed as the vestige of an earlier stage at which third person object pronouns were incorporated. To be precise, we suggest the following development. At Stage 1, third person pronouns are incorporated, as suggested above for Northern and Eastern Ostyak (repeated from (5)).

(12) **Stage 1: Pronoun incorporation (only in third person)**

V_{aff} $(\uparrow \text{OBJ}) = \downarrow$
$(\downarrow \text{PRED}) = \text{'pro'}$
$(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$
$(\downarrow \text{INDEX PERS}) = 3$
$(\downarrow \text{INDEX NUM}) = n$ where $n \in \{\text{SG}, \text{DU}, \text{PL}\}$

Then the PRED 'pro' specification was lost, leaving a topicality condition in its place.

(13) **Stage 2: Loss of PRED 'pro', retention of topicality condition**

V_{aff} $(\uparrow \text{OBJ}) = \downarrow$
~~$(\downarrow \text{PRED}) = \text{'pro'}$~~
$(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$
$(\downarrow \text{INDEX PERS}) = 3$
$(\downarrow \text{INDEX NUM}) = n$ where $n \in \{\text{SG}, \text{DU}, \text{PL}\}$

The idea that PRED 'pro' was lost in Hungarian is not entirely uncontroversial; den Dikken (2006) argues for an analysis of the objective conjugation as an incorporated clitic pronoun. Coppock and Wechsler (under revision) argue against this view, using evidence from binding, islands, and the fact that while clitic doubling is sensitive to semantic properties like definiteness, specificity and topichood, Hungarian object agreement is predictable solely based on the form of the object.[2] Coppock and Wechsler conclude that the objective conjugation does not represent an incorporated pronoun; rather that there is agreement (in definiteness) between the verb and the object (or that the choice of verb conjugation is *conditioned* by definiteness, as Corbett (2006) suggests).

What we have in Stage 2 is a system like Eastern Ostyak's. One piece of evidence in favor of the idea that Hungarian has such a stage in its history is that the Samoyedic languages have object agreement of the Eastern Ostyak type as well (agreement in number, restricted to third person), and the object agreement systems of Samoyedic are thought to be related to Hungarian's. Although the Samoyedic languages are genetically quite distant from Hungarian, Helimski (1982) reconstructs the objective conjugation in Hungarian to an areal feature shared by the

[2]This is not stictly speaking true, because there are cases where the same object can take either the objective or the subjective conjugation, with a concomitant change in meaning, e.g. *Ismerem/ismerek néhány titkodat* 'I know-DEF/-INDEF some secret-yours-ACC'; 'I know some of your secrets' (cf. Szabolcsi 1994, Bartos 2001). But the fact remains that it is not a semantic property that determines whether or not the objective conjugation may be used.

Samoyedic languages, carefully showing that the common features of their objective conjugations cannot have been present in proto-Uralic.

After PRED 'pro' was lost and replaced by a topicality condition, we propose that the number restriction was eliminated, but the person restriction was retained:

(14) **Stage 3: Loss of number**

V_{aff} $(\uparrow \text{OBJ}) = \downarrow$
~~$(\downarrow \text{PRED}) = \text{'pro'}$~~
$(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$
$(\downarrow \text{INDEX PERS}) = 3$
~~$(\downarrow \text{INDEX NUM}) = n$ where $n \in \{\text{SG, DU, PL}\}$~~

This represents a system quite similar to modern Hungarian, where the objective conjugation is restricted to third person objects. However, one difference is that the use of object agreement in Hungarian is conditioned by definiteness rather than topicality. We suggest that the topicality condition was reanalyzed as a definiteness condition, due to the large overlap in distribution between these two features (Givón 1976):

(15) **Stage 4: Reanalysis of topicality as definiteness**

V_{aff} $(\uparrow \text{OBJ}) = \downarrow$
~~$(\downarrow \text{PRED}) = \text{'pro'}$~~
~~$(\downarrow_{\sigma} \text{DF}) = \text{TOPIC}$~~ $(\uparrow \text{OBJ DEF}) =_c +$
$(\downarrow \text{INDEX PERS}) = 3$
~~$(\downarrow \text{INDEX NUM}) = n \in \{\text{SG, DU, PL}\}$~~

The [DEF] feature is a formal grammatical feature and not a semantic feature. As noted above, Hungarian object agreement is predictable based on the form of the object, but not its meaning: designated forms, such as certain determiners and possessives, are marked as [DEF]. (See Coppock and Wechsler (under revision) for detailed discussion.)

The Stage 4 scenario, depicted in (15), captures the facts of modern Hungarian very well except for one wrinkle: First and second person *reflexive* pronouns trigger the objective conjugation.

(16) (Én) szeret-em magam-at
I love-3SG.DEF myself-ACC
'I love myself'

(17) (Te) szeret-ed magad-at
You love-2SG.DEF yourself-ACC
'You love yourself'

Since they are not third person, it will not do to say that the objective conjugation requires a third person object. We propose that the third person restriction was reanalyzed such that reflexive pronouns of all person values count as [DEF]. Meanwhile, for non-reflexive pronouns, third person pronouns count as [DEF] and first and second person pronouns do not.

(18) *ő* $(\uparrow \text{PRED}) = \text{‘pro’}$
$(\uparrow \text{PERS}) = 3$
$(\uparrow \text{NUM}) = \text{SG}$
$(\uparrow \text{DEF}) = +$

(19) *engem* $(\uparrow \text{PRED}) = \text{‘pro’}$
$(\uparrow \text{PERS}) = 1$
$(\uparrow \text{NUM}) = \text{SG}$

Although this complicates the distribution of [DEF], it simplifies the grammar in another respect, by making the verb form sensitive to only one factor (formal definiteness) rather than two (definiteness and person). Hence the final stage has no person restriction equation at all:

(20) **Stage 5: Elimination of person restriction**
V_{aff} $(\uparrow \text{OBJ}) = \downarrow$
$\cancel{(\downarrow \text{PRED}) = \text{‘pro’}}$
$\cancel{(\uparrow \text{TOPIC}) = \downarrow}$ $(\uparrow \text{OBJ DEF}) =_c +$
$\cancel{(\downarrow \text{INDEX PERS}) = 3}$
$\cancel{(\downarrow \text{INDEX NUM}) = n \in \{\text{SG, DU, PL}\}}$

This is our proposed representation for modern Hungarian.

Before moving on, we must mention an alternative hypothesis regarding the development of the Hungarian objective conjugation, put forth by Dalrymple and Nikolaeva (to appear). They suggest that there is an earlier stage at which the objective conjugation is used for objects of all persons, as in Northern Ostyak, and then its usage is narrowed to third person objects. Assuming that the objective conjugation suffixes are derived from third person pronouns, this narrowing process involves loss of the pronoun-derived person specification, followed by a later step at which it is added back in. Our derivation has fewer steps, and we prefer it on these grounds.

To summarize, we posit that third person object pronouns were incorporated into verbs, and that feature loss is a mechanism of historical change that is active in the history of Uralic objective conjugations. These assumptions explain the lack of object person agreement in Northern Ostyak, the person restriction in Eastern Ostyak and Samoyedic languages, and the person restriction and lack of object number agreement in Hungarian.

4 Why were only third person pronouns incorporated?

A key assumption underlying our analysis of Uralic is that only third person pronouns were incorporated. This has allowed us to relate the third person restriction in Hungarian to the third person restriction in Eastern Ostyak and Samoyedic languages. Simply put, the objective conjugations are limited to the third person because it is only third person pronouns that were ever incorporated.

This raises the question of why incorporation was originally limited to third person pronouns. One possible explanation builds on the assumption that agreement derives from pronoun doubling constructions, like (21) (the *NP-detachment hypothesis*).

(21) I like him, that guy.

If the doubled NP were first or second person, it would be a pronoun, because first and second person analytic forms are always pronouns. Pronoun doubling of another pronoun may be possible in certain contexts, but it performs a more limited range of functions, as evidenced by the oddness of the following examples:

(22) #I like him, him.

(23) #I like you, you.

As Ariel (1999; 249, fn. 8) points out, "there is hardly any use for first/second persons in NP detachment constructions (cf. **I, I will do it*)."

As a general theory of how pronouns become agreement markers, the NP-detachment view is a bit too simple, because there are quite a few languages in which there is incorporation of first and second person pronouns, to the exclusion of third person pronouns. Incorporation of first and second person pronouns can also be shown to precede incorporation of third person pronouns in some cases (see Ariel 1999 and references cited therein). This motivates Ariel (1999) to posit an alternative hypothesis regarding the process by which pronouns become agreement markers, based on *Accessibility Theory:* high accessiblity leads to reduction of a pronoun, and first and second person pronouns are more highly accessible than third person pronouns. It may be, however, that the NP-detachment hypothesis is correct for some cases; indeed, Fuss (2005) argues that the set of processes that can lead from pronoun to agreement marker is heterogeneous. Cases such as ours, in which only third person pronouns incorporate (and the external NP is moreover topical), conform precisely to the predictions of the NP-detachment hypothesis.

To summarize, our explanation for the person restriction in Hungarian relies on the assumption that only third person pronouns were incorporated at an earlier stage of the language. We further speculate that incorporated pronouns may have been restricted to third person because those pronouns participated in NP-detachment constructions. In such constructions, the doubled NP is usually lexical, hence third person, so only third person pronouns incorporated.

5 Other accounts of the third person restriction

The third person restriction on the objective conjugation in Hungarian has intrigued many scholars of Hungarian, and several other explanations have been proposed.

Bartos (2001) suggests an account in terms of ergativity, noting that it is "reminiscent of an ergative-type split ... first and second person pronouns follow a

nominative-accusative pattern, as opposed to third person ones engaging in an ergative-absolutive pattern" (p. 322). There is, however, no evidence of ergativity in Hungarian; pronouns of all persons can be marked with the accusative marker *-t*, and the subject appears in nominative (unmarked) case regardless of the person of the object.

É. Kiss (2005) proposes that the person restriction can be understood using the notion of an inverse system. She posits the following animacy hierarchy for Hungarian:

$$1\text{SG} > 1\text{PL}/2 > 3$$

along with the following constraint:

(24) *Inverse agreement constraint* (É. Kiss 2005)
An object agreeing with a verb must be lower in the animacy hierarchy than the subject agreeing with the same verb, unless the subject represents the lowest level of the animacy hierarchy.

Kiss's proposal ties together two puzzling facts of Hungarian: the person restriction, and the fact that a special suffix is used with first person singular subjects and second person objects (*-lak/-lek*):

(25) Szeret-lek
love-1SG.OBJ:2
'I love you'

When the subject is first person singular and the object is second person, the subject outranks the object on the person hierarchy, so object agreement is allowed, according to Kiss's theory; this accounts for the existence of *-lak/-lek*. On our view, the existence of *-lak/-lek* is independent from the fact that first and second person objects do not count as definite; these two facts must be stipulated individually.

However, we are skeptical of Kiss's proposal for several reasons. First, in true inverse systems, the use of inverse morphology is determined by the relative rank of subject and object on an animacy hierarchy, but in Hungarian, the presence of object agreement is for the most part conditioned *only* by the features of the object. The rank of the subject does not affect the possibility of object agreement whenever the the object is third person, that is, whenever the objective conjugation is used. (This is captured by Kiss's exception clause "unless the subject represents the lowest level of the animacy hierarchy", along with the additional implicit condition that in that case, the object may be no higher than the subject.) The use of the objective conjugation is not conditioned by the subject's person value or indeed any other properties of the subject; only *-lak/-lek* is sensitive to both subject and object.

Furthermore, morphological marking (by an inverse marker) normally appears when the alignment between the animacy hierarchy and the grammatical relation

hierarchy is *non*-harmonic, that is, when there is a *violation* of the alignment constraint. In the case of Hungarian, the morphologically marked case occurs when the constraint is *satisfied*. This makes the putative inverse system in Hungarian a rather atypical exemplar of its category.

Moreover, we know of no other inverse agreement systems among the Uralic languages; Kiss's argument for its plausibility comes from the existence of inverse agreement systems in Chukchee, Koryak and Kamchadal (Comrie 1980), which are not genetically related to Hungarian, although "it does not seem implausible to hypothesize that Proto-Uralic, a distant ancestor of Hungarian, and Proto-Chukotko-Kamchatkan could belong to the same Sprachbund" (É. Kiss 2005; p. 115). However, if geographic influence on Proto-Uralic were responsible for the modern-day properties of Hungarian then we would expect at least some other Uralic languages to show traces of it.

A third explanation for the person restriction in Hungarian was given by Comrie (1977; 10), who claims that because first and second person pronouns are inherently definite, there is no need to mark them explicitly. Dalrymple and Nikolaeva (to appear) criticize this theory as follows:[3]

> This explanation is based on the premise that the primary function of object agreement in Eastern Uralic is the marking of definiteness. However, in all Eastern Uralic languages except modern Hungarian, information structure and not definiteness plays the primary role in patterns of object agreement. The Hungarian situation thus is likely to be secondary, as is also confirmed by the Old Hungarian data.

Dalrymple and Nikolaeva propose instead that it is their typical information structure status, rather than their inherent definiteness, which sets first and second person pronouns apart here. They claim that "the Samoyedic languages (Nenets, Selkup and Nganasan) and Old Hungarian have grammaticalised the tendency for first and second person pronouns to be likely primary topics and unlikely secondary topics." The *primary topic* is the topic that is most topical; a *secondary topic* is a topic that is less topical than the primary topic. Dalrymple and Nikolaeva assume that the subject position tends to align with the primary topic and that the object position tends to align with the secondary topic, following Nikolaeva (2001). Since first and second person pronouns tend to be primary rather than secondary topics, they will tend not to align with the object position. There is a problem with Dalrymple and Nikolaeva's explanation as well, however: the same reasoning could be used to explain the opposite distribution. It is generally the rare case that is more marked, rather than the other way around (Dixon 1994; 85ff).

From the above discussion, it should be clear that the person restriction is a phenomenon that has been tricky to account for. The feature loss view provides a

[3]The Old Hungarian data to which they refer consists of examples pointed out by Marcantonio (1985) in which the objective conjugation is used with an object that is not marked for definiteness, suggesting that topicality, rather than definiteness, was the factor governing the use of the objective conjugation at that stage.

simple and well-motivated explanation for this notorious puzzle. Another advantage of the feature loss view is that it simultaneously accounts for the lack of object person agreement in Northern Ostyak, to which we will now return.

6 Person agreement: structural versus historical accounts

Recall Baker's claim above that the facts of Northern Ostyak are difficult to account for under the Index/Concord (*dualist*) view of agreement, rightly pointing out that the endings in question should not be analyzed as Concord targets. What we have shown is that these facts are not problematic for the dualist view if Index agreement is characterized historically, rather than as a feature bundle that necessarily includes person. On the view being advanced here, the morphological endings in Northern Ostyak are not Concord targets, but rather Index targets whose person specification has been lost. This implies that sensitivity to the person feature is not a definitional feature of Index agreement; rather, Index agreement should be defined as agreement that historically derives from incorporation of personal pronouns.

An alternative hypothesis comes from Baker (2008). Baker has a *monist* view of agreement, according to which verbal and adjectival agreement are of the same ilk. To account for the special properties of the person feature, Baker proposes a universal structural condition on person agreement:

(26) **Structural Condition on Person Agreement (SCOPA)**
A functional category F can bear the features +1 or +2 if and only if a projection of F merges with an NP that has that feature, and F is taken as the label for the resulting phrase.

This amounts to saying that a target head can agree in person with a first or second person controller only if the controller is the complement or the specifier of the target head. It explains why person agreement is absent from predicative adjectives that lack a structural subject position, as in the French example (1), where a copula is needed to project the subject position.

It is important for Baker's monist approach that this condition apply not only to adjectives but also to verbs. In support of his theory, Baker cites languages where verb agreement normally includes the person feature, but that feature is lacking in special constructions in which the controller's structural relation to the target does not comply with the SCOPA. For example, agreement between a tensed verb and a nominative object in Icelandic is ruled out by the SCOPA, assuming that nominative objects are not in the specifier of the projection corresponding to the verb, as Baker does. And as the SCOPA predicts, Icelandic verbs do not agree in first or second person with nominative objects. While nominative objects can be third person, triggering verb agreement in number ((27)), nominative objects cannot be first or second person ((28) and (29)):

(27) a. Henni leiddust Þeir
her.DAT was.bored.by-3PL they.NOM
'She was bored with them'

b. Henni leið-ist bókin sín
her.DAT was.bored.by-3SG book self's
'She finds her own book boring'

(28) *Henni leidd-umst við
her.DAT was.bored.by-1SG we.NOM
'She was bored with us.'

(29) *Ég veit að honum lík-ið Þið
I know that him.DAT like-2PL you.NOM.PL
'I know that he likes you all.'

The case of Icelandic is slightly different from the case of predicate adjectives, however. With predicate adjectives, the target shows the same default person value across all controllers, while agreeing in number (this is *partial agreement*). In the case of Icelandic, the SCOPA-violating configuration with first and second person controllers is *ineffable*. Baker's (2008) theory does not specify which of these two outcomes will occur when the SCOPA is violated, although Baker (to appear) proposes that the variation is related to case assignment.

The dualist hypothesis, in contrast, is capable of predicting when partial agreement will occur. The dualist hypothesis does not allow for the type of partial agreement where the agreement features encoded by a target form can be selectively ignored depending on the syntactic structure. On the dualist view, ϕ-features are bundled on Index and Concord target morphemes, and agree in an 'all-or-none' fashion (c.f. Chomsky 2001, Béjar 2008). It allows for person insensitivity only when person distinctions are entirely absent from the target morpheme's paradigm. Two kinds of target morpheme can lack person: (i) Concord morphemes (for example, those on French adjectives); (ii) Index morphemes from which person distinctions have been historically lost (as in Northern Ostyak object agreement). Hence the dualist hypothesis predicts:

(30) If a given target morpheme reflects the person feature of the controller in any controller position, then that target morpheme will reflect the controller's person in every controller position.

The ineffability cases do not violate the prediction in (30), and are thus consistent with the dualist hypothesis, because the third person endings in (27) can be seen as displaying full number and person agreement with their third person controllers. Partial agreement, where person agreement morphology is available but the target morpheme agrees only in number, violates (30), because this constitutes a dissociation of the features that are hypothesized to be bundled together. According to (30),

Language	Target	Trigger	Pers. Morph.?	Result
Swahili, etc.	A	subjects	no	partial
English 'dialect'	V	Spec, CP	yes	both
Passamaquoddy	V	Spec, CP	yes	ineffable
Icelandic	V	nom. obj.	yes	ineffable
Gujarati	Aux	acc. obj.	yes	none
Chicasaw	V	nom. obj.	yes	ineffable
Nahuatl, etc.	V	2nd obj.	yes	ineffable
Lokaa	V	obj. of gerund	yes	ineffable
Basque	V	controlled subj.	yes	ineffable
Passamaquoddy	V	embedded subj.	yes	ineffable
Northern Ostyak	V	acc. obj.	no	partial
Sakha	V	Spec, CP	yes	partial

Table 4: SCOPA-violating configurations

partial agreement should be possible only when person agreement morphology is absent from the relevant paradigm.

Table 4 lists the SCOPA-violating agreement configurations that Baker (2008, to appear) discusses, and classifies each according to the category of the agreeing head ("Target"), the location of the controller ("Trigger"), whether the agreeing head can inflect for person ("Pers(on) Morph(ology)?"), and whether the SCOPA violation results in ineffability or partial agreement ("Result"). (In the case of the English 'dialect', there is partial agreement for some person/number values of the controller and ineffability for other ones, which is why it is assigned the value 'both'; in Gujarati, the auxiliary verb shows neither number nor person agreement with the target, so it is assigned 'none'.) The crucial columns are the last two: Baker's theory allows for languages in which person morphology is available but there is partial agreement, but the dualist theory predicts that this should not happen.

In all but one case, the prediction in (30) is straightforwardly met. For every line with a "yes" in the "Pers(on) Morph(ology)?" column, "partial" is *not* the value of the "Result" column— except for Sakha (Turkic). The case of Sakha requires closer inspection.

In Sakha ECM constructions, the predicate can lack the usual person/number inflection, registering plural number with the *-LAr* suffix, regardless of person (Baker to appear). Person agreement in these constructions is optional; example (31) shows the ECM construction with full person/number agreement; (32) shows the crucial variant with reduced agreement.

(31) Min ehigi/ehigi-ni bügün kyaj-yax-xyt dien erem-mit-im.
I you/you-ACC today win-fut-2PL COMP hope-PAST-1SG
'I hoped that you would win today.'

(32) Min ehigi-ni/*ehigi bügün kyaj-yax-tar-a dien erem-mit-im.
I you-ACC/you today win-past-PL-3 COMP hope-PAST-1SG
'I hoped that you would win today.'

It is not clear whether this is a real counterexample to the prediction in (30). The first question is whether the plural suffix *-Lar* (of which *-tar* in (32) is an allomorph) is in the same person/number paradigm as the first and second person morphemes. The suffix *-Lar* has a number of properties distinguishing it from the other agreement affixes. It marks plural on nouns, suggesting it is a Concord target. It is optional, in the sense that an unmarked element has general number, i.e. either singular or plural (Vinokurova 2005; p. 144); on the basis of such optionality, among other reasons, Hahm (2010) argues that the cognate morpheme in Turkish is not agreement at all, but rather an indicator of semantic plurality. Also, *-Lar* has a historical source distinct from that of person/number inflection (Hahm 2010, citing Adamović 1985; p. 27 and Good and Yu 2005). We tentatively suggest the plural suffix is either not agreement at all, or else a Concord target morpheme that does not derive from pronoun incorporation. The suffix *-a*, glossed as third person, may be an Index target, perhaps a default form arising in the absence of person agreement. But this would not be partial Index agreement, but rather the absence of any Index agreement at all— in which case *ehigi-ni* 'you-ACC' is not really a controller. Because number and person are encoded in separate morphemes, they may in principle have different controllers, or one may have a controller while the other lacks one. A clear counterexample to the dualist hypothesis would involve a single morpheme that simultaneously encodes number and person. A more detailed study of Sakha and its ECM construction is needed before we know definitively whether it falsifies the predictions of the dualist hypothesis.

Although Sakha requires further investigation, all of the cases listed in Table 4 corroborate the prediction that partial agreement is possible only when the target affix fails to encode person morphologically. Thus, the dualist view is capable of predicting when partial agreement, as opposed to ineffability, will occur. In this respect, our dualist theory is stronger than Baker's (2008) theory.[4]

7 Conclusion

The special distribution of person agreement can be largely explained as a consequence of its historical origin in incorporated pronouns. Synchronically, this is reflected in the distinction between Index targets and Concord targets; only the former are historically derived from pronouns, so only the former signal the person of their controller. Exceptions to that generalization arise when the person feature is lost over time, as we suggested for the historical development of object agreement in the Uralic languages. Synchronically, this means that Index targets do not

[4]Baker (to appear) does offer a way of explaining this variation based on case assignment, and further research is needed to distinguish between these two views.

always constrain the person feature of their controller. This view accounts for the main features of the Uralic objective conjugations, including the mysterious third person restriction in Hungarian.

We cannot offer any explanation for why person was lost and number was not in Northern Ostyak, but feature loss does not seem to be a process restricted to person; Toivonen's study of Finnish possessive affixes showed that feature loss can target both person and number. A typological investigation of feature loss and its constraints is currently being undertaken by the first author (Coppock 2010).

The dualist view of agreement, according to which Index and Concord are distinct feature bundles, leads to the prediction that person agreement can be dissociated from number agreement only when the feature specification has been lost from the target affix. In contrast to Baker's (2008, to appear) theory, it predicts that the presence of person agreement should not depend on the structural relationship between the controller and the target. We have shown preliminary typological evidence that this prediction is borne out.

References

Adamović, M. (1985). *Konjugationsgeschichte der türkischen Sprache*. E. J. Brill, Leiden.

Ariel, M. (1999). The development of person agreement markers: From pronouns to higher accessibility markers. In Barlow, M. and Kemmer, S., editors, *Usage-based Models of Language*, pages 197–260. CSLI Publications, Stanford.

Baker, M. C. (2008). *The Syntax of Agreement and Concord*. Cambridge University Press, Cambridge.

Baker, M. C. (to appear). When agreement is for number and gender but not person. *Natural Language and Linguistic Theory*.

Bartos, H. (2001). Object agreement in Hungarian: A case for minimalism. In Alexandrova, G. M. and Arnaudova, O., editors, *The Minimalist Parameter: Selected Papers from the Open Linguistics Forum, Ottowa, 21-23 March 1997*, volume 192 of *Current Issues in Linguistic Theory*, pages 311–324. John Benjamins, Amsterdam.

Béjar, S. (2008). Conditions on phi-agree. In Harbour, D., Adger, D., and Béjar, S., editors, *Phi Theory: Phi-Features across Modules and Interfaces*. Oxford University Press, Oxford.

Bresnan, J. and Mchombo, S. (1987). Topic, pronoun, and agreement in Chicheŵa. *Language*, 63:741–782.

Butt, M. (2007). The role of pronominal suffixes in Punjabi. In Zaenen, A., Simpson, J., and King, T. H., editors, *Architectures, Rules and Preferences*, pages 341–368. CSLI Publications, Stanford.

Chomsky, N. (2001). Derivation by phase. In Kenstowicz, M., editor, *Ken Hale: A Life in Language*, pages 1–52. MIT Press, Cambridge, MA.

Comrie, B. (1977). Subjects and direct objects in Uralic languages: A functional explanation of case-marking systems. *Études Finno-Ourgriennes*, 12:5–17.

Comrie, B. (1980). Inverse verb forms in Siberia: Evidence from Chukchee, Koryak and Kamchadal. *Folia Linguistica Hungarica*, 1:61–74.

Coppock, E. (2010). Pathways from pronoun to agreement and their destinations. Post-doctoral research project funded by the Swedish Research Council.

Coppock, E. and Wechsler, S. (under revision). The objective conjugation in Hungarian: Agreement without phi-features. University of Texas at Austin.

Corbett, G. (2006). *Agreement*. Cambridge University Press, Cambridge.

Dalrymple, M. and Nikolaeva, I. (to appear). Objects and information structure. Unpublished book manuscript.

den Dikken, M. (2006). When Hungarians agree (to disagree): The fine art of 'phi' and 'art'. CUNY Graduate Center.

Dixon, R. M. W. (1994). *Ergativity*. Number 69 in Cambridge Studies in Linguistics. Cambridge University Press.

É. Kiss, K. (1987). *Configurationality in Hungarian*. Akadémiai Kiado.

É. Kiss, K. (2005). The inverse agreement constraint in Hungarian: A relic of a Uralic-Siberian Sprachbund? In *A Festschrift for Henk van Riemsdijk*. John Benjamins, Amsterdam.

Fuss, E. (2005). *The Rise of Agreement: A Formal Approach to the Syntax and Grammaticalization of Verbal Inflection*. John Benjamins, Amsterdam.

Givón, T. (1976). Topic, pronoun and grammatical agreement. In Li, C. N., editor, *Subject and Topic*, pages 149–188. Academic Press, New York.

Good, J. and Yu, A. C. L. (2005). Morphosyntax of two Turkish subject pronominal paradigms. In Heggie, L. and Ordóñez, F., editors, *Clitic and Affix Combinations: Theoretical Perspectives*, pages 315–342. John Benjamins, Amsterdam.

Gulya, J. (1966). *Eastern Ostyak Chrestomathy*. University of Indiana Press, Bloomington.

Hahm, H.-J. (2010). *A Cross-linguistic Study of Syntactic and Semantic Agreement: Polite Plural Pronouns and Other Issues*. PhD thesis, University of Texas at Austin.

Hajdú, P. (1968). *Chrestomathia Samojedica*. Tankönyvkiadó, Budapest.

Helimski, E. A. (1982). *Vengersko-samodijskije lexičeskije paralleli*. Nauka, Moscow.

Honti, L. (1984). *Chrestomathia Ostiacica*. Tankönyvkiadó, Budapest.

Kálman, B. (1965). *Vogul Chresomathy*, volume 46 of *Indiana University Uralic and Altaic Series*. Indiana University Press, Bloomington.

Keresztes, L. (1989). *Chrestomathia Mordvinica*. Tankönyvkiadó, Budapest.

Kortvély, E. (2005). Verb conjugation in Tundra Nenets. *Studia Uralo-Altaica*, 46.

Lehmann, C. (1988). On the function of agreement. In Barlow, M. and Ferguson, C. A., editors, *Agreement in Natural Language*. CSLI Publications, Stanford.

Marcantonio, A. (1985). On the definite vs. indefinite conjugation in Hungarian: A typological and diachronic analysis. *Acta Linguistica Scientiarum Hungaricae*, 35:267–298.

Nikolaeva, I. (1999). Object agreement, grammatical relations, and information structure. *Studies in Language*, 23:331–376.

Nikolaeva, I. (2001). Secondary topic as a relation in information structure. *Linguistics*, 39:1–41.

Stassen, L. (1997). *Intransitive Predication*. Oxford Studies in Typology and Linguistic Theory. Clarendon Press, Oxford.

Szabolcsi, A. (1994). The noun phrase. In Kiefer, F. and É. Kiss, K., editors, *The Syntactic Structure of Hungarian*, volume 27, pages 179–274. Academic Press, New York.

Toivonen, I. (2001). Language change, lexical features, and Finnish possessors. In Butt, M. and King, T. H., editors, *Time Over Matter*, pages 209–225. CSLI Publications, Stanford.

Vinokurova, N. (2005). *Lexical categories and argument structure: a study with reference to Sakha*. PhD thesis, University of Utrecht.

Wechsler, S. (to appear). Mixed agreement, the person feature, and the Index/Concord distinction. *Natural Language and Linguistic Theory*.

Wechsler, S. and Zlatić, L. (2003). *The Many Faces of Agreement: Morphology, Syntax, Semantics, and Discourse Factors in Serbo-Croatian Agreement*. CSLI Publications, Stanford.

AGREEMENT PATTERNS AND COORDINATION IN LEXICAL FUNCTIONAL GRAMMAR

Mary Dalrymple and Bozhil Hristov
University of Oxford

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

Coordination and its interactions with agreement have been a focus of research in Lexical Functional Grammar over the past decade, though an account that captures the full range of agreement patterns in an elegant manner has proved elusive. Many previous proposals account for patterns of feature resolution but do not extend to single-conjunct agreement. Other proposals address single-conjunct agreement, but provide an account of standard resolution patterns that is less than satisfying. We provide a means of stating a typology of agreement patterns that handles resolution and single-conjunct agreement, as well as agreement requirements that apply in an across-the-board fashion to all of the conjuncts of a coordinate phrase.

1 Agreement in LFG

We begin with basic agreement patterns in Serbian/Croatian/Bosnian (SCB) and their treatment in LFG. In example (1), the determiner *moja* 'my' and the modifier *stara* 'old' show feminine singular agreement with the noun they modify, *knjiga* 'book', which is listed in the lexicon as a feminine singular noun. The predicative past participle *pala* 'fallen' also shows feminine singular agreement.[1]

(1) Modifier and verb agreement (Wechsler and Zlatić, 2003, 18):

Moja	stara	knjiga	je	pala.
my.F.Nom.Sg	old.F.Nom.Sg	book.F.Nom.Sg	Aux.3Sg	fall.PPart.F.Sg

'My old book fell.' (SCB)

As in most constraint-based linguistic theories, agreement in LFG is treated as multiple specification of feature values by a controller and target (Kaplan and Bresnan 1982; Bresnan 2001, chap. 8; Dalrymple 2001, chap. 5). In (1), the modifiers, noun, and predicate cospecify the number and gender features of the subject, as shown in (2), where the labels f and s are used as labels for the parts of the functional structure, and s=(f SUBJ):

[†]For helpful comments, we thank Anna Maykova, the audiences at CoBaLiSE 2009 (University of Essex) and LFG10, especially Louisa Sadler and Doug Arnold, and our informants Sandra Sćepanović and Anna Vlaisavljević.

[1]Abbreviations: M: masculine; F: feminine; N: neuter; Sg: singular; Pl: plural; PPart: past participle. Examples from other sources use the same transcription conventions as in the source.

(2) *moja* ‘my’: (s GEND) = F
(s NUM) = Sg

stara ‘old’: (s GEND) = F
(s NUM) = Sg

knjiga ‘book’: (s GEND) = F
(s NUM) = Sg

pala ‘fall’ (f SUBJ GEND) = F
(f SUBJ NUM) = Sg

$$f: \begin{bmatrix} \text{PRED} & \text{‘fall’} \\ \text{SUBJ} & s: \begin{bmatrix} \text{PRED} & \text{‘book’} \\ \text{SPEC} & \text{‘my’} \\ \text{ADJ} & \{[\text{PRED} \quad \text{‘old’}]\} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix}$$

2 INDEX vs. CONCORD agreement

In example (1), a single set of agreement features is proposed for agreement with determiners, prenominal modifiers, and predicates. However, the picture is in fact more complicated.

Traditional grammatical approaches have long recognised two types of agreement features, which Wechsler and Zlatić (2003) call INDEX and CONCORD features (see also Kathol 1999, Corbett 2001). Both are syntactic features, but as Wechsler and Zlatić show, INDEX features are more closely related to semantic features, while CONCORD features are more closely related to morphological declension. Wechsler and Zlatić explore agreement patterns with SCB nouns which have different INDEX and CONCORD features, with CONCORD features often relevant for noun phrase internal agreement, and INDEX features often relevant outside the noun phrase. Syncretism in the nominal paradigms means that determination of which features are involved in agreement is not completely straightforward.

The noun *deca* ‘children’ has mismatched CONCORD and INDEX features: it has feminine singular CONCORD but neuter plural INDEX. An example of *deca* with attributive and predicative agreement is given in (3):

(3) Ta dobr-a deca su došl-a
that.F.Sg good-F.Sg children Aux.3Pl come.PPart-N.Pl
‘Those good children came.’ (SCB; Wechsler and Zlatić, 2003, 51)

The *-a* ending on the determiner, the adjective, and the verb is ambiguous between F.Sg and N.Pl. However, it is easy to show that in general, agreement with determiners and attributive adjectives depends on the CONCORD feature, because all non-nominative attributive modifiers with *deca* are unambiguously F.Sg (Corbett, 1983, 81). Example (4) shows that the unambiguously feminine singular accusative adjective *dobru* ‘good’ is used to modify accusative ‘children’:

(4) ... svoje sestre, koja inače ima dobr-**u** dec-**u**.
own.Gen sister.Gen, who otherwise has good-**F.Sg.Acc** children-**Acc**
'... of my own sister, who, otherwise, has good children.' (SCB)
http://charolija.wordpress.com/2009/01/07/praviti-se-normalan-2/
(accessed on 08.12.09)

Example (5) is similar, with an unambiguously feminine singular accusative possessive determiner:

(5) Kad sam video njegov-**u** dec-**u**, znao sam ...
When Aux.1.Sg seen his-**F.Sg.Acc** children-**Acc**, realised Aux.1.Sg
'When I saw his children, I realised...' (SCB)
http://www.mojheroj.com/heroj.aspx?h=139 (accessed on 08.12.09)

Based on these patterns, we will analyse attributive agreement as involving the CONCORD feature, since — despite nominative attributive elements being ambiguous between F.Sg and N.Pl — non-nominative elements with *deca* 'children' and *unučad* 'grandchildren' are unambiguously F.Sg, and must be analysed as instances of CONCORD and not INDEX agreement.

The predicate in example (3) is also ambiguous between F.Sg and N.Pl. However, the auxiliary is unambiguously plural, and we take this as an indication that the main verbal predicate is plural (and therefore should be analysed as N.Pl rather than F.Sg), following Corbett's Agreement and Predicate Hierarchies (Corbett, 2006). We return to this issue in Section 6 below.

Having established a distinction between INDEX and CONCORD features, we enrich our representation of agreement features to take this distinction into account:

(6)

ta 'that':	(s CONCORD GEND) = F
	(s CONCORD NUM) = Sg
dobra 'good':	(s CONCORD GEND) = F
	(s CONCORD NUM) = Sg
deca 'children':	(s CONCORD GEND) = F
	(s CONCORD NUM) = Sg
	(s INDEX GEND) = N
	(s INDEX NUM) = Pl
došla 'came':	(f SUBJ INDEX GEND) = N
	(f SUBJ INDEX NUM) = Pl

$$f: \begin{bmatrix} \text{PRED} & \text{'come'} \\ \text{SUBJ} & s: \begin{bmatrix} \text{PRED} & \text{'children'} \\ \text{SPEC} & \text{'my'} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'good'} \end{bmatrix} \right\} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{N} \\ \text{NUM} & \text{Pl} \end{bmatrix} \\ \text{CONCORD} & \begin{bmatrix} \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

3 INDEX vs. CONCORD in coordination

King and Dalrymple (2004) provide a theory of agreement with coordinated nouns which relies on the distinction between **distributive** and **nondistributive** features

in coordination. Dalrymple and Kaplan (2000) present the following definition of function application for coordinate structures, taking into account the distributive/nondistributive distinction:

(7) For any *distributive* property P and set s, $P(s)$ iff $\forall f \in s.P(f)$.
For any *nondistributive* property P and set s, $P(s)$ iff P holds of s itself.

Features are individually classified as either nondistributive (properties of the coordinate structures as a whole) or distributive (properties of the conjuncts, or members of the coordinate set). King and Dalrymple (2004) provide evidence that INDEX is a nondistributive feature: the coordinate phrase has its own INDEX features, which may be different from the INDEX features of the conjuncts, as will be demonstrated in the next section. In contrast, the CONCORD feature is distributive: the CONCORD feature of a coordinate structure depends on the CONCORD value of each conjunct.

This provides a neat explanation for the otherwise puzzling English pattern shown in (8):

(8) This$_{\text{Sg}}$ boy and girl are$_{\text{Pl}}$ classmates.

English determiners show CONCORD agreement, in line with Wechsler and Zlatić's proposal for SCB. Since CONCORD is a distributive feature, the singular determiner *this* can combine with a coordinate phrase in which each conjunct has singular CONCORD, but is incompatible with plural conjuncts:

(9) *this boys and girl / *this boy and girls

Boy and girl meets the agreement requirements imposed by *this*, since each conjunct is singular, and the noun phrase *this boy and girl* is correctly predicted to be well-formed.

English verbs show INDEX agreement.[2] The coordinate phrase *this boy and girl* is semantically plural, and has plural INDEX. Thus, the phrase *this boy and girl* behaves like any other plural noun phrase, and requires plural verb agreement: the sentence *this boy and girl are classmates* is well-formed for the same reason that the sentence *the boys are classmates* is well-formed, namely that a phrase like *this boy and girl* or *the boys* has plural INDEX.

King and Dalrymple's classification of the INDEX and CONCORD feature allows, then, for the following patterns:

- INDEX agreement with the resolved INDEX features of the coordinate phrase as a whole

[2]This is a well-established generalisation, though research is needed on agreement patterns with certain determiners: both singular and plural agreement are attested with determiners such as *each* in examples like *Each boy and girl is/are required to....*

- CONCORD agreement with the CONCORD features of each conjunct

There are, however, other patterns to take into account, as we will see in the following.

4 Feature resolution in coordination

Resolution rules determine the features of a coordinate structure from the features of the individual conjuncts (Corbett, 1983, 2006; Dalrymple and Kaplan, 2000; Dalrymple et al., 2006). Wechsler and Zlatić (2003) provide the generalisation in (10) for gender resolution in SCB (see also Corbett 1991, chap. 9):

(10) Gender and number resolution in SCB (Wechsler and Zlatić, 2003, 174):

F + F ⇒ F.Pl
elsewhere ⇒ M.Pl

According to this rule, a coordinate phrase with two feminine conjuncts is feminine plural, while any other combination of conjuncts produces a masculine plural coordinate phrase. Wechsler and Zlatić (2003) illustrate this rule with the example in (11), where the predicate shows masculine plural agreement with the resolved features of the coordinate subject:

(11) [[Ogledalo] i [četka za kosu]] su bil-**i** na stolu
mirror.N.Sg and brush.F.Sg for hair Aux.Pl were-**M.Pl** on table
'The mirror and the hairbrush were on the table.' (SCB; Wechsler and Zlatić, 2003, 173)

King and Dalrymple (2004) show that resolution rules apply to the INDEX feature: the INDEX feature of a coordinate structure is computed by applying the resolution rule to the INDEX features of the conjuncts. Under these assumptions, a simplified functional structure for the subject in example (11) is shown in (12):

(12)
$$\begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{M} \\ \text{NUM} & \text{Pl} \end{bmatrix} \\ \left\{ \begin{bmatrix} \text{PRED} & \text{'mirror'} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{N} \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \text{PRED} & \text{'brush'} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix}$$

LFG models coordinate structures as sets: the set containing the f-structures for the nouns *ogledalo* 'mirror' and *četka* 'brush' represents the coordinate structure *ogledalo i četka* 'mirror and brush'. The f-structure in (12) shows that while the INDEX features of the individual conjuncts are F.Sg and N.Sg, respectively, the resolved INDEX features of the whole NP are M.Pl, as specified by the resolution

rule in (10). The participle *bili* ‘were’ shows INDEX agreement with the resolved masculine plural INDEX features of the coordinate phrase *ogledalo i četka* ‘mirror and brush’.

We will not address the question of precisely how the gender and number features of the coordinate structure are computed. Rather, we will simply assume that resolution rules of the general form discussed by Dalrymple and Kaplan (2000), Dalrymple et al. (2006) and Sadler (2006) ensure that the coordinate structure acquires the features specified by the gender resolution rule in (10).

Strikingly, the pattern in (10) is found even when both conjuncts are neuter:

(13) [[Ogledalo] i [nalivpero]] su bil-**i**/*bil-**a** na stolu
mirror.N.Sg and fountain.pen.N.Sg Aux.Pl were.**M.Pl**/*__N.Pl__ on table
‘The mirror and the fountain pen were on the table.’ (SCB; Wechsler and Zlatić, 2003, 174)

$$\begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{M} \\ \text{NUM} & \text{Pl} \end{bmatrix} \\ \left\{ \begin{bmatrix} \text{PRED} & \text{‘mirror’} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{N} \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \text{PRED} & \text{‘pen’} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{N} \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix}$$

5 Single-conjunct agreement

Wechsler and Zlatić (2003) discuss patterns of **single-conjunct agreement** in SCB, in which agreement is with one of the conjuncts of the coordinate phrase, not with the resolved features:

(14) [[Ova varošica] i [sva sela]] su poplavljen-**a** / ?poplavljen-**i**
this.F.Sg town.F.Sg and all.N.Pl village.N.Pl Aux.Pl flooded-**N.Pl** / ?**M.Pl**
‘This town and all the villages were flooded.’(SCB; Wechsler and Zlatić, 2003, 54)

Instead of the expected resolution pattern, according to which masculine plural agreement on the predicate would be expected, example (14) exhibits neuter plural predicate agreement: the predicate agrees with the closest conjunct, *sva sela* ‘all the villages’, not with the resolved features of the coordinate phrase. Such examples are common; example (15) is from our own fieldwork:

(15) [Nova [[kuča] i [kola]]] su koštal-**a** puno.
new.F.Sg house.F.Sg and car.N.Pl Aux.Pl cost-**N.Pl** a.lot
‘The new house and car cost a lot.’ (SCB; Ana Vlaisavljević, p.c.)

Again, the predicate agrees with the second conjunct, not with the resolved features of the subject noun phrase (we discuss agreement with the adjective *nova* ‘new’ in Section 7).

Sadler (1999) was among the first to discuss the treatment of single-conjunct agreement in an LFG setting (see also Sadler, 2003; Arnold et al., 2007), and the first to provide examples to show that single-conjunct agreement can coexist with resolved agreement patterns referencing the same coordinate structure. Kuhn and Sadler (2007) provide the following Welsh example to illustrate this point:

(16)
Gwelaist [[ti] a’th [frawd]] eich hunain.
saw.**2Sg** **2Sg** and-2Sg brother **2Pl** self
‘You and your brother saw yourselves.’ (Welsh; Kuhn and Sadler, 2007)

The subject of this sentence is *ti a’th frawd* ‘you and your brother’. The verb *gwelaist* ‘saw’ shows second person singular single-conjunct agreement with the first conjunct *ti* ‘you’, while the reflexive *eich hunain* ‘yourselves’ shows second person plural resolved agreement with the coordinate structure as a whole. This example shows that it is not possible to treat single-conjunct agreement as a kind of resolution; the two types of agreement must be distinguished, and a coordinate phrase must be able to show both kinds of agreement at the same time.

An exploration of the INDEX/CONCORD distinction in single-conjunct agreement has not so far been undertaken. SCB provides an ideal laboratory for such study, because of its rich patterns of agreement and the existence of nouns with mismatched INDEX and CONCORD features.

6 Single-conjunct INDEX agreement

Under our assumption that predicate agreement depends on the INDEX feature, the examples in (14) and (15) illustrate single-conjunct INDEX agreement. We can ensure that the relevant feature is INDEX and not CONCORD by examining coordinate structures with noun phrases having mismatched CONCORD and INDEX features. In this section, we examine coordinated mismatched nouns modified by a relative clause, and we show that the relative pronoun shows INDEX agreement with the closest conjunct.

To establish patterns of agreement with relative pronouns, we first examine constructions with noncoordinated nouns. The nominative relative pronoun in (17) can be analysed either as showing F.Sg agreement with the CONCORD features of *deca* ‘children’, or alternatively as showing N.Pl INDEX agreement. Non-nominative relative pronouns do not display this syncretism: the accusative relative pronoun in (18) is unambiguously F.Sg, and in the genitive (19) a choice exists between a relative pronoun which is unambiguously F.Sg and one which is unambiguously plural.

(17) deca koj-**a** su tada bil-a
children who-**F.Sg/N.Pl** Aux.**Pl** there were-F.Sg/N.Pl
'children who were there' (SCB Corbett, 1983, 78)

(18) deca koj-**u** vidite
children who-**F.Sg.Acc** you.see
'children who you see' (SCB; Corbett, 1983, 79)

(19) deca koje/kojih se svi boje
children who.**F.Sg.Gen**/who.**Pl.Gen** Refl all fear
'children whom all fear' (SCB; Corbett, 1983, 79)

We believe that the correct analysis for the *-a* ending on the nominative relative pronoun in (17) is N.Pl (INDEX agreement with the head noun) rather than F.Sg (CONCORD agreement). Our primary evidence for this is that the relative pronoun triggers plural agreement on the relative clause verb, and so must itself be plural. This analysis is supported by the possibility of an unambiguously plural genitive relative pronoun in (19), which must be analysed as INDEX agreement with the head noun. Note, however, that this pattern is not found with the accusative relative pronoun in (18), which unambiguously shows F.Sg CONCORD agreement. We assume that nominative relative pronouns agree in INDEX, while accusative relative pronouns agree in CONCORD, and genitive allows either type of agreement.[3]

The nouns in the coordinate phrases in examples (20) and (21) have feminine singular CONCORD and neuter plural INDEX. We examine patterns of agreement with nominative relative pronouns agreeing in INDEX with the nouns they modify. Recall that according to the resolution rules for SCB, we would expect the coordinate NP to have masculine plural resolved features. However, the relative pronoun instead shows neuter plural agreement, matching the INDEX features of the second

[3]This entails a difference in our treatment of attributive elements and relative pronouns in cases where it is difficult to tell whether CONCORD or INDEX agreement is involved. In constructions with the nouns *deca* 'children' and *unučad* 'grandchildren', nominative attributive modifiers and relative pronouns ending in *-a* can be analysed as either F.Sg or N.Pl, while non-nominative attributive modifiers and relative pronouns are unambiguously feminine singular, except for the genitive relative pronoun, which is either unambiguously F.Sg or unambiguously plural. Accusative relative pronouns are treated as F.Sg (CONCORD agreement) in accordance with their unambiguous morphological marking. By contrast, we treat nominative relative pronouns as N.Pl (INDEX agreement) because they trigger plural agreement on the relative clause verb. For attributive elements, there is no other agreement with them (comparable to the plural verb agreement within the relative clause which is triggered by nominative relative pronouns) to justify a split analysis whereby nominative attributive elements with *deca* 'children' and *unučcad* 'grandchildren' are treated as N.Pl and non-nominative as F.Sg. We therefore treat all attributive elements (including ambiguous nominative ones) as F.Sg. Thus, agreement with attributive elements involves CONCORD features, whereas relative pronouns can pick either the CONCORD (acc or gen) or INDEX (nom or gen) features of the head nominal. This is consistent with their dual role as NP-internal elements as well as pronouns at the same time (as noted by Wechsler and Zlatić 2003). For more discussion of this issue, see Wechsler and Zlatić (2003, 56) and Corbett (1983, 78).

conjunct.[4]

(20) njihov-a [[deca] i [unučad]] koj-**a** su rodjen-a u drugoj
their-F.Sg children and grandchildren who-**N.Pl** Aux.**Pl** born-**N.Pl** in another
državi
country
'their children and grandchildren who were born in another country' (SCB)
www.durmitorcg.com/forum/viewtopic.php?f=9&t=6064&start=0&st=0&sk=t&sd=a
(accessed on 20.03.2009)

(21) [[Tinejdžeri] i [deca]] koj-a preglasno i prečesto slušaju
Teenagers.M.Pl and children who-**N.Pl** too.loudly and too.often listen.to.3Pl
muziku ... riskiraju da ogluva 30 godina ranije od...
music ... risk to go.deaf 30 years earlier than
'Teenagers and children who listen to music too loudly and too often... risk going deaf 30 years earlier than...' (SCB)
www.zdraviilepi.com/Magazin-deca!P7.aspx (accessed on 17.11.2009)

Since *deca* 'children' and *unučad* 'grandchildren' in (20) have F.Sg CONCORD and N.Pl INDEX, it can be argued that the relative pronoun agrees with both of them. Example (21) shows that this is not the case; here, the first conjunct has M.Pl CONCORD and INDEX, whereas the second has F.Sg CONCORD and N.Pl INDEX. The relative pronoun therefore agrees only with the INDEX of the second conjunct.

7 Single-conjunct CONCORD agreement

We have so far encountered the following agreement patterns:

[4]Given the existence of mismatched nouns, we might claim that, in addition to the two homophonous relative pronouns *koja*, there is a third relative pronoun *koja* which has F.Sg CONCORD and N.Pl INDEX. This mismatched relative pronoun could then be claimed to agree in CONCORD with the F.Sg CONCORD of the antecedent ('grandchildren') and in INDEX with the plural auxiliary verb inside the relative clause. This would mean that (20) is actually not an instance of closest-conjunct agreement involving the INDEX feature. However, positing a third *koja* relative pronoun with mismatched features is not entirely empirically justified and violates Occam's Razor, which states that entities should not be multiplied beyond necessity.

We could also argue that SCB has only one relative pronoun with the form *koja* and it has F.Sg CONCORD and N.Pl INDEX. Under this analysis, the pronoun will always agree in CONCORD with feminine antecedents and in INDEX with neuter antecedents. What is more, the relative clause verb will have to sometimes pick the INDEX and sometimes the CONCORD of the relative pronoun. All of these are inelegant and theoretically undesirable consequences. Furthermore, and most importantly, we have examples of closest-conjunct INDEX agreement which is unmediated by relative pronouns (cf. the predicate 'be hungry' in (37), which agrees directly with the antecedent). Since the finite copula verb is plural, this resolves the ambiguity and the rest of the predicate is unambiguously N.Pl. Otherwise, we would have more semantically justified agreement on the verb than on the predicative adjective, which would be a violation of Corbett's otherwise robust Predicate Hierarchy (cf. Corbett (1983, 87)).

- Agreement with resolved INDEX features: examples (8), (11), (13)
- Agreement with CONCORD features of all conjuncts: example (8) (cf. example 9)
- Agreement with INDEX features of a single conjunct: examples (14), (15), (16), (20), (21)

We can also have single-conjunct agreement involving the CONCORD feature. Corbett (1991) provides example (22), in which the predeterminer *sve* 'all' and the possessive determiner *njegove* 'his' show agreement with the first conjunct of the coordinate phrase. Although the noun *molbe* 'prayers' does not have mismatched CONCORD and INDEX feature, we have established on the basis of agreement patterns with mismatched nouns that agreement with attributive elements involves CONCORD and not INDEX features (see Section 2 above and Corbett 1983, 81). Hence, this must be closest-conjunct CONCORD agreement.

(22) sve njegove [molbe i uveravanja] ni-su pomagali ništa
all.**F.Pl** his.**F.Pl** prayers.**F.Pl** and assurances.**N.Pl** Neg-Pl helped.M.Pl nothing
'All his prayers and assurances did not help at all.' (SCB; Corbett, 1991, 283)

Example (22) illustrates the following pattern:

- Agreement with CONCORD features of a single conjunct: example (22)

Kuhn and Sadler (2007) show that it is possible to have single-conjunct agreement with both the first and the last conjunct in the same coordinate structure – what they call **double-edged agreement**. The following example is from Brazilian Portuguese:

(23) os [mitos e lendas] brasileiras
the.M.Pl [myth.M.Pl and legend.F.Pl] Brazilian.F.Pl
'the Brazilian myths and legends' (Brazilian Portuguese; Kuhn and Sadler, 2007, (16b))

This example exhibits two instances of single-conjunct agreement. The first involves the determiner and the first conjunct, and the second involves the second conjunct and the adjective. Examples such as these are also likely to involve CONCORD agreement, though without the possibility of examining agreement patterns with mismatched nouns in Brazilian Portuguese, it is difficult to know for sure. Such examples also show the necessity of being able to refer to the features of both the initial and the final conjunct independently.

In sum, the patterns that our analysis must handle are: agreement with resolved INDEX features; agreement with distributive CONCORD features; agreement with the INDEX features of just one conjunct; and agreement with the CONCORD features of just one conjunct.

8 Our proposal: A typology of agreement patterns

Following Kuhn and Sadler (2007), we propose to handle these agreement patterns by defining functional metavariables to allow reference to peripheral conjuncts in a coordinate phrase. We adopt Kuhn and Sadler's notation f_L and f_R, where f can be replaced by any expression that refers to an f-structure, though we provide different definitions of these expressions: in particular, these expressions may refer to the Leftmost and Rightmost conjuncts in a coordinate structure, but may also refer to a noncoordinated phrase. As discussed below, we also do not adopt Kuhn and Sadler's proposed classification of features, since it does not allow a treatment of the full range of patterns discussed in the previous section; for example, it does not allow for both resolved and single-conjunct agreement with the INDEX feature.

Retaining the assumption that INDEX is a nondistributive feature and CONCORD is a distributive feature, our analysis allows us to state agreement requirements in terms of the following expressions, with f representing an arbitrary f-structure. For the INDEX feature:

- (f INDEX): the INDEX features of a noncoordinate phrase, or the resolved INDEX features of a coordinate phrase (the standard interpretation of this expression)
- ($f_{(L)}$ INDEX): the INDEX features of a noncoordinate phrase, the resolved INDEX features of a coordinate phrase, or the leftmost conjunct of a coordinate phrase (optional single-conjunct agreement with the leftmost conjunct)
- (f_L INDEX): the INDEX features of a noncoordinate phrase or the leftmost conjunct of a coordinate phrase (obligatory single-conjunct agreement with the leftmost conjunct)
- ($f_{(R)}$ INDEX): the INDEX features of a noncoordinate phrase, the resolved INDEX features of a coordinate phrase, or the rightmost conjunct of a coordinate phrase (optional single-conjunct agreement with the rightmost conjunct)
- (f_R INDEX): the INDEX features of a noncoordinate phrase or the rightmost conjunct of a coordinate phrase (obligatory single-conjunct agreement with the rightmost conjunct)

INDEX is a resolving feature, so in a coordinate structure, the INDEX value of the coordinate structure as a whole might be different from the INDEX of the individual conjuncts. We allow for agreement with the resolved INDEX features of a coordinate phrase as well as for agreement (either optionally or obligatorily) with a single conjunct.

For the CONCORD feature, there are fewer options:

- (f CONCORD): the CONCORD features of a noncoordinate phrase or each conjunct of a coordinate phrase (the standard interpretation of this expression)

- (f_L CONCORD): the CONCORD features of a noncoordinate phrase or the leftmost conjunct of a coordinate phrase

- (f_R CONCORD): the CONCORD features of a noncoordinate phrase or the rightmost conjunct of a coordinate phrase

When all conjuncts have the same CONCORD value, it is not possible to distinguish between CONCORD agreement with the closest conjunct and CONCORD agreement with all conjuncts: 'optional' closest-conjunct agreement is not distinguishable from exclusively closest-conjunct agreement. Hence, we do not express agreement constraints in terms of optional closest-conjunct CONCORD agreement: we distinguish only between the stronger requirement of distributive CONCORD agreement with all conjuncts ((f CONCORD)) and the weaker requirement of agreement only with the closest conjunct.

The expressions (f INDEX) and (f CONCORD) have their standard LFG meaning, with INDEX defined as a nondistributive (resolving) feature and CONCORD as a distributive feature. The expression $f_{(L)}$ is defined as follows:

(24) $f_{(L)} \equiv f \quad \begin{array}{c} \in^* \\ \neg[(\leftarrow \in) <_f \rightarrow] \end{array}$

This expression involves functional uncertainty (Kaplan and Zaenen, 1989), and makes use of the following notational conventions (for more discussion, see Dalrymple, 2001, chap. 5):

- The set-membership symbol $\in$ can be used as an attribute to nondeterministically pick out one of the conjunct members of a coordinate set. $\in^*$ picks out an arbitrarily deeply embedded member, to account for nested coordination: coordinate structures whose conjuncts are themselves coordinate structures.

- Constraints appearing under attributes in a functional uncertainty path are **off-path constraints**, regulating the f-structures through which the path may pass. The symbol $\leftarrow$ in an off-path constraint refers to the f-structure which contains the attribute on which the off-path constraint appears, and the symbol $\rightarrow$ refers to the value for that attribute.

- The symbol $<_f$ denotes the relation of functional precedence, a relation that holds between two f-structures if (roughly) a linear precedence relation holds between the constituent structure nodes that correspond to those f-structures. Functional precedence is formally defined as follows, where ϕ is the function that relates nodes of the constituent structure tree to their corresponding functional structures (Kaplan and Zaenen, 1989):
  f *f-precedes* g ($f <_f g$) if and only if for all $n_1 \in \phi^{-1}(f)$ and for all $n_2 \in \phi^{-1}(g)$, n_1 c-precedes n_2.[5]

[5] According to this definition, the functional precedence relation is irreflexive: an f-structure does not f-precede itself.

With these definitions in place, we can paraphrase the constraint in (24) as follows:

(25) $f_{(L)}$ is an f-structure possibly embedded within f as a conjunct in a coordinate set. If $f_{(L)}$ is embedded as a member of f, it must be the leftmost member: this is accomplished by the off-path constraint, which states that there may not be any (other) members of the coordinate structure that f-precede $f_{(L)}$.

The definition of $f_{(R)}$ is similar except for the f-precedence requirement on the other members of the same set, which is reversed:

(26) $f_{(R)} \equiv f \quad \in^* \quad \neg[(\leftarrow \in) >_f \rightarrow]$

These expressions refer to any leftmost or rightmost conjunct, and so with embedded coordinations, they could refer to a coordinate phrase which is leftmost or rightmost with respect to the other conjuncts, but which itself contains conjuncts. In other words, in a multiply embedded coordinate structure as in (27), the definition in (26) allows agreement with any of the underlined phrases:

(27) $[-\&\underline{[-\&\underline{[-\&\underline{-}]}]}]$

We do not have direct evidence for the possibility of agreement with closest conjuncts which are themselves coordinate structures — the relevant structures would be very complex — but we do not rule out this possibility in principle.

The definitions of f_L and f_R are similar, but add the requirement that the f-structure that is the controller of agreement must not itself be a coordinate structure:

(28) $f_L \equiv f \quad \in^* \quad : \neg(f_L \ \in) \quad \neg[(\leftarrow \in) <_f \rightarrow]$

f_R is defined similarly, but with the functional precedence constraints reversed, as in the definition of $f_{(R)}$.

We can also provide a definition of "closest-conjunct" agreement which relies on the linear relation between the agreement controller and the target.[6] This is useful in cases in which the controller can appear either before or after the target, as for example in free word order languages like Latin or Russian. The definition given in (29) assumes that the relevant relation is between the f-structure of the agreement target, which we assume is $\downarrow$, and the controller f; the closest conjunct is the leftmost one if the controller is on the left, and the rightmost one if the controller is on the right. The definition in (29) is for obligatory closest-conjunct agreement, and relies on the definitions of f_L and f_R given above:

(29) $f_c \equiv \{ \ f_L : \downarrow <_f f_L \ | \ f_R : f_R <_f \downarrow \}$

[6]Thanks to Anna Maykova for discussion of this point.

The definition in (30) is for optional closest conjunct agreement, and is stated by reference to the definitions of $f_{(L)}$ and $f_{(R)}$ given above:

(30) $$f_{(C)} \equiv \begin{array}{ll} \{ & f_{(L)} : \downarrow <_f f_{(L)} \\ \mid & f_{(R)} : f_{(R)} <_f \downarrow \} \end{array}$$

In the next section, we show how these are used in characterising the agreement patterns that we have seen so far.

8.1 Single-conjunct CONCORD agreement

We have seen that single-conjunct agreement with the CONCORD feature is attested in SCB (example 22) and perhaps in Portuguese (example 23). We provide the following lexical entry for the SCB possessive determiner *njegove* 'his', which shows CONCORD agreement with a noncoordinated noun, or with the closest conjunct of a coordinate structure:

(31) njegove 'his': $(\uparrow_C \text{ CONCORD GEND}) = \text{F}$
$(\uparrow_C \text{ CONCORD NUM}) = \text{Pl}$

This lexical entry assumes that the possessive determiner is an f-structure co-head with the noun it modifies. In (32), there is no coordinate structure: $\uparrow_C$ refers to the f-structure labelled p, and requires it to have feminine plural CONCORD. This constraint is satisfied, and the example is grammatical.

(32) njegove molbe
his prayers
'his prayers'

$$p : \begin{bmatrix} \text{PRED} & \text{'prayers'} \\ \text{SPEC} & \text{'his'} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{F} \\ \text{NUM} & \text{Pl} \end{bmatrix} \\ \text{CONCORD} & \begin{bmatrix} \text{GEND} & \text{F} \\ \text{NUM} & \text{Pl} \end{bmatrix} \end{bmatrix}$$

The subject phrase of example (22), repeated here, involves single-conjunct agreement with the closest (leftmost) noun in the coordinate structure:

(33) njegove molbe i uveravanja
his.F.Pl prayers.F.Pl and assurances.N.Pl
'(all) his prayers and assurances'

$$pa : \begin{bmatrix} \text{SPEC} & \text{'his'} \\ & \left\{ \begin{array}{l} \begin{bmatrix} \text{PRED} & \text{'prayers'} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{F} \\ \text{NUM} & \text{Pl} \end{bmatrix} \\ \text{CONCORD} & \begin{bmatrix} \text{GEND} & \text{F} \\ \text{NUM} & \text{Pl} \end{bmatrix} \end{bmatrix} \\ \begin{bmatrix} \text{PRED} & \text{'assurances'} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{N} \\ \text{NUM} & \text{Pl} \end{bmatrix} \\ \text{CONCORD} & \begin{bmatrix} \text{GEND} & \text{N} \\ \text{NUM} & \text{Pl} \end{bmatrix} \end{bmatrix} \end{array} \right\} \end{bmatrix}$$

In (33), the determiner agrees with the leftmost conjunct. Here, $\uparrow$ is instantiated to pa, and according to the definition in (28), $\uparrow_C$ must refer to the closest noncoordinate structure in pa, namely the f-structure for *molbe* ‘prayers’.

8.2 Single-conjunct INDEX agreement

Sadler (1999) and Kuhn and Sadler (2007) discuss single-conjunct agreement in Welsh, showing that single-conjunct agreement is obligatory with coordinate structures in which the first conjunct is pronominal; we assume that this is obligatory single-conjunct INDEX agreement. The subject noun phrase always appears to the right of the verb, and so the closest conjunct is always the leftmost one (if this were not the case, we would state the definition in terms of the “closest-conjunct” requirement f_C rather than the leftmost conjunct f_L). Kuhn and Sadler (2007) provide example (16), repeated here, for the verb *gwelaist* ‘was’:

(34) Gwelaist ti a’th frawd eich hunain.
saw.2Sg 2Sg and-2Sg brother 2Pl self
‘You and your brother saw yourselves.’ (Welsh; Kuhn and Sadler, 2007)

This verb is associated with the following constraints:

(35) gwelaist $((\uparrow \text{SUBJ})_L \text{ INDEX PERS}) = 2$
$((\uparrow \text{SUBJ})_L \text{ INDEX NUM}) = \text{Sg}$

These constraints disallow resolved INDEX agreement in coordinate structures, and require second person singular leftmost-conjunct agreement. The f-structure for example (34) is:

(36)
$$\begin{bmatrix} \text{PRED} & \text{‘see’} \\ \text{SUBJ} & yb: \begin{bmatrix} \text{PERS} & 2 \\ \text{NUM} & \text{Pl} \\ & \left\{ \begin{array}{l} y: \begin{bmatrix} \text{PRED} & \text{‘you’} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 2 \\ \text{NUM} & \text{Sg} \end{bmatrix} \\ \text{CONCORD} & \begin{bmatrix} \text{PERS} & 2 \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix} \\ \begin{bmatrix} \text{PRED} & \text{‘brother’} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{Sg} \end{bmatrix} \\ \text{CONCORD} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix} \end{array} \right\} \end{bmatrix} \\ \text{OBJ} & [\ldots\text{‘self’}\ldots] \end{bmatrix}$$

The expression $(\uparrow \text{SUBJ})_L$ refers to the f-structure labelled yb, if it is not a coordinate phrase, or to the leftmost conjunct in yb. Since yb is a coordinate structure, agreement is with the leftmost conjunct *ti* 'you.Sg', labelled y.

The SCB example in (37) shows that single-conjunct and resolved INDEX agreement patterns can be found in the same sentence:

(37)
[[Deca] i [unučad]] koj-a/koj-i su
children and grandchildren who-N.Pl/who-M.Pl Aux.3Pl
došl-a/došl-i su gladn-a/gladn-i
come-N.Pl/come-M.Pl be.3Pl hungry-N.Pl/hungry-M.Pl
'The children and grandchildren who came are hungry.' (SCB; elicited)

We focus here on the relative pronoun, which, like the Welsh verb, exhibits single-conjunct INDEX agreement. The noun *unučad* 'grandchildren' has neuter plural INDEX, and the coordinate noun phrase *deca i unučad* 'children and grandchildren' has resolved masculine plural INDEX. Either resolved INDEX agreement (masculine plural) or single-conjunct agreement referring to the INDEX features of 'grandchildren' (neuter plural) is possible:[7]

(38)
$$\begin{bmatrix} \text{PRED} & \text{'be.hungry'} \\ \text{SUBJ} & cg: \begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{M} \\ \text{NUM} & \text{Pl} \end{bmatrix} \\ & \left\{ \begin{array}{l} \begin{bmatrix} \text{PRED} & \text{'children'} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{N} \\ \text{NUM} & \text{Pl} \end{bmatrix} \\ \text{CONCORD} & \begin{bmatrix} \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix} \\ \begin{bmatrix} \text{PRED} & \text{'grandchildren'} \\ \text{INDEX} & \begin{bmatrix} \text{GEND} & \text{N} \\ \text{NUM} & \text{Pl} \end{bmatrix} \\ \text{CONCORD} & \begin{bmatrix} \text{GEND} & \text{F} \\ \text{NUM} & \text{Sg} \end{bmatrix} \end{bmatrix} \end{array} \right\} \\ \text{ADJ} & \left\{ \begin{array}{l} \begin{bmatrix} \text{TOPIC} & r: \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{rel} \end{bmatrix} \\ \text{PRED} & \text{'come'} \\ \text{SUBJ} & \text{———} \end{bmatrix} \end{array} \right\} \end{bmatrix} \end{bmatrix}$$

Agreement requirements for the neuter plural relative pronoun are expressed in (39). In (39), $\uparrow$ refers to the f-structure labelled r in (38), which appears as the TOPIC of the relative clause modifying the coordinate nouns. The expression

[7]Not all combinations of resolved and single-conjunct agreement are possible in this example; the relative clause verb must agree with the relative pronoun, and the predicate *gladn-a/i* 'hungry' is virtually certain to be masculine plural if the relative pronoun is masculine plural. Agreement patterns in such constructions obey Corbett's Agreement and Predicate Hierarchies (Corbett, 2006).

(MOD $\in$ TOPIC $\uparrow$) in (39) refers to the f-structure labelled cg in (38), which corresponds to the head noun which the relative clause modifies:

(39) koja (neuter plural relative pronoun) $((\text{ADJ} \in \text{TOPIC} \uparrow)_R \text{ INDEX GEND}) = \text{N}$
$((\text{ADJ} \in \text{TOPIC} \uparrow)_R \text{ INDEX NUM}) = \text{Pl}$

These constraints require the rightmost conjunct of the coordinated head nouns to be neuter plural; the constraint is satisfied, and example (37) is grammatical.

In (39), we have chosen to characterise the constraints associated with *koja* as involving obligatory rather than optional rightmost conjunct agreement. This is because resolved agreement can never be neuter: as shown by example (13) above, uniformly neuter conjuncts resolve to masculine, not neuter. Thus, neuter plural agreement in coordinate structures must be with a distinguished conjunct, and not with the resolved features.

For the masculine plural relative pronoun, the situation is different: if the rightmost conjunct is masculine plural, the resolved features must also be masculine plural. Therefore, we can treat masculine plural agreement as resolved agreement and not rightmost conjunct agreement in all cases. Finally, feminine plural agreement may be either with the rightmost conjunct (the resolved features may be either feminine plural or masculine plural, depending on the gender of the other conjuncts) or with the resolved feminine plural features of a coordinate phrase (the rightmost conjunct may be either feminine singular or feminine plural).

An alternative analysis is possible according to which all of the relative pronouns specify **optional** rightmost conjunct agreement: the feminine relative pronoun is correctly allowed to agree with either the rightmost conjunct or the resolved features; the neuter relative pronoun is allowed to agree with the resolved features (which, however, will never be neuter plural, due to the resolution rules of SCB); and the masculine relative pronoun is allowed to agree with the rightmost conjunct (though if the rightmost conjunct is masculine plural, the resolved features must also be masculine plural). This would produce uniformity of specification across the relative pronoun paradigm, though it would produce multiple analyses of apparently unambiguous structures; we prefer specifications which do not produce ambiguity.

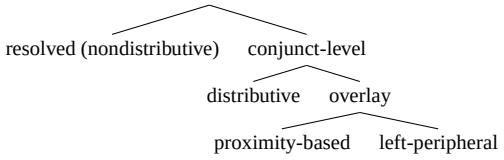
9 An alternative: Kuhn and Sadler 2007

Kuhn and Sadler (2007) provide a thorough discussion of previous proposals for agreement in coordinate structures in LFG, dividing them into **description-based approaches** and **representation-based approaches**. Our approach is description-based, since it works by introducing new functional vocabulary for expressing agreement constraints, rather than enriching the functional structure with additional structure. We are in full agreement with Kuhn and Sadler's criticisms of representation-based approaches, which we do not rehearse here. Kuhn and Sadler criticise existing description-based approaches on the grounds that they require

complex disjunctive statements of constraints, and cannot handle nested coordination. Our approach does not suffer from these difficulties, and so is immune to their criticisms.

Kuhn and Sadler propose the classification of features shown in (40):

(40)



Their approach requires features to be assigned to exactly one classification, and to behave uniformly as that classification requires. The main difficulty with this proposal is the existence of optional single-conjunct agreement. A classification of features entails that a feature will always behave in a certain way: always requiring resolved agreement, for example, or always requiring single-conjunct agreement. However, example (37) shows that the INDEX feature can participate in both single-conjunct agreement and resolved agreement in the same example, which is unexpected on Kuhn and Sadler's view.

10 Conclusion

We have presented an approach to agreement with coordinated structures which allows for the specification of agreement constraints with the entire coordinated phrase or with a single conjunct of the phrase. Our approach allows for the wide variation in agreement patterns exhibited in SCB, and we believe that it will extend unproblematically to agreement patterns in other languages as well. Our research turned up a number of interesting agreement patterns that remain for future work.

First, a basic assumption made by King and Dalrymple (2004) and others is that INDEX features resolve in coordination, but not CONCORD features. However, we have found some examples that seem to exemplify resolved CONCORD:

(41) [[žena] i [deca]] su mu bile u drugoj prostoriji.
wife.F.Sg and children.F.Sg Aux.Pl to.him been.F.Pl in other areas
'His wife and children were in other areas.' (SCB)
http://www.ana.rs/forum/viewtopic.php?t=24328&postdays=0&postorder=asc&start=15
(accessed on 12.06.2010)

The F.Pl agreement on the past participle 'been' above can be neither closest-conjunct nor distributive agreement, as neither conjunct is F.Pl. Therefore, the F.Pl features on 'been' could be analysed as resolved agreement. The resolution rules for SCB produce F.Pl only when both conjuncts are feminine. 'Wife' has F.Sg CONCORD and INDEX, whereas 'children' has F.Sg CONCORD and N.Pl INDEX. Therefore, it seems that it is the CONCORD feature that has resolved, and not the INDEX (a combination of feminine and neuter would resolve to masculine).

However, evidence from Wechsler (1999) suggests that it is the semantics and not the CONCORD that has produced the F.Pl agreement on the target in (41). This would mean that the children in the example above must be semantically feminine (although we currently have no native-speaker grammaticality judgements). The following example is provided in support of this hypothesis:

(42) Ova velika devoja i ovo malo devojče su se
This.F.Sg big.F.Sg girl.F.Sg and this.N.Sg little.N.Sg girl.N.Sg Aux.Pl Refl
lepo igrale/?igrali
well played-F.Pl/played-M.Pl
'This big girl (F) and this little girl (N) played well.' (SCB; Wechsler, 1999, 29)

Here, 'big girl' is unambiguously feminine both grammatically (in CONCORD and INDEX) and semantically. By contrast, 'little girl' has N.Sg INDEX and CONCORD but is semantically feminine. This shows that no feminine grammatical features are needed for this semantic resolution, and so in the absence of conclusive proof to the contrary, we conclude that the same generalisation is relevant for (41).

Second, it seems to be possible (though rare) for a single agreeing target to depend on different functional structure controllers; Arnold et al. (2007) provide examples of Portuguese adjectives which show closest-conjunct agreement for gender but resolved agreement for number, and we have found several similar examples (see also Corbett 1983, van Oirsouw 1987, Camacho 2003). These patterns would pose no problems for our formal model, but their rarity and the constraints on their distribution deserve investigation.

References

Arnold, Doug, Sadler, Louisa and Villavicencio, Aline. 2007. Portuguese: Corpora, Coordination and Agreement. In Sam Featherston and Wolfgang Sternefeld (eds.), *Roots: Linguistics in Search of its Evidential Base*, Studies in Generative Grammar, No. 96, pages 9–28, The Hague: Mouton de Gruyter.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell Publishers.

Camacho, José. 2003. *The Structure of Coordination: Conjunction and Agreement Phenomena in Spanish and Other Languages*. Dordrecht: Kluwer Academic Publishers.

Corbett, Greville G. 1983. *Hierarchies, Targets, and Controllers: Agreement Patterns in Slavic*. London: Croom Helm.

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

Corbett, Greville G. 2001. Agreement: Terms and Boundaries. In *The Role of Agreement in Natural Language: Proceedings of the 2001 Texas Linguistic Society Conference*.

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

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

Dalrymple, Mary, Dyvik, Helge and Sadler, Louisa. 2006. Gender Resolution in Coordination and Disjunction. Abstract in *On-line Proceedings of the LFG2006 Conference*, ed. Miriam Butt and Tracy Holloway King.

Dalrymple, Mary and Kaplan, Ronald M. 2000. Feature Indeterminacy and Feature Resolution. *Language* 76(4), 759–798.

Dalrymple, Mary, Kaplan, Ronald M., Maxwell, III, John T. and Zaenen, Annie (eds.). 1995. *Formal Issues in Lexical-Functional Grammar*. Stanford, CA: 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, Cambridge, MA: The MIT Press, reprinted in Dalrymple et al. (1995, 29–130).

Kaplan, Ronald M. and Zaenen, Annie. 1989. Long-distance Dependencies, Constituent Structure, and Functional Uncertainty. In Mark Baltin and Anthony Kroch (eds.), *Alternative Conceptions of Phrase Structure*, pages 17–42, Chicago University Press, reprinted in Dalrymple et al. (1995, 137–165).

Kathol, Andreas. 1999. Agreement and the Syntax-Morphology Interface in HPSG. In Robert Levine and Georgia Green (eds.), *Studies in Current Phrase Structure Grammar*, pages 223–274, Cambridge, UK: Cambridge University Press.

King, Tracy Holloway and Dalrymple, Mary. 2004. Determiner Agreement and Noun Conjunction. *Journal of Linguistics* 40(1), 69–104.

Kuhn, Jonas and Sadler, Louisa. 2007. Single Conjunct Agreement and the Formal Treatment of Coordination in LFG. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG2007 Conference*.

Sadler, Louisa. 1999. Non-Distributive Features and Coordination in Welsh. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG99 Conference*.

Sadler, Louisa. 2003. Coordination and Asymmetric Agreement in Welsh. In Miriam Butt and Tracy Holloway King (eds.), *Nominals: Inside and Out*, Stanford, CA: CSLI Publications.

Sadler, Louisa. 2006. Gender Resolution in Rumanian. In Miriam Butt, Mary Dalrymple and Tracy Holloway King (eds.), *Intelligent Linguistic Architectures: Variations on themes by Ronald M. Kaplan*, Stanford, CA: CSLI Publications.

van Oirsouw, Robert R. 1987. *The Syntax of Coordination*. New South Wales, Australia: Croom Helm.

Wechsler, Stephen. 1999. Gender Resolution in Coordinate Structures. In *Proceedings of the Workshop on the Structure of Spoken and Written Texts*, The University of Texas at Austin.

Wechsler, Stephen and Zlatić, Larisa. 2003. *The Many Faces of Agreement*. Stanford, CA: CSLI Publications.

AN UNMEDIATED ANALYSIS OF RELATIVE CLAUSES

Yehuda N. Falk

The Hebrew University of Jerusalem

Proceedings of the LFG10 Conference

Miriam Butt and Tracy HollowayKing (Editors)

2010

CSLI Publications

http://csli-publications.stanford.edu

The standard analysis of relative clauses takes the relativized element in the relative clause to be a pronoun, sometimes unexpressed. However, consideration of internally-headed relative clauses, the behavior of idioms, and the cross-linguistic forms of relative clauses argue for a constraint-based version of the alternative "raising" analysis.

1. Introduction

This paper[1] deals with the analysis of (restrictive) relative clauses. In this paper, I will argue that a consideration of the forms that relative clauses take suggests an alternative to the current consensus analysis, an alternative which has roots in the transformational literature of the 1970s and has been adopted in some current transformational analyses.

The standard view of relative clause constructions is that they consist of three parts: the head, the relative pronoun, and the clause.

(1)	the word processor	which	Bill prefers *e*
	head	**relative pronoun**	**clause**

This view holds the relative pronoun to be the central component of this construction, as it serves to link the other two elements of the construction: it is coreferential (or coindexed) with the head, and it has a grammatical function (and in transformational accounts originates) in the clause. The relation between the head and the in-clause function is indirect, mediated anaphorically by the relative pronoun. I will refer to this as the **anaphorically mediated** analysis of relative clauses.

Anaphorically mediated analyses of this kind represent a broad consensus in syntax, one which transcends disagreements about theoretical framework. From this perspective, it is very odd that there is an alternative form for relative clauses in English, one in which there is no relative pronoun:[2]

(2)	the word processor	(that) Bill prefers *e*
	head	**clause**

The existence of this kind of relative clause suggests a direct relation between the head and the in-clause position, what we might refer to as an **anaphorically unmediated** analysis. Yet, since Chomsky (1977) the standard analysis of this kind of relative clause has been an anaphorically mediated analysis, in which the mediating element is an unexpressed pronominal element. This is also the

[1]Thanks to Alex Alsina, Doug Arnold, Ash Asudeh, Joan Bresnan, Mary Dalrymple, Gabi Danon, Irit Meir, and Louisa Sadler for comments on various versions. This research is supported by the Israel Science Foundation (Grant No. 207/07).

[2]I am assuming, as is standard, that the optional *that* in finite relative clauses is the ordinary finite complementizer and not a relative pronoun.

standard analysis in LFG, with a [PRED 'PRO'] element functioning as TOPIC in the relative clause (Dalrymple 2001, Falk 2001).

This paper will argue for an anaphorically unmediated analysis of relative clauses. In §2 I will present new typologically based arguments for an anaphorically unmediated analysis and in §3 I will revisit some old arguments. This is followed with an analysis in §4. Finally, in §5 I return to what is standardly taken to be the paradigm case of relative clauses, the ones with relative pronouns. I will show that they also should be analyzed as anaphorically unmediated, and provide an explanation for the existence of this form of relative clause.

For the sake of style, I will drop the word "anaphorically" for the rest of this paper, and simply refer to "mediated" and "unmediated" analyses. The intention remains that the mediating element is an anaphoric element.

2. Typology

2.1. *Wh* Constructions

Relative clauses are one example of a class of constructions that has been referred to variously in the literature as *wh* movement (Chomsky 1977), long-distance dependencies (Bresnan 2001), unbounded dependencies (Pollard and Sag 1994), and Ā dependencies (Chomsky 1981), among other names. We will refer to them here as *wh* constructions.

What typifies *wh* constructions is that a single element has more than one function, potentially in distant clauses. Consider, for example, the following *wh* question.

(3) Who were you told that computer industry experts claim uses the TextMangler wordprocessor?

In this sentence, the DP *who* functions both as FOCUS in the main clause and as the SUBJ of *use* two clauses down in the sentence.

Wh constructions face what might be thought of as a realizational problem. If an element has two different grammatical functions, potentially in different clauses, in which position is it expressed overtly? Assuming some notion of economy, it is unlikely to be expressed in both positions. In principle, one would expect that either the higher or the lower position should be available for realizing the multifunctional element. And this is exactly what we find: some languages choose the structural position of the lower function, some choose the structural position of the higher function, and some allow either.

(4) Mandarin (low position only) (Huang 1982: 371)
Zhangsan xiangzin shei mai-le shu?
Zhangsan believe who bought books
'Who does Zhangsan believe bought books?'

(5) Egyptian Arabic (low position only) (Kenstowicz and Wahba 1983: 263)

a. Fariid ḥatt ʔeeh ʕala l- tarabeeza.
Fariid put what on the- table
'What did Fariid put on the table?'

b. Fariid ḥatt kitaab miin ʕala l- tarabeeza?
Fariid put book who on the- table
'Whose book did Fariid put on the table?'

c. Fariid ḥatt kitaab Mona ʕala ʔeeh?
Fariid put book Mona on what
'What did Fariid put Mona's book on?'

(6) English (high position only)
Which word processor do you think Bill uses?

(7) Kikuyu (either position) (Bergvall 1983: 247)

a. Oɣweʃiːria Goɣe oiɣirɛ mahɛirɛ keŋaŋi oː?
you.think Ngũgĩ said they.gave crab who
'Who do you think Ngũgĩ said they gave a crab to?'

b. Noo oɣweʃiːria Goɣe oiɣirɛ mahɛirɛ keŋaŋi?
FOC.who you.think Ngũgĩ said they.gave crab
'Who do you think Ngũgĩ said they gave a crab to?'

I will use the familiar but misleading term "in-situ" for the low realization construction.

For our purposes, the in-situ construction is important because it provides valuable information about the nature of *wh* constructions. While the "moved" construction, as in English, provides evidence for one function of the *wh* element, the in-situ construction provides evidence of the other one.

2.2. Internally Headed Relative Clauses

The "in-situ" construction associated with relative clauses is what is often referred to as the internally headed relative clause (IHRC).

(8) a. Dogon (Culy 1990: 21)
[Kandɔw nyan gɛ tɛgɔ]
just fire granary burn.PSTNARR.3SG.DEF
nɛ yu gaw to.
in millet a lot exist.3SG
'There was a lot of millet in the granary that the fire just burned.'

b. Imbabura Quechua (Cole 1982: 49)
[Wambra wagra- ta randi- shka] ali wagra- mi.
boy cow- ACC buy- NMNL good cow- FOC
'The cow which the boy bought is a good cow.'

c. Mooré (Culy 1990: 76)
[Yãmb sẽn yã dao ninga zamẽ wã] bee ka.
2PL AUX saw man INDEF yesterday DEF be there
'The man that you saw yesterday is here.'

d. Navajo (Platero 1974: 214)
[[[Hastiin łééchąą'í yiztał] nisin]
man dog 3.PERF.3.kick IMPFC.1.think
ni- (n)ęę] nahał'in.
IMPFC.3.say- REL IMPFC.3.bark
'The dog which he said he thought the man kicked is barking.'

As with in-situ questions, the IHRC is an alternative solution to the realization problem.

What is notable about IHRCs is the insight they provide into the nature of relative clauses. They show a direct connection between the external head position and the in-clause position: the choice of where to realize the relativized element is limited to these two positions. Agreement provides evidence that there is a direct connection. For example, in Dogon the verb agrees with the subject. If the relativized element is the subject of both the verb inside the relative clause (where it appears) and of the larger clause in which the relative construction is embedded (where it does not appear), both verbs must agree with it (Culy 1990: 83–84):

(9) a. Ya [[yaan pilli wɛn] gɔ] Moti
yesterday women White see.NPST.PL DEF Mopti
boliya.
go.PSTNARR.3PL
'The women who saw the White yesterday went to Mopti.'

b. Ya [[yaan pilli wɛ] gɔ] Moti
yesterday women White see.NPST DEF Mopti
boliya.
go.PSTNARR.3PL
'The women who the White saw yesterday went to Mopti.'

c. *Ya [[yaan pilli wɛn] gɔ] Moti
yesterday women White see.NPST.PL DEF Mopti
boli.
go.PSTNARR
'The women who saw the White yesterday went to Mopti.'

This shows that the relativized element is syntactically associated with both positions. Crucially, IHRCs provide no evidence for a relative pronoun, or for a mediated analysis of the relative clause construction.

It is instructive to consider what an in-situ mediated structure would look like. Consider (8d). Under the non-mediated analysis, 'dog' is both the

head of the construction and the clause-internal relativized element. In English it is realized in the position of the head, and in Navajo in the clause-internal position. However, under a mediated analysis, the head position is not part of the same functional unit as the clause-internal position. Instead, the clause-internal position is linked to the mediating relative pronoun. Under the mediated analysis, one would expect an in-situ relative clause to look not like the IHRC in (8d), which we can schematize as (10a), but rather like (10b).

(10) a. The [that he said he thought the man kicked dog] is barking.
b. The dog [(that) he said he thought the man kicked which] is barking.

Structures like (10b) appear not to exist, at least not in this form. (10b) resembles a resumptive pronoun construction (for competing LFG analyses, see Falk 2002 and Asudeh 2004), but resumptive pronouns crucially are not *wh* relative pronouns but ordinary personal pronouns. Resumptive pronoun relative clauses thus also provide no reason to accept a mediated analysis.

We conclude, therefore, that in-situ constructions provide evidence against the mediated analysis of relative clauses. They suggest very strongly that the correct analysis is the unmediated analysis.

2.3. Pronoun-less Relative Clauses Cross-Linguistically

The existence of pronoun-less relative clauses in English is not a quirk. Relative clauses which appear to be unmediated (with or without an invariant relative particle/complementizer) are quite widely attested.[3]

(11) a. Hebrew
meabed hatamlilim še Bill maadif
processor DEF.texts COMP Bill prefers
'the word processor that Bill prefers'
b. Japanese
Watasi wa sono otoko ga tataita inu o miru.
I TOP that man NOM struck dog ACC see
'I see the dog that the man struck.'
c. Maori
te taane i patu- a e te wahine
the man PST hit- PASS by the woman
'the man who was hit by the woman'

[3]All the examples other than the Hebrew come from Keenan and Comrie (1979).

d. Korean
hyənsik- i kɨ lä- lɨl ttäli- n maktäki
Hyensik- NOM the dog- ACC beat- REL stick
'the stick with which Hyensik beat the dog'

e. Persian
John mard- i râ ke zan zad mišenâsad.
John man- the ACC COMP woman hit knows
'John knows the man who the woman hit.'

f. Swedish
Jag har en bror, som talar tyska.
I have a brother REL speaks German
'I have a brother who speaks German.'

g. Yoruba
iṣu ti mo ra lana naa
yam REL I buy yesterday that
'that yam that I bought yesterday'

In fact, as shown by Maxwell (1979) for the 49 languages in the database of Keenan and Comrie (1979), languages with pronoun-less relative clause constructions are quite common. In some languages, such as Toba Batak and Japanese, these are the only kind of relative clause, while in others, such as Spanish and Czech, relative-pronoun relatives also exist.

The cross-linguistic distribution of relative-pronoun-less relative clauses is significant. If a mediated analysis were the right analysis of relative clauses, relative-pronoun-less constructions would be at best a highly marked construction. The fact that they are more common cross-linguistically than relative-pronoun relative clauses indicates that an unmediated analysis is called for.

3. Earlier Arguments Revisited

While mediated analyses of relative clauses are much more common in the literature, unmediated analyses have appeared as well. The alternative analysis in transformational theory is what is generally referred to as the raising analysis. The first appearance of this analysis in the literature is Schachter (1973: 31–35), who proposed the following underlying structure for relatives (where Δ indicates an empty position, *e* in more modern notation):

(12) $_{Nom}[\ _{Nom}[\Delta]\ S]$

The relativized NP then raises from the S into the empty position in the structure.

The unmediated analysis (under the guise of the raising analysis) was developed in Vergnaud (1974), and has been revived by Kayne (1994). While

various arguments have been given over the years for an unmediated analysis, the most compelling one comes from the behavior of idiom chunks, as in the following examples from Hulsey and Sauerland (2006: 114).

(13) a. Mary praised the headway that John made.
b. I was shocked by the advantage that she took of her mother.

Vergnaud (1974: 57) provides a similar example from French, using the French equivalent of the idiom *take part in*.

(14) Il est surpris de la part que Jean a prise aux débats.
he is surprised of the part that John has taken at.the debates
'He is surprised at the part that John took in the debates.'

In anybody's theory of idioms, idiom chunks like *headway*, *advantage*, and *part* are licensed by being the object of *make*, *take*, and *prendre*, respectively. Schematically (and abstracting away from specific theoretical frameworks), the mediated analysis provides the following analyses of these sentences (R=relative pronoun):

(15) a. Mary praised the headway [that John made R]
b. I was shocked by the advantage [that she took R of her mother]
c. Il est surpris de la part [que Jean a prise R aux débats]

The idiom chunks are not the objects of the licensing verbs in these cases; the relative pronouns are, and the idiom chunks appear in positions in which they are not licensed. This kind of anaphoric relation is not normally possible for idiom chunks:

(16) a. *Mary always praises headway when John makes it.
b. *I was shocked by the advantage when I saw her take it of her mother.

Under the unmediated analysis, on the other hand, the idiom chunks are correctly licensed:

(17) a. Mary praised the *x* [that John made *x*]; *x*=headway
b. I was shocked by the *x* [that she took *x* of her mother]; *x*=advantage
c. Il est surpris de la *x* [que Jean a prise *x* aux débats]; *x*=part

This provides a strong argument in favor of the unmediated analysis.

Previous versions of the unmediated analysis have been based on a derivational model of syntax, in which the head is taken to originate in the

relative clause and move to its surface position, as opposed to the non-derivational LFG account proposed here, in which the relativized element serves two functions simultaneously. The idiom chunk facts provide evidence to distinguish the derivational from the non-derivational account. Consider the following example.[4]

(18) Mary never made the headway that had been expected of her.

In this case, the verb of which *headway* is the surface object is the licensing verb. For a derivational raising analysis, the D-structure is (19a), while for a non-derivational multifunctionality analysis it is (19b).

(19) a. [Mary never made [the *e* [UNEXPR.SUBJ had been expected headway of her]]]
b. Mary never made the *x* [that *x* had been expected of her]; *x*=headway

The raising analysis provides the wrong account of this case, while the non-derivational multifunctional analysis has *headway* as the object of the licensing verb *made* here as well. The idiom-chunk facts thus provide evidence not only for an unmediated analysis, but for an LFG-style implementation.

A less convincing argument comes from the binding of reflexive anaphors (examples from Hulsey and Sauerland 2006: 115):

(20) a. I saw the picture of himself that John liked.
b. Mary discovered the book about himself that Bob wrote.

In these cases, the DP containing *himself* must be directly associated with the object position inside the relative clause for the coreference to be grammatical; such an association is provided by the unmediated analysis but not by the mediated analysis. On the other hand, if the binding of reflexives in picture noun phrases is not governed by syntactic principles (e.g. if the reflexives in picture NPs are logophors), this argument does not go through.

4. The Unmediated Analysis

4.1. Basics

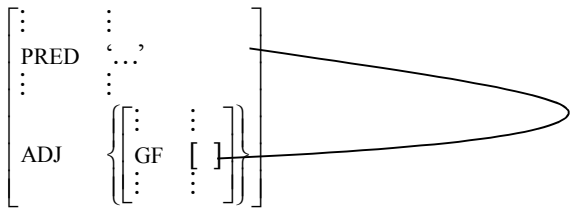
On the basis of the old arguments from idiom chunks and anaphoric

[4]I would like to thank Joan Bresnan (p.c.) for pointing this example out to me. Bresnan also notes that similar issues arise in pseudoclefts such as the following:
(i) What we have to make is more headway!
Here the object of *make* is *what*, not *headway*. I do not have anything to say about this case, as I do not have an analysis of pseudoclefts. It is possible that this should be taken as evidence for the direction that an analysis of pseudoclefts should go.

binding and the new arguments from the in-situ construction and the widespread distribution of relative-pronoun-less relatives, I take it that an unmediated analysis of relative clauses is the correct one. In this section, I will develop the analysis. As will become clear shortly, the analysis is not entirely straightforward, and could be construed as evidence against an unmediated analysis. Given the evidence that has been presented in the previous sections, I would consider this to be a mistake. The evidence for an unmediated analysis is clear; the challenge is to find the right descriptive tools to get the details right.

The core of an unmediated analysis of relative clauses is that the relativized element is a structure-shared element that has two functions: whatever function the nominal phrase has in the larger sentence and its function within the relative clause. (While it is convenient to speak of the head of the construction as the element that has these two functions, it is actually the larger nominal phrase of which the relative clause is an adjunct. The "head" is not an independent element.) In (21b), the structure-sharing functions are SUBJ of *leaves* and OBJ of *booked*.

(21) a.



b. [The trip to Mars that I booked [*e*]] leaves on Tuesday.

However, this is not quite accurate. A closer look reveals that the entities that fill the two functions do not have identical content.

The first difference is that the ADJ (i.e. the relative clause itself) is part of the SUBJ of *leaves* but not part of the OBJ of *booked*. The OBJ of *booked* is *trip to Mars*, not *trip to Mars that I booked*. The result is that while we want a direct relation between the larger nominal phrase and the in-clause element, they cannot be identical.

The next problem is case marking. In most languages, the relativized element always has the appropriate case for the position (head or in-clause) in which it is realized in c-structure.[5] There is no requirement that the same case be appropriate for the other position. This is true for both externally headed relatives and internally headed ones. In the case of IHRCs, the relative clause itself can be marked for case.

[5]Other treatments of case are also attested. Culy (1990: 268) discusses Cuzco Quechua case attraction, under which the larger nominal phrase can, for some speakers, be marked with the case of the in-clause element in both internally- and externally-headed relative clauses.

(22) a. **Hebrew**
externally headed
accusative head, nominative in-clause
Kaniti et ha- sefer še hicxik otxa.
I.bought ACC DEF- book that made.laugh you
'I bought the book that made you laugh.'

b. **Diegueño** (Gorbet 1977)
internally headed
inessive head, accusative (Ø suffix) in-clause
Tənay 'waː 'wuːw- pu- L^y 'čiyawx.
yesterday house I.see- DEF- INESSIVE I.sing.IRR
'I'll sing in the house I saw yesterday.'

This is a problem because a single structure-shared element is usually thought to have the same case everywhere. In this construction, however, the entities that bear the two grammatical functions are not completely identical; as we have already seen, one of them includes the relative clause and the other does not. It is thus not a completely structure-shared element. I hypothesize that the lack of complete identity allows the lack of case connectivity.

The most important element that is not shared is the DEF feature. This is not overtly visible in the case of externally headed relatives, as there is a gap in the position in the relative clause, and thus no overt marking of DEF. However, IHRCs, in which the construction and the internal head can be independently marked for a DEF value, clearly show that the head (or rather, the larger nominal phrase) and the in-clause position have different values.

Perhaps the most common pattern is that there is DEF marking on the larger nominal phrase but not on the internal head. This can be seen in the Diegueño example (22) above. Further examples are the following.

(23) Dogon[6] (Culy 1990: 20)
Iye kekeginɛ ya yaana poñ lagɔ agiya.
today crazy.person yesterday woman large.OBJ hit.DEF caught.3PL
'Today they caught the crazy person who hit a large woman.'

(24) Tibetan (Keenan 1985: 161)
Peemɛ thep khii- pa the nee yin.
Peem.ERG book.ABS carry- PART the.ABS I.GEN be
'The book Peem carried is mine.'

Since all nominal elements other than case appear on the internal head and not on the larger nominal phrase, this raises the suspicion that the DEF is specifically a property of the larger nominal phrase and not of the internal head.

[6]In this example, the determiner *gɔ* cliticizes onto the verb form *lagi*.

Other languages have a determiner marking a DEF value on the internal head that does not express the definiteness of the larger nominal phrase. In Bambara, there normally are no determiners, but there is a special determiner marking the internal relative head.

(25) Bambara (Keenan 1985: 162)
Tye ye ne ye so min ye san.
man PST I PST horse REL see buy
'The man bought the horse which I saw.'

This suggests that there is a special [DEF REL] value for the clause-internal element. More striking are languages like Lakhota and Mooré, in which two determiners are in evidence, one for the larger nominal phrase, and another one (indefinite) marking the internal head.

(26) Lakhota (Williamson 1987: 171)
a. Mary owiža wą kağe ki he ophewathų.
Mary quilt a make the DEM I.buy
'I bought the quilt that Mary made.'
b. Mary owiža wą kağa cha he ophewathų.
Mary quilt a make INDEF DEM I.buy
'I bought a quilt that Mary made.'
c. *Mary owiža ki kağe ki he ophewathų.
Mary quilt the make the DEM I.buy
'I bought the quilt that Mary made.'

(27) Mooré (Culy 1990: 76)
Yãmb sẽn yã dao ninga zamẽ wã bee ka.
2PL AUX saw man INDEF yesterday DEF be there
'The man that you saw yesterday is here.'

Williamson (1987) argues that the indefinite marking on the internal head is semantically motivated.

From the syntactic perspective, then, the in-clause element, although the same entity as the larger nominal phrase, is not feature-identical to it. This means that the analysis of relative clauses, although unmediated, will not involve a simple structure-sharing. Instead, the relative clause must contain a modified version of the larger nominal phrase. The functional equation licensing relative clauses specifies that the DEF value of the larger nominal phrase is replaced by the special REL value. This modified version lacks the relative clause adjunct and gets its own case-marking. These are the only differences between the two functional positions.

Expressing this formally presents some challenges. In the first place, we need to be able to selectively exclude features from one or more functions. This can be achieved formally through the use of the restriction operator (Kaplan and

Wedekind 1993), which will exclude the sharing of DEF and ADJ.[7] The other problem is correctly incorporating the restriction operator in an equation relating the larger element to an arbitrary in-clause function. This difficulty is compounded in those varieties of LFG which use inside-out functional uncertainty proceeding from the position of the in-clause function (inter alia Bresnan 1995, Falk 2007). My proposal is that the information flow between the larger nominal and the in-clause function is mediated by an OPERATOR (or OPER)[8] at the root of the relative clause. The f-structure of the English sentence (21b) with an externally headed relative clause is (28a), and the phrase structure rule required is (28b). For clarity, the shared elements are shown in normal type in the f-structure position associated with the c-structure position in which they actually occur, and grayed-out in other f-structure positions. This notation replaces the usual curved line.

(28) a.

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{DEF} & + \\ \text{PRED} & \text{'trip}\ \langle(\uparrow \text{OBL}_{\text{Goal}}\ \text{OBJ})\rangle\text{'} \\ \text{OBL}_{\text{Goal}} & \begin{bmatrix} \text{PCASE} & \text{OBL}_{\text{Goal}} \\ \text{OBJ} & [\text{PRED}\ \text{'Mars'}] \end{bmatrix} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{WHPATH} & [+\text{T}] \\ \text{OPER} & \begin{bmatrix} \text{DEF} & \text{REL} \\ \text{PRED} & \text{'trip}\ \langle(\uparrow \text{OBL}_{\text{Goal}}\ \text{OBJ})\rangle\text{'} \\ \text{OBL}_{\text{Goal}} & \begin{bmatrix} \text{PCASE} & \text{OBL}_{\text{Goal}} \\ \text{OBJ} & [\text{PRED}\ \text{'Mars'}] \end{bmatrix} \end{bmatrix} \\ \text{SUBJ} & [\text{"I"}] \\ \text{PRED} & \text{'book}\ \langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{OBJ} & \begin{bmatrix} \text{DEF} & \text{REL} \\ \text{PRED} & \text{'trip}\ \langle(\uparrow \text{OBL}_{\text{Goal}}\ \text{OBJ})\rangle\text{'} \\ \text{OBL}_{\text{Goal}} & \begin{bmatrix} \text{PCASE} & \text{OBL}_{\text{Goal}} \\ \text{OBJ} & [\text{PRED}\ \text{'Mars'}] \end{bmatrix} \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix} \\ \text{TENSE} & \text{PRES} \\ \text{PRED} & \text{'leave}\ \langle(\uparrow \text{SUBJ})\rangle\text{'} \\ \text{ADJ} & \{[\text{"on Tuesday"}]\} \end{bmatrix}$$

b.

$$\begin{array}{lcl} \text{NP} \rightarrow & \text{NP} & \text{CP} \\ & \uparrow = \downarrow & \downarrow \in (\uparrow \text{ADJ}) \\ & & (\downarrow \text{OPER}) \backslash \text{DEF} = \uparrow \backslash \text{DEF} \backslash (\text{ADJ} \in \rightarrow = \downarrow) \\ & & (\downarrow \text{OPER DEF}) = \text{REL} \\ & & (\downarrow \text{WHPATH}) = [+\text{T}] \end{array}$$

[7]And CASE, if it is an f-structure feature. For a contrary view of case, see Falk (2006).

[8]The name OPER is used in Falk (2001). Other names have been used in the LFG literature, such as Q for interrogative operators and RELPRO for relative operators in Kaplan and Bresnan (1982) and Dalrymple (2001).

Similarly, the Lakhota example (26a) will have the following f-structure.

(29) a.

$$\begin{bmatrix} \text{SUBJ} & [\text{“I”}] \\ \text{PRED} & \text{‘buy}\ \langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle\text{’} \\ \text{OBJ} & \begin{bmatrix} \text{DEF} & + \\ \text{PRED} & \text{‘quilt’} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{OPER} & \begin{bmatrix} \text{DEF} & \text{REL} \\ \text{PRED} & \text{‘quilt’} \end{bmatrix} \\ \text{WHPATH} & [+\text{T}] \\ \text{SUBJ} & [\text{“Mary”}] \\ \text{PRED} & \text{‘make}\ \langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle\text{’} \\ \text{OBJ} & \begin{bmatrix} \text{DEF} & \text{REL} \\ \text{PRED} & \text{‘quilt’} \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix} \end{bmatrix}$$

b.

$$\text{DP} \rightarrow \quad \underset{\substack{\downarrow \in (\uparrow \text{ADJ}) \\ (\downarrow \text{OPER})\backslash\text{DEF}=\uparrow\backslash\text{DEF}\backslash(\text{ADJ} \in \rightarrow=\downarrow) \\ (\downarrow \text{OPER DEF}) =_c \text{REL} \\ (\downarrow \text{WHPATH}) = [+\text{T}]}}{\text{CP}} \qquad \underset{\uparrow=\downarrow}{\text{DP}}$$

The f-structures are almost identical.[9]

OPER serves as an intermediate step in the flow of information between the in-clause and out-of-clause portions of the relative clause construction. It does not turn this into a mediated analysis in the sense defined here: the operator is not a pronominal element which is coreferential with the head. There is no anaphoric mediation, and the relation between the head and the in-clause position remains direct and unmediated. While OPER is being proposed here as a formal convenience, it will transpire (in §5) that it has syntactic consequences.

4.2. Internally-headed relative clauses in Choctaw

IHRCs in Choctaw are discussed by Broadwell (1985a, 1985b). Choctaw IHRCs differ from relative clauses in other languages in two important ways: they do not constitute islands for *wh* extraction, and they are marked by switch reference markers.

(30) a.

Joyce-	at	John-	at	ofi	aaipa	nota-	ma
Joyce-	NOM	John-	NOM	dog	table	under-	ACC
aa-	pisa-	tok-	ma		chopa-	tok.	
LOC-	see-	PST-	DIFF.SUBJ		buy-	PST	

‘Joyce bought the dog John saw under the table.’ or
‘Joyce bought the table John saw the dog under.’

[9]The f-structures and phrase structure rule annotations include the feature [WHPATH], motivated by Falk (2009). The proposal there is that a *wh* path is delimited by values of this feature, with the top f-structure bearing the value [+T] (for *top*) and others bearing the value [−T].

b. Katommah Joyce- at John- at ofi
where Joyce- NOM John- NOM dog
aa- pisa- tok- ma chopa- tok.
LOC- see- PST- DIFF.SUBJ buy- PST
'Where did Joyce buy the dog John saw.' (i.e. where did John see it?)

While this is not the place for a complete analysis of the Choctaw construction, Broadwell's basic conclusion seems well-founded. Broadwell analyzes Choctaw relatives as being clauses rather than nominal phrases. The presence of clausal marking (switch reference) rather than nominal marking (definiteness, case, nominalizing suffix) supports such an analysis. Sentence (30a) would be more accurately rendered:

(31) Joyce bought [John saw the dog under the table].

In more conventionally structured languages, like English, this would not be grammatical. Presumably the syntax-semantics mapping in Choctaw allows such sentences to be interpreted.

If this analysis is correct, Choctaw IHRCs are not *wh* constructions; this could be the reason for their non-islandhood. In Falk (2009), it is proposed that in-situ constructions in some languages are not *wh* constructions. That proposal was based on in-situ questions, but since IHRCs are in-situ constructions as well, it stands to reason that non-*wh* IHRCs also exist. Choctaw thus fills a typological gap.

5. Relative clauses with relative pronouns

We return now to the form of relative clause which the standard mediated analysis takes to be the basic form: the one with a relative pronoun. All the evidence for an unmediated analysis of relatives without relative pronouns points to an unmediated analysis even when a relative pronoun is present.

(32) a. Mary praised the headway which John made.
b. I was shocked by the advantage which she took of her mother.
(33) a. I saw the picture of himself which John liked.
b. Mary discovered the book about himself which Bob wrote.

In other words, contrary to the conventional wisdom, the relative "pronoun" is not a pronominal element which is coreferential with the head and mediates the relation between the head and the in-clause position.

If the relative pronoun is not truly pronominal, the question arises as to why relative pronoun constructions exist. To answer the question, consider the

distribution of relative constructions in English non-finite clauses.

(34) a. a word processor [(*which) to mangle the text]
b. a word processor [(*which) to hate with a passion]
c. a word processor [(*which) to crash the computer with]
d. a word processor [with which to crash the computer]

In infinitival relatives in English, relative pronouns are restricted to "pied-piping" contexts. If there is no pied-piping (relativizing subjects, objects, or even obliques with preposition stranding) the relative pronoun is not allowed.

The pattern that we find in English infinitival relatives can also be found in other languages. For example, the variety of Norwegian described by Taraldsen (1978) has relative clauses with and without relative pronouns.

(35) a. Mannen som / *hvilken ser urolig på oss, er sikkert
the.man REL / *which see uneasily at us is surely
svensk.
Swede
'The man that/who is looking uneasily at us is surely a Swede.'
b. Mannen som / *hvilken du ser der borte, må
the.man REL / *which you see there away must
komme fra Sverige.
come from Sweden
'The man you see over there must come from Sweden.'
c. Mannen som / *hvilken du snakker om, ser nervøs ut.
the.man REL / *which you speak about sees nervous out
'The man that you are talking about looks nervous.'
d. Mannen om hvilken du snakker, ser nervøs ut.
the.man about which you speak sees nervous out
'The man about whom you are talking looks nervous.'

Several of the languages surveyed by Maxwell (1979) display the same distribution: relative pronouns are only used in constructions involving pied piping. There thus appears to be a relationship between relative pronouns and pied-piping.

Wh relatives are licensed by the same phrase structure rule that licenses *wh* questions. Following Falk (2001), this is:

(36) $$\text{CP} \rightarrow \begin{array}{c} \text{XP} \\ (\uparrow \text{DF}) = \downarrow \\ (\uparrow \text{OPER}) = \downarrow \\ (\uparrow \text{OPER PRONTYPE}) =_c \text{WH} \end{array} \quad \begin{array}{c} \text{C}' \\ \uparrow = \downarrow \end{array}$$

Under the analysis proposed in this paper, relative clauses have an OPER (usually unexpressed). Nothing that has been said prevents this operator from being expressed overtly as a *wh* element.[10] If it is, (36) will specify that it has a discourse-related function (topic in the case of relative clauses). Since, unlike interrogative operators, the relative operator is not referential, the lexicon of the language needs to provide lexical *wh* elements that are not referential, but otherwise all the elements are already present to allow *wh* relatives. This results in the following:

(37) a. The trip to Mars which I booked leaves on Tuesday.

b.

$$\begin{bmatrix}
\text{SUBJ} & \begin{bmatrix}
\text{DEF} & + \\
\text{PRED} & \text{'trip}\ \langle(\uparrow \text{OBL}_{\text{Goal}}\ \text{OBJ})\rangle\text{'} \\
\text{OBL}_{\text{Goal}} & \begin{bmatrix} \text{PCASE} & \text{OBL}_{\text{Goal}} \\ \text{OBJ} & [\text{PRED}\ \text{'Mars'}] \end{bmatrix} \\
\text{ADJ} & \left\{ \begin{bmatrix}
\text{WHPATH} & [+\text{T}] \\
\text{OPER} & \begin{bmatrix} \text{DEF} & \text{REL} \\ \text{PRED} & \text{'trip}\ \langle(\uparrow \text{OBL}_{\text{Goal}}\ \text{OBJ})\rangle\text{'} \\ \text{OBL}_{\text{Goal}} & \begin{bmatrix} \text{PCASE} & \text{OBL}_{\text{Goal}} \\ \text{OBJ} & [\text{PRED}\ \text{'Mars'}] \end{bmatrix} \end{bmatrix} \\
\text{TOPIC} & \begin{bmatrix} \text{DEF} & \text{REL} \\ \text{PRED} & \text{'trip}\ \langle(\uparrow \text{OBL}_{\text{Goal}}\ \text{OBJ})\rangle\text{'} \\ \text{OBL}_{\text{Goal}} & \begin{bmatrix} \text{PCASE} & \text{OBL}_{\text{Goal}} \\ \text{OBJ} & [\text{PRED}\ \text{'Mars'}] \end{bmatrix} \end{bmatrix} \\
\text{SUBJ} & [\text{"I"}] \\
\text{PRED} & \text{'book}\ \langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle\text{'} \\
\text{TENSE} & \text{PAST} \\
\text{OBJ} & \begin{bmatrix} \text{DEF} & \text{REL} \\ \text{PRED} & \text{'trip}\ \langle(\uparrow \text{OBL}_{\text{Goal}}\ \text{OBJ})\rangle\text{'} \\ \text{OBL}_{\text{Goal}} & \begin{bmatrix} \text{PCASE} & \text{OBL}_{\text{Goal}} \\ \text{OBJ} & [\text{PRED}\ \text{'Mars'}] \end{bmatrix} \end{bmatrix}
\end{bmatrix} \right\} \end{bmatrix} \\
\text{TENSE} & \text{PRES} \\
\text{PRED} & \text{'leave}\ \langle(\uparrow \text{SUBJ})\rangle\text{'} \\
\text{ADJ} & \{[\text{"on Tuesday"}]\}
\end{bmatrix}$$

Here the relativized element bears an extra grammatical function, the grammaticized discourse function TOPIC. Pied-piping in *wh* questions is standardly licensed by dissociating the discourse function and the OPER

[10]The discussion here assumes that relative pronouns are always *wh* elements, i.e. related to interrogatives. They are obviously not always identical to interrogatives: English *which* as a relative pronoun has different properties than *which* as an interrogative. It has been argued (Vaillette 2000, Falk 2002) that Modern Hebrew has relative pronouns which are identical to anaphoric pronouns. Nothing prevents lexical items that are identical to anaphoric pronouns from having a *wh* feature, although this appears to be unusual. The Hebrew construction may have developed from a resumptive pronoun construction.

function:

(38) $$\text{CP} \rightarrow \quad \underset{\substack{(\uparrow \text{DF}) = \downarrow \\ (\uparrow \text{OPER}) = (\downarrow \text{GF}^{*}) \\ (\uparrow \text{OPER PRONTYPE}) =_{c} \text{WH}}}{\text{XP}} \quad \underset{\uparrow = \downarrow}{\text{C}'}$$

It is the existence of two grammatical functions, the DF and OPER, that allows for pied piping.

It should be noted that under this analysis, the existence of *wh* relatives is a consequence of the operator element proposed for relative clauses. While the operator in non-*wh* relatives is an abstract element with no overt realization, an operator can be included in a fronted *wh* phrase. The operator in relative clauses was originally proposed as a formal necessity, a way of expressing the relation between the head and the in-clause function; the fact that its existence is exploited to allow for relative-pronoun relative clauses is a fortuitous piece of evidence confirming the analysis.

6. Conclusion

It has been argued in this paper that relative clauses should be analyzed without a mediating relative pronoun. Such an analysis accounts for the existence of IHRCs as the in-situ variety of the construction, and for the widespread occurrence of relative clauses with no relative pronoun. The use of a relative pronoun is made available as a result of the formal properties of the construction, but the relative pronoun is not a true pronoun. The purpose of the relative pronoun is to make the construction more flexible by allowing pied-piping constructions. We have also seen that non-*wh* construction relative clauses exist; in particular, Choctaw fills a typological gap by having in-situ relative clauses in which there is no syntactic link with the head of the relative clause construction.

This analysis of relative clauses raises some interesting questions for LFG. One of them relates to the OPER function. It is not clear how OPER fits into the general LFG set of grammatical functions. While it is clearly related to the grammaticized discourse functions, it appears to be somewhat different from them.

Another question relates to the lack of case connectivity in relative clauses. This lack of case connectivity has some interesting parallels in the LFG literature, which includes other constructions with elements which appear to be related through structure-sharing but do not have the same case. For example, Asudeh (2004: 129) argues against a structure-sharing analysis of resumptive pronouns (as proposed by Falk 2002) in part on the grounds of lack of case connectivity. Similarly, Dalrymple and King (2000: 90) argue that the relationship between the SUBJ of a *tough* predicate and the element in the

subordinate clause in a *tough* construction cannot be one of structure sharing because the two elements have different cases. In both of these cases, an anaphoric analysis is proposed instead. In the present situation, this would be unfortunate; as we have seen, the evidence is for an unmediated analysis, not one with a relative pronoun. However, since (unlike Icelandic raising constructions discussed by Andrews 1982) the shared structures are not identical in relative clauses, there is no reason to expect the case to be identical. It can either be explicitly excluded (with the restriction operator), or we can adopt the proposal of Falk (2006). Under this proposal, case is not an f-structure feature value. Instead, every argument-taking predicate projects a grammatical-marking structure (g-structure), where case is realized in the form of an attribute that takes the argument as its value. Since the head and the in-clause element are not identical, they are distinct f-structure elements and would project distinct g-structures. I leave this question open here. I also leave open the question of whether the solution for relative clauses will also work for resumptive pronouns (in relative clauses) and *tough* constructions. In any case, lack of case connectivity is a weaker argument for an anaphoric analysis than it has generally been thought to be.

References

Andrews, Avery D. (1982) "The Representation of Case in Modern Icelandic." in Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 427–503.

Asudeh, Ash (2004) *Resumption As Resource Management*. Ph. D. dissertation, Department of Linguistics, Stanford University.

Bergvall, Victoria L. (1983) "Wh-Questions and Island Constraints in Kikuyu: A Reanalysis." in Jonathan Kaye, Hilda Koopman, Dominique Sportiche, and André Dugas, eds., *Current Approaches to African Linguistics*. Vol. 2. Dordrecht: Foris. 245–60.

Bresnan, Joan (1995) "Linear Order, Syntactic Rank, and Empty Categories: On Weak Crossover." in Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell, III, and Annie Zaenen, eds., *Formal Issues in Lexical-Functional Grammar*. Stanford, Calif.: CSLI Publications. 241–74.

Bresnan, Joan (2001) *Lexical-Functional Syntax*. Oxford: Blackwell.

Broadwell, George Aaron (1985a) "A-bar Anaphora and Relative Clauses." *NELS 16*: 47–56.

Broadwell, George Aaron (1985b) "Internally Headed Relative Clauses in Choctaw." *Kansas Working Papers in Linguistics* 10.2: 16–24.

Chomsky, Noam (1977) "On *Wh*-Movement." in Peter W. Culicover, Thomas Wasow, and Adrian Akmajian, eds., *Formal Syntax*. New York: Academic Press. 71–132.

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

Cole, Peter (1982) *Imbabura Quechua*. Amsterdam: North Holland.
Culy, Christopher D. (1990) *The Syntax and Semantics of Internally Headed Relative Clauses*. Ph. D. dissertation, Department of Linguistics, Stanford University.
Dalrymple, Mary (2001) *Lexical-Functional Grammar*. New York: Academic Press.
Dalrymple, Mary and Tracy Holloway King (2000) "Missing Object Constructions: Lexical and Constructional Variation." in Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG 00 Conference, University of California, Berkeley*. On-line: CSLI Publications. 82–103. http://cslipublications.stanford.edu/LFG/5/lfg00.html.
Falk, Yehuda N. (2001) *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Stanford, Calif.: CSLI Publications.
Falk, Yehuda N. (2002) "Resumptive Pronouns in LFG." in Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG 02 Conference, National Technical University of Athens, Athens*. On-line: CSLI Publications. 154–73. http://cslipublications.stanford.edu/LFG/7/lfg02.html.
Falk, Yehuda N. (2006) "On the Representation of Case and Agreement." in Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG 06 Conference, Universität Konstanz*. On-line: CSLI Publication. 184–201. http://cslipublications.stanford.edu/LFG/11/lfg06.html.
Falk, Yehuda N. (2007) "Do We Wanna (or Hafta) Have Empty Categories?" in Miriam Butt and Tracy Holloway King, eds., *Proceedings of LFG07*. On-line: CSLI Publication. 184–97. http://cslipublications.stanford.edu/LFG/12/lfg07.html.
Falk, Yehuda N. (2009) "Islands: A Mixed Analysis." in Miriam Butt and Tracy Holloway King, eds., *Proceedings of LFG09*. On-line: CSLI Publications. 261–81. http://cslipublications.stanford.edu/LFG/14/lfg09.html.
Gorbet, Larry (1977) "Headless Relatives in the Southwest: Are They Related?" *BLS* 3: 270–8.
Huang, C.-T. James (1982) "Move Wh in a Language Without Wh Movement." *The Linguistic Review* 1: 369–416.
Hulsey, Sarah and Uli Sauerland (2006) "Sorting out Relative Clauses." *Natural Language Semantics* 14: 111–37.
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*. Cambridge, Mass.: MIT Press. 173–281.
Kaplan, Ronald M. and Jürgen Wedekind (1993) "Restriction and Correspondence-Based Translation." *Proceedings of the 6th Conference of the Association for Computational Linguistics European Chapter*. Utrecht University 193–202.

Kayne, Richard S. (1994) *Antisymmetry in Syntax*. Cambridge, Mass.: MIT Press.

Keenan, Edward (1985) "Relative Clauses." in Timothy Shopen, ed., *Language Typology and Syntactic Description*. Vol. 2. Cambridge: Cambridge University Press. 141–70.

Keenan, Edward and Bernard Comrie (1979) "Data on the Noun Phrase Accessibility Hierarchy." *Language* 55: 333–51.

Kenstowicz, Michael and Wafaa Wahba (1983) "Wh-In-Situ Constructions in Egyptian Arabic." in Jonathan Kaye, Hilda Koopman, Dominique Sportiche, and André Dugas, eds., *Current Approaches to African Linguistics*. Vol. 2. Dordrecht: Foris. 261–81.

Maxwell, Daniel N. (1979) "Strategies of Relativization and NP Accessibility." *Language* 55: 352–71.

Platero, Paul R. (1974) "The Navajo Relative Clause." *International Journal of Americal Linguistics* 40: 202–46.

Pollard, Carl and Ivan A. Sag (1994) *Head-Driven Phrase Structure Grammar*. Stanford, Calif.: CSLI Publications.

Schachter, Paul (1973) "Focus and Relativization." *Language* 49: 19–46.

Taraldsen, Knut Tarald (1978) "The Scope of *Wh* Movement in Norwegian." *Linguistic Inquiry* 4: 623–40.

Vaillette, Nathan (2000) "Hebrew Relative Clauses in HPSG." in Dan Flickinger and Andreas Kathol, eds., *Proceedings of the 7th International HPSG Conference, UC Berkeley (22-23 July 2000)*. On-line: CSLI Publications. 305–24. http://cslipublications.stanford.edu/HPSG/1/hpsg00.html.

Vergnaud, Jean-Roger (1974) *French Relative Clauses*. Ph. D. dissertation, MIT.

Williamson, Janis S. (1987) "An Indefiniteness Restriction for Relative Clauses in Lakhota." in Eric J. Reuland and Alice G. B. ter Meulen, eds., *The Representation of (In)definiteness*. Cambridge, Mass: MIT Press. 168–90.

PARTICLE VERBS IN COMPUTATIONAL LFGS: ISSUES FROM ENGLISH, GERMAN, AND HUNGARIAN

Martin Forst†, Tracy Holloway King†, and Tibor Laczkó‡

Microsoft Corp.†, University of Debrecen‡

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

We present the ways in which particle verbs are implemented in two relatively mature computational grammars, the English and the German ParGram LFGs, and we address the issues that arise with respect to particle verbs in the development of a computational LFG for Hungarian. Considerations concerning the ParGram LFG implementation of productive Hungarian particle + verb combinations raise questions as to their treatment in the other two grammars. In addition to providing analyses for English, German, and Hungarian particle verbs, we use these phenomena to highlight how constraints on available lexical resources can affect the choice of analysis and how detailed implementations of related phenomena in typologically different languages can positively guide the analyses in all of the languages.

1 Introduction

In a number of languages, especially Germanic and Finno-Ugric, there are classes of verbs commonly called "particle verbs" (Ackerman, 1983; Piñón, 1992; Lüdeling, 2001; Toivonen, 2001; Booij, 2002).[1] Particle verbs are verbs whose meaning and argument structure depend on the combination of a (base) verb and a particle. Often the meaning and argument structure of a particle verb are not compositional, i.e. it is not predictable from the combination of its components, but it must be listed in the lexicon. An example of a meaning expressed by such a particle verb in English, German, and Hungarian[2] is shown in (1).

(1) a. He **gave up** the fight. (English)

b. Er **gab** den Kampf **auf**.
he gave the fight up
'He gave up the fight.' (German)

c. Ő **fel#adta** a küzdelmet.
he up#gave the fight
'He gave up the fight.' (Hungarian)[3]

[1]We would like to thank the audience of LFG10 for their comments on the abstract and presentation of this paper. Tibor Laczkó gratefully acknowledges that his research reported here has been supported by OTKA (Hungarian Scientific Research Fund), grant number: K 72983, by the Research Group for Theoretical Linguistics of the Hungarian Academy of Sciences at the Universities of Debrecen, Pécs and Szeged, and 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. The authors are listed alphabetically.

[2]In Hungarian, what we refer to here as "particles" are often referred to as "preverbs" in the linguistic literature. Since that term is not adequate for the particles in English phrasal verbs, and for ease of exposition, we use the "particle" terminology.

[3]The '#' sign is not part of the regular orthography. We use it to indicate the boundary between particles and base verb forms when these are spelled as one word, following a convention used in a number of computational language resources.

However, particle verbs can also be compositional, as shown in (2) for some directional particles, and highly productive, which is a challenge for the coverage of computational grammars (Villavicencio, 2003).

(2) a. **Push** them **up/in/out**. (English)
b. **Push up/in/out** the boxes. (English)

In this paper, we present the ways in which particle verbs are implemented[4] in two relatively mature computational grammars, the English and the German ParGram LFGs (Butt et al., 1997, 2002), and we address the issues that arise with respect to particle verbs in the subsequent development of a computational LFG for Hungarian. We will see that considerations concerning the ParGram LFG implementation of productive Hungarian particle + verb combinations raise questions as to the current treatment in the other two grammars, especially as regards the analysis of highly productive particle verbs. An additional interesting phenomenon brought to light by Hungarian is a set of particles which exhibit inflectional properties; we outline an LFG analysis of this phenomenon which is similar to that of incorporated pronouns, e.g. with Welsh prepositions (Sadler, 1999).

Thus, in this paper we provide analyses for English, German, and especially Hungarian particle verbs. We also use these phenomena and our experience creating computational grammars which account for them to highlight how constraints on available lexical resources can, sometimes negatively, affect the choice of analysis and how detailed implementations of related phenomena in typologically different languages can positively guide the analyses and implementations in all of the languages.

2 Particle verbs — syntactic or morphological objects?

English particle verbs are typically analyzed in such a way that the two components are separately inserted in their respective syntactic positions. This is not surprising given that particles are always written as separate words in English and short NPs can intervene between base verbs and particles, as in (3).

(3) a. They **threw** it/the trash **out**. (English)
b. They **cut** them/the onions **up**. (English)
c. They **bandied** it **about**. (English)

In German and Hungarian, however, particle + verb combinations are generally spelled as a single word when the particle immediately precedes the verb. However, variation with respect to the spelling as one or two words can be observed with semantically compositional particle + verb combinations. This particle + verb

[4]Throughout the paper we use standard LFG notation for rules and lexical entries. The implemented grammars use the XLE notation described in Crouch et al. (2010), which is a variant of the LFG notation which uses the ascii character set.

order is in a way the default order, since only clearly definable conditions (V1 and V2 in German; focus, negation, imperatives, etc. in Hungarian) cause particles to appear in positions other than the immediately preverbal one.[5] Even German and Hungarian verbs that do not exist on their own, but only in combination with particles, e.g. *aus#flippen* 'to flip/freak out' (German; **flippen*) and *be#fejez* 'to finish' (Hungarian; **fej-ez*), appear with the particle separate from the verb in these conditions, as in (4) and (5).

(4) a. . . . weil er immer so schnell **aus#flippt**.
. . . because he always so quickly out freaks
'. . . because he is always freaking out so quickly.' (German)

b. Er **flippt** immer so schnell **aus**.
he freaks always so quickly out
'He is always freaking out so quickly.' (German)

(5) a. János **be#fej-ez-te** a könyv-et.
John.NOM PV#head-VSUFF-PAST.3SG.DEF the book-ACC
'John finished the book.' (Hungarian)

b. János nem **fej-ez-te** **be** a könyv-et.
John.NOM not head-VSUFF-PAST.3SG.DEF PV the book-ACC
'John did not finish the book.' (Hungarian)

As a result of this behavior, there is substantial controversy in the linguistic literature concerning the status of particle + verb combinations as syntactic or morphological objects. We will argue for a uniformly syntactic treatment of particles along the lines of Piñón (1992) and É. Kiss (1992, 2005) (for Hungarian), and Lüdeling (2001) (for German) across the LFG implementations for the three languages, and offer analyses that nevertheless capture the lexical properties of particle verbs in a principled manner.

3 Current Implementations in the ParGram LFGs

In this section we present the current analyses used in the broad coverage English (Riezler et al., 2002) and German (Dipper, 2003; Rohrer and Forst, 2006) ParGram grammars. Both grammars aim to capture the often-idiosyncratic meaning of the particle verbs by forming a composite PRED, while allowing for the particle and verb to appear separated in the c-structure, an analysis which is enabled by the LFG projection architecture. However, due to differences in the morphologies of the two languages, in certain constructions, namely when the particle immediately precedes the verb in German, the analyses diverge at the lexical and hence at the c-structure level.

[5] For an overview of the most important instances of this separation in German and Hungarian, see Piñón (1992).

3.1 English

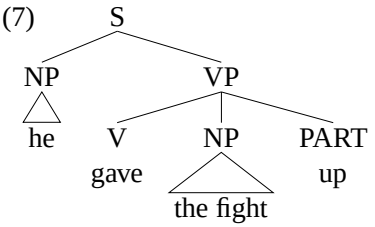
As verb particles are always spelled as separate words in English, particle verbs receive a syntactic analysis in the English ParGram LFG. The lexical entries of verb particles contribute a feature called PRT-FORM, which records the form of the respective particle. The lexical entries of base verbs introduce the semantic form of the particle verb with its argument structure. Finally, the lemma of the base verb and the form of the particle are concatenated via an implementational device (`CONCAT`) so that the combination of the two, rather than just the lemma of the base verb, is the PRED of the f-structure.

All particle verbs are listed with their argument structures in the verb lexicon of the grammar, and they appear under the corresponding base verb, but restricted to co-occurring with the appropriate particle. In (6) are the lexical entries involved in the analysis of the English example sentence *He gave the fight up*.[6] The verb *give* can appear with a SUBJ and OBJ only if there is a PRT-FORM *up*, which will be provided by the particle. The PRED value for the verb is formed by the `CONCAT` template which takes the lexical verb form (%stem = *give*), a hash mark (*#*), and the PRT-FORM (*up*) and concatenates them to create a new form (%NewPred = *give#up*). The resulting c- and f-structures are shown in (7).

(6) give V (↑ PRED) = ′%NewPred<(↑ SUBJ)(↑ OBJ)>′
(↑ PRT-FORM) $=_c$ up
@(CONCAT %stem # (↑ PRT-FORM) %NewPred).

up PART (↑ PRT-FORM) = up.

(7)



$$\begin{bmatrix} \text{PRED} & \text{′}\textbf{give\#up}\text{<SUBJ,OBJ>′} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{′he′} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{′fight′} \\ \text{SPEC} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{′the′} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

This analysis captures the syntactico-semantic facts in that the PRED reflects the potentially idiosyncratic particle verb meaning and the corresponding argument structure. However, it does not allow the system to construct productive particle verbs on the fly, nor does it differentiate between compositional and non-compositional uses of particle verbs.

3.2 German

In German V1 and V2 clauses, particle verbs are spelled as separate words, as shown in (8a). In these contexts, the German ParGram LFG thus treats them in

[6]The actual lexical entry calls a template which expands to (6). For more on templates in XLE and LFG see Crouch et al. (2010), Dalrymple et al. (2004), and Asudeh et al. (2008).

the same way as its English counterpart. In verb-final clauses and in headed VPs, however, particle verbs are usually spelled as single words, as in (9).

(8) a. Er **lud** seine Kusine **ein**.
he loaded his cousin in
'He invited his cousin.' (German)

b. *Er **ein#lud** seine Kusine.
he in#loaded his cousin
'He invited his cousin.' (German)

(9) Er wird seine Kusine **ein#laden**.
he will his cousin in#load
'He will invite his cousin.' (German)

This spelling difference is important because it is reflected in the finite-state morphologies used to process the lexical items (Kaplan et al., 2004). In particular, the DMOR finite-state morphology (Schiller, 1995) currently used by the German ParGram LFG outputs analyses like that in (10) for particle verbs written as a single lexical item.

(10) einlud $\Longleftrightarrow$ ein#laden +V .13 .Sg .Past .Ind

The hash mark indicates the boundary between the particle and the base verb and thus potentially disambiguates analyses involving a separable verb particle from analyses involving homophonous non-separable verb prefixes. However, the entire lemma is a single +V unit: the morphology does not analyze it as a particle followed by a verb.

In contrast, when the particle is separate from the verb, as in (8a), each form is passed separately to the finite-state morphology to be analyzed. The result is shown in (11), where the verb is analyzed as a +V unit and the particle as a +VPRE.

(11) a. lud $\Longleftrightarrow$ laden +V .13 .Sg .Past .Ind
b. ein $\Longleftrightarrow$ ein +VPRE

As a result, the grammar must analyze spelled-together particle verbs as morphological objects, while the spelled-apart ones comprise two morphological objects, like their English counterparts. This has the unfortunate consequence that the lexical information for the German particle verb *ein#laden* 'invite/load in' must be listed both under the base verb lemma (as in the English ParGram LFG), i.e. *laden*, and under the particle verb lemma, i.e. *ein#laden*. Fortunately, the `CONCAT` template makes it possible to project similar f-structures regardless of whether a given particle verb is spelled together or as separate words. So, the f-structure in (12) is the same for the sentences in (8) and (9), modulo the tense marking.

(12) $\begin{bmatrix} \text{PRED} & \text{'}\textbf{ein\#laden}\text{<SUBJ,OBJ>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'er'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'Kusine'} \\ \text{SPEC} & \begin{bmatrix} \text{POSS} & \begin{bmatrix} \text{PRED} & \text{'er'} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TENSE} & \text{past/future} \end{bmatrix}$

The German system exhibits the same limitations as the English one with respect to productively formed combinations and to the inability to differentiate compositional from non-compositional forms. Furthermore, the treatment of particle verb forms as syntactic atoms makes it necessary to use a lexically specified feature that records the fact that a given verb form contains (or does not contain) a particle, as verb forms with particles are disallowed in V1 and V2 position. Since this feature has no semantic relevance and only serves the purpose of ensuring morphosyntactic wellformedness, it is declared as a so-called CHECK feature; its name (and f-structure "path") is (CHECK _VMORPH _PARTICLE). Particle verb forms introduce this feature with the value ATTACHED, while finite verb forms in the Cbar rule are annotated with the following equation:

(13) Cbar → V[v,fin] VP
$\uparrow=\downarrow$ $\uparrow=\downarrow$
$(\downarrow \text{CHECK \_VMORPH \_PARTICLE}) \neq \text{ATTACHED}$

Consequently, verb forms that include a particle are excluded from the V1 and V2 positions. However, this exclusion is obtained via an otherwise unmotivated annotation in the Cbar rule, and it requires that this CHECK feature be introduced in the verb lexicon for each particle verb lemma.

4 A Uniformly Syntactic Analysis

In this section, we argue that a uniformly syntactic analysis as implemented in the English ParGram grammar is desirable in German and Hungarian, too, even though orthography and the semantic opacity of many particle verbs seem to suggest otherwise. We concentrate on why it is desirable to treat particles as separate c-structure nodes regardless of whether they are spelled as separate words or not.[7]

4.1 A Uniformly Syntactic Analysis for German Particle Verbs

In subsection 3.2, we described the current implementation of particle verbs in the German ParGram grammar, arguing that it is uneconomical in terms of lexical

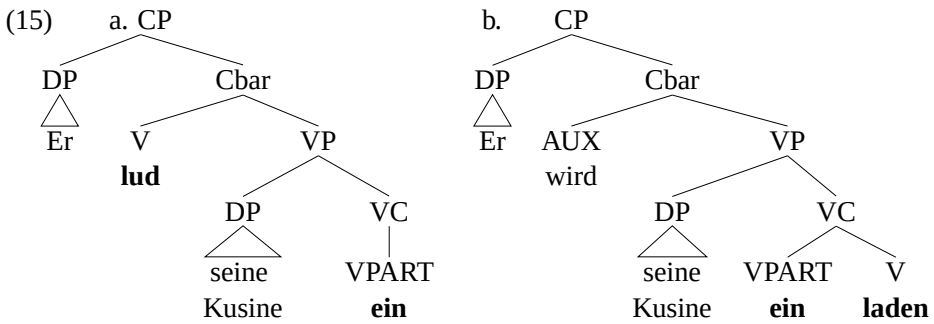
[7] We do not address the question of which category (or categories) the morphemes belong to. Many particles, or particles in general (see Toivonen (2001)), may not belong to a separate "particle" category but instead may be analyzed as adverbs, adjectives, nouns, or intransitive prepositions; here we simply assume a category called "(verb) particle".

entries and requiring a stipulative feature. Nevertheless, it is easy to understand why this analysis was chosen: With a morphology that outputs a single lemma for particle verb forms that are spelled as one word, other analyses are not an option. In order to allow for an implementation in which spelled-apart and spelled-together configurations receive more similar analyses, the output of the finite-state morphology would have to separate the particle from the verb. This is possible to do with finite-state morphologies. For example, SMOR (Schmid et al., 2004) encodes such an analysis; the resulting morphological analysis is shown in (14).[8]

(14) einlud $\Longleftrightarrow$ ein <VPART> laden <+V> <13> <Sg> <Past> <Ind>

With such a morphological analysis, it becomes possible to treat forms like *einlud* as two c-structure nodes rather than one, which in turn makes it possible to consider the particle part of the verb complex (but not of the verb itself). Under this assumption, the c-structures for the sentences in (8) and (9) look as follows:

(15)



While (15a) is basically the c-structure the German ParGram LFG produces for (8a), (15b) crucially differs from the c-structure that the grammar produces for (9) in that *ein* and *laden* are separate c-structure nodes. As a consequence of this, one of the lexical entries for the particle verb, namely the *ein#laden* form, becomes unnecessary; with the new analysis, only the subcategorization information listed under the base verb lemma is needed. Also, the stipulative feature (CHECK_VMORPH_PARTICLE) can be abolished because the distribution of verb particles is controlled by the c-structure rules: there is no slot for a VPART as a daughter of Cbar. Finally, the treatment of verb particles as separate c-structure nodes allows for a straightforward analysis of coordinations like (16).

(16) An der nächsten Haltestelle werden viele Leute [**ein-** und **aus**]#**steigen**.
at the next stop will many people in and out#step.
'At the next stop, there will be a lot of people getting on and off.' (German)

[8]The fact that the morphological tags are surrounded by angled brackets in this morphology instead of preceded by a full stop as in the other is unimportant. The difference in form is merely technical but is included for completeness.

A uniformly syntactic analysis of particle verbs that treats the spelled-apart and spelled-together variants similarly is thus more economical and systematic in as far as the lexicon is concerned, it gives a more parsimonious and less stipulative account of the word order facts observed with respect to particle verbs, and it makes it easy to account for coordinations of verb particles even when the second one is spelled as part of a complex verb.

4.2 A Uniformly Syntactic Analysis for Hungarian Particle Verbs

As with German particle verbs, Hungarian particle verbs are also sometimes written together and sometimes apart. As with the German case, we argue that regardless of which written form is used, the f-structure analysis should be similar. Since the finite-state morphology of the Hungarian ParGram LFG encodes an analysis that splits the particle from the base verb, as in (17), it is possible to have a more uniform c-structure analysis, thereby simplifying the Hungarian lexicon so that verbs are listed only under their stem form.

(17) belép $\Longleftrightarrow$ be +Prefix+ lép +Verb +PresInd +Indef +Sg +3P

The conditions under which Hungarian particles appear together and apart from the verb are discussed in Piñón (1992) and É. Kiss (1992, 2005). The conditions are not important for this paper, but we provide some examples in (18).

(18) a. János **fel#ad-ta** a küzdelm-et.
John.NOM up#give-PAST.3SG.DEF the fight-ACC
'John gave up the fight.'

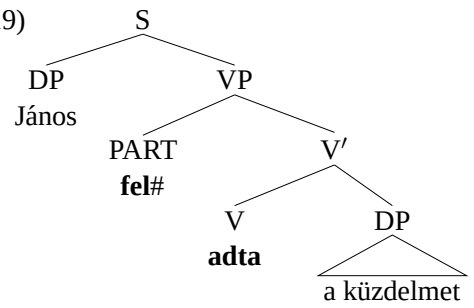
b. JÁNOS **ad-ta** **fel** a küzdelm-et.
John.NOM give-PAST.3SG.DEF up the fight-ACC
'It was JOHN who gave up the fight.'

c. János nem **ad-ta** **fel** a küzdelm-et.
John.NOM not give-PAST.3SG.DEF up the fight-ACC
'John didn't give up the fight.'

d. **Ad-d** **fel** a küzdelm-et!
give-IMPER.2SG up the fight-ACC
'Give up the fight!'

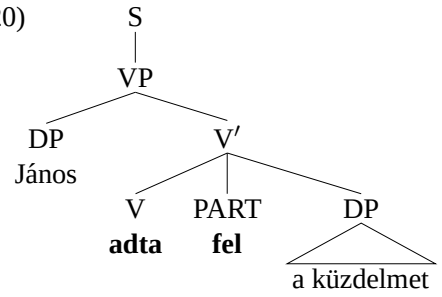
The c-structure and the f-structure for (18a) are given in (19), while those for (18b) are given in (20).

(19)



$$\begin{bmatrix} \text{PRED} & \text{'}\textbf{fel\#ad}\text{<SUBJ,OBJ>'} \\ \text{TOP} & \begin{bmatrix} \text{PRED} & \text{'János'} \end{bmatrix}_1 \\ \text{SUBJ} & [\ ]_1 \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'küzdelem'} \\ \text{DEF} & + \end{bmatrix} \end{bmatrix}$$

(20)



$$\begin{bmatrix} \text{PRED} & \text{'}\textbf{fel\#ad}\text{<SUBJ,OBJ>'} \\ \text{FOC} & \begin{bmatrix} \text{PRED} & \text{'János'} \end{bmatrix}_1 \\ \text{SUBJ} & [\ ]_1 \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'küzdelem'} \\ \text{DEF} & + \end{bmatrix} \end{bmatrix}$$

5 Compositional and Productively Formed Particle Verbs

The analyses described in Section 3 for English and German do not differentiate compositional particle + verb combinations from idiomatic particle verbs. This is a problem for the coverage of computational grammars because new combinations inevitably show up in texts and because the regular character of these combinations is not captured. For example, the particles *by* (English), *hinterher* 'after' (German) and *rá* 'onto' (Hungarian) can combine with a wide variety of verbs (mostly motion verbs), and generally contribute the same meaning. Following Toivonen (2001), we assume that particles in compositional particle + verb combinations can contribute aspectual information, thereby potentially affecting the telicity of the base verb, or that they can have a grammatical function subcategorized for by the base verb. In addition, we will see that certain particles can modify the argument structure of the base verb in a predictable way.

Productive, compositional particles play a number of roles relative to the verb they combine with. They can be adverbial ((21)), resultative ((22)), and aspectual ((23)). English, German, and Hungarian all have particle verbs of these types.

(21) a. She was rescued by Boris Johnson, who was **cycling by**. (English)

b. Einer Frau, die **vorbei#radelt**, johlen einige hinterher.
a.DAT woman who by#cycles hoot some after
'Some hoot after a woman who is cycling by.' (German)

c. Az ütközés után a fiú **tovább#kerékpároz-ott**.
the collision after the boy.NOM further#cycle-PAST.3SG
'After the collision the boy cycled on.' (Hungarian)

(22) a. He **sanded off** the paint from the timber ceiling. (English)

b. Er läßt sich die Rüsselnase **ab#operieren**.
He lets/makes himself the trunk nose off#operate
'He has his trunk-like nose surgically removed.' (German)

c. János **le#smirgliz-te** a festék-et a gerendá-ról.
John.NOM down#sand-PAST.3SG.DEF the paint-ACC the timber-FROM
'John sanded off the paint from the timber.' (Hungarian)

(23) a. She **painted on** for half an hour or so. (English)

b. Jetzt muß in der ganzen DDR **weiter#getrommelt** werden.
Now must in the entire GDR on#drummed be
'Now the drumming has to continue throughout the entire GDR.' (German)

c. János **meg#ír-ta** a level-et egy óra alatt.
John.NOM perf#write-PAST.3SG.DEF the letter-ACC one hour under
'John wrote the letter in one hour.' (Hungarian)

In the remainder of this section, we show how the different types of compositional and productively formed particle verbs can be accounted for in LFG. We demonstrate the technical details of the proposed analyses using mostly German examples, but these accounts naturally and straightforwardly carry over to Hungarian and English.

5.1 Adverbial Particles

Sentence (24) exemplifies a particle + verb combination where the particle adds adverbial information. Adverbials expressing a path tend to directly precede the verb in German; this may be the reason why a number of "path" adverbs that routinely occur in this position are considered particles by prescriptive grammars and hence spelled together with the verb, even though both spelled-together and spelled-apart variants can be found in corpora.

(24) Einer Frau, die **vorbei#radelt**, johlen einige hinterher.
a.DAT woman who by#cycles hoot some after
'Some hoot after a woman who is cycling by.' (German)

The analysis of this category treats the particle as a separate word with its own PRED, which either fills one of the argument positions subcategorized for by the base verb or acts as an ADJUNCT modifier of the verb. Here we do not concern

ourselves with the vexing question of whether the particle becomes an ADJUNCT or an OBL-DIR (see Zaenen and Crouch (2009) on differentiating adjuncts and obliques, especially in implemented grammars), but instead focus on how these analyses can be realized in LFG.

The lexical entries for the adverbial use of the particle *vorbei* 'by' and the lexical entry for the verb *radeln* 'cycle' are shown in (25).

(25) a. vorbei VPART (↑OBL-DIR PRED)='vorbei'
b. radeln V (↑PRED)='radeln<SUBJ, OBL-DIR>'

The VC rule, in which VPART is introduced, then combines these just like it combines verbs with particles that only introduce a PRT-FORM feature for an idiomatic particle + verb combination.[9]

(26) VC ⟶ (VPART) (V)
↑=↓ ↑= ↓

The f-structure this rule and these lexical entries produce is shown in (27).

(27)
$$\begin{bmatrix} \text{PRED} & \text{'Frau'} \\ \text{ADJUNCT} & \left\{ \begin{bmatrix} \text{PRED} & \text{'}\textbf{radeln}\text{<SUBJ, OBL-DIR>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \end{bmatrix} \\ \text{OBL-DIR} & \begin{bmatrix} \text{PRED} & \text{'}\textbf{vorbei}\text{'} \end{bmatrix} \end{bmatrix} \right\} \\ \text{SPEC} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'eine'} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Similarly, the adverbial particle *mit* 'with', illustrated in (28a), adds an ADJUNCT to the base verb. We therefore assume the lexical entries in (29) for the particle *mit* and the verb *fahren* 'go/drive'. These give rise to the analysis in (30).

(28) a. Wer will nach Norwegen **mit#fahren**?
who wants to Norway with#go
'Who wants to go to Norway with us?' (German)

b. Wer will **mit** nach Norwegen **fahren**?
who wants with to Norway go
'Who wants to go to Norway with us?' (German)

(29) a. mit VPART (↑ADJUNCT ∈ PRED)='mit'
b. fahren V (↑PRED)='fahren<SUBJ, OBL-DIR>'

[9] V and PART themselves are nodes with internal structure as dictated by the output of the morphology (see section 4). See Kaplan et al. (2004) on sublexical rules and the interaction of morphology and syntax in LFG.

(30)
$$\begin{bmatrix} \text{PRED} & \text{'wollen}\langle\text{XCOMP}\rangle\text{SUBJ'} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{'pro'}\end{bmatrix}_1 \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{'}\textbf{fahren}\langle\text{SUBJ, OBL-DIR}\rangle\text{'} \\ \text{SUBJ} & [\ ]_1 \\ \text{OBL-DIR} & \begin{bmatrix}\text{PRED} & \text{'nach}\langle\text{OBJ}\rangle\text{'} \\ \text{OBJ} & \begin{bmatrix}\text{PRED} & \text{'Norwegen'}\end{bmatrix}\end{bmatrix} \\ \text{ADJUNCT} & \left\{\begin{bmatrix}\text{PRED} & \text{'}\textbf{mit}\text{'}\end{bmatrix}\right\} \end{bmatrix} \end{bmatrix}$$

This analysis of the particle *mit* also has the advantage of projecting f-structures with identical verbal PREDs and argument structures for sentences (28a) and (28b), which are semantically equivalent. Finally, note that it is a directional PP that can intervene between the adverb/particle *mit* and the clause-final verb; the adverbial particle *mit* and the directional PP seem to compete for the position directly to the left of the verb.

5.2 Resultative Particles

Verb particles often take part in resultative constructions, as in (31).

(31) Er läßt sich die Rüsselnase **ab#operieren**.
He lets/makes himself the trunk nose off#operate
'He has his trunk-like nose surgically removed.' (German)

Resultative particle verbs are probably the category of (semi-)compositional particle + verb combinations that has received the most attention and that has motivated a number of Chomskyan syntacticians to claim that particles head small clauses.

In LFG terms, this corresponds to an XCOMP-PRED analysis, and so we posit the lexical entry in (32a) for the particle *ab* 'off' as used in (31). For the verb *operieren* 'operate', we assume the lexical entry in (32b).[10]

(32)	a.	ab	VPART	(↑XCOMP-PRED PRED)='ab⟨SUBJ⟩'
	b.	operieren	V	(↑PRED)='operieren⟨SUBJ,OBJ,OBJ$_\theta$,XCOMP-PRED⟩'
				(↑XCOMP-PRED SUBJ)=(↑OBJ)

These lexical entries in combination with the regular VC rule shown in (26) associate (31) with the following f-structure.

[10]The resultative subcategorization frame for *operieren* 'operate' is needed for particle and full phrase resultative predicates. The OBJ$_\theta$ in the subcategorization frame is a dative external possessor construction, which is also independent of the occurrence of a particle.

(33) $$\begin{bmatrix} \text{PRED} & \text{'lassen}\langle\text{SUBJ, XCOMP}\rangle\text{OBJ'} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{'pro'}\end{bmatrix}] \\ \text{OBJ} & [\text{ die Rüsselnase }]_1 \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{'}\textbf{operieren}\langle\text{NULL, SUBJ, OBJ}_\theta\text{, XCOMP-PRED}\rangle\text{'} \\ \text{SUBJ} & [\ ]_1 \\ \text{OBJ}_\theta & \begin{bmatrix}\text{PRED} & \text{'pro'}\end{bmatrix} \\ \text{XCOMP-PRED} & \begin{bmatrix}\text{PRED} & \text{'}\textbf{ab}\langle\text{SUBJ}\rangle\text{'} \\ \text{SUBJ} & [\ ]_1\end{bmatrix} \\ \text{PASSIVE} & + \end{bmatrix} \end{bmatrix}$$

This structure relies on a resultative subcategorization frame for *operieren* 'operate' that can be provided by a lexical rule in the spirit of Simpson (2006), for example. However, since the question of the resultative use of verbs needs to be addressed independently of their co-occurrence with particles, we believe that our implementational framework could, in theory, accommodate a purely syntactic treatment as well, for instance, along the lines of Alsina (1996). He proposes that the predicate and the resultative expression bring about their special joint argument structure in the syntax. This is based on his assumption that the argument structure of a predicate may be different from that of the clause it occurs in.

5.3 Aspectual Particles

For aspectual particles, examples of which were shown in (23), Toivonen (2001) suggests introducing aspectual features or a separate PRED for the particle. The lexical entry of an aspectual particle can thus look like (34). (35) shows the f-structure for *She painted on (for days and days)*. Here the aspectual information is contributed by the particle *on* as a simple feature for continuous aspect; see Toivonen (2001) and references therein for analyses of aspect within LFG.

(34) on PART (↑ TNS-ASP ASPECT)=continuous
(↑ PRT-FORM)=on

(35) $$\begin{bmatrix} \text{PRED} & \text{'}\textbf{paint}\langle\text{SUBJ}\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{'she'}\end{bmatrix} \\ \text{TNS-ASP} & \begin{bmatrix}\text{ASPECT} & \textbf{continuous} \\ \text{PRT-FORM} & \textbf{on}\end{bmatrix} \end{bmatrix}$$

5.4 Argument-Changing Particles

As seen above, particles can combine with verbs in a compositional way without altering the argument structure. However, at least in German and Hungarian, there are also productive particles that can add arguments, as pointed out, e.g. by

Stiebels (1996). An example is the German particle *hinterher* ‘after’, which in (36) subcategorizes for the dative argument *dem Glück* ‘the happiness’.[11]

(36) **Lauf** dem Glück nicht länger **hinterher**!
run.IMP.2SG the-DAT happiness not longer after
‘Don’t run after happiness any longer!’ (German)

This type of behavior where the addition of a morpheme, here the particle, results in a predictable modification of the argument structure is reminiscent of causatives and complex predicates. Butt et al. (2003) and Butt and King (2006) use restriction to analyze these constructions and provide a way to implement them in XLE for the ParGram LFG grammars. We propose that certain compositional particle verbs should be analyzed similarly.

The lexical entries for a complex-predicate analysis of (36) are shown in (37).

(37) a. hinterher VPART (↑ PRED)='hinterher<%ARG1,(OBJ$_\theta$)>'
b. laufen V (↑PRED)='laufen<SUBJ>'

The rules that require additional annotations for proper predicate composition are the VC rule, whose “regular” version was presented in (26), as well as the Cbar rule, which introduces the finite verb of a clause when it is in V2 position.

(38) a. VC ⟶ (VPART) (V)
↑=↓ ↑/PRED/OBJ$_\theta$ = ↓/PRED
(↑PRED ARG1) = (↓ PRED)

b. Cbar ⟶ V (VP)
↑/PRED/OBJ$_\theta$ = ↓/PRED ↑=↓
(↑PRED ARG1) = (↓ PRED)

These rules and lexical entries produce the f-structure in (39) for (36). Note that the particle and the verb together make up the PRED of the top f-structure, both contributing to the argument structure of the combination.

(39)
$$\begin{bmatrix} \text{PRED} & \text{'}\textbf{hinterher}\text{<'}\textbf{laufen}\text{<SUBJ>',OBJ}_\theta\text{>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \end{bmatrix} \\ \text{OBJ}_\theta & \begin{bmatrix} \text{PRED} & \text{'Glück'} \\ \text{SPEC} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'die'} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{ADJUNCT} & \left\{ \begin{matrix} \begin{bmatrix} \text{PRED} & \text{'nicht'} \end{bmatrix} \\ \begin{bmatrix} \text{PRED} & \text{'lang'} \end{bmatrix} \end{matrix} \right\} \end{bmatrix}$$

[11]The fact that the sentential negation intervenes between *dem Glück* ‘the happiness’ and *hinterher* ‘after’ makes it impossible to analyze the particle as a postposition. Also compare (24), where *hinterher* is separated from its argument by even more material.

5.5 Corpus Frequency of Compositional Particle Verbs

To quantify the importance of a productive rather than a list-based treatment of compositional particle + verb combinations, we did a small-scale corpus study of German particle verbs. Knowing that the particles *entgegen* 'towards' and *hinterher* 'after' combine with base verbs almost exclusively in a compositional fashion, we extracted all verb forms whose lemma starts with one of these particles from the so-called Huge German Corpus of the University of Stuttgart, and sorted them by frequency. Then we determined which of these particle verbs are in the German ParGram verb lexicon.[12]

The result in terms of tokens is the following: Of the 11,652 verb forms starting with *entgegen*, 11,067 correspond to a lexical entry in the grammar's verb lexicon; 585 or 5.0% of the verb tokens starting with the particle *entgegen* are not covered by the verb lexicon. For the particle *hinterher*, the situation is even clearer. Of the 1,164 verb forms starting with *hinterher*, only 604 correspond to a lexical entry in the verb lexicon, which leaves 542 or 47.3% of the tokens of this type uncovered.

Looking at types reveals that, for both kinds of particle + verb combinations, more than half of the types correspond to a single token, which suggests that these combinations are truly productive. In the case of verb forms starting with *entgegen*, the proportion of hapax legomena (words which occur only once in a given corpus) is 117/223=52.5%; in the case of verbs starting with *hinterher*, it is 78/148=52.7%. Only two of the *entgegen* and one of the *hinterher* hapax legomena are listed in the verb lexicon; overall, 85.7% of the *entgegen* + verb combination types and 91.2% of the *hinterher* + verb combination types are unknown to the verb lexicon.

5.6 Compositional vs. Non-compositional Particle Verbs

To conclude this section, we return to the issue of how best to distinguish compositional from non-compositional particle verbs in an LFG analysis. We propose that compositional particle verbs be analyzed by means of a lexical entry for the particle where, rather than a PRT-FORM feature, it contributes a PRED that, in certain cases, may subcategorize for an argument it introduces. A predicate composition rule involving restriction then fills argument slots of argument-changing particles.

Although we argue that compositional particle verbs are best accounted for using the analyses proposed above, the non-compositional particle verbs should be captured by an analysis similar to that outlined in Section 3. Under this analysis, the verb has a lexical entry which lists the particle it co-occurs with as well as its argument structure. In addition, the predicate is formed by concatenation, i.e. prt#verb, so that a unique PRED is formed, reflecting the non-compositional nature of the particle verb combination. In this way, compositional and non-compositional particle verbs are easily distinguished both in the grammar implementation and in the resulting f-structures.

[12] We only verified whether there was any lexical entry at all for a given particle + verb combination; we did not verify whether the lexical entry was adequate.

(40) **Non-compositional:**

$$\begin{bmatrix} \text{PRED} & \text{'prt\#verb}\langle\text{GFS}\rangle\text{'} \\ \text{GFS} & [\ ] \\ \text{PRT-FORM} & \text{prt} \end{bmatrix}$$

(41) **Compositional: adv. oblique**

$$\begin{bmatrix} \text{PRED} & \text{'verb}\langle\text{GFS,OBL}\rangle\text{'} \\ \text{GFS} & [\ ] \\ \text{OBL} & \begin{bmatrix} \text{PRED} & \text{'prt'} \end{bmatrix} \end{bmatrix}$$

Compositional: adv. adjunct

$$\begin{bmatrix} \text{PRED} & \text{'verb}\langle\text{GFS}\rangle\text{'} \\ \text{GFS} & [\ ] \\ \text{ADJUNCT} & \left\{\begin{bmatrix} \text{PRED} & \text{'prt'} \end{bmatrix}\right\} \end{bmatrix}$$

Compositional: resultative

$$\begin{bmatrix} \text{PRED} & \text{'verb}\langle\text{GFS,XCOMP-PRED}\rangle\text{'} \\ \text{GFS} & [\ ] \\ \text{XCOMP-PRED} & \begin{bmatrix} \text{PRED} & \text{'prt}\langle\text{SUBJ}\rangle\text{'} \end{bmatrix} \end{bmatrix}$$

Compositional: aspectual

$$\begin{bmatrix} \text{PRED} & \text{'verb}\langle\text{GFS}\rangle\text{'} \\ \text{GFS} & [\ ] \\ \text{TNS-ASP} & \begin{bmatrix} \text{ASPECT} & \text{prt} \end{bmatrix} \end{bmatrix}$$

Compositional: predicate composition

$$\begin{bmatrix} \text{PRED} & \text{'prt}\langle\text{'verb}\langle\text{GFS}\rangle\text{',PRTGF}\rangle\text{'} \\ \text{GFS} & [\ ] \\ \text{PRTGF} & [\ ] \end{bmatrix}$$

As a final point, these productive particles may still be semantically restricted so that they cannot occur with verbs with incompatible meanings (e.g. argument structure and aspectual incompatibilities).

6 Hungarian Inflected Particles

In addition to the uninflected particles found in Germanic, Hungarian has inflected particles. Many Hungarian particles are etymologically related to postpositions (*alá* 'to.under') or oblique case suffixes (-*ra* 'onto'). Under normal circumstances, when such elements are used in a pronominal context, the standard Hungarian morphological strategy is to take these elements as 'stems' and to add the 'pronominal content' inflectionally, as in (42).

(42) a. alá 'to.under' b. alá-m 'to.under-1SG'
c. -ra 'onto' d. rá-m 'onto-1SG'

When the argument of the particle verb would be pronominal, Hungarian does not use the particle in its 'neutral' form and express the pronominal oblique argument by a separate inflected element, as shown by the ungrammatical (43).

(43) *János **rá#lép-ett** rá-m.
John.NOM onto#step-PAST.3SG onto-1SG
'John stepped onto me.'

Instead, the ‘pronominally inflected’ particle alone is used to encode this meaning.

(44) János **rá-m** **lép-ett**.
John.NOM onto-1SG step-PAST.3SG
‘John stepped onto me.’

Furthermore, when the oblique argument is 3rd person and non-pronominal, the corresponding inflected particle cannot be used; compare the forms in (45).

(45) a. Mari **rá** **lép-ett** a toll-ak-ra.
Mari onto step.PAST.3SG the pen-PL-SUBL
‘Mari stepped onto the pens.’ (Hungarian)

b. Mari **rá-juk** **lép-ett**.[13]
Mari onto-3PL step.PAST.3SG
‘Mari stepped onto them.’ (Hungarian)

We propose that these inflected particles are straightforwardly analyzed in LFG by the classic pronoun incorporation analysis. All these facts can be captured by assuming that with inflected particles the inflectional morphology obligatorily introduces the PRED=′pro′ feature.[14] An example lexical entry is shown in (46) and the resulting f-structure for (45b) is shown in (47).

(46) rájuk (↑ OBL PRED)=′rá′
(↑ OBL OBJ PRED)=′pro′
(↑ OBL OBJ PERS)=3
(↑ OBL OBJ NUM)=pl

(47)
$$\begin{bmatrix} \text{PRED} & \text{′}\textbf{lép}\text{<SUBJ,OBL>′} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{′Mari′} \end{bmatrix} \\ \text{OBL} & \begin{bmatrix} \text{PRED} & \text{′}\textbf{rá}\text{<OBJ>′} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{′pro′} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \end{bmatrix} \end{bmatrix} \\ \text{TENSE} & \text{past} \end{bmatrix}$$

The fact that the PRED ′pro′ is obligatory in inflected particles, instead of optional as is the case in many analyses of pro-drop, ensures that the inflected particle cannot be doubled by an overt argument. This is shown by the ungrammaticality of (48) where the inflected particle *rá-juk* is doubled by an overt oblique NP.

(48) *Mari **rá-juk** **lépett** a toll-ak-ra.
Mari onto-3PL step.PAST.3SG the pen-PL-SUBL
‘Mari stepped onto them the pens.’ (Hungarian)

[13]Note that conventional Hungarian orthography has the uninflected particle + verb combination as one word and the inflected particle + verb combination typically as two words.

[14]For an inventory of Hungarian inflecting particles and a lexicalist analysis, see Ackerman (1990).

7 Conclusions

We established that the implemented LFG analysis of particle verbs for English and German is appropriate and feasible for non-compositional particle constructions. However, the existing analyses had two problems. First, the analysis did not distinguish between idiomatic and compositional particle-verb constructions: both types were listed in the lexicon and created PREDs composed of the particle and the verb. Second, the German analysis required two different c-structure analyses and, as a result, repetition in the lexicon.

We proposed that idiomatic particle-verb constructions be listed in the lexicon and have PRED values which are composed of the particle and the verb. These idiomatic particle verbs may have argument structures which differ significantly from the verb's non-particle counterpart. This analysis is similar to that usually assumed in the LFG literature and implemented in XLE grammars.

We then argued that compositional particle verbs be composed in the syntax. In many cases, the particle fills an argument slot of the base verb, e.g. a resultative XCOMP-PRED or an adverbial OBL, it modifies the base verb by functioning as an adjunct, or it adds aspectual information. In other cases, the particle introduces an additional argument, and restriction is used to create a new PRED for the verb which differs from the original PRED only in the addition of the new argument. In both types, the additional f-structure information is provided by the lexical entry of the particle. Note that although the compositional particle-verb construction can be very productive, there are semantic constraints on the allowable combinations; we leave the investigation of these constraints for future work.

The issue with the repetition of lexical items in the German lexicon was solved by incorporating a different morphological analysis whereby, even when particle-verb combinations are written as a single word, they comprise two different tokens in the c-structure. That is, their sublexical analysis involves the particle, the verb, and the morphemes encoding inflectional information. The c-structure rule for verbs is then minimally modified to allow for this construction.

Hungarian particle verbs can be straightforwardly accounted for with this division between idiosyncratic and compositional particle verbs, as well as a morphological analysis similar to the one proposed for German. In addition, Hungarian has inflected particles. We argued that LFG, and its implementation via XLE, allows for a straightforward pro-drop style analysis of Hungarian inflected particles.

An orthogonal issue to those addressed here is that of how particle verbs participate in derivational morphology and how best to implement this (e.g. English *bystanders*, German *Einladung* 'invitation'). An additional derivational morphology issue involving particle verbs is discussed in Booij (2002): he provides a construction grammar analysis of particles used to create verbs out of nouns and adjectives in Dutch and German. We leave these areas for future work.

References

Ackerman, Farrell. 1983. *Miscreant Morphemes: Phrasal Predicates in Ugric*. Ph. D. thesis, UC Berkeley.

Ackerman, Farrell. 1990. The Morphological Blocking Principle and Oblique Pronominal Incorporation in Hungarian. In Katarzyna Dziwirek, Patrick Farrell and Errapel Meijas-Bikandi (eds.), *Grammarical Relations: A Cross-Theoretical Perspective*, pages 1–19, CSLI Publications.

Alsina, Àlex. 1996. Resultatives: A Joint Operation of Semantic and Syntactic Structures. In *Proceedings of the First LFG Conference*, Grenoble, France: Rank Xerox Research Centre.

Asudeh, Ash, Dalrymple, Mary and Toivonen, Ida. 2008. Constructions with Lexical Integrity: Templates as the Lexicon-Syntax Interface. In *Proceedings of LFG08*, pages 68–88, CSLI On-Line Publications.

Booij, Geert. 2002. Separable Complex Verbs in Dutch: A Case of Periphrastic Word Formation. In *Verb-Particle Explorations*, pages 21–42, Berlin, Germany: Mouton de Gruyter.

Butt, Miriam, Dyvik, Helge, King, Tracy Holloway, Masuichi, Hiroshi and Rohrer, Christian. 2002. The Parallel Grammar Project. In *Proceedings of the COLING 2002 Workshop on Grammar Engineering and Evaluation*.

Butt, Miriam and King, Tracy Holloway. 2006. Restriction for Morphological Valency Alternations: The Urdu Causative. In *Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan*, pages 235–258, CSLI Publications.

Butt, Miriam, King, Tracy Holloway and Maxwell, III, John T. 2003. Complex Predication via Restriction. In *Proceedings of the LFG 2003 Conference*, pages 92–104, CSLI On-line Publications.

Butt, Miriam, King, Tracy Holloway, Niño, María-Eugenia and Segond, Frédérique. 1997. *A Grammar-Writer's Cookbook*. CSLI Publications.

Crouch, Dick, Dalrymple, Mary, Kaplan, Ron, King, Tracy, Maxwell, John and Newman, Paula. 2010. XLE Documentation. Technical Report, Palo Alto Research Center, Palo Alto, CA, www2.parc.com/istl/groups/nltt/xle/doc/xle_toc.html.

Dalrymple, Mary, Kaplan, Ronald M. and King, Tracy Holloway. 2004. Linguistic Generalizations over Descriptions. In *Proceedings of LFG05*, pages 199–208, CSLI On-Line Publications.

Dipper, Stefanie. 2003. *Implementing and Documenting Large-scale Grammars – German LFG*. Ph. D. thesis, IMS, University of Stuttgart, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), Volume 9, Number 1.

É. Kiss, Katalin. 1992. Az egyszerű mondat szerkezete [The structure of the simple sentence]. In Ferenc Kiefer (ed.), *Strukturális Magyar Nyelvtan 1. Mondattan [Structural Hungarian Grammar 1. Syntax]*, pages 79–177, Akadémiai Kiadó.

É. Kiss, Katalin. 2005. First Steps towards a Theory of the Verbal Particle. In Christopher Piñón and Péter Siptár (eds.), *Papers from the Düsseldorf Conference. Approaches to Hungarian 9*, pages 57–88, Akadémiai Kiadó.

Kaplan, Ronald M., Maxwell, III, John T. Maxwell, King, Tracy Holloway and Crouch, Dick. 2004. Integrating Finite-state Technology with Deep LFG Grammars. In *Proceedings of the Workshop on Combining Shallow and Deep Processing for NLP (ESSLLI)*, pages 11–20.

Lüdeling, Anke. 2001. *On Particle Verbs and Similar Constructions in German*. CSLI Publications.

Piñón, Chris. 1992. The Preverb Problem in German and Hungarian. In *Proceedings of BLS*, pages 395–408.

Riezler, Stefan, King, Tracy Holloway, Kaplan, Ronald, Crouch, Dick, Maxwell III, John T. and Johnson, Mark. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the ACL*.

Rohrer, Christian and Forst, Martin. 2006. Improving Coverage and Parsing Quality of a Large-scale LFG for German. In *Proceedings of the Language Resources and Evaluation Conference (LREC-2006)*, Genoa, Italy.

Sadler, Louisa. 1999. Non-Distributive Features in Welsh Coordination. In *Proceedings of LFG99*, CSLI On-Line Publications.

Schiller, Anne. 1995. DMOR Benutzerhandbuch. Technical Report, IMS, University of Stuttgart.

Schmid, Helmut, Fitschen, Arne and Heid, Ulrich. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Simpson, Jane. 2006. Resultatives. In *Lexical Semantics in LFG*, Stanford, CA: CSLI Publications.

Stiebels, Barbara. 1996. *Lexikalische Argumente und Adjunkte: Zum semantischen Beitrag von verbalen Präfixen und Partikeln*. Berlin, Germany: Akademie Verlag.

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

Villavicencio, Aline. 2003. Verb-Particle Constructions and Lexical Resources. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

Zaenen, Annie and Crouch, Dick. 2009. OBLs Hobble Computations. In *Proceedings of LFG09*, CSLI On-line Publications.

MULTIPLE QUESTIONS IN FRENCH AND IN HUNGARIAN AN LFG ACCOUNT

Anna Gazdik
Université Paris 7 / ELTE Budapest

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

In this paper we propose an LFG account of multiple questions in French and in Hungarian. It will be argued that (together with clausal coordination) six main types of multiple questions can be identified in the two languages, which can be associated with different readings. Multiple questions in which both question words are in the same clause can be ambiguous between a pair list and a single pair reading, whereas in the case of clausal coordination, which is analyzed as an elliptical structure, only the single pair reading is available. We identify some problems, like that of two preverbal question words in Hungarian, or the ambiguity concerning D-linkedness in French. We then propose an LFG analysis in which information structure is projected at a different level of analysis, containing semantic information as well, and in which the different question words can belong to different sets, *TOPIC* and *FOCUS*, respectively.

1 Introduction

In this paper I discuss multiple questions in French and in Hungarian. Multiple questions have received much attention in the literature, mostly in transformational frameworks. The main challenge was to account for the different types occurring in languages, such as multiple fronting in Bulgarian or Romanian, single fronting in English, or *in situ wh*-questions in Chinese. The aim of the present paper is twofold. First, it concentrates on two typologically different languages and attempts to provide a coherent analysis for both. Second, it will argue that an analysis should rely on both syntactic and discourse information: this is why an account in the framework of Lexical-Functional Grammar will be proposed. In section 2, we introduce some generalities about multiple questions, such as pair-list *vs.* single pair readings and *D-linkedness*. In the next section, we present the data, associating the different syntactic structures with possible interpretations. In section 4, we introduce the LFG approach to information structure, which proves to be crucial in the analysis, then we go on to the proposed analysis, which we illustrate with examples in the subsequent section.

2 Generalities

In a multiple question we find more than one information gap in a sentence. In syntactic terms, it means that a sentence contains more than one question word, like in the following French example:

(1) *Qui a dit quoi ? [French]*
who has said what
Who said what?

In the example, the question refers to the subject and the object at the same time.

It is commonly accepted that a multiple question can be answered in two ways. Some of them license a pair-list, others a single-pair answer, and some are ambiguous between the two. To decide which answer is appropriate, in certain cases only the context provides the clue, but some languages express this difference explicitly in syntax. Let us examine such examples from French (2)-(3).

(2) Q:
Qui est parti quand ?
who is left when
Who left when?

[†] I would like to thank Anne Abeillé and András Komlósy for their valuable comments on earlier versions of this paper, as well as my fellow students at the Université Paris 7 for the discussion of the data. The usual disclaimers apply.

A: John left in the morning, Mary in the afternoon, and Jane in the evening.

(3) Q:
Qui est parti et quand ?
who is left and when
Who left and when?

A: John left in the morning.

In French, the example in (2) asks for a pair-list answer, whereas the equivalent (3) with coordination is more naturally answered with a single pair. In other languages (like Romanian), it is possible to accumulate *wh*-words in sentence-initial-position, which license a pair-list answer, whereas the coordination of *wh*-words licenses a single-pair answer. As we will see later, the pair-list interpretation can be analyzed as a function applying between the two *wh*-words (Krifka (2001)), in which each element (or partition) of the set denoted by the first is paired up with one element denoted by the other *wh*-word.

An important factor about pair-list questions is the observation that one of the question words denotes a contextually given set, which the locutor and the interlocutor can partition in the same way (Comorovski (1996)), and the range of felicitous answers is limited by this set (Pesetsky (1987)).[1] Pesetsky (1987) and Comorovski (1996) call this set *D(iscourse)-linked*, referring to the fact that the set has already been introduced into the discourse.[2] Ginzburg and Sag (2000) argue, on the other hand, that neither (none) of the sets has to be contextually determined, like in (4).

(4) Which recently published reports should be made required reading for which government departments? (Ginzburg and Sag (2000), p. 248)

According to the authors, any new public official can ask (4), without knowing a defined set of reports. We believe, on the other hand, that without the context the issue is very difficult to judge. However, the modifier *recently published* makes it possible to establish a context and to restrict the possible reports to those published recently, which is, in a way, a contextually determined set, even if the new public official cannot list them by title. Note that without the above mentioned modifier the sentence is less acceptable, if we suppose that the locutor does not know any of the reports (5):

(5) #Which reports should be made required reading for which government departments?

In this article we will therefore keep the term *D-linked* to refer to *wh*-words denoting sets that are contextually determined or salient in the discourse. Just like the type of answer expected to the question, D-linkedness can also determine syntactic structure. In some languages, D-linked question words tend to precede non-D-linked ones, and similarly, the constituents corresponding to the D-linked question word precede their non-D-linked counterparts in the answer. Let us consider the following Hungarian examples (6)-(7):

(6) Q:
Ki mit hozott a buli-ra?
who what brought the party-to
Who brought what to the party?

A:

[1]This means that, for example, in order to give a pair-list answer to a multiple question referring to *the students*, the interlocutor does not have to know them individually, but they can list them by their nationalities.

[2]Pesetsky (1987) claims, in a transformational framework, that D-linked *wh*-phrases do not have to move to Comp at LF, but can take scope, whereas non-D-linked ones must move, since the former are no quantifiers, but the latter are.

János bort, Mari pedig sütiket hozott.
John wine.ACC Mary as for cookies brought
John brought wine, and Mary cookies.

(7) Q:
Mit ki hozott a buli-ra?
what who brought the party-to
*What did who bring to the party?

A:
A bort János, a sütiket pedig Mari hozta.
the wine.ACC John the cookies as for Mary brought
The wine was brought by John, the cookies by Mary.

In example (6), the question asks for the enumeration of all the people present at the party and then about the thing each of them brought. In (7), on the other hand, all the things brought to the party are listed and then identified with a person. In the English translation, apart from the passive, the definite article in the answer to (7) also expresses this change of perspective. This function of the D-linked *wh*-phrase is referred to by Kuno and Takami (1993), as the *Sorting Key Hypothesis*, where the linearly first question word determines the organization of the information in the answer.

In the next section, our aim will be to investigate how the pair-list/single-pair difference is expressed in the two languages. As we will see, some syntactic structures are ambiguous between the two readings, whereas others have clearly only one possible interpretation. Then, we will also examine the problem of how the D-linked question word can be identified in French.

3 The Data

3.1 Possible syntactic structures

Based on the syntactic structure, we can identify five main types of multiple questions in the two languages. A sixth structure also exists, but it is argued to contain clausal coordination and thus does not qualify as a true multiple question.

3.1.1 All question words extracted (only in Hungarian)

As we have seen above, all question words can appear preverbally in Hungarian, both in main and in embedded clauses: [3]

(8) *Ki hogyan oldotta meg az előző nap eseményei-nek tálalását?*
who how solved VM the preceding day events-POSS presentation
Who could find which way of presenting the events of the preceding day?

(9) *Cikksorozat-unk-ból megtudhatják végre, mi mi-t jelent.*
series of articles-POSS-from learn-can.3PL at last, what what-ACC means
From our series of articles you can learn, at last, what means what.

[3] The examples come from the Hungarian National Corpus

(10) *A szerző pontosan megfogalmazta, ki mikor megy át a szín-en (és mit csinál).*
the author precisely said, who when go across the scene-on (and what does)
The author gave precise instructions about who should cross the scene when (and what they should do).

These questions all license a pair-list reading, determined by the structure of the question. This means, that the answer to (8) pairs up a set of people with ways of presenting, the answer to (9) a set of things with things, and the answer to (10) a set of people with time slots.

3.1.2 One question word extracted, the other(s) *in situ*

This structure is possible in both languages. Let us consider Hungarian first:

(11) *Ki hívott meg kit a bulira?*
who invited PRT whom the party-to
Who invited whom to the party?

This type of question, contrary to the previous one, usually expects a single-pair answer. We should note, however, that acceptability judgements show great variation and some speakers would even answer such questions with a pair-list. Interestingly, for some speakers, this structure is grammatical, only if the question words denote the same type of set:

(12) *?? Ki mondott mit?*
who said what
Who said what?

(13) *Honnan utazol hova?*
from where travel.2SG where
From where to where are you travelling?

(14) *Melyik lány táncolt melyik fiúval?*
which girl danced which boy-with
Which girl danced with which boy?

According to a plausible explication Kálmán (2001), the expected answer to such a question can be a direction, or, more precisely, the direction of the relation between the denotations of the two question words and not the denotations themselves. In other words, this means that the question contains two (or more) forms of the interrogative word ki (who), mi (what) or melyik (which), or some locative question word *hol, honnan, hová* (where, from where, to where).

In French, for this type the pair-list is the preferred reading, but the single pair answer is not excluded either. We can find examples for both, especially if both question words are arguments. Let us consider the following examples:

- Arguments

  (15) *Qu'est-ce que tu as donné à qui ?*
  what is it that you have given to whom
  What did you give to whom?
  The question is compatible with both answers:
  A1 (single-pair):

J'ai donné un livre à Marie.
I have given a book to Mary.
I gave a book to Mary.

A2 (pair-list):
J'ai donné un livre à Marie, et un stylo à Paul...
I have given a book to Mary, and a pen to Paul...
I gave a book to Mary, and a pen to Paul.

- Argument and adjunct

(16) Q:
En famille, qui dort où ? (google)
in family, who sleeps where
In the family (home) who sleeps where?

A (pair-list):
Les parents dorment au premier étage et les enfants au deuxième.
the parents sleep on the first floor and the children on the second
The parents sleep on the first floor and the children on the second.

(17) Q:
Quand est-ce que tu as vu qui ?
when is-it that you have seen whom
When did you see whom?

A (single-pair):
J'ai vu Jean ce matin.
I have seen John this morning
I saw John this morning.

- Adjuncts

(18) Q:
Où Jean a dormi quel jour ?
where John has slept which day
Where did John sleep on which day?

A (pair-list):
Lundi, Jean a dormi à Rome, mardi à Nice, mercredi à Cannes.
Monday John has slept in Rome, Tuesday in Nice, Wednesday in Cannes
On Monday, John slept in Rome, on Tuesday in Nice, on Wednesday in Cannes.

(19) Q:
Quand est-ce qu'il est arrivé avec quelle intention?
when is it that he is arrived with what kind of intention
When did he arrive and what was his intention?

A (single pair):
Il est arrivé lundi pour rencontrer Jean.
he is arrived Monday to meet John
He arrived on Monday to meet John.

However, in some examples, one of the readings is excluded on pragmatic grounds. For instance, in the case of unique events, only the single-pair reading is available:

(20) **Qui a tué Henri IV quand ?*
who has killed Henri IV when
*Who killed Henri IV when?

The ungrammaticality of this example shows that the preferred reading, especially in cases where one of the question words is an argument and the other is an adjunct, is the pair-list. The tendency can be explained by the fact that there is another structure available (clausal coordination), which, in turn, licenses only single pair answers.

3.1.3 All question words *in situ* (only in French)

Similarly to the previous type, both readings are available in these questions. This type is, as we have shown above, the informal equivalent of the second structure. Let us have a look at the following examples:

(21) Q:
Tu vas chercher qui à quelle heure ?
you go pick up whom at which hour
Whom are you going to pick up when?

A (pair-list):
Je vais chercher Max à 17 heures et Léa à 18 heures.
I go pick up Max at 17 hours and Léa at 18 hours
I'm going to pick up Max at 5 pm and Léa at 6 pm.

(22) Q:
Tu es allé où avec quelle intention?
you are gone where with what kind of intention
Where did you go (and) what was your intention with it?

A (single-pair):
Je suis allé chez Jean pour lui dire la vérité.
I am gone to John to him tell the truth
I went to John to tell him the truth.

3.1.4 Constituent coordination in initial position

In Hungarian, almost all kinds of question words can be coordinated. This structure typically triggers a single-pair answer:

(23) *Ki és mikor ment moziba?*
who and when went cinema-to
Who went to the cinema and when?

This structure seems to complement the second type in that it is not felicitous if the two question words denote the same type of set:

(24) **Ki és ki-be szeretett bele?*
who and who-into loved PRT
*Who and with who fell in love?

However, in some contexts, the pair-list reading is also acceptable:

(25) *Korábban csak a tejesembertől vagy a postástól lehetett információkat*
before only the milkman-from or the postman-from could information
megtudni arról, hogy éppen ki és mikor nyaral, vagy általában
get about, that at the moment who and when on holiday, or usually
melyik napszakban nem tartózkodik otthon, manapság viszont a
which part of the day not be at home, nowadays on the other hand the
betörők internetes kutatómunká-val készülnek fel a kiszemelt házak, lakások
burglars internet research work-with prepare VM the chosen houses, flats
kifosztására.
burglary.POSS
Before one could get information only from the milkman or the postman about who went on holiday and when, or usually who is not at home in which part of the day. Nowadays, on the other hand, burglars prepare for the burglary of the chosen houses and flats with some research on the internet.

In French, only question words with the same grammatical function (for instance: subject, object, adjunct) can be coordinated:

(26) (adjuncts)

Quand et pourquoi est-il parti ?
when and why is he left
When and why did he leave?

(27) (different arguments)

**Qui et quoi fait ?*
who and what do
*Who and what is doing?

(28) (argument and adjunct)

?? Qui et quand est parti ?
who and when is left
*Who and when left?

The coordination of two arguments with the same function, on the other hand, cannot always be considered as a multiple question, since it does not contain two information gaps in the sentence. Its function is rather the enlargement of the domain set of the question, for instance, to cancel the restriction imposed by *qui* (who) that the answer has to be animate.

(29) *Qui ou quoi a favorisé l'allaitement maternel ?*
who or what has approved the breastfeeding maternal
Who or what approved of maternal breastfeeding?

In other cases, however, it is indeed a multiple question:

(30) *Quand et où a eu lieu le concert ?*
when and where has had place the concert
When and where did the concert take place?

(31) *Quel âge et quel grade a M Martin?*
which age and which rank has M Martin
How old is Mr. Martin and what is his rank?

Concerning the interpretation, similarly to Hungarian, the single-pair reading is preferred, but the pair-list reading is also possible, depending on the context (single or general event):

(32) (single-pair reading)

Pourquoi et quand avez-vous décidé d'arrêter vos études universitaires?
why and when have-you decided to stop your studies university
Why and when did you decide to stop your academic studies? (Rochefort, Christiane (1978) Ma vie revue et corrigée par l'auteur à partir d'entretiens avec Maurice Chavardès. Stock: Paris. p. 308.)

(33) (pair-list reading)

Quand et pourquoi voit-on circuler des trains sans voyageurs? (SNCF)
when and why see-we circulate ART trains without passengers
When and why can we see trains without passengers?
(http://www.infolignes.com/article.php3?id_article=3505)

The possibility of the pair-list reading suggests that the question words are in the same clause, and that the structure contains constituent and not clausal coordination with ellipsis in the first conjunct. As we will see, the pair list reading is not available in the sixth type, which, in turn, will be considered as clausal coordination.

Interestingly, at first sight, it is also possible in French to coordinate *wh*-words with different functions, if the verb is in the infinitive. Consider the following example:

(34) *Qui et quand consulter quand un couple rencontre des problèmes pour avoir un enfant ?*
whom and when to consult when a couple meets problems for have a child
Whom should we consult and when, in case a couple has problems having a child?

According to Anne Abeillé (p.c.), however, this is only possible if the verb has an optional argument (in both examples, the object is optional) and thus the sentences are elliptical structures with right-node-raising of the infinitive.

3.1.5 Constituent coordination *in situ*

In spoken French, the coordinated interrogative constituent can also appear *in situ*.

(35) *La conférence a eu lieu où et quand ?*
the conference has had place where and when
When and where did the conference take place?

In case if the second/sentence-final question word is an adjunct, this type is difficult to distinguish from clausal coordination.

3.1.6 Clausal coordination

In this sixth type of multiple questions, the structure consists of two coordinated clauses with ellipsis in the second clause. In this case, only the single pair reading is available in both languages, which we can consider as an argument for the biclausal analysis. We will see see other arguments supporting this view. Le us have a look at some examples:

(36) *À qui as-tu parlé et pourquoi ? [French]*
to whom have you spoken and why
To whom did you speak and why?

(37) *Ti persze biztosan nagy tudorok vagytok, de pontosan mit hazudott volna*
you of course certainly big scientists are, but precisely what lied AUX
és miért? [Hungarian]
and why
You certainly know everything, but what precisely would he have said when he lied and why did he lie?

The arguments supporting the biclausal analysis are the following:

First of all, the second part of the sentence can presuppose that the first question is already answered and can thus refer only to one of the possible answers to the question, like in the following invitations to a party:

(38) *Dites-nous si vous venez et à combien, pour qu'on puisse*
tell us if you come and PREP how many, in order that we could us organize
s'organiser en fonction.
in function
Tell us if you come, and if so, how many of you, so that we can take it into consideration during the organization.

(39) *Léci, léci, jelezzen, aki még nem tette, hogy jön-e és hányan!!!*
please please write, who yet not did, that comes-CL and how many
Please please, tell me if you come and if so, how many of you!

In this case, in order to answer the question *à combien* or *hányan* (how many), it has to be accommodated Ginzburg (1997) that the person invited is actually coming to the party.

Another argument for the biclausal analysis is the fact that when both question words are arguments, the second one must be optional. This is possible only in the case of verbs that have both a transitive and an intransitive use:

(40) *Qui va parler et de quoi ?*
who FUT talk and about what
Who is going to give a talk and about what?

In (40) *parler* (talk) cannot appear in its transitive use. In Hungarian, with these verbs only the transitive reading is possible when the interrogative words are in a preverbal position and one of them refers to the object:

(41) *Mit és miért olvasott?*
what and why read.PAST
What did s/he read and why did s/he read it?

(42) *Miért és mit olvasott?*
why and what read.PAST
What did s/he read and why did s/he read it?

On the other hand, in the structure that contains sentence-final coordination, sometimes two readings are possible depending on the verb (Gracanin, 2007): the first is called the *it-reading*, referring to the fact that the verb is understood as transitive in both clauses, like in the following example:

(43) *Mit olvasott és miért?*
what read.PAST and why
What did s/he read and why did s/he read it?

However, in the second case, two readings are possible: the first is the above mentioned *it-reading*, and the second is the *at all reading*.

(44) *Miért olvasott és mit?*
why read.PAST and what
Why did s/he read (what s/he was reading) and what did s/he read?/Why did s/he read at all and what did s/he read?

The *at all* reading is only possible if the transitive and the optionally transitive use of the same verb can be coordinated in the same sentence, which means that the above structure contains clausal coordination.

The examples with obligatory arguments are ungrammatical:

(45) **Qui va faire et quoi ?*
who FUT do and what
*Who is going to do and what?

However, when both interrogative words are obligatory arguments, the preverbal coordination is grammatical.

The structure is perfectly possible with an argument and an adjunct (46) or with two adjuncts (50), if the argument is in the first clause:

(46) *Qui a encore décroché le récepteur et pourquoi ? (J. Genet) [French]*
who has again picked up the receiver and why
Who has again picked up the receiver and why?

(47) *Pourquoi voulait-il l'aider et comment ? [French]*
why wanted he him help and how
Why did he want to help him and how?

(48) *Ki járt itt és mikor? [Hungarian]*
who came here and when
Who was here and when?

(49) **Mikor járt itt és ki? [Hungarian]*
when came here and who
*When was here and who?

From this it follows that examples (36)-(37) are indeed the coordination of two clauses. If it is an adjunct that is coordinated sentence-finally, or both are in the preverbal field, the sentence is grammatical. This supports the view that in this example constituents and not clauses are coordinated. ((48) and (49) are grammatical with preverbal coordination.)

Concerning Hungarian, Lipták (2001) points out that the conjugation type of the finite verb differs from what we would expect based on the supposed non-elliptical counterpart of the sentence:

(50) *Nem érdekel, hogy mit készítesz és hogyan készíted.*
not interests, that what make.2SG.INDEF and how make.2SG.DEF
I am not interested in what you make and how (you make it).

(51) *Nem érdekel, hogy [mit és hogyan] készítesz.*
not interests, that [what and how] make.2SG.IND
I am not interested in what you make and how (you make it).

She argues that the structure cannot be elliptical, since then the verb should be definite. However, Bánréti (2007) shows that the rule of ellipsis in Hungarian is not so strict with respect to the agreement with the definiteness of the object, as it is in the case of tense and mood endings, i.e. even if the overt verb and the one falling under ellipsis are of different conjugations, the sentence can be grammatical. These are the reasons why we analyze structure 4 and 5 as constituent and structure 6 as sentential coordination. In what follows we will not deal with sentential coordination, since they do not qualify as true multiple questions according to the definition we presented above. After considering the data, we identify some problems that have to be accounted for by the analysis proposed.

3.2 Problems

3.2.1 Two preverbal question words in Hungarian

In Hungarian, only one preverbal focus is permitted:

(52) **JÁNOS TEGNAP ESTE ment moziba.*
János yesterday evening went cinema-to
JOHN went to the cinema YESTERDAY EVENING.

If both *JÁNOS* and *TEGNAP ESTE* are focussed, one of them (in this case *TEGNAP ESTE*, since *JÁNOS* is the subject), has to appear in a postverbal position (we leave it now open if this position is *in situ* or sentence-final):

(53) *JÁNOS ment moziba TEGNAP ESTE.*
János went cinema-to yesterday evening
JOHN went to the cinema YESTERDAY EVENING.

Nevertheless, if question words are considered as a subtype of focus, it is surprising why two of them can appear in the preverbal position in Hungarian, which can accommodate only one focus in a declarative sentence. One type of analysis would be to assume that question words, in spite of all the prosodic, syntactic and semantic similarities, are not a subtype of focus, and what makes them similar is that they play similar roles in the discourse, i.e. that they constitute the most prominent parts of question-answer pairs. In this case, the information structure architecture should not contain the focus as a primitive category, but more neutral ones that can be filled differently in declaratives and interrogatives. Another way is to claim that interrogative words constitute a special type of focus that are subject to different restrictions in Hungarian. This is what Mycock (2006) proposes: she argues that

this is why two of them can precede the finite verb in Hungarian. This way, focus can be kept as an information structure primitive, but it has to be signaled whether its interrogative or not. However, the different categorization itself does not explain the phenomenon. A third possibility is to claim that only one of these preverbal interrogative words can be considered as focus. This is what we will propose in this article, keeping in mind that an analysis following either of the two other directions would also be possible.

3.2.2 *D-linkedness* in French

Contrary to the Hungarian data, where the D-linked or Sorting key question word is identifiable from the syntactic position, this difference is not manifested in the syntax in French. Since word order is more rigid in French, the order, or the syntactic position of the question words, does not change according to the discourse-status of the question words. Consider the following examples:

(54) Q:
Quel groupe est allé voir quel monument ?
which group is gone see which monument
Which group went to see which monument?

A:
Les linguistes sont allés voir la Tour Eiffel, les psychologues l'Arc de Triomphe...
the linguists are gone see the Tower Eiffel, the psychologists the Arc de Triomphe
The linguists visited the Eiffel Tower, the psychologists the Arc de Triomphe...

Both the question and the answer are ambiguous between a contextually determined set of tourists and a contextually determined set of monuments. However, the interlocutor is aware of the fact that s/he has to enumerate exhaustively all the tourist groups or all the monuments and pair them up with an element of the other set. What helps, in this case, is the context, and not syntactic or prosodic information. The contextual difference means that the difference between the question words is encoded in the information structure and not at the other levels of linguistic analysis. The parallel architecture of Lexical-Functional Grammar will make it possible to account for these differences, since it dissociates discourse functions from syntax and prosody.

3.2.3 Question words denoting the same type of set

We have seen above that in structure 3, not all question words can appear in Hungarian, but only those that denote the same type of set. Again, this is a semantic restriction that cannot easily be captured in syntax. As we will see, according to a recent proposal (Dalrymple, 2010), in LFG semantic information is integrated into the information structure, and both are related to, but dissociated from the syntax. Again, what we see is that the information structure architecture that a certain theory assumes plays a crucial role in the analysis of the phenomena. After looking at the last problematic point, we continue with the introduction of the concept of information structure in LFG.

3.2.4 Coordination of different functions

In French, coordinated constituents have to share all their functions (e.g. subject and topic), whereas in Hungarian, almost all preverbal question words can be coordinated (if they do not denote the same type of set):

(55) *Ki és mikor ment moziba?*
who and when went cinema-to
Who went to the cinema and when?

The question is then, what is the common function that constitutes the basis of the coordination. Lipták (2001) and Skrabalova (2006) claim that this function is that they are focussed. This means that in Hungarian (and in some other languages, like in Czech), the identity of at least one function the conjuncts share is enough to obtain a grammatical coordination. In other languages, like in French, all functions have to be shared by the conjuncts. However, two non-interrogative foci cannot be coordinated in the preverbal position:

(56) **JÁNOS és TEGNAP ment moziba.*
János and yesterday went cinéma-to
JÁNOS went to the cinema YESTERDAY.

On the other hand, it is possible to coordinate other, prosodically prominent/focussed elements, like in the following example:

(57) *Ide MINDENKI és MINDIG be-jö-het.*
here everybody and always PRT-enter-can
EVERYBODY can ALWAYS come in here.

Although the universal quantifiers are syntactically not in the designated focus position (this follows from the fact that the verbal particle precedes the verb and in the presence of a focussed constituent it follows it), their prosody and use[4] makes them similar to focussed constituents, syntactically they differ from them. This example clearly shows that *focus* as a semantic/information structure concept cannot always be associated with fixed syntactic positions. Rather, focussed elements can appear at different parts of the sentence in Hungarian. It seems thus that in Hungarian the common (grammatical or discourse) function has to be completed by some common lexical feature: an interrogative or universal quantifier. Again, this shows that levels other than the syntactic play a crucial role in the grammar of multiple questions.

4 The information structure in LFG

In earlier versions of the LFG framework, discourse functions were integrated in the functional structure, linked via functional uncertainty (one syntactic unit was associated with two functions at the same time, for instance *topic* and *subject*). The projection of the information structure as a separate level of representation was motivated by the following problems.

First of all, this meant that different kinds of information were represented in the same structure. This goes against the LFG way of representing information at different levels of representation (Choi (1999), Dalrymple (2010)).

Secondly, King (1997) argues that encoding discourse functions in the f-structure leads to circularity, in the case where it is only the verb, without its arguments, that is focussed. This is why she proposes an independent level of representation, where discourse functions are encoded, with their bare predicate value (without their arguments).

Another reason why a separate level of information structure is necessary is that syntactic constituents do not correspond systematically to constituents of information structure, like in the following example:

[4]This sentence cannot be uttered out of the blue. Just like in the case of focus, this sentence is also a correction or the answer to a question.

(58) It was the RED shirt that Mary gave to John, not the blue one.

In (58) the whole constituent *the RED shirt* that is clefted; however, it is only the element *RED* that is focussed. This difference can be captured if clefting and focalization are represented at different levels.

Butt and King (1996) propose that the information structure consists of four sets, which are defined by the combination of two features: *new* +/− and *prominent* +/−. The *TOPIC* set contains elements that are prominent, but not new, the *(Information) FOCUS* set contains new and prominent elements, whereas old and not prominent elements belong to *BACKGROUND* and new but not prominent ones to *COMPLETIVE INFORMATION*. Although this classification simplifies the definition of discourse functions, for instance, foci are not always new, and prominence is, in some cases, difficult to define (Krifka (2006)), we will base our analysis on this architecture, with some precision of the sets mentioned above. We will consider as *FOCUS* elements that are answers to questions.[5] These are not necessarily new, since some of the answers can already be introduced in the discourse, for instance in the case of questions which require the choice between two possibilities already present in the question:

(59) -Who did you invite to the party, Mary or John?
-John.

In our analysis, the *TOPIC* set will contain elements that are contextually determined and salient in the discourse, bearing at the same time prosodic and/or syntactic prominence (the latter meaning, in most cases, a position in the left periphery of the sentence). It is important to remark that *topic* here corresponds to *sentential topic* and not to *discourse topic*, referring to one element of the sentence and not to the question under debate in the whole discourse (Beyssade et al. (2004)), although the two are related: sentence topics constitute subtopics or subquestions of the *Question under Discussion/Discourse Topic*. All other elements will be placed in the set of *BACKGROUND INFORMATION*. In the next section we examine how the LFG architecture can account for the problems mentioned above.

Dalrymple (2010) proposes that the information structure categories contain the semantic description of the particular elements. Our data confirm this. As we will see in the concrete examples, some phenomena are best accounted for by referring to semantic features within an information structure category. In addition, the pair-list reading can also be derived from such a representation.

In the next section we examine how the LFG architecture can account for the problems mentioned above.

5 The proposed analysis

5.1 Two preverbal question words in Hungarian and *D-linkedness* in French

Both of the problems mentioned above are related to the fact that, in pair-list questions, one of the question words differs from the other in some respects (semantics, prosody, etc.). Since this difference plays an important role in the answer, we propose that (in the already presented information structure architecture) the *D-linked/Sorting key* question words belong to the *TOPIC* set, and the other(s) to the *FOCUS* set.

It has already been suggested in the literature that D-linked, or in some languages, initial interrogative words, share some properties with topics, in that they are both contextually determined, given, and salient in the discourse. Surányi (2006) argues that Hungarian high *wh*-phrases, although not moved by topicalization, are interpreted at interfaces as topics, since, like topics, they invariably quantify over

[5] A more precise definition that takes account of all focus-related phenomena would also include parallelism and corrections, in which the focussed constituents are the parts of, the sentence that are in parallel with, or correct parts of previous sentences. The common property with answers to questions is the fact that sentences containing foci cannot be uttered out of the blue, but always as a reaction to the preceding context. These other uses of focus are, however, beyond the scope of the present study.

presupposed sets, and constituents that correspond to these *wh*-phrases appear in the topic, and not in the focus position in Hungarian. Unfortunately, the claim that an element is not moved by topicalization, but is interpreted at interfaces as a topic, is difficult to formalize in a minimalist framework that aims to account for all phenomena in syntax, via movement to various functional projections, which, in turn express discourse or semantic information. It seems, therefore, that an LFG approach, in which information structure is treated independently from syntax, is a more adequate framework to account for this phenomenon.

Another argument supporting this claim is that constituents corresponding to *D-linked/Sorting key* question words in the answer are contrastive topics (Büring (2003)). In Hungarian pair-list questions the question words cannot be preceded by (other) contrastive topics:

(60) **Jánosról, arról mikor mit mondtak?*
János about that about when what said.3PL
John, what did they say about him when?

In French, in the answers to pair-list questions, only one XP gets prosodically distinguished, the one corresponding to the non-D-linked question word (Marandin (2006)). The others bear a certain type of accent that Beyssade et al. (2004) call C-accent and relate to the so-called B-accent in English (Jackendoff (1972), Büring (2003)). Both are supposed to encode contrastive topics. Let us consider the following example:

(61) *Quels étudiants étudient quoi dans ce département ?*
which students study what in this department
Which students study what in this department?

(62) *Les étudiants de première année étudient la syntaxe, ceux de seconde année la sémantique.*
the students of first year study the syntax, those of second year the semantics
The first year students study *syntax*, those of the second year *semantics*.

Hungarian is not the only language in connection with which the claim about the topic status of high *wh*-phrases has been made. According to Jaeger (2003) topic-fronted objects are clitic-doubled in colloquial Bulgarian. He also notices that *wh*-phrases in multiple questions are also subject to clitic-doubling, in which case superiority effects are cancelled, and argues for the topicality of high *wh*-phrases in Bulgarian.

Furthermore, Grohman (2006) shows that only topicalizable elements can intervene between two *wh*-phrases in German multiple questions and proposes that therefore all *wh*-phrases undergo topicalization. It is not our aim here to contest Grohmann's analysis (for instance the discourse status and semantics of lower *wh*-phrases is clearly different from that of initial ones), but to show that the topicality of *wh*-words has been claimed in connection with various languages, based on a variety of criteria (syntactic position, cliticization, intervention).

In addition, some semantic similarities can also be observed between these *wh*-phrases and topics: both tend to be given, referential, salient in the discourse and to denote a contextually determined set of entities.

The analysis proposed is thus the following. At the level of information structure, *D-linked/Sorting key* question words belong to the *TOPIC* set, but they can be associated with different syntactic positions, depending on the language. In Hungarian they precede non-D-linked question words, whereas in French they can be sentence-initial, just like sentence-final ones. The advantage of this approach is that it does not suppose that question words are exceptional in that there can be two (interrogative) preverbal foci

in Hungarian, and that this cancels the apparent ambiguity in pair-list questions in French. Consider the analysis of Hungarian (6) and (7):

Functional structure

$$\begin{bmatrix} \text{PRED} & \text{'hoz}\left\langle\left(\uparrow \text{SUBJ}\right),\left(\uparrow \text{OBJ}\right),\left(\uparrow \text{OBL}\right)\right\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'}\left(\text{KI}\right) \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'}\left(\text{MIT}\right) \end{bmatrix} \\ \text{OBL} & \begin{bmatrix} \text{PRED} & \text{'a bulira'} \end{bmatrix} \\ \text{TNS} & \begin{bmatrix} \text{PAST} \end{bmatrix} \end{bmatrix}$$

Information Structures

Ki mit hozott a bulira?
(Who brought what to the party?)

$$\begin{bmatrix} \text{TOP} & \text{ki} \\ \text{FOC} & \text{mit} \\ \text{BACKGR} & \text{hozott a bulira} \end{bmatrix}$$

Mit ki hozott a bulira?
(What was brought by whom to the party?)

$$\begin{bmatrix} \text{TOP} & \text{mit} \\ \text{FOC} & \text{ki} \\ \text{BACKGR} & \text{hozott a bulira} \end{bmatrix}$$

The two questions have the same f-structure, but they have different c-structures and different information structures. In the first case *ki* (who) belongs to the topic set and *mi* to the focus set, whereas in the second, it is the other way round. These information structure categories are associated with syntactic positions in Hungarian, but other encodings are also possible in other languages. The French example is represented in a similar way:

Functional Structure

Quel groupe a visité quel monument ?
(Which group visited which monument?)

$$\begin{bmatrix} \text{PRED} & \text{'visiter}\left\langle\left(\uparrow \text{SUBJ}\right),\left(\uparrow \text{OBJ}\right)\right\rangle\text{'} \\ \text{Q} & \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'quel groupe'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'quel monument'} \end{bmatrix} \end{bmatrix}$$

Information structure

$$\begin{bmatrix} \text{TOP} & \begin{bmatrix} \text{QUEL GROUPE} \end{bmatrix} \\ \text{FOC} & \begin{bmatrix} \text{QUEL MONUMENT} \end{bmatrix} \\ \text{BACKGR} & \begin{bmatrix} \text{VISITER} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{TOP} & \begin{bmatrix} \text{QUEL MONUMENT} \end{bmatrix} \\ \text{FOC} & \begin{bmatrix} \text{QUEL GROUPE} \end{bmatrix} \\ \text{BACKGR} & \begin{bmatrix} \text{VISITER} \end{bmatrix} \end{bmatrix}$$

In these French examples, the questions have not only the same f-structure, but the same c-structure as well (the presentation of this latter is beyond the scope of this paper). What makes the interpretational difference between them is the information structure.

5.2 Restrictions on sentence-final question words in Hungarian

Concerning the formalization of this phenomenon, we can suppose that the lower interrogative word in the structure must share some (semantic) feature with the higher one (animacy, specificity or location) and this accounts for the fact that they belong to the same lexeme and denote the same type of set. These features are distributive, characterizing the whole set of interrogative words at the level of information structure, which means that the unification fails if they have different values.

Functional structure — Information structure

Ki hívott meg kit a bulira?

(Who invited whom to the party?)

$$\begin{bmatrix} \text{PRED} & \text{'meghív}\left\langle\left(\uparrow \text{SUBJ}\right),\left(\uparrow \text{OBJ}\right),\left(\uparrow \text{OBL}\right)\right\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'}\left(\text{KI}\right) \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'}\left(\text{KIT}\right) \end{bmatrix} \\ \text{OBL} & \begin{bmatrix} \text{PRED} & \text{'a bulira'} \end{bmatrix} \\ \text{TNS} & \begin{bmatrix} \text{PAST} \end{bmatrix} \end{bmatrix} \qquad \begin{bmatrix} \text{TOP} & \\ \text{FOC} & \left\{ \begin{bmatrix} \text{KI} \\ \text{Q} \\ \text{ANIM} + \\ \text{SPEC-} \end{bmatrix}, \begin{bmatrix} \text{KIT} \\ \text{Q} \\ \text{ANIM} + \\ \text{SPEC-} \end{bmatrix} \right\} \\ \text{BACKGR} & \text{hívott meg a bulira} \end{bmatrix}$$

In the LFG architecture, since semantic information is integrated into the information structure, it is possible to indicate the distributive features of the set there. The treatment of interrogative words in this case is similar to that of coordination, which is also represented in LFG as a set. Let us now have a look at coordination and the emerging problems.

5.3 Coordination

Coordination is analyzed in LFG as a set (Dalrymple (2001)), in which the elements have to share certain features. The general rule of coordination is the following:

Rule:

$$\begin{array}{cccccc} \text{S} & \rightarrow & \text{XP} & \text{Conj} & \text{XP} \\ & & \downarrow\in\uparrow & \uparrow=\downarrow & \downarrow\in\uparrow \end{array}$$

From the data we have seen above it seems (in LFG terms) that in Hungarian sharing a function at one level (f- or i-structure) is enough for the coordination to be grammatical: this contrasts with French, where the conjuncts have to share all their functions at all levels. However, besides the common information structure set, the conjuncts also share some lexical features in Hungarian (for instance the fact that they are interrogative). Again, this also supports the view that semantic information is integrated into the information structure. Consider the following example from French =(30):

Où et quand a eu lieu la conférence ? (Where and when did the conference take place?)

Functional structure

$$\begin{bmatrix} \text{PRED} & \text{'avoir lieu}\left\langle\left(\uparrow \text{SUBJ}\right),\left(\uparrow \text{OBL}\right),\left(\uparrow \text{OBL}\right)\right\rangle\text{'} \\ \text{SUBJ} & \left[\text{PRED} \quad \text{'la conférence'}\right] \\ \text{OBL} & \left\{\begin{matrix}\left[\text{PRED} \quad \text{'pro'}\left(\text{OÙ}\right)\right] \\ \left[\text{PRED} \quad \text{'pro'}\left(\text{QUAND}\right)\right]\end{matrix}\right\} \\ \text{TNS} & \left[\text{PAST}\right] \end{bmatrix}$$

Information Structure

$$\begin{bmatrix} \text{TOP} & \\ \text{FOC} & \left\{\begin{matrix}\text{Q} \\ \text{OÙ} \\ \text{QUAND}\end{matrix}\right\} \\ \text{BACKGR} & \text{la conférence a eu lieu} \end{bmatrix}$$

A similar Hungarian example:
Ki és mikor ment moziba? (Who went to the cinema and when?)

Functional structure

$$\begin{bmatrix} \text{PRED} & \text{'megy}\left\langle\left(\uparrow \text{SUBJ}\right),\left(\uparrow \text{OBL}\right)\right\rangle\text{'} \\ \text{SUBJ} & \left[\text{PRED} \quad \text{'pro'}\left(\text{KI}\right)\right] \\ \text{OBL} & \left[\text{PRED} \quad \text{'moziba'}\right] \\ \text{ADJ} & \left[\text{PRED} \quad \text{'pro'}\left(\text{MIKOR}\right)\right] \\ \text{TNS} & \left[\text{PAST}\right] \end{bmatrix}$$

Information Structure

$$\begin{bmatrix} \text{TOP} & \\ \text{FOC} & \left\{\begin{matrix}\text{Q} \\ \text{KI} \\ \text{MIKOR}\end{matrix}\right\} \\ \text{BACKGR} & \text{ment moziba} \end{bmatrix}$$

6 Conclusion

In this paper we have proposed an LFG account of multiple questions in French and in Hungarian. We have identified five (or, together with clausal coordination, six), different structures in the two languages and seen that they are (except for the first Hungarian structure) compatible with both a pair-list and a single-pair reading, if the question words are in the same clause. However, the following tendency can be observed: when the interrogative words are arguments, both the pair-list and the single pair readings are possible, whereas the pair-list reading is preferred in the case of adjuncts (if at least one of the question words is an adjunct), since an alternative structure also exists (clausal coordination), which is, in turn, only compatible with the single-pair reading. We identified four problems and argued that the LFG framework is suitable for handling all of them, because of its modular architecture, separating the different levels of linguistic information. Concerning the puzzle of two preverbal (interrogative) foci in Hungarian and the syntactic ambiguity of D-linkedness in French we proposed that D-linked/Sorting key question words belong to the topic set at information structure, since they share many properties with (contrastive) topics. Concerning the restrictions on preverbal and postverbal question words, and the coordination of question words with unlike functions, we have proposed that they are due to semantic features that question words are supposed to share in the focus set of the information structure. In Hungarian, apart from sharing some lexical features, conjuncts have to belong to the same set at at least one level of representation, whereas in French these are shared at all levels.

References

Bánréti, Z. (2007). *A mellérendelés és az ellipszis nyelvtana a magyarban*. Tinta könyvkiadó, Budapest.

Beyssade, C., Delais-Roussarie, E., Doetjes, J. S., Marandin, J.-M., and Rialland, A. (2004). Prosody and information. In Corblin, F. and de Swart, H., editors, *The Handbook of French Semantics*, pages 477–500. CSLI Publications, Stanford.

Büring, D. (2003). On d-trees, beans, and b-accents. *Linguistics & Philosophy*, 26(5):511–545.

Butt, M. and King, T. H. (1996). Structural topic and focus without movement. In Miriam, B. and King, T. H., editors, *Proceedings of the LFG96 Conference*, Rank Xerox, Grenoble. CSLI.

Choi, H.-W. (1999). *Optimizing Structure in Context: Scrambling and Information Structure.* Dissertations in Linguistics. CSLI Publications, Stanford, CA.

Comorovski, I. (1996). *Interrogative Phrases and the Syntax-Semantics Interface*. Kluwer Academic Publishers, Dordrecht, Boston, London.

Dalrymple, M. (2001). *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press, London.

Dalrymple, M. (2010). Glue and information structure. Paper presented at the LFG Meeting, SOAS, London.

Ginzburg, J. (1997). On some semantic consequences of turn taking. In Dekker, P., editor, *11th Amsterdam Colloquium*.

Ginzburg, J. and Sag, I. (2000). *Interrogative Investigations. The Form, Meaning and Use of English Interrogatives*. CSLI, Stanford.

Gracanin, M. (2007). *About Sharing*. PhD thesis, MIT.

Grohman, K. K. (2006). Top issues in questions: Topics-topicalization-topicalizability. In Cheng, L. L.-S. and Corver, N., editors, *Wh-Movement Moving On*. MIT, Cambridge MA, London, England.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge MA.

Jaeger, F. T. (2003). Topicality and superiority in Bulgarian wh-questions. In Arnaudova, O., Browne, W., Rivero, M. L., and Stojanovic, D., editors, *FASL-12: The Ottawa Meeting 2003*, pages 207–228., Stanford. Michigan Slavic Publications.

Kálmán, L., editor (2001). *Magyar leíró nyelvtan. Mondattan 1.* Tinta Könyvkiadó, Budapest.

King, T. H. (1997). Focus domains and information structure. In Miriam, B. and King, T. H., editors, *LFG97*, University of California, San Diego. CSLI Publications.

Krifka, M. (2001). For a structured meaning account of questions and answers. In Caroline, F. and Sternefeld, W., editors, *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, pages 287–319. Akademie Verlag, Berlin.

Krifka, M. (2006). Basic notions of information structure. In Caroline, F., Fanselow, G., and Krifka, M., editors, *Interdisciplinary Studies on Information Structure*.

Kuno, S. and Takami, K. (1993). *Grammar and Discourse Principles: Functional Syntax and GB Theory.* Chicago University Press, Chicago.

Lipták, A. (2001). *On the Syntax of Wh-items in Hungarian*. PhD thesis, Utrecht University.

Marandin, J.-M. (2006). Focus informationnel. In Godard, D., Roussarie, L., and Corblin, F., editors, *Sémanticlopédie: dictionnaire de sémantique*. GDR Sémantique & Modélisation, CNRS, http://www.semantique-gdr.net/dico/.

Mycock, L. (2006). *A New Typology of Wh-Questions*. PhD thesis, Manchester University.

Pesetsky, D. (1987). Wh-in-situ: Movement and unselective binding. In Reuland, E. and ter Meulen, A., editors, *The Representation of (In)definites*. MIT Press, Cambridge MA.

Skrabalova, H. (2006). Parataxe apparente et coordination des interrogatifs en tchèque. *Faits de langue*, 28.

Surányi, B. (2006). Mechanisms of wh-saturation and interpretation in multiple wh movement. In Cheng, L. L.-S. and Corver, N., editors, *Wh-Movement Moving On*, pages 289–318. MIT Press, Cambridge MA.

CLOSING THE GAP BETWEEN STOCHASTIC AND RULE-BASED LFG GRAMMARS

Annette Hautli
Universität Konstanz

Özlem Çetinoğlu Josef van Genabith
CNGL, School of Computing
Dublin City University

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

Developing large-scale deep grammars in a constraint-based framework such as Lexical Functional Grammar (LFG) is time-consuming and requires significant linguistic insight. Recently, treebank-based constraint-grammar acquisition approaches have been developed as an alternative to hand-crafting such resources. While treebank-based approaches are wide coverage and robust and achieve competitive evaluation results for many languages, the granularity of the linguistic analyses provided by treebank-based resources tends to be less fine-grained than what is offered by state-of-the-art hand-crafted grammars. This paper presents an approach to extend the English DCU LFG annotation algorithm with more detailed f-structure information to provide probabilistic treebank-based LFG grammars with rich feature information comparable to that implemented by the hand-crafted English XLE grammar, while maintaining the robustness and the coverage of treebank-based stochastic grammars.

1 Introduction

Robustly parsing natural language has been the focus of research for the last decades, with frameworks evolving that attempt to go beyond the analysis of constituency and hierarchical order to provide a more abstract level of linguistic analysis. Lexical Functional Grammar (LFG) (Bresnan and Kaplan, 1982; Dalrymple, 2001), among others, is one approach that combines two levels of representation, namely constituent structure and functional structure, which are related by a projection architecture where functional information is encoded in terms of functional descriptions annotated on constituents in phrase-structure rules. By employing both representations, LFG provides insight into the surface as well as the deeper, more abstract properties of natural language syntax. In addition, LFG has proven to be a theory that can serve as the backbone for a computational analysis of natural language. These features have led to the development of computational LFG grammars that allow for an automatic syntactic analysis of natural language.

Over time, various methodologies of developing computational LFG grammars have evolved. One approach that has proven highly successful is the employment of a rule-based XLE LFG parser (Crouch et al., 2010) where the manual encoding of syntactic rules and functional descriptions provides a deep and highly detailed syntactic analysis that forms the input to the computation of a semantic representation at a subsequent processing step (Crouch and King, 2006).

In general, manual development of large-scale deep grammars faces a number of challenges. First, language data is complex and varied and some perfectly legitimate constructions may be outside the coverage of the grammar. Second, "real" input may contain typos and disfluencies not envisaged by the hand-crafted

[†]This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

grammar. Third, even highly efficient LFG parsing architectures operating on detailed hand-crafted grammars can be significantly slower than some state-of-the-art treebank-based stochastic parsers. Because of this, full coverage on unrestricted language data can often only be achieved by taking short-cuts: either combining fragment analyses in parser output or constraining the amount of computation to only partially explore sections of the full parsing search space.

An alternative DCU LFG approach (Cahill et al., 2004) uses robust treebank-based stochastic parsers that automatically produce Penn-II Treebank style (Marcus et al., 1993) trees. Functional information is provided by an f-structure annotation algorithm (Burke, 2006) that automatically annotates treebank-style trees with f-structure information. In this approach, both trees and f-structure information are provided automatically, reducing human grammar development effort. Up to now, treebank-based DCU LFG f-structures encoded substantially less information compared to the rich and detailed f-structures produced by the hand-crafted XLE LFG grammars, even though the treebank-based approach outperforms the hand-crafted grammars on the restricted core (preds-only and slightly extended) feature sets in the PARC700 evaluation gold standard (Cahill et al., 2008). Because of the lack of detail, however, to date treebank-based f-structure output could not be used for further semantic processing as implemented by Crouch and King (2006).

This paper attempts to combine the best of the two worlds: the linguistic depth and detail of the f-structures provided by hand-crafted XLE LFG grammars with the coverage, robustness and parse quality provided by the automatic grammar acquisition methodology of the treebank-based DCU LFG approach. In order to achieve this, we extend the f-structure annotation algorithm of the treebank-based DCU LFG grammar for English with more detailed f-structure information, approaching the feature granularity and linguistic sophistication of the state-of-the-art hand-crafted XLE LFG grammar for English.

We currently achieve an f-score of 80.20 against the full feature set produced by the XLE parser for a test suite of 720 sentences designed to evaluate the semantic representation of the XLE grammar, and 73.70 against the corresponding semantic representations taking the treebank-based f-structure output as input to the XFR-based semantic construction, substantially improving on earlier rewriting-based approaches to map treebank and annotation algorithm-based f-structures to XLE output. Furthermore, evaluating the f-structure output of the treebank-based parsing pipeline with the extended f-structure annotation algorithm against the features of the PARC700 gold standard, we achieve an f-score of 83.59, outperforming both earlier XLE- and treebank-based LFG parsing results.

The paper is structured as follows: section 2 reviews the state-of-the-art in computational LFG grammars for English, followed by a particular linguistic example and how the hand-crafted XLE LFG and treebank- and annotation algorithm-based DCU LFG grammars attempt to provide a linguistically motivated analysis for it. Section 3 presents how the gap between the two LFG grammar development architectures can be closed. Section 4 presents evaluation results, followed by remarks on future work and the conclusion in section 5.

2 State-of-the-Art

The aim of parsing natural language has resulted in computational grammars in various constraint-based frameworks, among them Lexical-Functional Grammar (LFG) (Bresnan and Kaplan, 1982; Dalrymple, 2001).

LFG is able to provide cross-linguistically valid analyses by employing levels of representation that abstract away from the surface structure of sentences. Furthermore, due to its computational and mathematical tractability, it has also proven an excellent theory for use in computational linguistics (Maxwell and Kaplan, 1996), supporting a combination of theoretically-founded, deep, linguistic and computationally efficient analyses.

For the purpose of this paper, we restrict ourselves to discussing two state-of-the-art LFG grammars for English: the hand-crafted XLE grammar in section 2.1 and the DCU LFG treebank- and annotation algorithm-based approach in section 2.2, followed by an exemplary linguistic issue, namely a non-local dependency, and how the two approaches cope with it.

2.1 The English XLE grammar

XLE is an efficient rule-based grammar development platform, developed at Palo Alto Research Center (PARC) and consisting of cutting-edge algorithms for parsing and generation using LFG grammars, along with a user interface for writing and debugging LFG grammars and lexical resources (Crouch et al., 2010).

XLE provides the shared technology platform within the ParGram effort (Butt et al., 1999, 2002) that aims at developing parallel LFG grammars for various languages such as English, German, French, Norwegian, Japanese, Turkish and Urdu. Besides providing computationally efficient analyses for natural language, a main focus of ParGram lies on the cross-linguistically valid analysis of natural language, making the LFG analyses as parallel and informative as possible across languages.

The English XLE grammar is part of a larger system that maps text to an Abstract Knowledge Representation (AKR) which can then be used for applications such as search, question answering (Bobrow et al., 2007) and redacting text (Bier et al., 2009). Figure 1 shows the basic system pipeline.

In order to break text into sentences and sentences into tokens, finite-state transducers (FSTs) are applied. These are also used for the morphological component, where each word is analyzed and morphological information is passed on to the XLE grammar. Hand-coded syntax rules in the XLE grammar pick up the morphological information and add constituent structure and functional information. Like all rule-based LFG grammars, the output of the syntax is a c-structure (constituent structure) (Figure 2) that encodes constituency and linear order, and an f-structure (functional structure), encoding functional information such as the predicate-argument structure and semantically important features such, e.g. tense and number (see Figure 3). In cases of syntactic ambiguity, such as PP-attachment, the XLE grammar outputs a packed representation of all possible analyses, which

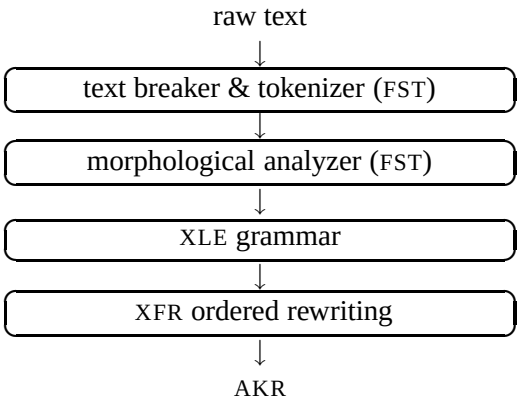


Figure 1: XLE pipeline

allows optimality marks and a stochastic disambiguation component to choose between them in further processing. Optimality Theory marks in the syntax rules indicate which analyses are (dis)preferred (Frank et al., 1998).

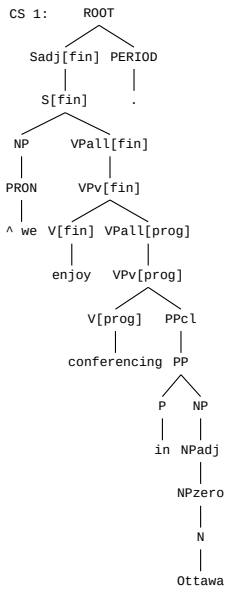


Figure 2: C-structure for *We enjoy conferencing in Ottawa.*

In a later step, the syntactic information contained in the f-structure is passed on to XFR ordered rewrite rules that map f-structure information to a semantic representation by using external resources to replace words with concepts and grammatical functions with semantic roles (Crouch and King, 2006). Further rewriting by XFR rules converts the semantic representation into an Abstract Knowledge Representation.

```
"We enjoy conferencing in Ottawa."

      [PRED      'enjoy<[21:we], [67:conferencing]>'                                                    ]
      [                [PRED  'we'                                                                      ]
        21             [                                                                                ]
      [SUBJ    365     [NTYPE [NSYN pronoun]                                                            ]
      [        370     [CASE nom, HUMAN +, NUM pl, PERS 1, PRON-TYPE pers]                              ]
      [                [PRED    'conferencing<[21:we]>'                                                 ]
      [                [SUBJ    [21:we]                                                                 ]
      [                [                 [PRED    'in<[136:Ottawa]>'                                    ]
      [                [                 [                [PRED  'Ottawa'                                ]
      [                [                 [                [CHECK [_LEX-SOURCE morphology]                ]
      [                [                 [          136   [                                            ]
      [                [ADJUNCT          [OBJ  833        [NSEM [PROPER [LOCATION-TYPE city, PROPER-TYPE location]]]
      [XCOMP           [                 [          874   [NTYPE                                       ]
      [                [                 [         1007   [NSYN proper                                 ]
      [                [                 [         1017   [CASE obl, NUM sg, PERS 3                    ]
      [                [          108    [                                                            ]
      [                [         1186    [PSEM    ⟨loc⟩                                               ]
      [                [         1199    [PTYPE   sem                                                 ]
 157  [                [                 [                                                            ]
 187  [         67     [CHECK   [_LEX-SOURCE guesser, _SUBCAT-FRAME V-SUBJ]                           ]
  41  [        573     [                                                                              ]
2091  [       1036     [TNS-ASP [PERF -_, PROG +_]                                                    ]
2106  [       1433     [PASSIVE -, VTYPE main                                                         ]
2178  [CHECK            [_SUBCAT-FRAME V-SUBJ-XCOMPprog]                                               ]
2182  [                                                                                               ]
2187  [TNS-ASP          [MOOD indicative, PERF -_, PROG -_, TENSE pres]                                ]
1100  [CLAUSE-TYPE decl, PASSIVE -, VTYPE main                                                         ]
```

Figure 3: F-structure for *We enjoy conferencing in Ottawa.*

In cases where the parser produces a large number of analyses, XLE uses a stochastic disambiguation model (Riezler et al., 2002) trained on LFG analyses for a sub-set of Penn-II Treebank sentences using some information from the Penn-II Treebank trees to guide the XLE analyses. For sentences where the XLE grammar cannot produce a spanning parse, i.e. a complete analysis for the entire input, XLE is able to produce a sequence of largest well-formed fragment analyses. This provides a degree of robustness for language data outside the coverage of the grammar and also allows for "ungrammatical" or fragmented language, including typos etc.

2.2 The DCU LFG annotation algorithm

Annotation algorithms and wide-coverage treebank-based LFG systems have been developed in the Dublin City University (DCU) GramLab project for English, Chinese, French, Spanish, Arabic and German. The treebank-based LFG annotation algorithm for English (Cahill, 2004; Cahill et al., 2008), generates c- and f-structures for sentences in a different manner than the XLE grammar: making use of reliable treebank-parsers, the f-structure annotation algorithm annotates the nodes in the tree with f-structure equations.

In general, the parsing pipeline comprises the following parts: first, a text breaker splits running text into sentences. These are then parsed by a treebank-trained stochastic parser (Charniak and Johnson, 2005; Bikel, 2002), which creates trees in the Penn-II Treebank style (Marcus et al., 1993). The treebank function label tagger (Chrupała et al., 2007) (also trained on the Penn-II treebank) enriches the bare CFG parser output trees with further information by assigning treebank function labels (where possible), e.g. adding SBJ to subject noun phrases (NP-SBJ), and LOC to locative prepositional phrases (PP-LOC). This information helps the

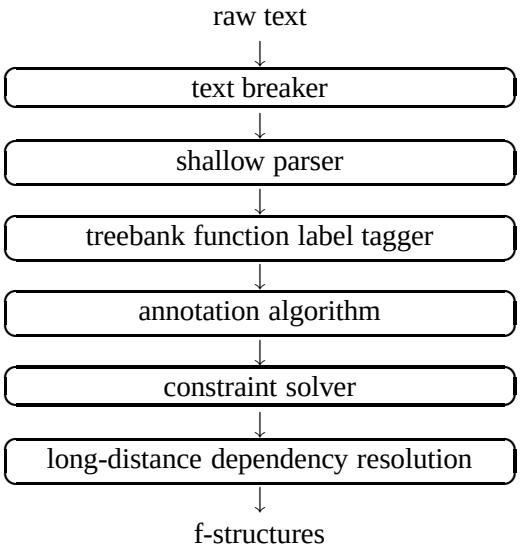


Figure 4: DCU pipeline

f-structure annotation algorithm to assign f-structure equations to the tree nodes (including the terminal nodes). A constraint solver collects and resolves the f-structure equations and produces an f-structure. The last step, the long-distance dependency resolution module, resolves non-local dependencies using automatically acquired subcategorisation frames and finite approximations of functional-uncertainty equations (all automatically extracted from the f-structures automatically generated from the training set of the original Penn-II treebank trees). See Figure 4 for a schematic overview of the DCU pipeline.

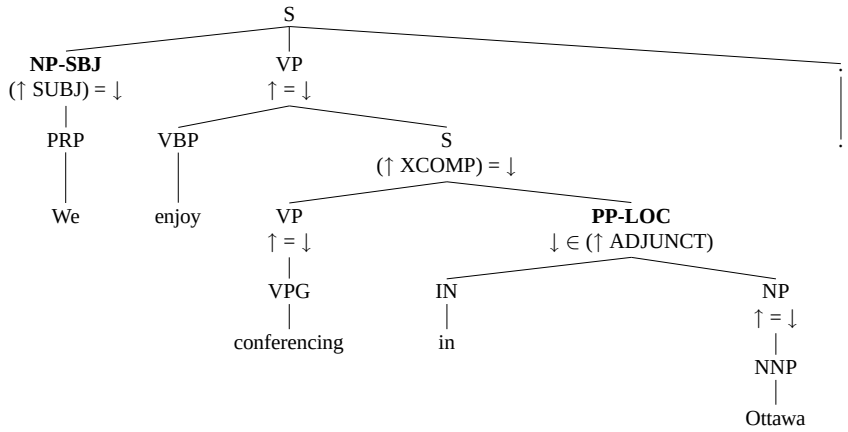


Figure 5: DCU tree with f-structure equations

Figure 5 shows the Penn-II Treebank-style tree for *We enjoy conferencing in Ottawa* with the functional equations added by the f-structure annotation algorithm.[1] As the CFG-parser is trained on the Penn-II Treebank, tree nodes are named according to Penn-II conventions and may not necessarily correspond to usual LFG textbook or XLE-style phrase-structure labels.

The DCU LFG parsing pipeline including the f-structure annotation algorithm has been successfully evaluated on gold standards such as the PARC700 (King et al., 2003), outperforming the rule-based English XLE grammar and parser by more than 2% f-score absolute (Cahill et al., 2008).

This provides an excellent basis for further development of the DCU processing pipeline, with a particular focus on the f-structure annotation algorithm. However, as the original DCU system was tuned to produce only basic LFG representations concentrating on a restricted set of core syntactic and semantic features, it lacks the detail and linguistic sophistication of the f-structures produced by the hand-crafted English XLE grammar. In order to close this gap between the treebank-based and the hand-crafted grammars, we need to extend the restricted feature space of the DCU annotation algorithm to obtain f-structures as detailed as XLE f-structures.

2.3 An example linguistic issue for computational grammars

A recurring problem for computational grammars that goes beyond the analysis of pure surface structure and provides a linguistically motivated analysis is the treatment of non-local dependencies. Consider the example '*We enjoy conferencing in Ottawa.*'. Here we have an unexpressed argument in the subordinate clause, with the subject of the main clause also being the subject of subordinate clause.

In the case of the hand-crafted LFG grammar, we can express this information by writing a lexical entry for the verb *enjoy*, as given in Figure 6, which encodes this dependency information:

```
enjoy V * (^ PRED) = 'enjoy ⟨(^ SUBJ) (^ XCOMP)⟩'
          (^ SUBJ) = (^ XCOMP SUBJ)
```

Figure 6: The LFG lexical entry for *enjoy*

Whenever the grammar finds a construction with the word *enjoy* that subcategorizes for a SUBJ and an XCOMP, the subject of the main clause is automatically made the subject of the XCOMP.

For the treebank- and annotation algorithm-based DCU parsing pipeline, the situation is different. Instead of manually encoding information on the non-local dependencies between the main and the subordinate clause in the lexical entry, all that is available is a parser output simplified Penn-II treebank-style tree structure as in Figure 7.

[1] Lexical equations are not shown.

```
(S (NP (PRP We))
   (VP (VBP enjoy)
   (S (VP (VBG conferencing)
          (PP (IN in)
              (NP (NNP Ottawa))))))
(.  .))
```

Figure 7: Parser output tree

Treebank-trained CFG parsers such as Charniak and Johnson (2005) and Bikel (2002) do not capture non-local dependencies. The parser output tree for the example sentence contains no information on the missing argument in the subordinate clause and does not record the non-local dependency between the subject of the main clause and the subordinate clause. From this tree alone, the basic f-structure annotation algorithm cannot recover the non-local dependency.

The long-distance dependency resolution module (Cahill et al., 2004) in the DCU LFG parsing pipeline employs statistical methods and works as follows: unlike the simplified parser output trees, the original Penn-II Treebank contains co-indexed paths for long-distance dependencies, that mark the missing subject in the subordinate clause with an empty node (`NP-SBJ` (`-NONE-` *-1)) and relate it to the subject in the main clause.

Cahill et al. (2004) apply the f-structure annotation algorithm to the full Penn-II treebank trees producing fully non-local dependency-resolved f-structures which record non-local dependencies as corresponding reentrancies in f-structure. From these non-local dependency-resolved f-structures, Cahill et al. (2004) learn subcategorization frames and finite approximations of functional uncertainty equations. These are used in an algorithm to resolve non-local dependencies in the f-structure output and the f-structure with the highest non-local dependency resolution probability is chosen.

So instead of making use of lexical information in the verb entries as in the XLE grammar, the long-distance dependencies are learned from the Penn-II Treebank.

3 Closing the gap

The overall aim of this paper is to show that the gap in detail and granularity of linguistic representation between the hand-crafted English XLE grammar and the treebank-based DCU LFG approach for English can be closed by extending the DCU annotation algorithm.

The proof of concept has been shown by earlier experiments reported in Hautli and King (2009). However this paper present a different approach that substantially outperforms previous experiments. This section presents the two approaches, with evaluations following in the next section.

3.1 Using XFR rewrite rules

A recent approach (Hautli and King, 2009) attempted to overcome the differences between the hand-crafted XLE grammar and the treebank-based DCU LFG approach by using a set of XFR rewrite rules that added missing f-structure information to DCU LFG output, effectively treating the DCU system as a black box. This means that information about the CFG tree structure is not taken into account, and the XFR rewrite rules solely operate on the DCU f-structure output. For a schematic view of the experimental layout see Figure 8.

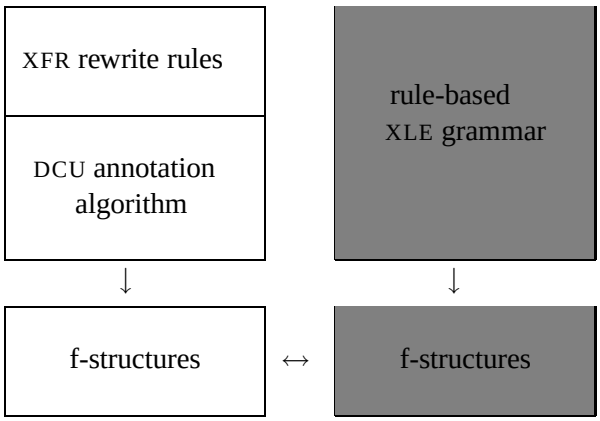


Figure 8: Schematic view of Hautli and King (2009)

In particular, the DCU LFG pipeline as described in the previous section and its output was used as is. After reformatting the DCU LFG output for it to be readable by XLE, XFR rewrite rules were defined and applied, adding information on the level of f-structure that the DCU LFG system had not yet provided. Additionally, existing information was modified, such as feature names and values, to make it XLE-compatible.

Below, we provide a very simple example of the kind of ordered XFR rewrite rule that is employed in this approach, and although being a somewhat artificial example, the principle remains the same for more complicated constructions. Figure 9 shows a DCU LFG f-structure for the subject pronoun 'we' that is rewritten via an XFR rewrite rule. The rule works as follows: the facts on the left hand side of the rewrite rule (before the arrow) constitute the input f-structure facts that must be matched for the rule to apply. If they cannot be matched, the rule does not apply. In cases where the rule fires, the facts on the left side are rewritten to the facts on the right hand side of the rule (after the arrow). Forms beginning with a percent sign (%) are variables that can be instantiated by f-structures.

Consider the XFR rule in Figure 9: the variable `%X` is the f-structure which contains a `subj`; this `subj` is then referred to by the variable `%Subj`. This `%Subj` f-structure must have a `pred` attribute with value `pron`, a `num` attribute with value `pl` and a `pron_form` attribute with the value `we` in order for the XFR rule to

input:

$$\begin{bmatrix} \text{subj} & \begin{bmatrix} \text{pred} & \text{pron} \\ \text{num} & \text{pl} \\ \text{pron\_form} & \text{we} \end{bmatrix} \end{bmatrix}$$

XFR **rule**:

```
subj(%X,%Subj), pred(%Subj,pron), num(%Subj,pl),
pron_form(%Subj,we)
==>
SUBJ(%X,%Subj), PRED(%Subj,we),
NTYPE(%Subj,%Ntype), NSYN(%Ntype,pronoun),
CASE(%Subj,nom), HUMAN(%Subj,+), NUM(%Subj,pl), PERS(%Subj,1),
PRON-TYPE(%Subj,pers).
```

output:

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{pron} & & & \\ \text{NTYPE} & \begin{bmatrix} \text{NSYN pronoun} \end{bmatrix} & & & \\ \text{CASE nom} & \text{HUMAN} + & \text{NUM pl} & \text{PERS } 1 & \text{PRON-TYPE pers} \end{bmatrix} \end{bmatrix}$$

Figure 9: Rewriting of the pronoun *we*

match. The input f-structure in Figure 9 matches the facts required by the XFR rule, and is rewritten to the output f-structure shown at the bottom of Figure 9. In total, the XFR system mapping from DCU to XLE f-structures consists of 162 manually coded rewrite rules that add and modify information from the DCU f-structures.

Evaluation shows that adding information to the DCU LFG f-structures by XFR rules and then evaluating these against the f-structures produced by the hand-crafted XLE grammar proves successful. However, issues remain where lexical entries would have to be listed in the XFR rules in order to make their analysis parallel to the XLE analysis, e.g. in the case of distinguishing first and last names.

An alternative approach based on extending the DCU LFG annotation algorithm directly without the intermediate step of XFR rewriting will be discussed in the following section.

3.2 Extending the DCU annotation algorithm

Instead of applying XFR rewrite rules as in Hautli and King (2009), the approach explored in this paper is to directly enrich the DCU annotation algorithm with the full feature inventory, detail and sophistication of the hand-crafted XLE grammar, as shown in Figure 10. There is no intermediate step between DCU and XLE output,

and they are compared directly. In addition, we retain a version of the original annotation algorithm (Cahill, 2004) and its feature space.

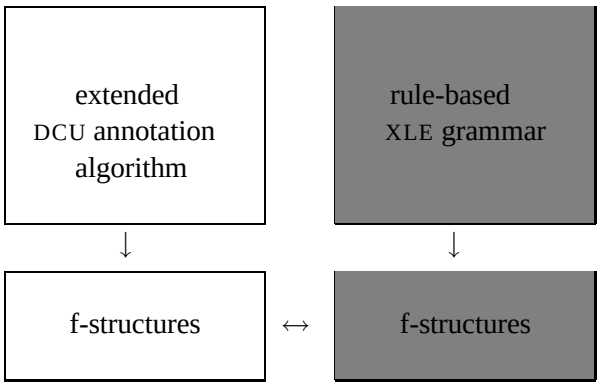


Figure 10: Schematic view of 2010 experiment

The overall DCU pipeline is not changed for this paper, only the f-structure annotation algorithm is extended, as shown in Figure 11.

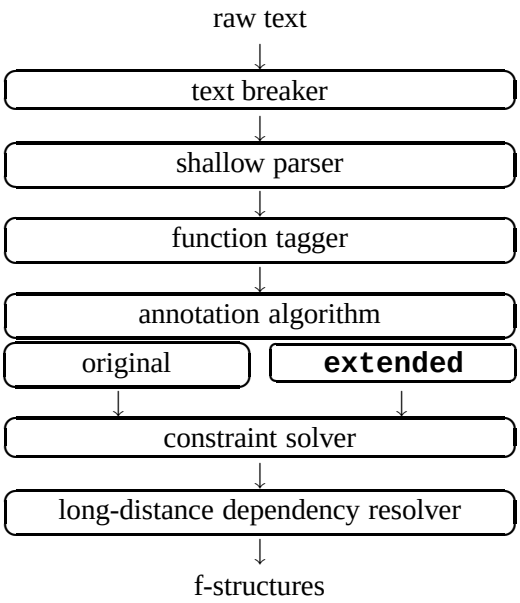


Figure 11: Extended DCU Annotation Algorithm

Apart from extending the feature space, some existing features were also renamed and their representation in the f-structure changed. To give an overview of the feature space of the DCU LFG annotation algorithm, Table 1 lists the original DCU features (36 in total) with the newly added f-structure features in bold (30 added). Besides adding features, we also extended the range of feature values to

make the DCU f-structures more informative, for instance the values ‘first_name’ and ‘last_name’ of the feature NAME-TYPE.

original features	**added** features
adegree, adjunct, aquant, comp, conj, coord-form, det-form, focus-int, mod, num, number, number-type, obj, obj-th, obl, obl-ag, obl-compar, passive, pcase, perf, poss, prog, pron-form, pron-int, pron-rel, proper, prt-form, quant, stmt-type, subj, subord-form, tense, topic-rel, xcomp	adjunct-type, adv-type, atype, case, clause-type, coord, common, deg-dim, degree, deixis, det, focus, gend-sem, human, inf-type, mood, name-type, nsem, nsyn, ntype, part, preco-ord_form, proper-type, psem, ptype, spec, time, tns-asp, vtype, xcomp-pred
extended features	

Table 1: Original and extended feature space of the DCU annotation algorithm

Figures 12 and 13 show the f-structures for *We enjoy conferencing in Ottawa.*, exemplifying tense and aspect as it is represented in the original and the extended DCU LFG annotation algorithm, respectively. The `tense` feature, which is on the top level of the f-structure in the original DCU representation, is moved to the TNS-ASP f-structure that is complemented by the aspectual features MOOD, PERF and PROG.

```
    [subj  [num pl, pred pro, pron_form we]                                          ]
    [                                                                                ]
    [xcomp [adjunct [1 [obj   [num sg, pers 3, pred Ottawa, proper location]]]]]
    [      [        [  [pform in                                                  ]]]]
    [      [pred conference, prog +, stmt_type declarative                           ]
-1  [pred enjoy, stmt_type declarative, tense pres                                   ]
```

Figure 12: Tense and aspect in the original DCU f-structures

Following the general methodology of making the DCU f-structures as closely resembling the hand-crafted XLE grammar as possible, the representation of tense and aspect in the extended DCU f-structures is now equivalent to the representation in the XLE grammar. This enables the output of the treebank- and annotation algorithm-based DCU LFG pipeline to serve as valid input to the XFR semantics. As the correct XLE-style representation of tense and aspect is crucial for getting valid semantic representations, it is necessary to capture this information as completely and precisely as possible.

Another important point, in particular for the corpora we evaluated on, was the precise XLE-style analysis of first and last names (as in *John Smith*). In order to add this information, the relation of the CFG parse tree nodes is taken into consideration. If two (or more) proper names are sisters and the lemma of the leftmost node can be found in a list of first names, we annotate it with the information that it is a first name. This is recursively done until one of the sisters to the right is not

```
[PRED     'enjoy<[-1-XCOMP:conference]>[-1-SUBJ:we]'
          [PRED  'we'
SUBJ      NTYPE [NSYN pronoun]
          CASE nom, HUMAN +, NUM pl, PERS 1, PRON-TYPE pers]
CHECK     [_SUBCAT-FRAME V-SUBJexpl-XCOMPinf]
TNS-ASP   [MOOD indicative, PERF -_, PROG -_, TENSE pres]
          [PRED     'conference'
          SUBJ      [-1-SUBJ:we]
          ADJUNCT { [PRED  'in<[-2-OBJ:Ottawa]>'
                    OBJ    [PRED  'Ottawa'
                            NTYPE [NSEM [PROPER [PROPER-TYPE location]]
                                   NSYN proper
                            CASE obl, NUM sg, PERS 3
                    -2 PTYPE sem ]}
XCOMP     CHECK     [_SUBCAT-FRAME V-SUBJ]
          TNS-ASP   [MOOD indicative, PERF -_, PROG +_]
          CLAUSE-TYPE decl, PASSIVE -, VTYPE main
-1 CLAUSE-TYPE decl, PASSIVE -, VTYPE main ]
```

Figure 13: Tense and aspect in the extended DCU f-structures

a first name, annotating this node with the feature for last names. This is a substantial improvement over the approach with XFR rewrite rules, because in addition to listing first names, last names also have to be listed. By considering node information, this can be automatically done, with middle names also being detected (e.g. *John Adam Smith*). As a result, the precision for any corpora is substantially increased, but in particular for those containing newspaper text, as is the case for the PARC700.

4 Evaluation

The evaluation covers two aspects we are concerned with: the quality of the extended DCU f-structures and the overall coverage of the DCU annotation algorithm.

With respect to the quality of the f-structures, the purpose of the evaluation is to see how closely the extended treebank- and annotation algorithm-based DCU f-structures resemble the hand-crafted XLE grammar f-structures. For this we use evaluation measures from information retrieval, namely precision ("How accurate are the extended DCU f-structures?"), recall ("How complete are the extended DCU f-structures?") and f-score, a weighted average of precision and recall. These measures are calculated on a per-feature basis, i.e. every feature of either the DCU or the XLE side is checked whether it appears on the other side. The number of found and not found features is calculated and divided by the total number of features present. The overall matching results lie between 0 (no feature matches) and 1 (all features match). The same methodology is employed when matching the semantic representations that the DCU and the XLE f-structures generate.

Concerning the evaluation of the coverage of the DCU annotation algorithm, the aim is to detect constructions where either the stochastic parser or the annotation algorithm fails to produce a valid representation.

4.1 Quality of representations

4.1.1 F-Structure matching

In the first set of experiments we measure the similarity of treebank- and annotation algorithm-based DCU and hand-crafted XLE grammar f-structures and compare our results to the architecture proposed in Hautli and King (2009) with XFR rewrite rules. Both architectures (XFR rewrite rule approach vs. extended DCU annotation algorithm) are tested on a testsuite of 720 sentences constructed by PARC and designed to test the semantic representation of the English XLE grammar.

Given the difference that the XLE grammar can produce multiple f-structures for a sentence and the DCU grammar cannot, we match the single best DCU analysis against each XLE analysis and choose the best matching result.

	Hautli and King (2009)			**extended DCU system**		
	precision	recall	f-score	precision	recall	f-score
sem_test	70.31	67.69	68.98	**83.44**	**77.22**	**80.20**

Table 2: Evaluation results for semantics testsuite

The results in Table 2 show that the features can be reconstructed successfully in both architectures. However, extending the DCU algorithm is more effective than using the set of XFR rewrite rules, because we can allow for operations that are unavailable to the rewrite approach, such as taking into account the relation of nodes in the tree, as shown for the annotation of first and last names.

In order to test the quality of the extended DCU LFG annotation algorithm more independently, we evaluate the DCU pipeline against the PARC700 gold standard which contains 700 randomly extracted sentences from Section 23 of Penn-II Treebank (WSJ section) with an average length of 19.8 words per sentence. The sentences are split into a development set (140 sentences) and a test set (560 sentences). In addition, we compare our evaluation results of the extended DCU LFG approach with the PARC700 evaluation results of the original DCU LFG approach and the XLE grammar (Cahill et al., 2008).[2] Evaluation on the test set generates the results in Table 3.

This shows that the extended DCU annotation algorithm and parsing pipeline outperform the XLE grammar and the original DCU annotation algorithm. In addition to improving the evaluation results for the extended DCU version, we also improved the f-score for the original DCU version from 82.73% in Cahill et al. (2008) to 83.88%. This is due to further refinement of the annotation algorithm.

[2]Cahill et al. (2008) only provide the f-scores for the evaluations.

PARC700			
	XLE grammar	**original DCU**	**extended DCU**
precision	—	—	85.45
recall	—	—	81.81
f-score	80.55	82.73	83.59

Table 3: PARC700 evaluation

In order to show in detail how the extended DCU system is performing, Table 4 gives the breakdown by dependency relation of the evaluation against PARC700.

Dependency	Precision	Recall	F-score	Dependency	Precision	Recall	F-score
adegree	81	79	80	pcase	91	77	83
adjunct	73	72	73	perf	95	88	92
aquant	33	77	47	poss	88	90	89
comp	69	75	72	precoord_form	0	0	0
conj	82	79	80	prog	97	75	85
coord_form	74	90	81	pron_form	91	84	88
det_form	98	98	98	pron_int	0	0	0
focus_int	0	0	0	pron_rel	68	56	62
mod	80	69	74	proper	87	90	88
num	91	89	90	prt_form	78	78	78
number	89	89	89	quant	84	73	78
number_type	95	92	94	stmt_type	90	82	86
obj	91	87	89	subj	88	69	77
obj_theta	42	45	43	subord_form	84	74	79
obl	51	70	59	tense	96	93	95
obl_ag	87	87	87	topic_rel	39	64	49
obl_compar	57	27	36	xcomp	86	78	82
passive	85	69	77				

Table 4: Breakdown by dependency relation of extended DCU annotation algorithm against PARC700

The results show that as far as the f-structure matching is concerned, the extended DCU annotation algorithm performs well. Especially for features of tense and aspect (e.g. TENSE, PROG, PERF), where recall and precision are above 95%, we get the right analysis for most sentences, except in cases where the treebank-trained CFG tree parser produces a wrong tree. The same holds for other features such as NUMBER-TYPE and DET-FORM with matching figures around 95%. Remaining issues are notorious cases such as the adjunct and oblique distinction, where only around 60% of annotations are correct, as well as the annotation of OBJ-TH with a matching precision of 43%.

4.1.2 Matching of semantic representations

To validate our claim that we can provide treebank- and annotation algorithm-based DCU LFG output that can serve as input to the XFR semantic representation, we match the semantic representations that are generated using the extended DCU LFG

f-structures as input with semantic representations based on using the f-structures generated by the hand-crafted XLE grammar as input. As a testsuite, we use the 720 sentences that were already used in Section 4.1, achieving an f-score of 73.7.

This is a promising result and although the f-score is lower compared to matching f-structures, the information we capture in the extended DCU f-structure is comprehensive enough that the semantic representations generated from them are useful.

4.2 Coverage evaluation

In addition to evaluating the quality of the extended DCU f-structures, we also measure the coverage of the extended DCU annotation algorithm (i.e. the percentage of sentences for which at least one analysis is found). By using the PARC700 gold standard, we count the number of parsed sentences, divided by the total number of sentences.

For both the training and the test set, we get full coverage. Preliminary experiments of running the DCU annotation algorithm on all of the Penn-II Treebank have resulted in coverage of above 98%.

5 Future work and conclusion

In this paper we have presented the extension of the DCU LFG annotation algorithm so that its output can serve as input to the XFR semantic representation which requires detailed f-structure information. The experiments show that the gap in the richness of the feature space and detail of the f-structure representations between the treebank-based DCU LFG approach and the hand-crafted XLE grammars can be closed, which means that there is a possibility of generating rich and deep LFG grammars on the basis of treebanks and annotation algorithms. This technique can be used not only for English but also for other languages, benefiting from the fact that the CFG trees are automatically created with a treebank-trained robust parser and have good coverage for general unrestricted text or even fragmented text such as *John. sings*.

One aspect that remains for future work is to test the coverage and performance of the extended (as compared to the original) DCU annotation algorithm on larger corpora, for instance all of the Penn-II treebank training set. This would allow for a more comprehensive assessment of the coverage of the extended DCU annotation algorithm and could also provide a large-scale rich f-structure bank that can be used for other natural language processing tools.

Another area of future work is the long-distance dependency module that has to be retrained to get more exact probabilities for the resolution of the non-local dependencies. Having more exact probabilities will most likely improve our evaluation results.

A more long-term goal is the development of an integrated DCU-XLE system, which would use the current XLE system for in-coverage sentences, but the extended DCU system for out-of-coverage and fragmenting sentences. Given the positive results for the matching of the semantic representations, this would potentially help in getting a more wide-coverage semantic representation.

The paper also shows that by extending the DCU LFG system, we can outperform the earlier approach where the DCU LFG system was employed as a black box and rewrite rules were used to modify the f-structures. The approach here results in a single DCU LFG system, which is more efficient and also has higher accuracy due to the extra information that is available by looking at the node relations in the tree. By being able to take into account tree information which previously could not be done, we allow for more linguistic insight that can be captured in the final f-structure representation.

As more researchers wish to build meaning-sensitive applications, we can contribute a robust, deep syntactic analysis that can be used for further levels of abstraction.

References

Bier, Eric, Chow, Richard, Gollé, Philippe, King, Tracy Holloway and Staddon, Jessica. 2009. The Rules of Redaction: Identify, Protect, Review (and Repeat). In *IEEE Volume on Access Control*, pages 46–53, IEEE Computer and Reliability Societies.

Bikel, Daniel M. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In *The Proceedings of HLT 2002*.

Bobrow, Daniel G., Cheslow, Bob, Condoravdi, Cleo, Karttunen, Lauri, King, Tracy Holloway, Nairn, Rowan, de Paiva, Valeria, Price, Charlotte and Zaenen, Annie. 2007. PARC's Bridge and Question Answering System. In *Grammar Engineering Across Frameworks*, pages 46–66, CSLI Publications.

Bresnan, Joan and Kaplan, Ronald M. (eds.). 1982. *The Mental Representation of Grammatical Relations*. The MIT Press.

Burke, Michael. 2006. *Automatic Treebank Annotation for the Acquisition of LFG Resources*. Ph. D.thesis, Dublin City University.

Butt, Miriam, Dyvik, Helge, King, Tracy Holloway, Masuichi, Hiroshi and Rohrer, Christian. 2002. The Parallel Grammar Project. In *Proceedings of COLING2002, Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Butt, Miriam, King, Tracy Holloway, Niño, María-Eugenia and Segond, Frédérique. 1999. *A Grammar Writer's Cookbook*. CSLI Publications.

Cahill, Aoife. 2004. *Parsing with Automatically Acquired, Wide-Coverage, Robust, Probabilistic LFG Approximations*. Ph. D.thesis, School of Computing, Dublin City University.

Cahill, Aoife, Burke, Michael, O'Donovan, Ruth, Riezler, Stefan, van Genabith, Josef and Way, Andy. 2008. Wide-Coverage Deep Statistical Parsing Using Automatic Dependency Structure Annotation. *Computational Linguistics* 34(1), 81–124.

Cahill, Aoife, Burke, Michael, O'Donovan, Ruth, van Genabith, Josef and Way, Andy. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 320–327.

Charniak, Eugene and Johnson, Mark. 2005. Coarse-to-Fine n-Best-Parsing and MaxEnt Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 173–180, Ann Arbor.

Chrupała, Grzegorz, Stroppa, Nicolas, van Genabith, Josef and Dinu, Georgiana. 2007. Better Training For Function Labeling. In *International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, pages 133–138, Bulgaria.

Crouch, Dick, Dalrymple, Mary, Kaplan, Ron, King, Tracy, Maxwell, John and Newman, Paula. 2010. XLE Documentation, http://www2.parc.com/isl/groups/nltt/xle/doc/.

Crouch, Richard and King, Tracy Holloway. 2006. Semantics via F-structure Rewriting. In *Proceedings of LFG06*, CSLI Publications.

Dalrymple, Mary. 2001. *Lexical Functional Grammar (Syntax and Semantics)*. Academic Press.

Frank, Anette, King, Tracy Holloway, Kuhn, Jonas and Maxwell, John T. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In *Proceedings of LFG98*, CSLI Publications.

Hautli, Annette and King, Tracy Holloway. 2009. Adapting Stochastic LFG Input for Semantics. In *Proceedings of LFG09*, CSLI Publications.

King, Tracy Holloway, Crouch, Richard, Riezler, Stefan, Dalrymple, Mary and Kaplan, Ronald M. 2003. The PARC700 Dependency Bank. In *Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*.

Marcus, Mitchell, Santorini, Beatrice and Marcinkiewicz, Mary Ann. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.

Maxwell, John T. and Kaplan, Ronald M. 1996. Unification-based Parsers that Automatically Take Advantage of Context Freeness. In *Proceedings of LFG96*, CSLI Publications.

Riezler, Stefan, King, Tracy Holloway, Kaplan, Ron, Crouch, Dick, III, John T. Maxwell and Johnson, Mark. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 271–278.

THE SYNTAX OF LEXICAL RECIPROCAL CONSTRUCTIONS

Peter Hurst

University of Melbourne

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King
(Editors)

2010

CSLI Publications

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

1 Abstract[1]

Many languages are able to productively form two types of reciprocal constructions from a transitive base: a monadic construction which groups the participants in the relation in the subject NP while losing an object NP; and a dyadic construction which creates a symmetric situation by placing one participant in a subject NP and another in a comitative phrase (also with the corresponding loss of an object). I show that the syntax of these constructions in Swahili (as well as in many other languages) can be understood as a natural consequence of speakers reanalysing a comitative phrase in a monadic reciprocal construction as an argument. My analysis builds upon recent work within LFG and is not only sensitive to these constructions' diachronic development from a basic transitive verb, but also provides insight into a variety of features these constructions share cross-linguistically. Furthermore, this analysis predicts the syntactic behaviour of many naturally symmetric verbs in English (such as *dance*, *quarrel etc.*).

2 Background

Many verbally marked reciprocal constructions have two forms; a basic monadic form (1b) and a dyadic alternation (1c). The examples below are from Swahili, but the monadic/dyadic alternation has been attested in many other Bantu languages (*e.g.,* Chichewa, Mchombo (1991) and Luganda, McPherson (2008)) as well as Hungarian (Rákosi (2008)), Hebrew (Siloni 2001), Icelandic, German (Dimitriadis 2004) *etc.*:

(1a) Swahili - Basic transitive construction[2]
Juma a-li-m-pig-a Halima
Juma 3sgS-PST-3sgO-hit-FV Halima
"Juma hit Halima"

(1b) Monadic reciprocal construction
Juma na Halima wa-li-pig-an-a
Juma and Halima 3plS-PST-hit-REC-FV
"Juma and Halima hit/fought each other"

(Both participants share roughly equal participation in this event.)

---

1 Thank you to Mr Hillary Kissassi, my Swahili teacher, for his aid in the collection of data presented here. For helpful comments and discussions, I would also like to thank Rachel Nordlinger, Averil Grieve and György Rákosi. All errors are mine.

2 3sgS – 3rd person singular subject, REC – reciprocal morpheme, FV – final vowel, MV – middle voice, APP – applicative.

(1c) Dyadic reciprocal construction

Juma	*a-li-pig-an-a*	*na*	*Halima*
Juma	3sgS-PST-hit-REC-FV	COM	Halima

"Juma and Halima hit each other"
lit. "Juma hit.rec/fought with Halima"

(Unlike in (1b), Juma in this situation is usually the instigator of the event and Halima's participation does not have to match his – she may be a reluctant participant, for example.)

The monadic reciprocal construction (1b) is formed by marking the verb with a reciprocal morpheme (*-an-*) and grouping the participants of the event into the subject NP. The resulting reciprocated verb no longer has an object. The dyadic construction (1c) is formed using the same reciprocal morphology on the verb, but the subject is singular and in place of the object there is now a prepositional phrase.

In previous work, Dimitriadis (2004) has convincingly argued for other languages that the comitative phrase in these constructions is not part of a discontinuous subject and that it is syntactically (and semantically) distinct from the subject NP. In the case of Swahili, despite the fact that the comitative preposition *na* shares the same form as the coordinator, the phrase *na Halima* in (1c) can be shown to be distinct from the subject by verbal agreement. Verbs in Swahili agree in number with their subject – irrespective of whether the subject is discontinuous or not. So, unlike the dyadic construction, the discontinuous form of the monadic construction has plural subject agreement:

(2) Discontinuous form of the monadic reciprocal construction

Juma	***wa**-li-pig-an-a*	*na*	*Halima*
Juma	**3plS**-PST-hit-REC-FV	and	Halima

"Juma and Halima hit each other"

That (2) is distinct from the dyadic reciprocal construction in (1c) is also demonstrated by the semantic differences between them. A dyadic reciprocal construction allows the two participants to have unequal participation in the event – however, (2) above has the same meaning as the monadic construction in (1b) where both Halima and Juma are equal participants in the event.

That the comitative phrase in (1c) is an argument (as opposed to an adjunct) is demonstrated by the fact it is obligatory and that its existence is presupposed by the verb:

(3) **Juma a-li-pig-an-a*
Juma 3sgS-PST-hit-REC-FV
*"Juma hit each other"

As far as I am aware, there is no account in LFG that unifies the productive alternation between a transitive verb with its monadic and dyadic reciprocal counterparts as occurs in Swahili. Previous analyses of the dyadic construction have for the most part been limited to European languages where the construction is restricted to a small set of verbs. Dimitriadis (2004:4) for example, bases his analysis on the concept of irreducible symmetry but he explicitly recognises that it cannot be applied to the Bantu languages. Likewise, Rákosi (2008) provides an analysis that is synchronically well-justified in the case of Hungarian monadic/dyadic constructions – but it has limited cross-linguistic applicability because it treats the verb in dyadic constructions as a basic lexeme that is no longer derived from its transitive base.

What is needed for Swahili and the other Bantu languages is an account of the monadic and dyadic reciprocal constructions that explains their productive formation from a transitive base. I argue that the dyadic reciprocal construction is formed by the reanalysis of a comitative phrase in a monadic construction as a core argument of the verb. I capture this phenomenon in LFG by building upon the work of two researchers; Rákosi (2008) and Webb (2008). As a consequence of this analysis, not only do I provide an account of the alternation of the Swahili monadic and dyadic constructions above with respect to their transitive base, but in section 4 I also show how several intriguing cross-linguistic properties of these constructions are predicted, including:

1. Why the dyadic reciprocal construction is typically formed from a comitative phrase.

2. Why these constructions have subtly different semantics with respect to the participation of the entities in the symmetric relation.

3. Why the clitic marked reciprocal construction in Romance languages, and periphrastic reciprocal construction in English (i.e., *they saw each other*) are not able to form the dyadic reciprocal construction.

4. Why there exist in English a small set of so called "naturally symmetric" verbs (such as *dance*, *argue*, *fight etc.* – see Kemmer 1993) which despite not being marked with reciprocal morphology still allow the formation of a dyadic construction.

3 Analysis

In order to have cross-linguistic applicability, this analysis of the monadic and dyadic reciprocal constructions in Swahili is built upon a new understanding of comitative phrases. First though, I present an analysis of the monadic construction as being formed from a transitive base. The dyadic construction is in turn formed from the monadic reciprocal construction:

lexical base →	monadic construction →	dyadic construction
(1a) *pig-a*	(1b) *pig-an-a*	(1c) *pig-an-a* *na* ...
hit	hit.rec	hit.rec / fight with

3.1 Formation of the monadic reciprocal construction

This analysis of the Swahili monadic construction ultimately draws from work by Alsina (1996) in his account of reciprocal constructions in Catalan. However, I differ from him by using the formalism introduced by Rákosi (2008).[3] The difference between these analyses is that Alsina maps two argument slots to one grammatical function at the level of f-structure, whereas Rákosi uses a process of "argument unification" to bundle two arguments into a single slot at the level of a-structure, and this slot is subsequently mapped to f-structure via the usual mapping principles. This process occurs at the level of the lexicon – and can be understood as one that takes a lexeme as its input, and produces another lexeme as its output. For example, the a-structure of the verb *piga* – "hit" in (1a) above is:

(4) *piga* < [P-A][P-P] >[4]

After undergoing argument unification, the two argument roles are assigned a single argument slot. This produces the new lexeme in (1b) above, *pigana* – "hit_each_other", which like other lexemes forms a part of the mental lexicon of the speaker:

(5) *pigana* < [[P-A][P-P]] >

The outermost square brackets above indicate that the argument bundle of proto-agent and proto-patient have been unified to a single argument slot.

For the mapping between a-structure and f-structure, I follow Rákosi in

3 Unlike Rákosi (2008), who forms the monadic reciprocal construction from the dyadic construction, I will use this analysis to explain how the monadic construction is formed from a transitive base.

4 I follow Alsina (1996) in using Dowty's (1993) analysis of thematic roles as belonging to two prototypical categories: proto-agent (P-A) and proto-patient (P-P).

employing Ackerman's (1992) version of lexical mapping theory. Ackerman developed his version of mapping theory to incorporate Dowty's concept of proto-roles.[5] For reasons of space, I will only briefly outline how Ackerman's theory differs from the LMT as described by Bresnan (2001):

Firstly, arguments are intrinsically classified by assigning them one of two possible syntactic role features, [-r] or [-o]. The arguments are classified by the following principles:

(1) The argument with the most heavily weighted proto-patient properties is intrinsically classified as [-r].
(2) The argument with the most heavily weighted proto-agent properties is intrinsically classified as [-o].
(3) All other arguments are intrinsically classified as [-o].

Ackerman (1992:64)

Secondly, arguments are assigned a default classification where the highest ranking argument receives a [-r] feature and all other arguments receive [+r], unless they have already received an intrinsic classification of [-r]. The arguments are then mapped to grammatical functions according to the standard principles of *function-argument bi-uniqueness* and the *subject condition* (see Bresnan 2001:311 and the references contained within). For example, consider the mapping between arguments and grammatical functions for the transitive verb *piga* - "hit" and monadic reciprocal verb *pigana* - "hit_each_other":

Transitive *hit* from (1a)					Monadic *hit.rec* from (1b)	
piga <	[P-A][P-P] >		→	*pigana* <	[[P-A][P-P]] >	
intr.	-o	-r		intr.	-o	
def.	-r			def.	-r	
bi-uniq.	SUBJ	SUBJ/OBJ		bi-uniq.	SUBJ	
result:	SUBJ	OBJ		result:	SUBJ	

The resulting analysis of the monadic reciprocal construction predicts that it be intransitive, where the verb at the level of f-structure selects only a subject. This prediction is borne out in (6) below where the clause is ungrammatical because its associated f-structure violates the coherence condition (see Bresnan 2001:63):

(6) **Juma na Halima wa-li-pig-an-a* *Fatuma*
Juma and Halima 3plS-PST-hit-REC-FV Fatuma
"Juma and Halima hit each other Fatuma"

5 Rákosi (2008) provides a very clear summary of the important features of Ackerman's (1992) work, some of which are repeated here.

The analysis I present here for the formation of the monadic reciprocal construction is similar to that of Rákosi's – the key difference being that it is applied to a transitive verb rather than one in a dyadic reciprocal construction (note in section 3.2.2 below I will return to Rákosi's work on Hungarian and show how it can be encapsulated in this analysis of the dyadic reciprocal construction). Insofar as this analysis presents the monadic reciprocal construction as being formed via a lexical operation on a transitive base it is akin to the work of Siloni (2008) and Reinhart and Siloni (2005).

3.2 Formation of the dyadic reciprocal construction

The most striking feature of dyadic reciprocal constructions from a cross-linguistic perspective is their formation from a comitative phrase. In fact, as we will see in section 4, many of the unusual properties that dyadic reciprocal constructions share in the world's languages actually arise as a consequence of their relationship with comitative phrases. The analysis of the dyadic reciprocal construction presented is able to explain these phenomena because it builds upon a new understanding of comitative phrases in reciprocal constructions which extends Webb's (2008) work on instrument phrases. In section 3.2.1 below I first present my analysis of comitative phrases in LFG before turning to the analysis of the Swahili dyadic reciprocal construction in section 3.2.2.

3.2.1 Analysis of typical comitative phrases

A typical comitative phrase adds an optional entity to an event. This comitative entity participates in the event in a manner very similar to another participant in the event carrying a thematic role. For example, in (7) below, the PP "with Abel" is a comitative phrase and Abel participates in the event of *catching* in a manner similar to that of Cain:

(7) *Cain caught fish* <u>*with Abel*</u>

The terminology I use here is that the comitative entity (Abel above) is *linked* to another participant in the event. In this case Abel is linked to Cain and his participation in the event is similar to that of Cain's. A further examination of comitative phrases like those in (7) above reveal that they typically have properties of both adjuncts and arguments:

Adjunct properties

- Comitative phrases are optional and their presence is not part of the verb's meaning. In "*Cain ate (spam) (*<u>*with Abel*</u>*)*" the patient's meaning is implied, even if not present, whereas the comitative's meaning is wholly optional.

Argument properties

- Comitative phrases participate in the event described by the verb – acting with another entity, and upon another entity, if present. For example, in "Cain caught fish <u>with Abel</u>", Abel has very similar participation in this event as Cain.
- Comitative phrases are restricted in their number – unlike adjuncts. So in (8) below, only one comitative phrase can appear after the verb whereas in (9) many adjuncts may modify an event:

(8) **Cain caught fish <u>with Abel</u> <u>with Eve.</u>*

(9) *Cain caught fish <u>in summer</u> <u>in the mornings.</u>*

As comitative phrases share properties of both adjuncts and arguments, how should they be analysed? In fact, an analysis of phrases which sit in the middle of the argument/adjunct continuum has been explored by many linguists but most notably by Grimshaw (1990). She defines the term an "a-adjunct" (argument/adjunct) as a phrase which has an intermediate status between an argument and an adjunct and analyses it as:

(i) not being assigned a theta-role by the verb.

(ii) being licensed by the argument structure of the verb.

(adapted from Grimshaw 1990:108)

Her first property accounts for why the existence of comitatives is not presupposed by the verb when they are not present (and hence their adjunct-like optionality). The second property accounts for why the number of comitatives is restricted.

Webb (2008) analyses instruments as a-adjuncts and proposes that verbs can have two tiers of argument structure; arguments on the first tier and a-adjuncts on the second tier. I propose that comitative phrases are also usually a-adjuncts and so should be analysed in a manner parallel to Webb's analysis of instruments. For example, in (10) below, the 1st tier argument structure has two slots for the two thematic roles selected by the verb *catch*. However, an optional 2nd tier of argument structure is also defined for the a-adjuncts:

(10) *Cain caught fish with Abel*
catch: 1st tier < [P-A][P-P] > 2nd tier < [ ] >

Note that in Webb's analysis, the instrument a-adjunct receives the theta-role of agent. I depart from his analysis here and follow Grimshaw (1990) in supposing that comitative a-adjuncts are not directly assigned a theta-role. Under the system of proto-roles adopted here (from Dowty 1993), this in turn means that the comitative phrase is unspecified for a thematic role – instead

its participation in the event is understood to be the same as (or similar to) that of another entity.[6] I represent this below using empty square brackets [ ]. This analysis of a comitative entity as having an argument description of [ ] ends up being equivalent to that which Rákosi gives for a *partner* thematic role in his analysis of the dyadic reciprocal construction in Hungarian (see Rákosi 2008:424 and the references contained within).

The mapping from argument structure to f-structure occurs via the standard mapping rules described above, but with the provision that the arguments in the first tier are mapped before any of the a-adjuncts in the second tier:

catch:	1st tier < [P-A]	[P-P] >	2nd tier < [ ] >
intrinsic	-o	-r	-o
default	-r		+r
bi-uniq.	SUBJ	SUBJ/OBJ	OBL
result:	SUBJ	OBJ	OBL

Below I outline some of the predictions this analysis makes which will have relevance to my analysis of dyadic reciprocal constructions.

3.2.1.1 Comitative phrases as arguments

The first implication of this analysis is that we might expect to find verbs which select a comitative phrase as an argument rather than as an a-adjunct, especially given that a-adjuncts already appear at a lower tier of a-structure. In this situation, the comitative entity is part of the verb's first tier of argument structure – making it a prototypical argument. This is seen in English with the verb *quarrel*:

(11) *Bob quarrelled with Fiona.*

quarrel:	1st tier < [P-A]	[ ] >
intrinsic	-o	-o
default	-r	+r
bi-uniq.	SUBJ	OBL
result:	SUBJ	OBL

Note that this definition of *quarrel* is precisely that which Rákosi (2008:444) gives for *veszekedik* - "quarrel" in Hungarian. This analysis also raises the possibility of comitative phrases appearing as both a 1st tier argument and a 2nd tier a-adjunct. Rákosi (2003) observes that verbs which

6 Although a comitative entity is typically linked to the highest thematic role the verb selects, this does not have to be the case. Consider "The man was beaten up on the bus with his wife". See Rákosi (2006:107-112) for further discussion.

treat the comitative phrase as being an argument optionally allow a comitative a-adjunct as well – a possibility that is not available to verbs which do not select a comitative phrase as an argument:[7]

(12a) **Péter-rel (együtt) ritkán fut-ott-am Kati-val*
Peter-with together rarely run-pst-1Sg Kate-with
*"I rarely ran with Kate with Peter"

(12b) *Péter-rel (együtt) ritkán veszeked-t-em Kati-val*
Peter-with together rarely quarrel-pst-1Sg Kate-with
"I rarely quarrelled with Kate, together with Peter"
Rákosi (2003:3)

Under this analysis of comitative phrases, the verb *veszekedtem* – "quarrel" in (12b) now selects a comitative entity in both 1st and 2nd tier argument structure:

veszekedtem	1st tier <	[P-A][ ] >	2nd tier <	[ ] >
intrinsic		-o -o		-o
default		-r +r		+r
bi-uniq.		SUBJ OBL		OBL
result:		SUBJ OBL_{arg}		$OBL_{a\text{-}adj}$

3.2.1.2 Unequal participation in the event

In the analysis of comitative phrases above, we see that the comitative entity does not receive the same thematic role as that of the entity to which it is linked. This means we might expect that the comitative entity is able to engage in the event in a manner different to that of its linked entity. This behaviour is in fact observed in the day-to-day usage of comitative phrases. For example, consider Maria's participation in the event in (13) below:

(13) *Maria selected photos for the gallery with Mark.*

Although it is possible Maria and Mark have identical participation in the event, my understanding is that Mark is more likely to be providing assistance to Maria – perhaps looking over her shoulder. Fortunately, we do not need to rely on intuitions alone to argue that there are differences between these comitative entities and the entities to which they are linked – for a restricted set of verbs, these intuitions can be formalised. For example, consider the verb "collide":[8]

7 An analogous example in English is "Michel fought [with Satan]$_{Arg}$ [with God on his side]$_{A\text{-}Adj}$"

8 Also *dance* – "Don danced with the broomstick", *fight* – "The fishermen fought with the elements" and others.

(14a) *The car collided with the tree.*
(14b) *#The car and the tree collided.*
(14c) *##The tree collided with the car.*

It has been noted by many authors (*e.g.,* Rákosi (2008), Dimitriadis (2004), Gleitman *et al.* (1996), Dowty (1993) *etc.*) that sentences like (14b,c) potentially throw doubt upon any analysis where the car and the tree must participate identically in the predicate. Simply put, if both participants were receiving identical thematic roles from the predicate, we would expect that the participants could be swapped with little or no change in meaning – and clearly that is not the case. Dimitriadis offers a counterargument to this point, arguing that the differences in acceptability of (14b) and (14c) are caused by "structural differences between the two argument positions" (Dimitriadis 2004:37). The evidence he uses to support this view comes from Gleitman *et al.* (1996) who argue that even for predicates which must logically assign the same thematic roles to two participants, there are still differences in interpretation which seem dependent upon structural factors. For example Gleitman *et al.* (1996) argue that both China and North Korea must have the same thematic roles in (15) below – and so the differences in interpretation between (15a) and (15b) are due to structural differences:

(15a) *China is similar to North Korea.*
(15b) *North Korea is similar to China.*

(from Dimitriadis 2004, based upon examples from Gleitman *et al.* 1996.)

Here, the *ground* (the oblique argument) is the source of comparison, and the *figure* (the subject) is compared to it. For example, in (15b) North Korea might be similar to China insofar as, like China, it has a communist government. However, while structural effects might play some role in the contrast in meaning between two structurally distinct arguments, the change in acceptability of sentences like (14a) and (14b,c) above is factored around agency – *i.e.*, it is a thematic distinction. Consider again the contrast in meaning between (14a) and (14c). Dowty (1993) points out that changes in context change the admissibility of the sentences. For example, imagine the following contexts: the tree has been torn from the ground by a hurricane and slams into a stationary car, or the tree is animate (such as in a children's story) and while running through the forest, trips over an unseen car. In these contexts (14c) – "The tree collided with the car" is now the appropriate description of the event. Examining the contexts in more detail we see that the acceptability of (14c) is dependent upon whether the subject NP is moving and/or animate. These are both properties that Dowty argues contribute to the characterisation of an agent proto-role (Dowty 1993:572). In other words, the tree and car in (14) have differing degrees of agency, and so

cannot carry identical thematic roles.

3.2.1.3 Concluding remarks on the comitative construction

This analysis of comitative phrases parallels that of Webb's analysis of instruments. The typical comitative phrase is an a-adjunct which is unspecified for a thematic role by the verb but whose presence is licensed in a 2nd tier of argument structure. With respect to the analysis of the dyadic reciprocal constructions, two properties of this analysis are of particular importance:

1. That some predicates can select a comitative phrase as an argument rather than an a-adjunct. This requires that the comitative entity be selected in the first tier of argument structure rather than the second.
2. That the comitative entity need not have identical participation in the event as that of the entity to which it is linked. This is as a result of the comitative entity not being assigned a thematic role, rather its participation in the event is inferred.

3.2.2 Analysis of the dyadic reciprocal construction in Swahili

I propose that in Swahili a comitative a-adjunct in the monadic reciprocal construction has been reanalysed as an argument. Recall that Swahili speakers productively form the monadic reciprocal construction from a transitive base:

Transitive *visit*		Monadic *visit.rec*
tembelea < [P-A][P-P] >	→	*tembeleana* < [[P-A][P-P]] >

A verb in a monadic construction, like other verbs, can take a comitative a-adjunct. For example, in (16) the comitative phrase *na Daktari* – "with the doctor" is an a-adjunct and the a-structure of the verb *tembeleana* – "visit" is given below:[9]

(16) *Juma na Halima wa-li-tembel-e-an-a na Daktari*
Juma and Halima 3plS-PST-visit-APP-REC-FV COM doctor
"Juma and Halima visited each other, with the doctor"

Monadic *visit.rec* with a-adjunct
tembeleana: 1st tier < [[P-A][P-P]] > 2nd tier < [ ] >

Now consider the motivations for the reanalysis of the comitative a-

9 Unless given a specific context, this adjunct reading of (16) is now dispreferred in favour of the argument reading of the comitative phrase where the doctor visited Juma and Halima and they visited the doctor.

adjunct as an argument. The comitative entity adds a participant to the event which engages in the same activity as another entity. Symmetric and lexically derived reciprocal verbs have an interpretation that is very similar in its entailments – two entities engage in an event with the same participation, but because reciprocal verbs are symmetric, the verb now *requires* the presence of a partner (*e.g.,* for the verb *hit* in (1) above, both participants are hitting, they are both being hit, and they are necessarily engaged in the event together). Crucially then, those verbs which have an interpretation that is symmetric at the level of the lexicon allow the possibility for a comitative a-adjunct to be reanalysed as an argument in the 1st tier of argument structure.

Monadic *visit.rec* with a-adjunct			Dyadic *visit.rec*
tembeleana:	1st tier < [[P-A][P-P]] >	→	1st tier < [[P-A][P-P]] [ ] >
	2nd tier < [ ] >		

Given that comitative phrases may already appear as 1st tier arguments, it is unsurprising that in the case of symmetric verbs a comitative a-adjunct can be reanalysed in this way. Now that the comitative entity is an argument of the verb, the subject NP no longer requires two participants as the symmetric relation occurs between the subject NP and the comitative phrase:

(17) *Juma a-li-tembel-e-an-a na Halima*
Juma 3sgS-PST-visit-APP-REC-FV COM Halima
"Juma and Halima visited each other"

This representation of reciprocal verbs in Swahili was built upon the notion that speakers reanalyse a comitative a-adjunct in a monadic construction as being an argument. The resulting form of the dyadic reciprocated verb correctly predicts the syntax of these constructions. For example, the a-structure of the dyadic verb *pigana* - "hit_each_other" from (1c) maps the comitative entity to an oblique argument:

(18) Dyadic *pigana* - "hit each other with" from (1c)

pigana:	1st tier < [[P-A][P-P]] [ ] >	
intrinsic	-o	-o
default	-r	+r
bi-uniq.	SUBJ	OBL
result:	SUBJ	OBL

(1c) *Juma a-li-pig-an-a na Halima*
Juma 3sgS-PST-hit-REC-FV COM Halima
"Juma and Halima hit each other"
lit. "Juma hit.rec/fought with Halima"

Note that this final analysis of the Swahili dyadic construction is able to provide a diachronic explanation for Rákosi's account of the same construction in Hungarian. He argues that the dyadic reciprocal verb is basic, or underived, with the monadic form of the verb being formed from the dyadic base through the process of argument unification described above:

<u>Rakosi's (2008) analysis for *veszekedik* - "quarrel"</u>
basic dyadic lexeme monadic form

veszekedik < [P-Agent] [ ] >→ < [[P-Agent] [ ]] >

The most convincing evidence for his position is that the transitive counterpart of many reciprocal verbs in Hungarian no longer exists or has a meaning which is no longer related to the reciprocal form of the verb. The limitation of this account is that it cannot (nor seeks to) explain the origin of the dyadic verb and why it has a reciprocal morpheme. The analysis I present here shows how this situation comes about. Recall that the reanalysis of a comitative phrase as being an argument must occur at the level of the lexicon. From a purely formal point of view, the dyadic reciprocation operation takes a symmetric lexeme as its input and creates a new lexeme from it which selects a comitative entity as an argument:

$\text{Verb}_{\text{rec,monadic}}$ 1st tier < [[P-A][P-P]] > → $\text{Verb}_{\text{rec,dyadic}}$ < [[P-A][P-P]] [ ] >
2nd tier < [ ] >

As these dyadic verbs are prototypically used with just two participants, the complex structure of the first argument slot is a good candidate for simplification during a process of lexicalisation:

$\text{Verb}_{\text{rec,dyadic}}$ < [[P-A][P-P]] [ ] > → $\text{Verb}_{\text{rec,dyadic,lexicalised}}$ < [P-A] [ ] >

This final version of the verb is now a bonafide new lexeme whose new a-structure reflects that fact that speakers no longer consider it productively formed from a transitive base.

3.3 Remarks

In this section I have argued that a comitative a-adjunct attached to a reciprocated verb has been reanalysed as an argument. This reanalysis stems from the fact that a reciprocated verb meets all the preconditions for productive formation with a comitative entity, but with the additional condition that the comitative entity no longer be optional. I have extended the work of Rákosi (2008) and Webb (2008) to capture this phenomenon and I have shown that this analysis correctly predicts the syntax of the monadic and dyadic reciprocal constructions and the relationship they have with their

transitive base. However, the process I have described above is not specific to just Swahili. In section 4 below I examine the implications this analysis has from a cross-linguistic perspective and show that many properties of dyadic reciprocal constructions can be now understood.

4 Cross-linguistic implications

4.1 Why the dyadic reciprocal construction is typically formed from a comitative phrase

The analysis of dyadic reciprocal constructions presented above is built upon the notion that a comitative a-adjunct is reanalysed as an argument. Under this analysis, we would expect that a similar operation could occur in different languages. More specifically though, it predicts that the dyadic reciprocal construction be formed from a comitative phrase and overwhelmingly, this is the case:

(19) Greek (Hellenic) - Dimitriadis 2004:25
O Yanis filithike me ti Maria
the John kissed-REC COM the Maria
lit. "John kissed each other with Maria"
"John and Maria kissed (each other)"

(20) Hebrew (Afro-Asiatic) - Siloni 2008
Ha-yeladim hitnaešku im ha-yeladot
the-boys kissed.REC COM the-girls
lit. "The boys kissed each other with the girls"
"The boys and girls kissed each other"

(21) Hungarian (Uralic) - Rákosi 2008:443,444
A katoná-k vesz-eked-t-ek az őrmester-rel
the soldier-pl quarrel-REC-PST-3pl the sergeant-COM
"The soldiers quarrelled with the sergeant"

(22) Icelandic (Germanic)
Þeir börðu-st við víkinga
3pl.m.nom fight.pl.PST-MV COM viking.m.pl.acc
"They fought (from *berja* - 'hit') with the vikings"

4.2 That the dyadic reciprocal construction is formed from the monadic construction

Dimitriadis (2008, 2004) and Rákosi (2008) argue that the monadic

reciprocal construction is formed from its dyadic equivalent whereas the analysis above predicts the opposite. Now, in the literature I have examined, the monadic reciprocal constriction is either more prevalent than its dyadic counterpart, or equally prevalent. Recall that both these constructions contain reciprocal morphology, so any analysis which posits that the dyadic construction is the base of formation for the monadic construction needs to provide an additional mechanism to account for the existence of the monadic constructions which have no dyadic counterpart. I think this is an unlikely historical path for the formation of monadic reciprocal constructions.

4.3 Weakening of symmetric semantics

In this section I discuss the implications this analysis of dyadic reciprocal constructions has on their associated semantics. In particular, I contrast the predications it makes with Dimitriadis' (2004) work on irreducible symmetry and show that some unexpected data revealed by Rákosi (2008) can be explained by the analysis presented here.

4.3.1 Irreducible symmetry

Dimitriadis' notion of irreducible symmetry and its relation to dyadic reciprocal constructions can be understood as the requirement that the symmetric event be conceived as being atomic and that the symmetry be absolute insofar as the entities are required to have identical participation within the event. A predicate has the property of irreducible symmetry if:

(a) it expresses a binary relationship, but
(b) its two arguments have necessarily identical participation in any event described by the predicate.

(Dimitriadis 2008:378)

For example, for the predicate *kiss* to be irreducibly symmetric requires that both participants kiss is the same way in a single event. Situations where one participant has a different participation in the event are ruled out, as are situations which are only symmetric when seen as a collection of events (*e.g.,* Henry kissed Sally on the hand at night, and Sally kissed Henry on the cheek at dawn). Under the analysis presented here, the dyadic construction is conceived of as arising from a reanalysis of a comitative a-adjunct as a comitative argument. What semantic features might we expect in the resulting construction – and are they compatible with Dimitriadis' conception of irreducible symmetry? I characterised the comitative construction as introducing an entity which participates in the same event as the entity to which it is linked, but optionally with a less active participation. The implications such an analysis has on our understanding on the formation of dyadic reciprocal constructions revolve around two concepts; atomic events

and symmetry.

4.3.2 Atomic events

An atomic event is one where the participants engage in single event. For example, in (23) below there is one event of cooking, and Olaf is doing it:

(23) *Olaf cooked some fish.*

With a plural entity, we can often conceive of an event in two ways, as being atomic or non-atomic. For example the event of cooking below can be either atomic (24a) or non-atomic (24b):

(24a) *Olaf and Sally cooked some fish together.*
(24b) *Olaf and Sally cooked some fish, Olaf in the morning and Sally in the evening.*

Comitative phrases introduce an entity to the event which is linked to another participant. Above I argued that this linkage does not require that the two entities have identical participation in the event, however, it does usually require an atomic conception of the event. For example, like (24a) above, the comitative construction in (25) requires an atomic interpretation of the event:

(25) *#Olaf cooked some fish with Sally, Olaf in the morning and Sally in the evening.*

With respect to reciprocity, unlike analytic (or phrasal) reciprocal constructions, most monadic reciprocal constructions describe atomic events because they are formed at the level of the lexicon. The addition of a comitative entity further hardens this tendency because it is a participant which shares in an atomic event. This view is in line with Dimitriadis – and is supported by the massive amounts of linguistic evidence he provides.

4.3.3 Symmetric participation

Symmetric participation occurs when every participant in a relation acts as both the initiator and endpoint of that relation. For example, we can say that Olaf and Sally have symmetric participation in the event of seeing if we can say: "Olaf saw Sally and Sally saw Olaf". Symmetric participation is independent of whether the event is atomic or not. For example, if Olaf saw Sally in the morning, and Sally saw Olaf in the evening, their participation in the relation of seeing is still symmetric. Given my analysis of comitative entities, how might symmetry be affected in dyadic reciprocal constructions? I argued above that comitative entities do not need to have identical participation in the event with the entity to which they are linked – and often have less agentive properties. Therefore we might expect dyadic

constructions to have weaker symmetric entailments between the two participants. This predication is at odds with Dimitriadis (2008, 2004) who argues that participation of the two entities in a dyadic construction must be identical. However, my prediction is in line with the data where we do see this weakening of symmetry – even to the point of allowing the symmetry to be cancelled. For example, Rákosi (2008:423) illustrates this very clearly with the following example:

(26) *Én num veszeked-t-em János-sal ő veszeked-ett vel-em*
I not quarrel-pst-1sg John-with he quarrel-pst with-1sg
"I was not quarrelling with John, he was quarrelling with me"

I have found similar examples for Icelandic (*slást* – "brawl" from the middle voice of *slá* - "hit") and this pattern is particularly clear in Swahili where the symmetry can be cancelled for most verbs in the dyadic reciprocal construction. For example:

(27) Juma a-na-pend-an-a na Halima,
Juma 3sgS-pres-love-rec-fv with Halima
lakini Halima ha-m-pend-i Juma
but, Halima neg-3sgO-love-fv Juma
"Juma loves each other with Halima,
but Halima does not love Juma"

4.3.4 Concluding remarks

With respect to whether the symmetric situation should be seen as an atomic event or as a collection of events, the analysis I present is in line with work by Dimitriadis in predicting that dyadic reciprocal constructions be atomic. However, contra Dimitriadis, I predict that the dyadic reciprocal construction can have looser entailments with regard to symmetry when compared to its monadic equivalent. This prediction is borne out in some situations in Swahili and Hungarian where there are examples of the comitative entity not participating in the symmetric event to the same extent of the subject entity.

4.4 Lexical versus syntactic reciprocal constructions

Siloni (2008) notes that French, although having a superficial resemblance to the Swahili and Hungarian reciprocal constructions, does not allow the formation of a dyadic reciprocal construction:

(28) **Jean s'-est embrassé avec Marie.* (From Siloni 2008:482)
Jean se-is kissed with Marie
*"Jean kissed each other with Marie"

Also, English:

(29) **John saw each other with Marie.*

The analysis I presented for the dyadic construction occurs at the level of the lexicon where a comitative entity associated with a verb in a monadic reciprocal construction is reanalysed as an argument. Because the reciprocal construction in English is formed via a phrase, the symmetric sense of the situation is introduced at the level of syntax – not at the level of the lexicon. This means that the predicate in a reciprocal construction (*see* in (29) above) is asymmetric at the level of the lexicon and consequently the comitative phrase is not able to be reanalysed as an argument. Given that the dyadic construction in French is ungrammatical, it would imply that the French reciprocal construction is also formed at the level of syntax. Siloni (2008) investigates these constructions in detail and this is her conclusion too. The evidence she provides is based upon whether a reciprocal event can be understood as a collection of asymmetric events or not. Her argument is that lexemes can only denote atomic events, and the pluralisation of events must occur at the level of syntax. Examples like (30) below demonstrate that the French clitic reciprocal construction can describe a collection of asymmetric events in the context of a game where Jean kisses Marie five times and Marie kisses Jean five times (i.e., there were ten separate kisses):

(30) *Jean et Marie se sont embrassés cinq fois.*
Jean and Marie se are kissed five times
(From Siloni 2008:481)

Contrast this with the Swahili lexical reciprocal construction where the only possible reading is that there were three shared kisses:

(31) *Juma na Halima wa-li-bus-i-an-a mara tatu*
Juma and Halima 3plS-PST-kiss-APP-REC-FV times three
"Juma and Halima kissed each other three times"

The analysis I present here fits with Siloni (2008) and Reinhart and Siloni's (2005) work on reciprocals which treats the French reciprocal construction as being syntactically derived, and which associates the dyadic reciprocal construction with an operation that occurs at the level of the lexicon.

4.5 Naturally Symmetric Verbs

In English there exist a set of "naturally symmetric" verbs (such as *quarrel, argue etc.*) which appear to participate in a dyadic construction,

despite not having any reciprocal morphology. For example, consider (32) below:

(32a) *Bill and Mary argued/quarrelled/danced.*
(32b) *Bill argued/quarrelled/danced with Mary*

This alternation is in fact predicated as a consequence of the formation of the dyadic construction. Recall that this construction takes a symmetric verb as its base and reanalyses an associated a-adjunct as an argument. Thus far we have examined reciprocated (monadic) verbs which became symmetric as a result of their formation. However, some verbs are already symmetric as basic lexemes (such as *quarrel etc.*). These verbs then are suitable candidates for participation in a dyadic construction because, like the derived symmetric verbs in Swahili, they can serve as a base for the dyadic lexical operation.

5 Conclusion

The motivation for this analysis of the Swahili reciprocal constructions is driven by a reanalysis of a comitative a-adjunct as an argument. Because comitative constructions form part of the a-structure associated with a predicate, this reanalysis necessarily occurs at the level of the lexicon. Additionally, in order for the comitative a-adjunct to be reanalysed as an argument, a candidate predicate requires the participation of two entities and that these entities must engage in a relationship which is compatible with comitative semantics. Both naturally symmetric verbs and symmetric verbs that are formed in the lexicon fulfil these criteria; a symmetric verb requires (prototypically) two participants with similar participation in the event (recall for the Swahili verb *pigana* – "hit_each_other", both participants are hitting, they are both being hit, and they are engaged in the event together). This understanding of comitative a-adjunct reanalysis predicts the central features that lexically derived reciprocal constructions share cross-linguistically:

- Why the entity represented by the oblique grammatical function is typically marked as being a comitative phrase.
- Why these constructions are so strongly associated with high degrees of symmetry within a single event.
- Why languages with reciprocal constructions formed in syntax are not compatible with the dyadic reciprocal construction.

Finally, this explanation provides a natural account for why other symmetric verbs which are not marked with reciprocal morphology (*e.g., quarrel, bicker etc.* in English or *ongea* - "chat" in Swahili) allow the same alternations between a monadic and dyadic form.

6 References

Ackerman, F. 1992. Complex predicates and morphological relatedness: locative alternation in Hungarian. In Sag, I. & Szabolcsi, A. (eds.), *Lexical Matters*, 55-83. Stanford: CSLI Publications.

Alsina, A. 1996. *The role of argument structure in grammar: Evidence from Romance.* Stanford: CSLI Publications.

Bresnan, J. 2001. *Lexical-functional syntax.* Blackwell Pub. Mass.

Dimitriadis, A. 2008. *Irreducible symmetry in reciprocal constructions.* In König & Gast (eds.) 2008.

Dimitriadis, A. 2004. *Discontinuous reciprocals*. Manuscript: Utrecht Institute of Linguistics OTS.

Dowty, D. 1993. Thematic proto-roles and argument selection. *Language* 67.3: 547-619.

Gleitman, L., Gleitman, H., Miller, C. & Ostrin, R. 1996. Similar, and similar concepts. *Cognition* 58: 321-376.

Grimshaw, J. 1990. *Argument structure.* Cambridge, MA: The MIT Press.

Kemmer, S. 1993. *The middle voice.* John Benjamins. Amsterdam/Philadelphia.

König, E. & Gast, V. (eds.) 2008. *Reciprocals and reflexives: theoretical and typological explorations.* Mouton de Gruyter: Berlin, Germany.

Mchombo, S. 1991. Reciprocalization in Chichewa: A lexical account. *Linguistic Analysis* 21:3-22 1991.

McPherson, L. 2008. *A descriptive and theoretical account of Luganda verbal morphology*. BA thesis, Scripps College. Available online at http://www.lcs.pomona.edu/McPherson_thesis.pdf.

Rákosi, G. 2008. *The argument structure of inherently reflexive and reciprocal predicates*. In König & Gast 2008.

Rákosi, G. 2006. *Dative experiencer predicates in Hungarian.* PhD. Dissertation. Utrecht. Published as volume 146 of the LOT Dissertation Series. http://www.lotpublications.nl/publish/issues/Rakosi/index.html

Rákosi, G. 2003. Comitative arguments in Hungarian. In: Heeren, Willemijn, Dimitra Papangeli and Evangelia Vlachou (eds.), *UiL-OTS Yearbook 2003*, 47-57. Utrecht: Utrecht Institute of Linguistics OTS.

Reinhart, T. & Siloni, T. 2005. The lexicon-syntax parameter. *Linguistic Inquiry* Vol 36, Num 3, 2005.

Siloni, T. 2008. *The syntax of reciprocal verbs: an overview.* In König & Gast 2008.

Siloni, T. 2002. Active lexicon. *Theoretical Linguistics,* 28:383-400.

Siloni, T. 2001. Reciprocal verbs. In: Falk, Y. (ed.), *Proceedings of the Israel Association of Theoretical Linguistics.*

Webb, J. 2008. *Instruments in LFG's argument structure.* Manuscript: Master's thesis, University of Oxford. Available online at: http://ucl.academia.edu/documents/0009/9899/Webb__MPhil_Thesis.pdf

RIGHT NODE RAISING IN PARSING AND GENERATION

Jonas Kuhn, Christian Rohrer and Sina Zarrieß

Universität Stuttgart

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

We discuss a grammar engineering account of non-constituent coordination in the clausal syntax of German as it occurs in real corpus data, generalizing the approach of Forst and Rohrer (2009) to right node raising cases not captured so far. We compare our account to the unimplemented metagrammatical technique by Maxwell and Manning (1996) and point out that many of their ideas can be captured using only the standard LFG formalism and making minimal adjustments to an existing large-scale grammar. We show that information-structural constraints on right node raising can be simulated in f-structures and note that they are well-behaved in a generation scenario.

1 Introduction

Syntactic coordination phenomena above the straightforward levels of NPs and PPs continue to be a challenge for broad-coverage parsing (and generation) systems. As the corpus study in Forst and Rohrer (2009) showed, a non-negligible proportion of German corpus instances involving coordination at the verbal/clausal level are cases of "non-constituent coordination", including a fair number of right node raising instances. (1) is a list of examples from Maxwell and Manning (1996) indicating the different constructions typically assumed under the umbrella term non-constituent coordination. This paper focusses on the phenomena in (1a-b).

(1)	a.	*Conjunction reduction*	Bill gave the girls spades and the boys recorders.
	b.	*Right node raising*	Bill likes, and Joe is thought to like cigars from Cuba.
	c.	*Gapping*	Bill gave a rhino to Fred, and Sue a camera to Marjorie.
	d.	*Ellipsis*	Bill likes big cars, and Sally does too.
	e.	*Non-symmetric coordination*	Bill went and took the test.

Overall, each variant of these coordination constructions is in the low frequency spectrum in real corpus data, which explains in part why broad-coverage parsing approaches tend to ignore the problem. For statistical approaches, the training data for the numerous variants can be expected to be too sparse to support the picking up of effective generalizations; moreover, the effort to make a parser capable of dealing with the phenomena is unlikely to pay off in terms of a significant gain in evaluation results on standard data sets.

Rule-based approaches on the other hand could in principle address these issues at the right level and capture the morphosyntactic constraints that drive the

*The work reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) in SFB 732 *Incremental specification in context*, project D2 (*Combining Contextual Information Sources for Disambiguation in Parsing and Choice in Generation*; principal investigators: Jonas Kuhn and Christian Rohrer).

phenomena. LFG, with its elegant set-based representation of coordination phenomena (Kaplan and Maxwell, 1989/95), is a particularly promising framework for such an approach. Based on a straightforward c-structure treatment of coordinations, each conjunct contributes its f-structure information in the form of an element of a set, following the annotation in the schematic coordination rule in (2); information from "outside" the coordination construction is then simply distributed over the set elements, making sure, for instance, that *Bill* in (3) ends up as the f-structural subject of the two f-structures for *sing* and *play*.

(2) $$\begin{array}{ccccc} \text{XP} & \rightarrow & \text{XP} & \text{CONJ} & \text{XP} \\ & & \downarrow\in\uparrow & & \downarrow\in\uparrow \end{array}$$

(3) Bill sings and plays the guitar.

Moreover, with Maxwell and Manning (1996), there exists the outline of a meta-grammatical approach (on top of standard LFG parsing) to dealing with certain non-constituent coordination phenomena, which extends ideas of the standard set-based coordination analysis. (Compare Milward (1994) for a discussion of different computational strategies for non-constituent coordination.)

But although LFG provides a good basis for tackling non-trivial coordination cases from real corpus data in a broad-coverage approach, it turns out that it is very hard to tell ahead of time whether a treatment of certain low-frequency non-constituent coordination phenomena will be more helpful than harmful. This may sound counterintuitive, but it has to do with the trade-off one has to make in grammar engineering between goals that are not fully compatible in practice, as sketched in Figure 1. Linguistic quality will normally warrant an advantage in engineering goals (e.g., it is easier to maintain, extend or adapt a systematic analysis; hence, one can over time achieve higher coverage if the analyses chosen are linguistically valid and general). With low-frequency phenomena however, adding a linguistically appropriate analysis to the grammar typically has the effect that the search space for many other constructions is also increased. This can lead to a loss of efficiency – not just for the critical data, but potentially for every input sentence. In any practical system, this will have an indirect negative effect on the system's

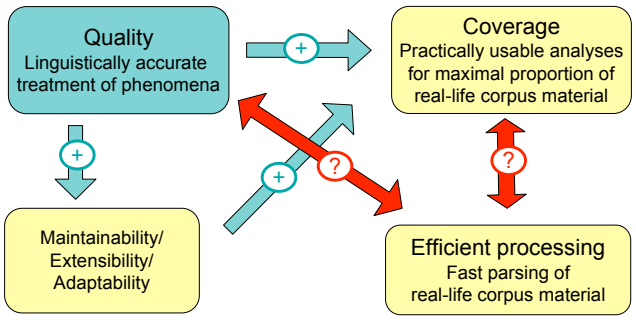


Figure 1: Trade-off between grammar engineering goals

coverage, for the following reason: in parser application, there is a maximal time that the application can wait for a parsing result, the timeout period. If a change in the grammar increases the search space, this can mean that more sentences will time out. This way, the gain from increased linguistic quality in the analysis of a low-frequency phenomenon may be outweighed by the increase of timeouts in unrelated material.

The treatment of coordination (above the NP/PP level) is a particularly tricky area in grammar engineering, because any extension or relaxation of the c-structure rules will readily interact with a large number of other analyses: any added ambiguity may extend the search space dramatically. In the present study, we discuss recent advances that we made by careful conservative extensions to the rule set of the German ParGram grammar (Butt et al., 2002; Dipper, 2003) in order to provide an appropriate analysis for the bulk of right node raising examples occurring in real corpus data. This work is an extension of Forst and Rohrer (2009).

Our analysis is inspired by Maxwell and Manning's (1996) stack-of-automata approach, but is formulated in terms of standard LFG c-structure rules. We take advantage of the fact that an account of German clausal structure following the endocentric principles of Bresnan (2001) goes along with fairly compact rule networks, such that the effect of Maxwell and Manning's idea can be emulated reasonably well on the basis of independently motivated c-structure rules.

We believe that the complex of interacting factors that have to be dealt with in an implementation of non-constituent coordination is typical of the situation in advanced grammar engineering work. The existence of a highly developed engineering platform like XLE and long-term efforts like the ParGram project have brought about relatively mature broad-coverage grammars for which the trade-off sketched in Figure 1 can be addressed at a more realistic level than it was possible a few years ago. The present contribution is meant to convey some of the considerations that have to be taken in this trade-off. We make no claims that the proposed analysis is linguistically justified outside the grammar engineering context; but we do believe that the view from the grammar engineering perspective can give some inspiration for a realistic linguistic account.

Section 2 introduces the technique of Maxwell and Manning's (1996) account in some detail, since it is important for the discussion of our approach. Section 3 introduces the strategy of Forst and Rohrer's (2009) approach, extending it to cover a broader set of non-constituent coordination cases. In section 4, we briefly discuss an overgeneration issue and the role of information structure in restricting the choices. Section 5 offers a short conclusion.

2 Maxwell and Manning's (1996) approach

The coordination constructions usually called "conjunction reduction" and " right node raising" (or "left-deletion") can be characterized as follows: Material which could in principle be realized twice and which would occur at the left or right

periphery in both conjuncts, occurs only a single time (on the left edge of the left-most conjunct, or on the right edge of the right-most conjunct).

There are (at least) two views on this type of data. They could be seen as a result of deleting identical material from the left or right conjunct (or in less transformational terms, as the result of some economy principles), as sketched in (4). Or they can be viewed as coordination of special units that are not normally considered constituents, illustrated in (5), and that can be completed by the identical (left or right) material (which does not have to form a canonical constituent either).

(4) a. Bill [[gave the girls spades] and [~~gave~~ the boys recorders]].

b. [[Bill likes ~~cigars from Cuba~~] and [Joe is thought to like cigars from Cuba]].

(5) a. Bill gave [[the girls spades] and [the boys recorders]].

b. [[Bill likes] and [Joe is thought to like]] cigars from Cuba.

Putting aside the issue of the unjustified non-constituent units, the latter approach is very much in the spirit of the standard LFG treatment of coordination (Kaplan and Maxwell, 1989/95). Since the material occurring just once (shown underlined in (5) and all following examples) is outside the technical coordination, the usual distribution mechanism will fire, in the same way as the subject is distributed for (3).

2.1 C-structure coordination as a meta-grammatical "process"

Creating rules for all possible non-constituent units just for the purpose of coordination is of course an unattractive prospect. This is where the main idea of Maxwell and Manning's (1996) "meta-grammatical" account comes in: they posit a general mechanism, living outside the grammar proper, which provides the required units for the purposes of non-constituent coordination, as needed.

We use the graphical illustration in Figure 2 to suggest that partial versions of the regular constituents are involved when analyzing such coordinations. A partial constituent is required both on the outside of the coordination (the beginning of the VP triangle above "gave"), and inside the coordination – here in two copies, the two hatched parts of the VP triangle for "the girls spades" and "the boys recorders".

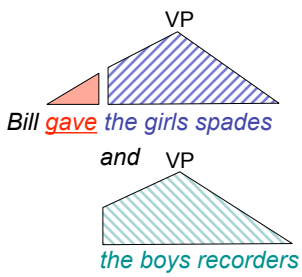


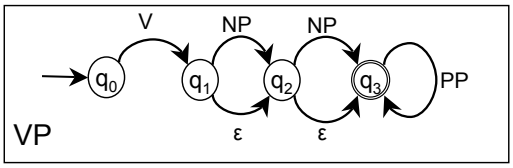
Figure 2: Non-constituent analysis for (4)

C-structurally, the outside and (each of the) inside parts have to complement one another to form a regular constituent. In the LFG parser, the right-hand side

(RHS) of a c-structure rule (which is a regular expression that may contain a Kleene star, optionality brackets etc.) is defined as a transition network or finite-state automaton. Maxwell and Manning ensure the correct assembly of partial constituents in the parser by keeping track of how much of the RHS of a c-structure rule has been consumed in terms of the states of the RHS automaton. The automaton for the VP rule in (6) is given in (7).

(6) VP → V (NP) (NP) PP*

(7)



In Figure 2, the coordination comprises the c-structure material starting after the initial V, so the subautomaton from the initial state q_0 to q_1 is used only once, outside of the coordination, whereas each of the conjuncts covers the subautomaton from q_1 to the final state q_3.

Technically, the analysis of non-constituents is assumed to be a meta-grammatical "process" (which can be perfectly viewed in a non-derivational sense, just like automata are not a derivational model in plain LFG) that can pass a subautomaton of a rule twice (or more times), using the automaton states to keep track of the rule coverage. The meta mechanism will also ensure that the conjuncts are separated by (commas and) a conjunction like *and*/*or*. The abstract process driving the analysis in Figure 2 is sketched in the flow-chart in Figure 3, suggesting a "parallel" analysis of the two conjuncts, covering the stretch of the automaton between q_1 and q_3.

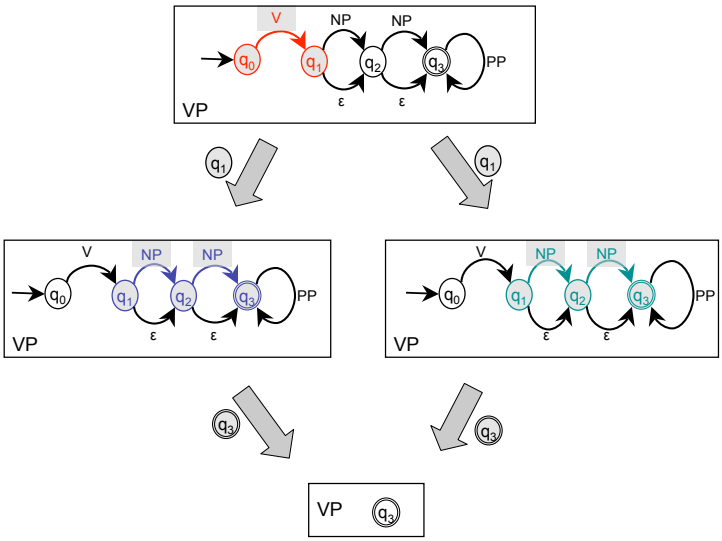
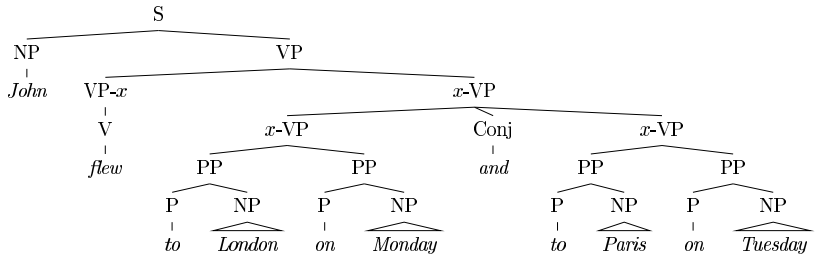


Figure 3: "Split" parsing process for Figure 2

Maxwell and Manning use the notation in (8) to draw the resulting c-structures. The internal automaton states are indicated by variables $x, y, \ldots,$ which have to match across conjuncts and with the material outside of the coordination. In the resulting c-structures, subtrees do not necessarily correspond to constituents licensed by the LFG base grammar – so the coordination of "non-constituents" can be derived in an elegant way, without altering the grammar proper.

(8)

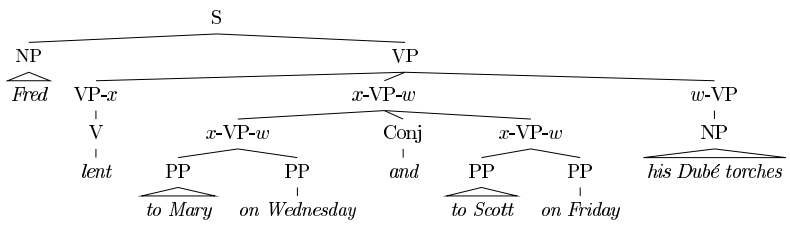


Since the mechanism keeps track of subautomata covered, the two conjuncts do not have to include the same categorial material, nor the same number of (sub-)constituents, as long as the alternatives are each licensed by a rule automaton. This licenses cases like (9).

(9) You can call me [[directly] or [after 3pm through my secretary]].

Monitoring of partial rule consumption can also occur on both sides of the coordinated sub-structures, as is shown in (10):

(10)



This example could not be derived with the VP rule/automaton in (6)/(7), since it involves a shift of the object NP *his Dubé torches*. But based on a more general VP rule automaton as in (11), the example could be analyzed as in Figure 4. Note that on both ends of the original VP rule, material occurs only a single time (and is thus underlined in our notation), whereas the middle part between q_1 and q_4 is duplicated.

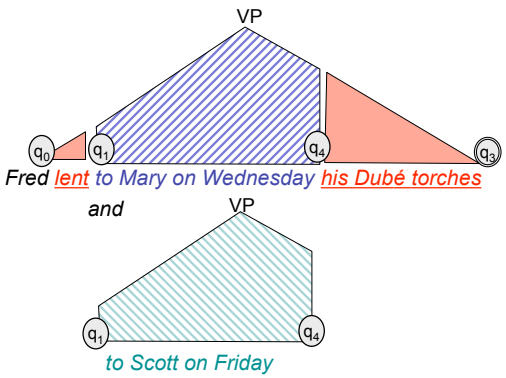
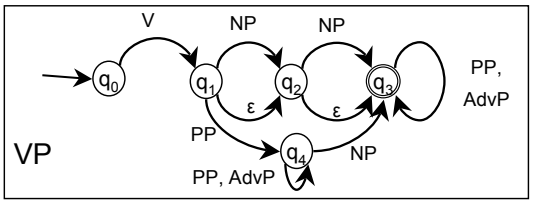


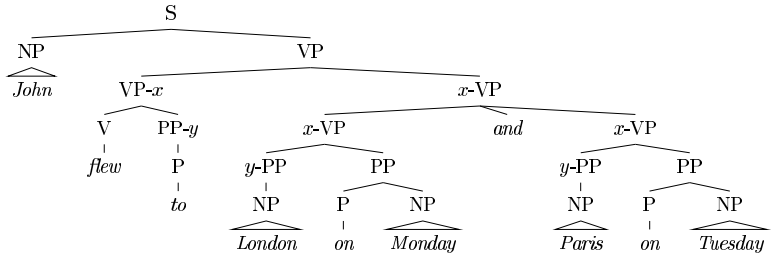
Figure 4: Analysis of (11) keeping track of states

(11)



Examples like (12) suggest that the mechanism is even more general. Here, the non-coordinated material *flew to* does not originate from a single rule automaton, but essentially cuts across hierarchically embedded LFG rules.

(12)



The flow-chart of automaton calls in Figure 5 illustrates how the mechanism proceeds. The dotted box at the top displays the parsing status after reading *flew to*, having consumed the beginning of the VP, plus part of the PP. The rest of the PP and the continuation of the VP (using the PP loop from q_3) is again processed in parallel for the two "conjuncts" *London on Monday* and *Paris on Tuesday*, before

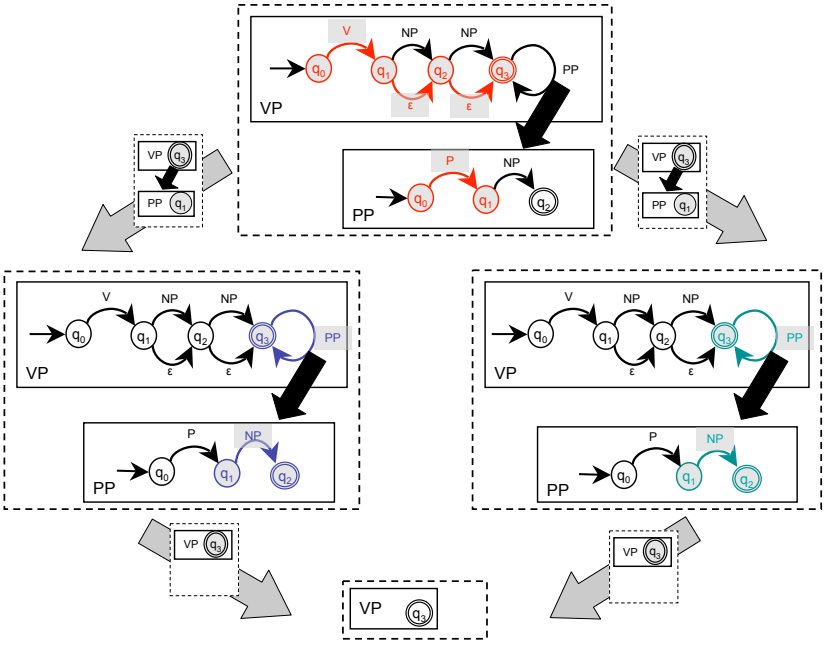


Figure 5: Coordination meta-mechanism for (12)

jumping back to the final state of the VP rules.

Note that for such hierarchically embedded cases, more than a single automaton state is needed for keeping track of the consumption status. In Figure 5, this is indicated by the boxes labelling the gray flow-chart arrows, which can include state recordings from several embedded rule automata. Maxwell and Manning assume a *stack* of automaton states from (potentially) different rules as the representation controlling the meta mechanism.

2.2 F-structure treatment and discussion

So far, we did not mention the f-structure treatment that goes along with the meta-grammatical mechanism of non-constituent coordination. The mechanism was designed the way it was so that the standard LFG coordination f-annotation ($\downarrow\in\uparrow$ on each of the conjuncts) can be directly used. So, in addition to keeping track of c-structure rule consumption, the meta-mechanism will simply add this annotation to each conjoined "pseudo constituent" arising from the process.

We note that Maxwell and Manning's (1996) approach provides a very elegant solution to the main problems that conjunction reduction and the right-node raising phenomena pose: no unjustified pseudo constituents have to be added to the grammar since the meta-grammatical mechanism generates the partial constituents on the fly, as needed – both for the coordinated material and the "outside" material occurring only once. F-structural distribution of the outside material follows the standard LFG coordination treatment: since the $\downarrow\in\uparrow$ annotations coerce the f-structures involved in the coordination to be set elements, the standard "across the board" distribution mechanism is triggered (at least in the simple configurations).

To our knowledge, Maxwell and Manning's coordination approach has never been implemented on a larger scale (Zarrieß and Seeker (2008) present an experimental implementation which uses finite-state technology to compile an XLE grammar into a version following Maxwell and Manning's approach). This presumably has to do with a fundamental practical issue of the approach: when the coordination mechanism is assumed to interact with the rule automata as they are used in large-scale LFG grammars, a combinatorial explosion of potential "rule splitting" points will arise – in particular through the hierarchical rule stacking. To see this intuitively, note that for a typical attachment ambiguity example like *John read a book and an article on Semantics*, the meta-grammatical mechanism would add the possibility of coordination of partial higher-level constituents; in this case, *a book* could be a partial VP, possibly open for taking in more material at the embedded NP level (namely, the PP *on Semantics*), besides numerous other splitting alternatives.

Moreover, the combinatory possibilities of rule splitting are multiplied by alternatives in the plain c-structure rules. The VP automaton in (11), which would be required to capture example (10), is still a very reduced sample automaton. Yet it gives an idea of the degree of optionality that broad-coverage rules have to offer at the level of c-structure (compare Kuhn and Rohrer (1997)). With the constraining

role of f-structure this is not a practical problem in standard LFG, as the impressive parsing speed of the XLE system shows. However, pairing this c-structure level optionality up with a highly liberal, c-structure based non-constituent coordination approach is likely to be computationally intractable.

In fact there seems to be a tendency – at least for the broad-coverage grammar of German – that due to optionality of most rule parts (essentially following the principle-based approach of Bresnan (2001)), many argument phrases and sequences of argument phrases can legally be analyzed, for instance, as VPs (without the verb being realized). It is the broader c-structure rule context for using such VPs, plus f-structure that constrains this c-structural freedom, striking a balance between coverage of real corpus material and efficient processing. Allowing for arbitrary partial constituents to capture non-constituent coordination would almost definitely break this balance.

Independent of these complexity problems, there is an issue with the apparent simplicity of Maxwell and Manning's coordination approach that tends to escape attention when one first looks at it. What we said above about the f-structure treatment and the automatic effect of distributing "outside material" over the conjunct f-structures is actually not quite true. In a hierarchical-stacking situation, the outside material would not receive the correct f-structural interpretation if the standard f-annotations were not changed. In case of asymmetrically "deep" coordinations the f-annotation will even need to contain a functional uncertainty equation. Consider example (13):

(13) [[John drove to] and [Sue is planning to bike to]] San Francisco.

In order for *San Francisco* (as part of a partial x-VP with an embedded partial y-PP) to receive the correct interpretation, the f-annotation has to introduce an optional XCOMP embedding in order to capture the situation in both conjuncts. Similar examples with different embeddings are possible too.

Since the standard f-annotation of PP in the VP rule does not come with this functional uncertainty, it turns out that the meta-grammatical coordination mechanism does have to make some more substantial alternations to the plain LFG grammar, which are motivated in a purely technical way.[1]

3 Generalized constituent approach

Since the direct implementation of Maxwell and Manning's approach is not an option for broad-coverage grammar engineering, we discuss here a feasible alternative strategy for dealing with non-constituent coordination in the context of the German ParGram grammar. The goal is to find an analysis that avoids introducing new special "constituents" as much as possible, but instead to use independently

[1]This particular aspect of the f-structure distribution problem is not discussed in the (Maxwell and Manning, 1996) paper.

motivated rules for the coordination case, making only very few special additions to rules.

The treatment of a number of VP coordination cases was already proposed in Forst and Rohrer (2009) (see Section 3.1). In the present paper, we discuss the underlying strategy and extend it to a number of additional cases of right node raising (Section 3.2).

3.1 Coordination configurations in the verbal projection spine

In the remainder of this paper, we discuss modifications of the German ParGram grammar. Since a discussion of changes in the grammar code would be hard to convey, we use schematic illustrations of the c-structure analysis of the clausal spine to highlight the most important aspects of the analysis.

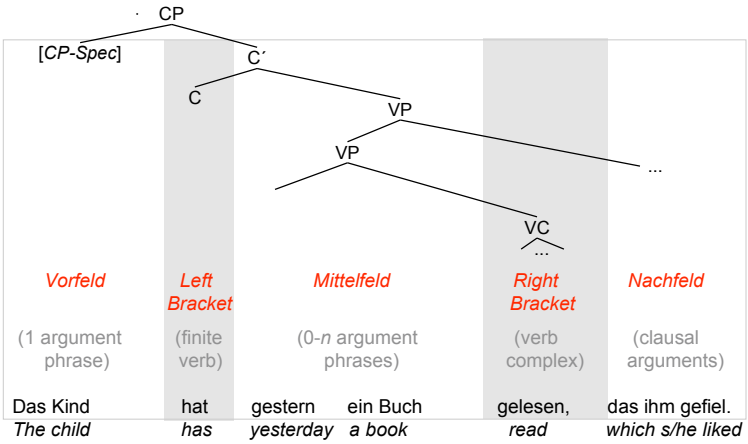


Figure 6: Schematic illustration of German clausal syntax

Figure 6 illustrates the basic clause structure of German, overlaying the classical descriptive field model over a c-structure scheme which by and large follows Bresnan's (2001) endocentric schema. We use the shaded areas to highlight (i) the left verbal bracket, containing the complementizer in embedded clauses and the finite verb in main clauses, and (ii) the right verbal bracket, containing the verb complex, with the least embedded verb appearing at the right edge. (In main clauses containing just a single verb, the verb complex will be empty.) The two verbal brackets delimit the *fields* in the German clausal structure: *Vorfeld, Mittelfeld, Nachfeld*.

The problem posed by typical VP coordination cases in corpus data is illustrated by the two examples in Figure 7: Under the analysis chosen for the German clause structure in the grammar, both cases involve conjuncts not corresponding to constituents in the grammar. In both cases, it would be possible to modify the clausal analysis in a way that would turn it into a constituent coordination: for the first example, a strictly right-branching VP analysis could be assumed; for the second example, the verb complex analysis that bundles up all verbs under VC could be replaced with a recursive VP embedding analysis. The problem is however that

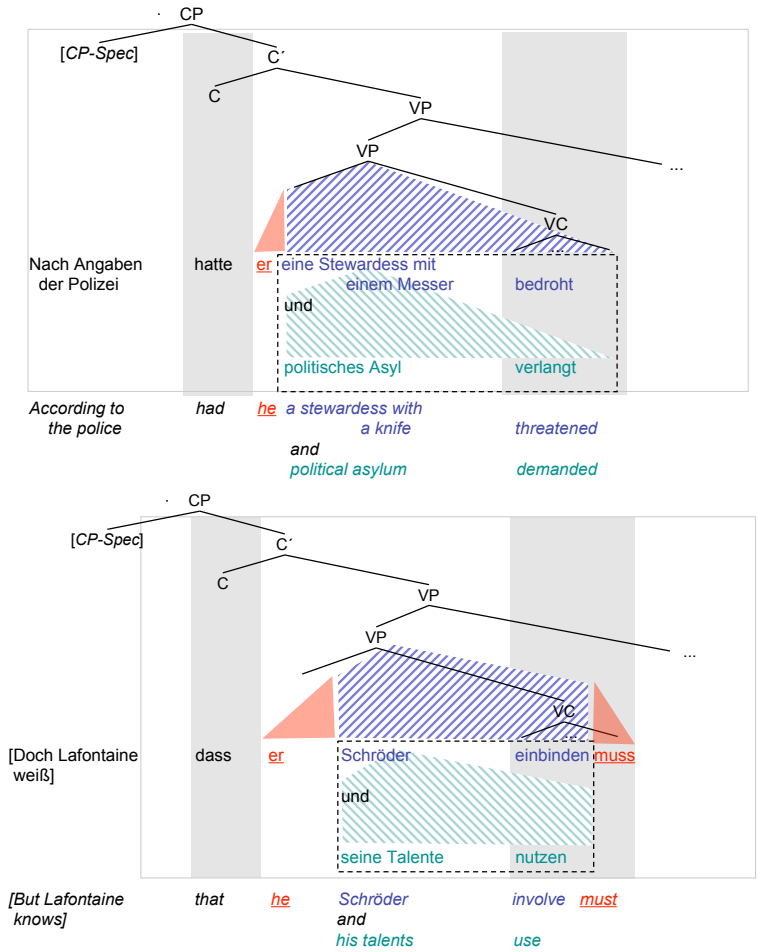


Figure 7: VP coordination issues involving coordinated verbs

the analyses would lead to spurious ambiguity in all non-coordination examples, if they were allowed along with the standard analysis. Figure 8 is a case where the coordinated material does not include any verb; hence, none of the above options would work.

Forst and Rohrer (2009) propose a solution for all three cases, following the strategy of making minimal modifications to the standard clausal backbone, specifically for the coordination situation. So, the rules are not relaxed in arbitrary situations but only in the presence of the coordinating conjunction. This guarantees that the search space is extended very carefully and the analysis stays computationally manageable.

The generalized constituent analysis simply allows the conceivable variants of the standard clausal analysis mentioned above *for the coordination case*. This way, the two "non-constituent coordination" cases from Figure 7 receive the constituent coordination analysis in Figure 10 with minimal grammar adjustment. The coordination-specific use of categories (here, VP) is indicated by the dotted box.

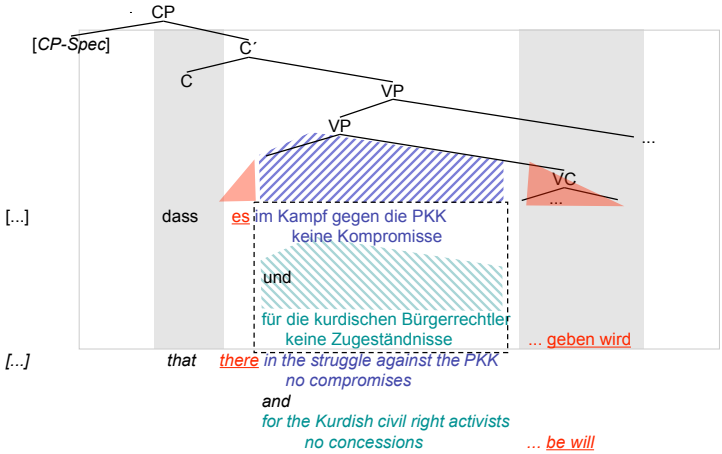


Figure 8: A VP coordination issue with a “raised” verb

Figure 8 is the only type of construction for which a new special category has to be posited – combining a sequence of two or more verbal arguments (or adjuncts) into a “pseudo-constituent” VPargs, as shown in Figure 9. (One could also view this pseudo-constituent as a VP in which the optional V is not realized – but calling it by a special name has the advantage that its use can be heavily restricted.)

When we compare the “generalized constituent approach” to Maxwell and Manning’s, we note that since (a) the combination of verbal constituents typically makes them f-structure co-heads, and (b) most material in c-structure rules is actually optional, the effect of combining partially consumed rules (from Maxwell and Manning’s mechanism) can be simulated quite well by combining several instances of “full” verbal constituents in which the optional subparts are left unrealized, just as required. In particular, the material outside of the coordination uses canonical rules with standard f-annotations in all cases that we have seen so far (obtaining

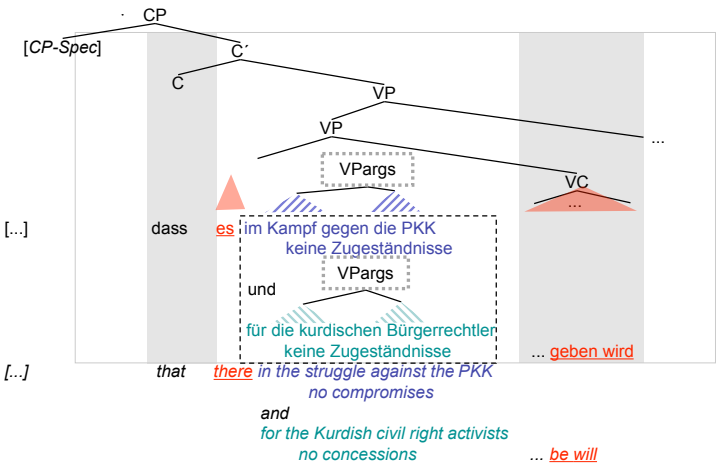


Figure 9: Forst and Rohrer’s (2009) solution for Figure 8

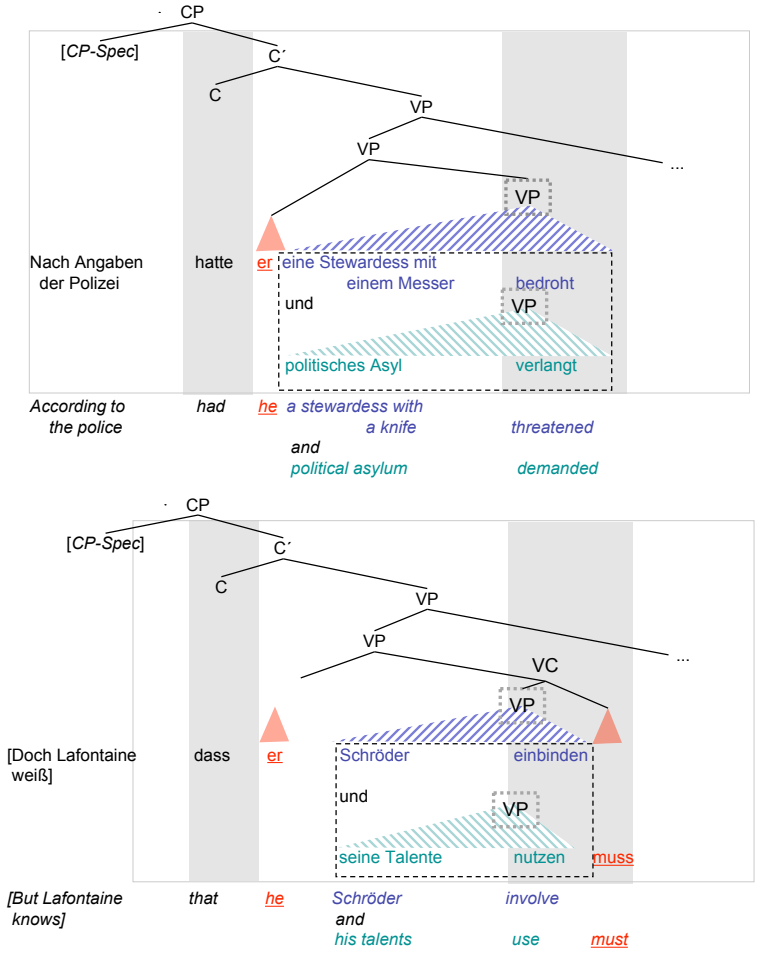


Figure 10: “Generalized constituent analysis” for Figure 7

the f-structure distribution effect with the standard LFG mechanism of coercing the coordinated f-structure into a set).

The strict bookkeeping of c-structural splitting points of Maxwell and Manning (through the stack of states) is lost in this simulation. However, with rules similar to the rule scheme “XP $\rightarrow$ { YP | ZP | . . . }* (X)” this is not too critical in practice: There are many alternative paths between the rule automaton states, and f-structural constraints play the most important role in constraining the options of combination.

Of course the generalized constituent approach will only cover cases of non-constituent coordination that were foreseen in the coordination-specific relaxation of the rules. Given the subtle balance between coverage and efficiency (through limited search space), this is actually a desirable property of the account.

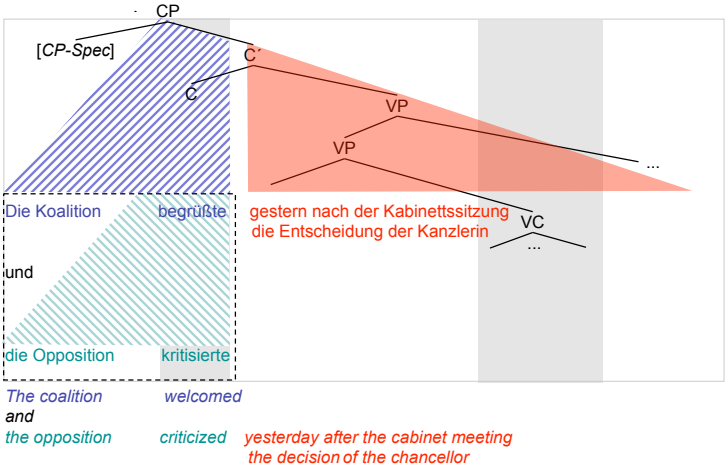


Figure 11: Coordination problem involving "low" non-coordinated material

3.2 Extending the generalized constituent strategy

In all cases we discussed so far, the material "outside" the coordinated stretch of the clause was either the head (verb) or a "high" argument phrase like the subject. Is it possible to generalize the idea further to also allow for cases in which the "outside material" appears to be much "lower", as in Figure 11?

It turns out that we can take Forst and Rohrer's philosophy of minimal grammar extension one step further. Figure 12 shows a canonical case of an extraposed clause, appearing in the *Nachfeld* region of the clause. Since the verb selecting it may be embedded under a modal or some raising verb, the f-annotation for CPs in this position has to contain a functional uncertainty equation, providing an optional embedding under XCOMP's.

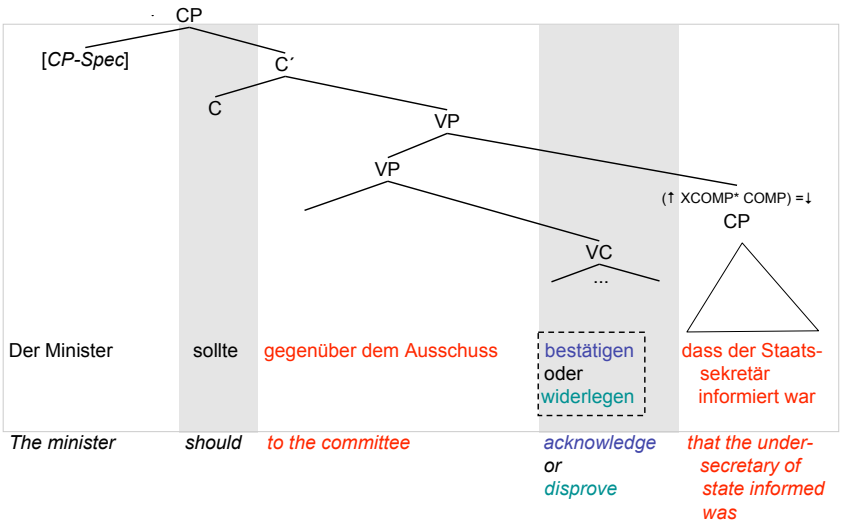


Figure 12: Standard *Nachfeld* analysis

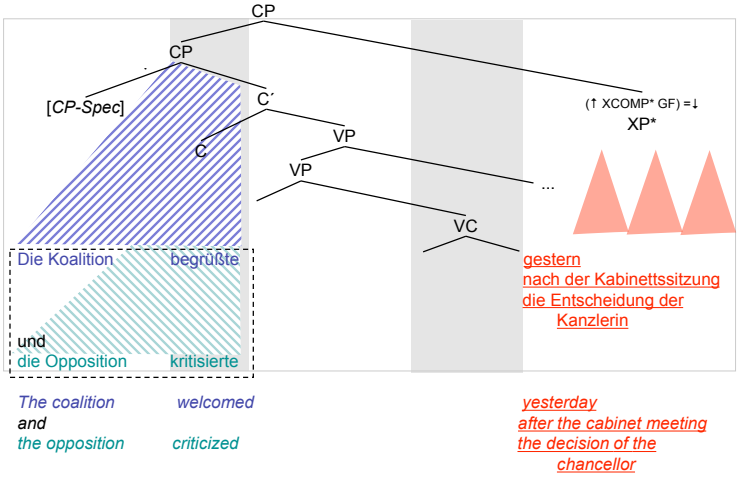


Figure 13: Extended *Nachfeld* analysis for Figure 11

Now, it turns out that the existing *Nachfeld* position can very easily be used in order to capture cases like Figure 11 in terms of a constituent coordination, as is seen in Figure 13. If we assume that the adjunction of the *Nachfeld* can be at the root CP level and we relax the categorial restriction to also allow for nominal arguments (as they also appear extraposed in spoken German), the material that appeared to occur "low" inside the constituents can quite naturally be analyzed as appearing "high", outside the coordinated CP constituent. The fact that the entire *Mittelfeld* and the right verbal bracket can be empty is nothing new: these parts of the clausal structure have to be optional anyway in order to capture sentences like *Maria lacht* ("Maria laughs") as instances of verb second, where *Maria* appears in Spec-CP, and *lacht* in C. The variant of Figure 13 in Figure 14 shows that the "splitting point" can easily be moved around.

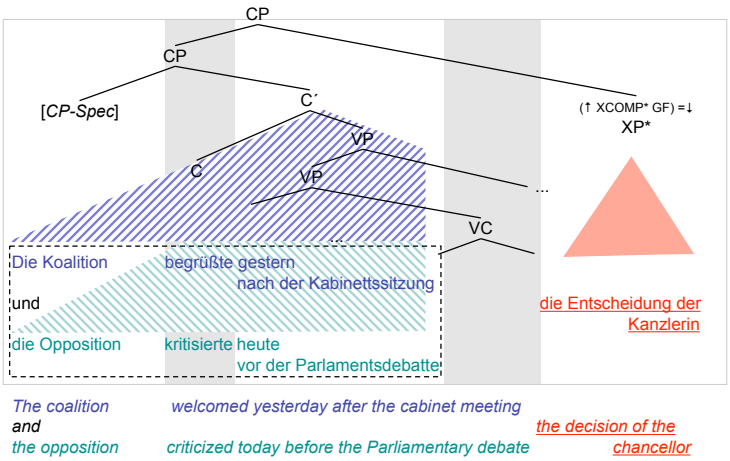


Figure 14: Variant of Figure 13 with a smaller overlap in conjuncts

In our analysis, the unwanted duplication of such optional constituents (generating them both inside the coordination and as part of in the "outside material") or the complete skipping of required material is excluded on the basis of standard f-structure mechanisms like uniqueness and completeness/coherence.

However, we have to enforce that the "gaps" in the constituents (resulting from optionality) are indeed peripheral. In Maxwell and Manning's account this is ensured by keeping track of the automaton state up to which material was covered. Without further provisions, our account would overgenerate, for instance allowing the ungrammatical example (14) in which a complex verb form *hat . . . begrüßt* is used, which requires the VC to be filled, such that the *Nachfeld* material is no longer adjacent to the *Mittelfeld* "gap" in the conjuncts. In our grammar, this restriction is implemented as a simple c-structural constraint that forbids right node raising constructions in contexts where the VC is filled.

(14) *[[Die Koalition hat gestern nach der Kabinettssitzung begrüßt] und
The coalition has yesterday after the cabinet meeting welcomed and
[die Opposition hat heute vor der Parlamentsdebatte kritisiert]] die
the opposition has today before the Parliamentary debate criticized the
Entscheidung der Kanzlerin
decision of the chancellor

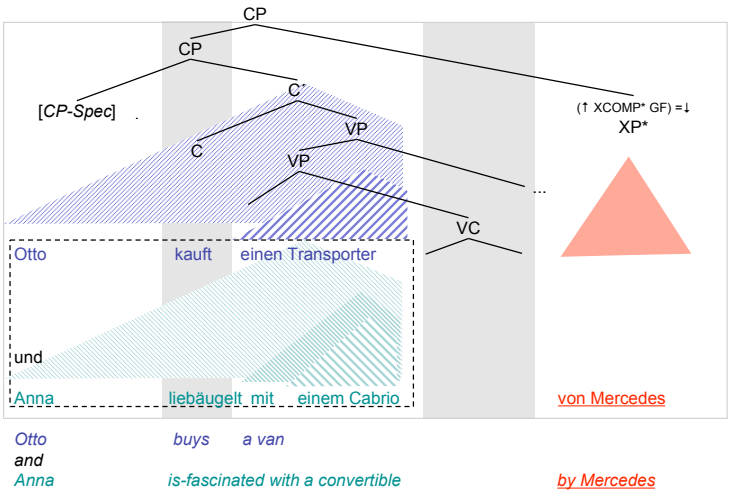


Figure 15: Right node raising involving asymmetric hierarchical embeddings

The extended *Nachfeld* approach is also directly compatible with a situation where the "raised material" originates from constituents involving a complex hierarchical embedding (Figure 15), replicating Maxwell and Manning's idea that involves stacks – with the proviso that any "partial" rules one would like to use for this either have to coincide with available full categories usable in the *Nachfeld* or must be especially added.[2]

[2]Note that the application of such hierarchically embedded partial rules can no longer be con-

We note that the philosophy of applying minimal extensions to the canonical grammar can also be applied to cover right-node raised argument phrases. The functional uncertainty path needed to ensure distribution into embedded structures is even independently motivated from the canonical *Nachfeld* treatment (contrary to the situation in Maxwell and Manning's approach discussed at the end of Section 2.2, where the functional uncertainty annotations needed to be stipulated).

However, there are limitations to what can be realistically covered without opening up the search space too much. In particular, the addition of otherwise unmotivated embedded "partial constituents" as would be required for examples like (12) (which are however not too common in German) leads to serious efficiency problems. As a general strategy, we did not include hierarchical embeddings which would descend more than two constituent levels in our large-scale LFG. This is typically the case in right node raising constructions which involve partial PPs or DPs, illustrated in example (15).

(15) Bei den Wahlen von 1990 verlor [$_{VP}$ [$_{VP}$ [$_{DP}$ die CDU] [$_{DP}$ die rechten]]
In the elections of 1990 lost the.NOM CDU the right.ACC
und [$_{VP}$ [$_{DP}$ die SPD] [$_{DP}$ die linken]]] [$_{DP}$ Wähler].
and the.NOM SPD the left.ACC voters

"In the 1990 elections, the CDU lost the right-wing voters and the SPD lost the left-wing voters."

4 Information structure, prosody and generation

In our generalized constituent approach, the "partial" constituents required in cases of non-constituent coordination are in fact using up entire rules, mostly by not realizing one or more optional subconstituents at the edge. As mentioned above, our approach cannot enforce an "automatic" c-structural periphery condition as it is included in Maxwell and Manning's meta mechanism, which is entirely controlled by the c-structure rule consumption.

Considering the linguistic constraints at play, it appears however that it is not c-structural adjacency alone that drives the construction. As Maxwell and Manning point out too, prosody and information structure place important additional constraints. Right node raising examples typically involve a pronounced contrastive focus on the two (right-peripheral subconstituents of) the conjuncts, which is marked by pitch accents and the intonational phrasing. Furthermore, the two focus elements have to be parallel (typically at the level of grammatical functions), excluding cases like (16).

strained straightforwardly, since the rule in which they are called does not directly include the coordination. One could in principle use the parametrized rule mechanism of XLE to propagate the information of being inside a coordination (compare for instance Kuhn (1999)), but this would involve highly complex grammar engineering, unless one could automatize rule compilation based on a cross-classification of rule meta-principles.

(16) ?[[Peter schenkte seiner Tochter] und [ihrem Sohn schenkte
Peter.NOM gave his.DAT daughter and her.DAT son gave
Maria]] ein Buch
Maria.NOM a.ACC book

"Peter gave to his daughter and to her son gave Mary a book."

Following the main ideas of Féry and Hartmann (2005), it is in fact relatively easy to simulate these information-structural constraints in an LFG account using only f-structure, by introducing a special discourse function RNR-FOCUS (standing in for a full account involving an information structural projection), as indicated in Figure 16.

The coordination rules can generally be formulated in a way so they enforce (a) that each of the conjuncts ends in an element introducing the RNR-FOCUS (leaving the appropriate "gap" to its right), and (b) that the two (or more) peripheral elements introduce the same grammatical function.

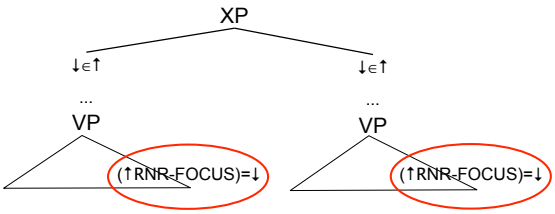


Figure 16: Information structural constraints

In the grammar engineering context, these parallelism constraints are particular relevant for using the grammar in generation tasks. For instance, Cahill and Riester (2009) apply the German LFG in a surface realisation ranking task, where they exploit the reversible XLE engine to generate all possible sentences from a given input f-structure. In such a task, a grammar without the above focus constraint on RNR would generate a lot of infelicitous word order variations that independently permute the constituents in the conjuncts, as in example (16).

The RNR focus restriction constrains the generation of word order variations to cases where the constituents inside the conjuncts are permuted in parallel to each other. Thus, if the generator is supplied with the f-structure for the sentence in (17) (without fixing the discourse functions to particular grammatical functions), it produces, among others, the output alternations in (18). Note that the order of the nominative and accusative arguments in the conjuncts is always the same. Infelicitous cross-over configurations are not generated due to the grammatical constraint.

(17) Gestern hat [[der Attentäter einen Polizisten] und [der
Yesterday has the assassin.NOM a police man.ACC and the
Demonstrant einen Politiker]] mit einer Pistole bedroht
demonstrator.NOM a politician.ACC with a gun threatened.

(18) *Surface alternations generated for (17)*

a. Gestern hat [[einen Polizisten der Attentäter] und [einen Politiker der Demonstrant]] mit der Pistole bedroht.

b. Mit der Pistole hat [[einen Polizisten der Attentäter] und [einen Politiker der Demonstrant]] gestern bedroht.

c. Bedroht hat [[der Attentäter einen Polizisten] und [der Demonstrant einen Politiker]] gestern mit der Pistole.

However, the generation constraints on word order become more difficult to formulate in the case of right node raising that involves partial DPs or PPs. As an example, consider the following word order alternations of varying fluency/acceptability that the grammar generates for sentence (15), where the "raised" noun figures before the coordination (the latter being an instance of the Split NP construction (Kuhn, 2001)):

(19) *Surface alternations generated for (15)*

a. ?Bei den Wahlen von 1990 verlor Wähler [[die CDU die rechten] und [die SPD die linken]].

b. Wähler verlor bei den Wahlen von 1990 [[die CDU die rechten] und [die SPD die linken]].

5 Conclusion

We discussed a practical account of non-constituent coordination cases in the clausal syntax of German as they occur in real corpus data, generalizing the approach of Forst and Rohrer (2009). In particular, we compared the account to the meta-grammatical technique proposed by Maxwell and Manning (1996), which captures most aspects of the construction in a very elegant way, but is presumably impossible to implement on a larger scale.

(20) [[[Bíll]$_T$ [lìkes]$_F$ ~~cigars from Cuba~~] and [[Jóe]$_T$ [is thòught to like]$_F$ ~~cigars from Cuba~~]] cigars from Cuba.

The main ideas behind our account – sketched in an inutive notation in the English example (20), a repetition of (4) – are the following: (i) to use canoncial rules as much as possible and take advantage of optionality to generate a variant of the constituent involving what looks like "deletion" in the periphery of the conjuncts; (ii) to extend/relax existing rules (such as extraposition rules) in order to host the raised or "outside material", using the standard LFG mechanism of f-structural distribution and in particular taking advantage of the independently motivated functional uncertainty annotation in order to capture asymmetric hierarchical embeddings in right node raising examples; (iii) to approximate information-structure constraints behind right node raising, which enforce an opposition of two pairs of contrastive topic/contrastive focus, along with the appropriate prosodic (phrasal and pitch) marking.

We make no claims that the approach captures the linguistic constraints in the most appropriate way since the overarching motivation comes from grammar engineering considerations, i.e., the goals of maximizing (1) coverage and (2) quality of analysis, relative to corpus frequency of the phenomena.

However, we believe that linguistically grounded grammar engineering work, as in the LFG-based research paradigm, is now at a point where the efficiency of the parsing system (XLE) and the broad coverage of the grammars allow us to address questions that are interesting both from a linguistic and an engineering point of view, and find engineering answers that can at least be related in an interesting way to linguistic considerations.

References

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, Taipei, Taiwan.

Cahill, Aoife, and Arndt Riester. 2009. Incorporating information status into generation ranking. In *Proceedings of ACL-IJCNLP 2009*, pp. 817–825, Singapore.

Dipper, Stefanie. 2003. *Implementing and Documenting Large-Scale Grammars — German LFG*. PhD thesis, IMS, University of Stuttgart. appeared in Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), Vol. 9(1).

Féry, Caroline, and Katharina Hartmann. 2005. Focus and prosodic structures of German Gapping and Right Node Raising. *The Linguistic Review* 22:67–114.

Forst, Martin, and Christian Rohrer. 2009. Problems of German VP Coordination. In *Proceedings of the LFG '09 Conference*, pp. 297–316, Cambridge, England.

Kaplan, Ronald, and John Maxwell. 1989/95. Constituent coordination in Lexical-Functional Grammar. In M. Dalrymple, R. M. Kaplan, J. T. Maxwell, and A. Zaenen (eds.), *Formal Issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.

Kuhn, Jonas. 1999. Towards a simple architecture for the structure-function mapping. In M. Butt and T. H. King (eds.), *Proceedings of the LFG99 Conference, Manchester, UK*, CSLI Proceedings Online.

Kuhn, Jonas. 2001. Resource sensitivity in the syntax-semantics interface and the German split NP construction. In D. Meurers and T. Kiss (eds.), *Constraint-Based Approaches to Germanic Syntax*, pp. 177–215. Stanford: CSLI Publications.

Kuhn, Jonas, and Christian Rohrer. 1997. Approaching ambiguity in real-life sentences – the application of an Optimality Theory-inspired constraint ranking in a large-scale LFG grammar. In *Proceedings of DGfS-CL Heidelberg*.

Maxwell, John, and Christopher Manning. 1996. A theory of non-constituent coordination based on finite-state rules. In M. Butt and T. H. King (eds.), *On-line Proceedings of the First LFG Conference*, Grenoble.

Milward, David. 1994. Non-constituent coordination: Theory and practice. In *Proceedings of the fifteenth International Conference on Computational Linguistics (COLING-94)*, pp. 935–941.

Zarrieß, Sina, and Wolfgang Seeker. 2008. Finite-State Rule Deduction for Parsing Non-Constituent Coordination. In *Finite-state methods and natural language processing : 6th International Workshop, FSMNLP 2007*, Potsdam.

WALMAN *AND*-VERBS AND THE NATURE OF WALMAN SERIALIZATION

Lewis C. Lawyer
University of California Davis

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

http://csli-publications.stanford.edu

Abstract

The so-called *and*-verbs of Walman, described recently in Brown and Dryer (2008), are of a crosslinguistically unusual mixed category. They are morphologically verbs, but they serve to coordinate noun phrases and their maximal projections are noun phrases. This paper uses a Lexical-Functional Grammar (LFG) framework to compare Walman *and*-verbs to Walman transitive verbs, demonstrating the similarity between these two categories. Evidence for this similarity comes from the fact that Walman transitive verbs participate in inclusory serial constructions, in which the subject of one verb is the combined subject and object of the previous verb. It is argued therefore that all Walman transitive verbs must be able to coordinate their arguments in f-structure just as Walman *and*-verbs do. Thus *and*-verbs, while unexpected crosslinguistically, fit into the grammar of Walman relatively comfortably. The paper concludes with a brief commentary on the utility of an LFG-like framework as a tool for descriptive linguistics.

Introduction

In a recent article Brown and Dryer (2008) discuss the phenomenon of Walman *and*-verbs, words which are morphologically transitive verbs but syntactically serve as coordinators coordinating noun phrases. In the present paper I take the phenomenon described by Brown and Dryer and examine it more formally, focusing on Walman *and*-verbs and Walman transitive verbs. I wish to demonstrate that the functional[1] nature of Walman *and*-verbs is parasitic on the functional nature of Walman transitive verbs in general. The evidence for this similarity can be found in a particular type of serial verb construction (SVC) found in Walman. This type of SVC is called inclusory serialization (following Crowley 2002, 41), and is characterized by the subject of one verb being the aggregate of the subject and object of the previous verb. (See 9 below.) I will show that the functional structure present in the maximal projections of transitive verbs in this kind of serialization is the same type of structure found in noun phrases coordinated by *and*-verbs.

1 By "functional" I mean relating to functional structure (f-structure) in the LFG sense. See Dalrymple (2006) for a brief introduction. This paper is not concerned with syntactic functionalism in the tradition of André Martinet, John Hawkins, etc.

The grammatical framework used in this paper is classical Lexical-Functional Grammar (LFG).[2] I hope this paper will serve to demonstrate the utility of an LFG-like model to descriptive linguistics. The relationship between inclusory serialization and *and*-verbs may be difficult to discern in a tree-theoretic framework in which constituent structure is the sole determinant of functional roles and functional structure. In an LFG-like framework, where constituency and functional structure are modeled separately, the functional similarity between *and*-verbs and inclusory serial constructions is readily apparent.

This paper is organized as follows: Section 1 gives a brief introduction to aspects of Walman grammar. Section 2 describes Walman *and*-verbs briefly and then shows how they appear in f-structure. The f-structure projections of *and*-verbs are also compared to the f-structure projections of ordinary coordinators. Section 3 shows how Walman SVCs appear in f-structure, arguing that Walman SVCs require a different f-structure analysis than the kind that has been given for SVCs in other languages. Drawing from sections 2 and 3, section 4 notes the unexpected f-structural similarity between Walman *and*-verbs and Walman transitive verbs. Section 5 describes the c-structural complexity of Walman *and*-verbs. Section 6 concludes the paper by noting that an LFG-like framework allows the descriptive linguist to make observations about functional structure directly, without the analysis being complicated by issues of constituency and category affiliation.

1 Background information about Walman grammar

Walman is a language in the Torricelli family spoken on the northern coast of Papua New Guinea. Verbs in Walman obligatorily carry pronominal affixes for core arguments subject (SUBJ) and object (OBJ). These affixes indicate person, number and gender of their respective arguments. Free noun phrases (NPs) representing the arguments are optional. The subject pronominal is always a prefix and the object pronominal is usually a suffix, unless the object is 1st or 2nd person (Brown and Dryer 2008, 531–2).

2 Thanks to Dr. Raúl Aranovich for illuminating commentary on LFG leading up to this paper.

Like in many other Papuan languages, SVCs are common in Walman, as in (1).[3]

(1) Ako runon n-orou n-arau
then 3sg.m 3sg.m-go 3sg.m-go.up

n-an nakol nngkal mnon.
3sg.m-be.at house small GEN.3sg.m

'Then he went up to his own little house.'
(Brown and Dryer 2008, 531)

(1) is a single sentence with multiple verbs -*orou* 'go', -*arau* 'go up', and -*an* 'be at'. Interpreted collectively in the SVC, these verbs mean approximately the same thing as the English predicate *go up to*. There are sentence-level particles in Walman which can modify the entire SVC, typically occurring between the first verb and that verb's subject. These include the negative particle *mon*, the perfective particle *tu*, and the future particle *ampa* (Brown and Dryer 2008, 546). When these particles modify an entire SVC, the feature values associated with their lexical entries are interpreted as feature values for the entire sentence. I will return below to the theoretical treatment of SVCs in LFG. I will call properties of the entire SVC "sentence-level" properties.

2 *And*-verbs

2.1 *And*-verbs descriptively

The main topic of Brown and Dryer (2008) is Walman's unusual *and*-verbs. These words are morphologically transitive verbs, bearing obligatory subject and object pronominal affixes, and optionally taking overt NPs as arguments. However their meaning is the basic meaning of the English word *and*, the meaning of the logical operator &. *And*-verbs do the work of NP coordinators in Walman.

3 Walman examples are written in the practical orthography used by Brown and Dryer (2008). See their footnote 2 for details. Following their convention, the first word in every complete sentence is capitalized, as are proper names.

(2)	To	[[ru]	**w-aro-n**	[na]]	y-anan.
	then	3sg.f	**3sg.f-and-3sg.m**	son	3pl-go.down

'Then she and the son went down.'

(Brown and Dryer 2008, 551)

In (2), the *and*-verb *-aro-* is coordinating the free pronoun *ru* (3rd-person singular feminine) and the noun *na* 'son'. Pronominal morphology on *-aro-* agrees with the coordinands as if they were its subject and object. Hereafter I will refer to the first and second coordinands of an *and*-verb as its subject and object, respectively, since they behave in every way as if they were its subject and object. In (2), the pronominal prefix *y-* on the sentence-level predicate *-anan* 'go down' agrees with the entire coordinated structure *ru waron na* 'she and the son'.

Walman *and*-verbs can only coordinate NPs, not clauses or adjectives.[4] Non-NP constituents are coordinated with the particle *o*, which can also coordinate NPs. There are two *and*-verbs in Walman, *-aro-* and *-a-*. *-Aro-* can only take 3rd-person objects, but *-a-* can take any NP object, and both *and*-verbs can take any NP subject. Brown and Dryer find no meaning distinction between the two *and*-verbs, or between the *and*-verbs and the coordinating particle *o*. They report that these are intersubstitutable without change of meaning in every instance where selectional restrictions are not violated (Brown and Dryer 2008, 537–8).

2.2 *And*-verbs in LFG

It is felicitous to assume that *and*-verb stems are of the same category as transitive verb stems in Walman, and that they take a subject and an object argument. Under this assumption they will naturally participate in the general lexical rules which assign pronominal affixes to verb stems. For instance, rules (3) and (4) assign the 3rd-person plural object pronominal and the 1st-person singular subject pronominal respectively.

4 Category-specificity is not uncommon for coordinators crosslinguistically. Languages with category-specific coordinators include Chechen, Chinese, Hausa, Japanese, Somali, Upper Kuskokwim Athabaskan, and Xârâcùù (Haspelmath 2004, 11–12).

(3) -y : [$V_{\text{trans stem}}$ __]$_{V(\text{trans stem})}$, (↑OBJ PRED) = 'PRO'
(↑OBJ INDEX PERS) = { }
(↑OBJ INDEX NUM) = pl

(4) m- : [__ V_{stem}]$_{V(\text{fin})}$, (↑SUBJ PRED) = 'PRO'
(↑SUBJ INDEX PERS) = {S}
(↑SUBJ INDEX NUM) = sg

These rules apply to all transitive verb stems, including *and*-verb stems. They do not apply to the coordinating particle *o*, which is of a distinct lexical category. Note that the value of the person feature (PERS) is a set. This set may contain S ("speaker") or H ("hearer"), or any combination thereof (including nothing, as in 3). If a PERS value includes S, it is 1st person. If it includes H and not S, it is 2nd person. Otherwise, it is 3rd person. This system for person values I adopt from Dalrymple and Kaplan (2000, §6). It is a good reflection of the semantic property that the feature PERS encodes, and proves to be a useful system for deriving PERS values in NP coordination.

An *and*-verb also must have the function of a coordinator for its arguments. Coordinators have two major functional roles crosslinguistically. First, they must distribute across-the-board properties to their arguments. Since there is no case marking in Walman, there are for NP coordinations no visible across-the-board properties — that is, properties which like case must be distributed to each member of a set of coordinands. Second, and importantly for this paper, coordinators must index properties of the coordinated structure that are non-distributive — that is, not across-the-board. For instance, consider *maro*, a fully inflected *and*-verb coordinating a 1st-person singular subject and a 3rd-person singular feminine object:

(5) m-aro-Ø
1sg-and-3sg.f

The PERS and NUM features of the coordinated structure are called non-distributive because their values are not necessarily shared with the PERS and NUM values of the individual coordinands. For instance, neither coordinand in (5) is 1st-person plural, although the entire coordinated structure is. Following King and Dalrymple (2004; via Kuhn and Sadler 2007, 6–7), I will refer to the non-distributive features of coordinations as

INDEX features. In f-structure, INDEX features occur in a feature labeled INDEX. The value of INDEX for a coordinated NP will always have the NUM value plural, and its PERS value will be the combined set of values from the individual coordinands' PERS features. The appropriate INDEX feature could be accounted for with the annotations in (6).[5]

(6) $(\uparrow \text{GF INDEX PERS}) \subseteq (\uparrow \text{INDEX PERS})$
$(\uparrow \text{INDEX NUM}) = \text{pl}$

Note that since the coordinands of the *and*-verb are its subject and object, they can be collectively referred to as its GFs. The f-structure for (5) can be represented as in (7), this structure being derived using the lexical rules in (3) and (4), the annotations in (6), and a lexical entry for *-aro-* specifying that its PRED value is 'and ⟨(SUBJ)(OBJ)⟩'.

(7)
$$\begin{bmatrix} \text{PRED} & \text{'and}\ \langle(\text{SUBJ})(\text{OBJ})\rangle\text{'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\text{S}\} \\ \text{NUM} & \text{pl} \end{bmatrix} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\text{S}\} \\ \text{NUM} & \text{sg} \end{bmatrix} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\ \} \\ \text{NUM} & \text{sg} \\ \text{GEN} & \text{f} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

It will be instructive to compare the structure in (7) to the structure of an ordinary coordinated NP structure. (8) represents the English coordinated NP *me and her*, given the kind of analysis used in Kuhn and Sadler (2007).

5 Writing a formal LFG grammar of Walman, the linguist would have to decide whether these annotations appear in the lexical entries for the *and*-verbs or whether they are instead annotations on the c-structure nodes containing the *and*-verbs. For the descriptive purposes of this paper, this distinction is immaterial. Importantly, *and*-verbs are associated with these annotations such that their f-structure projections look like the structure in (7).

(8) English coordinated NP: *me and her*

$$\begin{bmatrix} \text{CONJ-FORM} & \text{AND} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\text{S}\} \\ \text{NUM} & \text{pl} \end{bmatrix} \\ & \left\{ \begin{array}{l} \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\text{S}\} \\ \text{NUM} & \text{sg} \end{bmatrix} \end{bmatrix} \\ \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\ \} \\ \text{NUM} & \text{sg} \\ \text{GEN} & \text{f} \end{bmatrix} \end{bmatrix} \end{array} \right\} \end{bmatrix}$$

The major difference between the *and*-verb structure in (7) and the ordinary coordinated structure in (8) is that the latter has no GFs. There is no reason to suppose that coordinators have GFs in languages without *and*-verbs. Instead of appearing as grammatical functions, coordinands of ordinary coordinators appear in f-structure as a set, like that indicated in (8) by curly brackets. It is possible that Walman *and*-verb f-structures have such a set also, in addition to having their coordinands represented as GFs. The existence of such a set is not material to the observations I wish to make about the *and*-verb, so I will ignore the issue hereafter. I will represent the f-structures of *and*-verbs without the set, simply to save space. Importantly, Walman *and*-verbs *do* have GFs, as evidenced by their taking regular pronominal morphology like transitive verbs.

Another difference between (7) and (8) is that the former has a PRED feature where the latter has a CONJ-FORM feature. I do not know a particular reason to posit a form feature CONJ-FORM for Walman *and*-verbs, as Kuhn and Sadler (2007) do for ordinary coordinators. The PRED feature seems a more natural choice for Walman *and*-verbs, since they subcategorize for grammatical functions, a common property of verbal PRED values. On the other hand, a logical operator is a strange meaning for a predicate. Again this is orthogonal to the purpose of this paper, so I will assume without further argument that PRED is the correct feature type.

Both (7) and (8) have an INDEX feature representing the non-distributive features of the entire coordinate structure. In both structures, this INDEX feature is derived from the INDEX features of the individual coordinands.

3 Walman verb serialization

Walman SVCs are of at least two types in Crowley's (2002) and Aikhenvald and Dixon's (2006) SVC typologies, specifically with regard to argument sharing. There are some constructions in which the SUBJ function of each individual verb picks out the same referent — these are subject-sharing serial constructions, such as in (1) above. In (1), all three verbs share the same subject. Walman also has what Crowley (2002) calls inclusory serialization constructions. In this type, the subject of one verb is the subject and object of the preceding verb taken together. An example is in (9), where the referent of the subject of *kesi* 'we go out' is the referent of *kum* 'me' and *pelen* 'dog' together (1sg + 3sg = 1pl).

(9)	Kum	m-rachere-Ø	pelen	**k**-esi	nakol.
	1sg	1sg-chase-3sg.f	dog	**1pl**-go.out	house

'I chased the dog out of the house.'

(Brown and Dryer 2008, 551)

3.1 Previous LFG treatments of verb serialization

Modeling verb serialization in LFG, Bodomo (1996) (following Alsina 1994) makes a strong assumption of feature-sharing between verbs in an SVC. Their PRED features "compose" into a single predicate-chain feature labeled PREDCHAIN, which is the predicate of the entire SVC. Their other features, such as tense, aspect, and GFs, simply unify at the sentence level.[6]

6 Bodomo (1996) uses the c-structure annotation ↑ = H↓ to indicate the special type of co-heads whose PRED values compose while their other features unify. The details of Bodomo's PRED-composing process are unclear, though it is clear that the outcome is different from that of the predicate composition process in Arka et al. (2009) or Butt et al. (2003). These studies are concerned with the composition of predicates with unequal syntactic status, such that one predicate is comfortably analyzed as being contained within the other's set of arguments in f-structure:

PRED1 ⟨ (ARG1), … , PRED2 ⟨ … ⟩ ⟩

Since GFs unify in Bodomo (1996), the PREDCHAIN subcategorizes for a single subject argument SUBJ, and likewise for any other arguments present in a sentence. Following is the example Bodomo uses from Dagaare to illustrate predicate composition and sentence-level feature unification.

(10) Dagaare

Bayuo	da	ngmE-Ø	la	a	gan	lOO-Ø.
Bayuo	PAST	knock-PERF	FACT	DEF	book	cause.fall-PERF

'Bayuo knocked the book down.'

(Bodomo 1996, 11)

(11) Dagaare: f-structure for (10) (adapted from Bodomo 1996, 13)

$$\begin{bmatrix} \text{PREDCHAIN} & \text{'knock-cause.fall } \langle(\text{SUBJ})(\text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{ASPECT} & \text{PERF} \\ \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Bayuo'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{sg} \\ \text{GEND} & \text{m} \end{bmatrix} \end{bmatrix} \\ \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'book'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{sg} \\ \text{GEND} & \text{n} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

A similar f-structure (12) could be posited for the Walman shared-subject serial construction in (1), ignoring the connective *ako* 'then'.

In Bodomo's analysis the serialized PREDs fuse together to form a single predicate feature labeled PREDCHAIN. This seems a more comfortable analysis for Dagaare, since the verbs in Dagaare SVCs do not seem to be of unequal syntactic status such that one would be contained within the other in f-structure.

(12) F-structure for (1): preliminary

$$\begin{bmatrix} \text{PREDCHAIN} & \text{'go-go.up-be.at } \langle(\text{SUBJ})(\text{OBL})\rangle\text{'} \\ \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\ \} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{m} \end{bmatrix} \end{bmatrix} \\ \\ \text{OBL} & \begin{bmatrix} \text{PRED} & \text{'house'} \\ \text{POSS} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\ \} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{m} \end{bmatrix} \end{bmatrix} \\ \text{ADJ} & \{\ [\text{PRED 'small'}]\ \} \end{bmatrix} \end{bmatrix}$$

Bodomo (1996) is not the only LFG study to assume shared arguments in SVCs. Bodomo et al. (2003) and Beermann and Hellan (2002) both assume that every verb in an SVC shares a subject, and Seiss (2009) assumes that at least one argument is shared between verbs in an SVC.

3.2 Walman verb serialization in LFG

The studies listed above all capture the data in their target languages well. However, their assumptions of shared arguments do not hold for all Walman SVCs — for instance, those of the inclusory type. In this type, there is not a single SUBJ or OBJ for the entire sentence that is shared by all of the verbs in the SVC. Thus it cannot be said that GFs always unify in a straightforward way in Walman SVCs. In fact, since tense and aspect are often unmarked in Walman, there is often nothing that unifies across verbs in an SVC at all. Instead, the SVC appears as simply a series of verbs, each with its own set of GFs which is not necessarily identical to the GF sets of its sisters.[7]

7 Since each verb in the SVC is relatively autonomous, at least insofar as its set of GFs is concerned, it might be appropriate to think of the Walman SVC as a "clause chain" rather than a typical serial verb construction. Clause chaining is common in the Papuan languages. It usually consists of a series of simple clauses with non-finite verbs ("medial verbs") all partially subordinated ("cosubordinated" per Foley and Van

Therefore I propose the following f-structure-theoretic treatment of SVCs in Walman: Each SVC is a set of independent clausal f-structures, each with its own PRED and its own set of GFs, these GFs having a scheme of inter-clausal coindexation.[8] Arguments of separate verbs in an SVC thus may be linked or remain independent. This is a certainly simplistic and probably overly permissive treatment of Walman SVCs; however, the paucity of data currently available about Walman grammar[9] does not permit a more subtle analysis. It is therefore appropriate here to work under a simplistic assumption about the f-structure of Walman SVCs that is in keeping with their general nature, rather than attempting a more complicated analysis based on too little data. The treatment I propose here reflects the relative functional autonomy of each verb in a Walman SVC, while also allowing the possibility of argument sharing.[10]

Valin 1984) under one finite verb (called a "final verb", since these generally come at the end of the chain). Like SVCs, clause chains are single sentences. There is an indeterminate boundary between SVCs and clause chains, and I believe Lehmann (1988) is correct in identifying the difference as one of degree of grammaticalization, the SVC being a more highly grammaticalized clause chain. Thus I think it is not necessary to posit inherently different syntactic mechanisms for SVCs and clause chains. It is also worth noting that Walman SVCs are not canonical clause chains, since no verb in the SVC is (overtly) more finite than the rest.

8 I assume here a mechanism of coindexation rather than structure-sharing, though I see no immediate argument for choosing one over the other in this case.

9 Dryer's website indicates that a descriptive grammar is forthcoming.

10 A complete analysis of Walman SVCs would require an explicit account of how the system of inter-clausal coindexation is generated by the grammar. Such an account is not possible however without more information about Walman grammar. Only some very general comments are possible: (i) Presumably only features whose INDEX values are non-contradictory may be coindexed with one another. (ii) The relative order of the clauses is presumably important for determining what is coindexed with what. This information need not be represented in f-structure however, since it is available from c-structure information. In formal LFG, information about c-structure order is available to f-structure via f-precedence relations and head-precedence relations (Crouch et al. 2005, ref.#N4.2.9; Zaenen and Kaplan 1995, 226). (iii) Some interplay with semantic or pragmatic structures may also need to play a role, if it is found that c-structure and f-structure cues are inadequate for completely determining the scheme of coindexation.

The revised f-structure for (1) will thus appear as in (13).[11] Note that separate reflexes of the shared subject *runon* (3sg.m) appear in the local f-structures of each verb, and these are coindexed with one another.

(13) F-structure for (1): revised

$$\left[\left\{\begin{array}{l}
\left[\begin{array}{ll}
\text{PRED} & \text{'go}\ \langle(\text{SUBJ})\rangle\text{'} \\
\text{SUBJ}_1 & \left[\begin{array}{ll}
\text{PRED} & \text{'PRO'} \\
\text{INDEX} & \left[\begin{array}{ll}
\text{PERS} & \{\ \} \\
\text{NUM} & \text{sg} \\
\text{GEN} & \text{m}
\end{array}\right]
\end{array}\right]
\end{array}\right] \\
\left[\begin{array}{ll}
\text{PRED} & \text{'go.up}\ \langle(\text{SUBJ})\rangle\text{'} \\
\text{SUBJ}_1 & \left[\begin{array}{ll}
\text{PRED} & \text{'PRO'} \\
\text{INDEX} & \left[\begin{array}{ll}
\text{PERS} & \{\ \} \\
\text{NUM} & \text{sg} \\
\text{GEN} & \text{m}
\end{array}\right]
\end{array}\right]
\end{array}\right] \\
\left[\begin{array}{ll}
\text{PRED} & \text{'be.at}\ \langle(\text{SUBJ})(\text{OBL})\rangle\text{'} \\
\text{SUBJ}_1 & \left[\begin{array}{ll}
\text{PRED} & \text{'PRO'} \\
\text{INDEX} & \left[\begin{array}{ll}
\text{PERS} & \{\ \} \\
\text{NUM} & \text{sg} \\
\text{GEN} & \text{m}
\end{array}\right]
\end{array}\right] \\
\text{OBL} & \text{“his little house”}
\end{array}\right]
\end{array}\right\}\right]$$

A further question arises with regard to the f-structure of inclusory SVCs. In an inclusory SVC, the SUBJ of one verb is the aggregate of the SUBJ and OBJ of the previous verb. Assuming that the relationship between the arguments of the two verbs in the SVC is to be represented in the syntax, how can this relationship be represented using the proposed scheme of inter-clausal coindexation? In other words, with what f-structure entity is the SUBJ of the second verb coindexed? It would appear that the first verb must construct a coindexable entity in f-structure for this purpose, whose INDEX features are the aggregate of the INDEX features of its arguments. The

11 As in (12), any f-structure reflex for *ako* 'then' is omitted in (13). The NP *nakol nngkal mnon* 'his little house' is represented as an argument only of the final verb, mostly in the interest of space, though in principle (and without having consulted with a Walman speaker or expert) it could be an oblique argument for more than one of the verbs.

f-structure of (9) can thus be represented as in (14), with the INDEX of *mrachere* 'chase' coindexed with the SUBJ of *kesi* 'go out'.[12,13]

(14)
$$\begin{bmatrix} \left\{ \begin{array}{l} \begin{bmatrix} \text{PRED} & \text{'chase } \langle(\text{SUBJ})(\text{OBJ})\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\text{S}\} \\ \text{NUM} & \text{sg} \end{bmatrix} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'dog'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\ \} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{f} \end{bmatrix} \end{bmatrix} \\ \text{INDEX}_1 & \begin{bmatrix} \text{PERS} & \{\text{S}\} \\ \text{NUM} & \text{pl} \end{bmatrix} \end{bmatrix} \\ \\ \begin{bmatrix} \text{PRED} & \text{'go.out } \langle(\text{SUBJ})(\text{OBL})\rangle\text{'} \\ \text{SUBJ}_1 & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{PERS} & \{\text{S}\} \\ \text{NUM} & \text{pl} \end{bmatrix} \end{bmatrix} \\ \text{OBL} & \text{“house”} \end{bmatrix} \end{array} \right\} \end{bmatrix}$$

12 I assume that the INDEX node is directly attached at the local f-structure of the verb which constructs it.

13 Alert readers may notice an inconsistency in my treatment of pronominals in (14). There is a 3rd-person singular feminine object pronominal suffix on *-rachere-* 'chase', and this suffix generally supplies the value 'PRO' for the PRED feature of its GF. However, the value of PRED here is supplied by the free NP *pelen* 'dog' instead. A more consistent theoretical treatment here would be to consider the f-structure projection of the NP *pelen* to be an adjunct (ADJ) of *-rachere-*, coindexed with its OBJ. In this way the PRED values of the pronominal and the NP would not conflict, yet *pelen* could still be structurally identified with the OBJ of *-rachere-*. This would be an f-structural equivalent of Jelinek's (1984) tree-theoretic analysis of nonconfigurationality in Warlpiri.

It seems to me that representing this kind of scheme in (14) would distract the reader from the coindexation most relevant to the inclusory serialization, namely that between the SUBJ of *-esi* and the INDEX of *-rachere-*. I have therefore (inconsistently) treated the object-pronominal on *-rachere-* as if it were an agreement marker in this case, in an attempt to maintain focus on the central descriptive issues of this paper. This treatment is reminiscent of Hale's (1983) original analysis of Warlpiri, in which overt nominals are “linked” to verbal argument positions if and only if there are overt nominals. (See Hale's Linking Rule, p. 14.)

4 Inclusory serialization and Walman *and*-verbs

In a language like Walman that allows inclusory serialization, any transitive verb must have the capacity to construct an INDEX feature composed of the coordinated INDEX features of its GFs. This is because any transitive verb may in principle be the first verb in an inclusory SVC, and the second verb will need some entity in f-structure with which its SUBJ may be coindexed. This entity can be accounted for with the following argument-coordinating annotations:

(↑GF INDEX PERS) ⊆ (↑INDEX PERS)
(↑INDEX NUM) = pl

Note that these annotations are identical to the annotations for *and*-verbs proposed above in (6).[14]

It is therefore not a special property of *and*-verbs that they coordinate their arguments in f-structure. Since all Walman transitive verbs can coordinate their arguments in f-structure, Walman *and*- verbs coordinate their arguments simply by virtue of the fact that they are Walman transitive verbs. The functional nature of the crosslinguistically unusual Walman *and*-verb is thus parasitic on the functional nature of Walman transitive verbs in general.

5 *And*-verbs in c-structure

Given the claim that *and*-verbs have functional properties identical to ordinary transitive verbs in languages with inclusory serialization, it might be expected that *and*-verbs should be more common than they are. Inclusory serialization has been widely documented,[15] but *and*-verbs have only been documented in the small Torricelli family.[16] I propose that the crosslinguistic

14 Writing a formal grammar of Walman one would have to decide whether these annotations appear in the lexical entries of transitive verbs or as annotations on the c-structure nodes containing them. The latter may be preferable, so that the annotations could appear only in syntactic environments where they are necessary (i.e. inclusory SVCs). For the descriptive purposes of this paper, it suffices to say that the annotations are somehow associated with Walman transitive verbs such that they project a coordinator-like f-structure, like the projection of *mrachere* 'chase' in (14). Cf. footnote 5, above.

15 Crowley (2002, 41) notes its presence in Paamese; and chapters from Aikhenvald and Dixon (2006) demonstrate inclusory SVCs in Ewe (p.130), Dumo (214), Thai (167), and Mwotlap (231).

16 Similar categories have been observed in some Austronesian languages of Timor

rarity of *and*-verbs may be due to the complexity inherent in their categorial affiliation.

Walman *and*-verbs in themselves behave like ordinary Walman transitive verbs. They take two NP arguments, and pronominal morphology agrees with these as subject and object. They share the entire pronominal paradigm of ordinary transitive verbs. Thus morphologically, and with regard to the category of their arguments, *and*-verbs are typical of the category transitive verb. Also, sentence-level particles such as the negative, perfect, and future particles — which typically occur between the first verb in a sentence and its subject — may occur between the *and*-verb and its subject, if the *and*-verb is the first verb in the sentence (Brown and Dryer 2008, 546–7). See (15). In this way *and*-verbs are distributed in c-structure as if they are ordinary transitive verbs.

(15)	Rita	ampa	w-aro-Ø	Millie	y-orou	Achapei.
		FUT	3sg.f-and-3sg.f		3pl-go	Aitape

'Rita and Millie will go to Aitape.'

(Brown and Dryer 2008, 547)

Not all c-structure evidence indicates that the *and*-verb is an ordinary transitive verb, however. Strangely, the maximal projection of the *and*-verb in c-structure is a noun phrase. It can participate in possessive constructions (16), it can be a nominal predicate (17), the object of an adposition (18), it can be modified by the NP-modifying adjunct *alpa* 'only' which cannot modify verb phrases (19), it cannot stand on its own as a grammatical sentence (20), and it can even be coordinated by other *and*-verbs, which must have NP coordinands (21).

(16)	Kompowaley	w-kum	[m-aro-y]	w-orou	kelki	pa.
	story	GEN-1sg	1sg-and-3pl	3sg.f-go	end	DEM

'Our ([my and their]) story has come to an end.'

Brown and Dryer (2008, 541)

(Nishiyama and Kelen 2007, §§11.6, 12.1; van Klinken 2000, §§3.2, 4.2), and Broadwell (2006) describes a Choctaw verb which can be used to mean 'and.'

(17) ... chu-tey w-ri lasi [Kampail w-aro-Ø Slim].
wife-pl GEN-3pl name 3sg.f-and-3sg.f
'... their wives' names were [Kampail and Slim].'
Brown and Dryer (2008, 541)

(18) Nyi y-roul [Rita w-aro-Ø Millie] wor.
lamp 3pl-hang 3sg.f-and-3sg.f above
'The lamp is hanging above [Rita and Millie].'
Brown and Dryer (2008, 543)

(19) [Rita w-aro-Ø Millie] alpa-y y-orou Achapei.
3sg.f-and-3sg.f only-pl 3pl-go Aitape
'Only [Rita and Millie] went to Aitape.'
Brown and Dryer (2008, 543)

(20) *[Runon n-aro-n au].
3sg.m 3sg.m-and-3sg.m elder.brother
'He is with elder brother.' / 'He and elder brother.'
Brown and Dryer (2008, 550)

(21) [[Steve n-aro-Ø Mary] y-p-a] k-orou tesin.
3sg.m-and-3sg.f 3pl-1.OBJ-and 1pl-go town
'[[Steve and Mary] and I] went to town.'
Brown and Dryer (2008, 544)

An interesting piece of evidence for the categorial complexity of the *and*-verb may be found in the details of the distribution of the negative particle *mon* (see Brown and Dryer 2008, 551–2). In its unmarked use, *mon* appears immediately before the first verb in a sentence, after its subject if the subject is overt. When an *and*-verb occurs with no NP arguments at the beginning of a sentence, *mon* may appear in one of two places. Sometimes

mon is positioned first, as if the *and*-verb is a verb (22); sometimes *mon* appears after the *and*-verb, as if the *and*-verb is the NP subject of the following verb (23). Thus whatever c-structure rule determines the placement of *mon* is ambiguous in this type of sentence. The position of the negative particle in (22) provides evidence that although the maximal projection of an *and*-verb is an NP, the *and*-verb itself is still interpretable in c-structure as a verb.

(22) Mon n-aro-n y-ara.
NEG 3sg.m-and-3sg.m 3pl-come
'They ([he and him]) didn't come.'
Brown and Dryer (2008, 552)

(23) N-aro-n mon y-ara.
3sg.m-and-3sg.m NEG 3pl-come
'They ([he and him]) didn't come.'
Brown and Dryer (2008, 552)

The categorial complexity of the Walman *and*-verb is thus evident from its behavior in c-structure: It is a transitive verb whose maximal projection is an NP. This presents a difficult problem for a c-structure analysis of the *and*-verb. Apparently an unusual rewrite rule such as (24) is necessary.

(24) NP → (NP) V (NP)

This rule would have to be restricted to instances where V is an *and*-verb, and in any case it violates the common assumption that a phrase must contain a head of its own category.[17,18] The reason for the crosslinguistic rarity of the *and*-verb may therefore be the complexity of its categorial affiliation.

17 Though see Dryer (2004) for a discussion of NPs without nouns crosslinguistically.

18 Broadwell (p.c.) points out that (24) may be present in the grammar independently if Walman has head-internal relative clauses. From the brief description in Brown and Dryer (2008, 533), it appears that Walman relative clauses are not head-internal.

6 The utility of an LFG-like model to descriptive linguistics

In this section, the content of the preceding sections will be briefly discussed at a meta-analytical level to demonstrate the utility of an LFG-like model for descriptive linguistics. By "LFG-like" I mean a model in which different kinds of grammatical relationships are modeled separately — especially relevant here is that f-structure and c-structure are modeled separately. This kind of model will be contrasted with a "tree-theoretic" model[19] in which constituent structure is taken to be primary, and all other grammatical relationships are ultimately derived from constituency.

Sections 2.2 and 3.2 are analyses of Walman *and*-verbs and Walman serialization. Drawing from these analyses, section 4 presents the observation that a seemingly unusual property of *and*-verbs (that they coordinate their arguments) is in fact a property of all Walman transitive verbs. This observation is a simple one, and it sheds some light on the nature of Walman *and*-verbs and their position within Walman grammar. These analyses and the resulting observation are all conducted within a limited domain — that domain which LFG models as f-structure. The similarity between Walman *and*-verbs and Walman transitive verbs lies in the way that they coordinate their arguments (GFs) in f-structure such that the resulting coordination is available externally, for arguments of other verbs to be identified with. The resulting comparison of *and*-verbs and transitive verbs in Walman thus sidesteps the difficult issues of categorial affiliation discussed in section 5.

In contrast, a tree-theoretic model of grammar cannot sidestep the issues of category and constituency, since these are taken to be primary. A linguist using such a model would therefore be forced to confront the difficult issue of the *and*-verb's category and maximal projection before proceeding to make any observation about its functional similarity to transitive verbs. It remains for such an analysis to be attempted to discover the extent to which this issue would be problematic in practice. However it is clear that an LFG-like model provides a more direct avenue for describing phenomena in the realm of f-structure, in isolation from issues of category and constituency.

19 This terminology is from Ackerman (2009).

Summary

Due to the presence of inclusory SVCs in Walman, all Walman transitive verbs must be able to construct an INDEX feature composed of the INDEX features of their arguments. In this respect Walman transitive verbs have a coordinative function, and Walman *and*-verbs are in fact typical Walman transitive verbs. *And*-verbs however do not behave like ordinary transitive verbs in c-structure. In a tree-theoretic approach, the functional relationship between Walman *and*-verbs and Walman transitive verbs may be difficult to discern, since the c-structural complexity of the *and*-verb would be a distraction. An LFG-like framework in which f-structure is modeled separately from c-structure thus allows for a more straightforward description of phenomena like the Walman *and*-verb.

References

Ackerman, Farrell (2009) A Construction-theoretic approach to possessive relatives. UCD Linguistics Colloquium, November 2009.

Aikhenvald, Alexandra Y., and R. M. W. Dixon (2006) *Serial Verb Constructions*. Oxford University Press.

Alsina, Alex (1994) A Theory of complex predicates: Evidence from causatives in Bantu and Romance. unpublished Ms., Stanford University.

Arka, I Wayan, Mary Dalrymple, Meladel Mistica, Suriel Mofu, Avery Andrews, and Jane Simpson (2009) A linguistic and computational morphosyntactic analysis for the applicative -i in Indonesian. *Proceedings of the LFG09 Conference*. CSLI Publications, pp. 85–105.

Beerman, Dorothee A., and Lars Hellan (2002) VP-chaining in Oriya. *Proceedings of the LFG02 Conference*. CSLI Publications, pp. 40–56.

Bodomo, Adams B. (1996) Complex predicates: The case of serial verbs in Dagaare and Akan. *Proceedings of the First LFG Conference*. CSLI Publications, pp. 89–103.

Bodomo, Adams B., Olivia S. C. Lam, and Natalie S. S. Yu (2003) Double object and serial verb benefactive constructions in Cantonese. *Proceedings of the LFG03 Conference*. CSLI Publications, pp. 59–74.

Broadwell, George Aaron (2006) *A Choctaw Reference Grammar*. University of Nebraska Press.

Brown, Lea, and Matthew S. Dryer (2008) The verbs for 'and' in Walman, a Torricelli language of Papua New Guinea. *Language* 84 (3), 528–565.

Butt, Miriam, Tracy Holloway King, and John T. Maxwell III (2003) Complex predicates via restriction. *Proceedings of the LFG03*

Conference. CSLI Publications, pp. 92–104.

Crouch, Dick, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman (2005) XLE Documentation: Grammatical notations. Palo Alto Research Center (PARC). Available online: http://www2.parc.com/isl/groups/nltt/xle/doc/notations.html

Crowley, Terry (2002) *Serial Verbs in Oceanic: A descriptive typology*. Oxford University Press.

Dalrymple, Mary (2006) Lexical Functional Grammar. in Keith Brown, ed. *Encyclopedia of Language and Linguistics*. Elsevier, pp. 82–94.

Dalrymple, Mary, and Ronald M. Kaplan (2000) Determiner agreement and noun conjunction. *Journal of Linguistics* 40 (1), 69–104.

Dryer, Matthew S. (2004) Noun phrases without nouns. *Functions of Language* 11, 43–76.

Foley, William A. and Robert D. Van Valin, Jr. (1984) *Functional Syntax and Universal Grammar*. Cambridge University Press.

Hale, Kenneth (1983) Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory* 1 (1), 5–47.

Haspelmath, Martin (2004) Coordinating constructions: An overview. in Haspelmath, ed. *Coordinating Constructions*. Benjamins, pp. 3–40.

Jelinek, Eloise (1984) Empty categories, case, and configurationality. *Natural Language and Linguistic Theory* 2 (1), 39–76.

King, Tracy H. and Mary Dalrymple (2004) Determiner agreement and noun conjunction. *Journal of Linguistics* 40, 69–104.

van Klinken, Catharina (2000) From verb to coordinator in Tetun. *Oceanic Linguistics* 39 (2), 350–363.

Kuhn, Jonas and Louisa Sadler (2007) Single conjunct agreement and the formal treatment of coordination in LFG. *Proceedings of the LFG07 Conference*. CSLI Publications, pp. 302–322.

Lehmann, Christian (1988) Towards a typology of clause linkage. in John Haiman and Sandra A. Thompson, eds. *Clause Combining in Grammar and Discourse*. Benjamins, pp.181–225.

Nishiyawa, Kunio, and Herman Kelen (2007) *A Grammar of Lamaholot, Eastern Indonesia: The morphology and syntax of the Lewoingu dialect*. LINCOM Europa.

Seiss, Melanie (2009) On the difference between auxiliaries, serial verbs and light verbs. *Proceedings of the LFG09 Conference*. CSLI Publications, pp. 501–519.

Zaenen, Annie, and Ronald M. Kaplan (1995) Formal devices for linguistic generalizations: West Germanic word order in LFG. in Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, eds. *Formal Issues in Lexical-Functional Grammar*. CSLI Publications, pp. 215–240.

RUSSIAN VERBAL AFFIXES IN THE PROJECTION ARCHITECTURE

Jean-Philippe Marcotte and Kateryna Kent
University of Minnesota

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

Russian perfectivity paradigms raise a complex network of formal issues for the projection architecture of LFG, including the structure of morphological representation and its relationship to the c-, f- and a-structures, with some consequences that appear to favor description-by-analysis over co-description for semantic interpretation. This paper presents the data and navigates its formal implications, suggesting in the end that Paradigm Function Morphology allows a clear description of the Russian facts that is equally compatible with both co-description and description-by-analysis, while permitting the elimination of m-structure.

1 Introduction

A foundational idea of Lexical-Functional Grammar is that different types of grammatical information may belong in different, related structures. Originally, Kaplan and Bresnan (1982) endowed the framework with the familiar c(onstituent)-structure and f(unctional)-structure, with the c-structure containing information necessary for stating generalizations about constituency, syntactic category, and linear order, and the f-structure containing information necessary for stating generalizations about grammatical function, agreement, control and raising, and so on. The architecture of the grammar assumed was as sketched in Figure 1, where ϕ is the correspondence function mapping from pieces of c-structure to pieces of f-structure.

$$\text{c-structure} \xrightarrow{\phi} \text{f-structure}$$

Figure 1: The original LFG architecture

Kaplan (1987) generalized the notion of correspondence function, arguing that the fundamental grammatical mapping between form and meaning, conceptualized as the function Γ, can be decomposed in an arbitrary number of component correspondence functions, each between a domain structure and its range structure projection. The array of structures is to be motivated "descriptively or linguistically", on the basis of "sound theoretical argumentation" (Kaplan 1987: 363), and the correspondence functions between these structures can be composed to recover Γ.

Since this work, LFG grammars have been understood as being stated within this projection architecture. The approach explicitly requires engagement with issues like: the kinds of linguistic generalizations that need to be stated, the kinds of structures that should be assumed in order for these generalizations to be formulated perspicuously, and how these structures are to be projected one from another.

This last issue can be reframed as an issue of interface: a correspondence function between two structures is a direct interface between them. Since some linguistic generalizations are interface generalizations, it stands to reason that the set of correspondence functions assumed should reflect the linguistic generalizations that

need to be stated. Furthermore, the type of any particular correspondence function should be primarily driven by the properties of the structures between which it interfaces.

We argue that a close look at Russian perfectivity paradigms can help shed light on these formal issues. These complex datasets have properties that place them at the intersection of several different proposed structure types: though we will mention in passing their interaction with argument structure, semantic structure, and information structure, we focus here on their morphological and morphosyntactic properties. In particular, we show that a simple sublexical-rule and m-structure treatment of these properties imposes constraints on the syntax–semantics interface; but a treatment within Paradigm Function Morphology (PFM; Stump 2001) does not. Since the syntax–semantic interface is generally viewed as the correspondence function σ between f- and s(emantic)-structure, a morphological treatment orthogonal to this interface is preferable. Our PFM morphology, unlike that of Sadler and Nordlinger (2004), also obviates the need for m-structure altogether.

We review some theoretical prerequisites in Section 2, present the Russian data in Section 3, and in Section 4 illustrate some constraints imposed on the syntax–semantics interface by a sublexical rules and m-structure analysis. In Section 5 we provide an alternative PFM account, which allows us to dispose of m-structure in Section 6.

2 Structures and correspondences: Theoretical overview

It is now generally accepted that there is a correspondence function σ from f-structure to s-structure (Dalrymple 1999), and that the original correspondence function ϕ from c- to f-structure is best understood as the composition of α from c- to a(rgument)-structure, and λ from a-structure to f-structure (Butt et al. 1997). Some recent work has also addressed i(nformation)-structure (Butt and King 2000; O'Connor 2006) and p(honological/prosodic)-structure (Butt and King 1998; Mycock 2006).[1] But there is little agreement as to how these structures project relative to one another: the architecture of LFG grammars is an issue that remains unresolved.

One manifestation of this is the debate over m(orphosyntactic)-structure, concerning among other things its formal type, its placement within the projection architecture, and the type of correspondence function that links it to the rest of the grammar. These issues are addressed separately below.

Placement of m-structure Miriam Butt originally conceived of the m-structure as a "junk structure" (p.c.), in which were placed attribute–value pairs necessary

[1] P-structure is envisioned by Butt and King (1998) as the syntax–phonology interface and intended to contain prosodic information "which feeds into a further phonological component". Their p-structure attribute P-FORM has a phonemic form as its value. Projection to p-structure for us is limited to this attribute and involves no prosodic information.

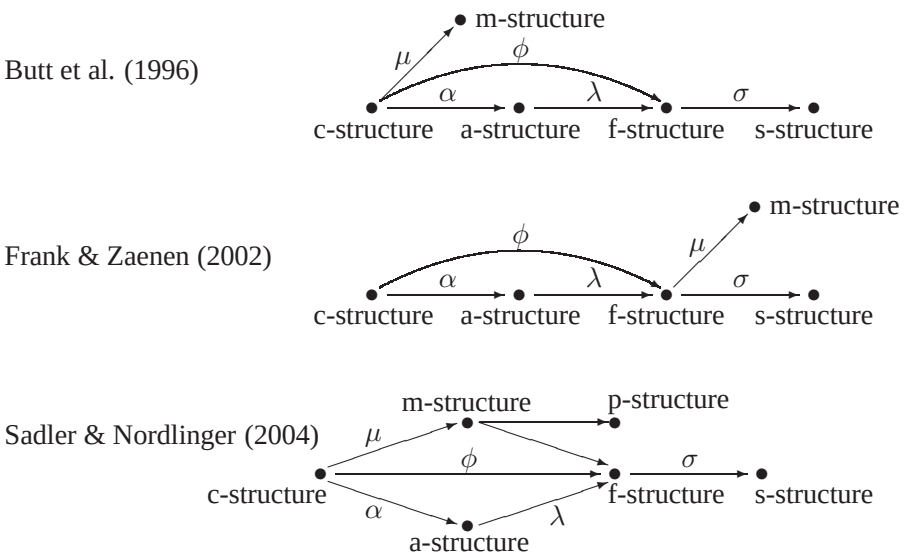


Figure 2: Proposals for placement of m-structure in the projection architecture

to an analysis of auxiliary constructions (Butt et al. 1996) but deemed to clutter the f-structure. In this conception, m-structure is a projection of the c-structure, as at the top of Figure 2, though it does not itself project another structure. In contrast, Frank and Zaenen (2002) argue that the m-structure is better thought of as a projection of the f-structure; this is shown in the middle of Figure 2. Finally, though it is implicit in their presentation, Sadler and Nordlinger (2004) envision m-structure as a projection of the c-structure, and as projecting to both the f-structure and the p-structure — see the bottom of Figure 2.

Formal type of m-structure These treatments of the m-structure also differ in the type of structure assumed. Both Butt et al. (1996) and Frank and Zaenen (2002) assume that m-structures have the same formal type as f-structures: they are functions from attributes to values, where values can be other such functions. In contrast, Sadler and Nordlinger (2004), working within a Paradigm Function Morphology framework (Stump 2001), assume that m-structures are sets of morphosyntactic properties, where such a property can also be a set.

Correspondence function type For Butt et al. (1996), the m-structure is projected from the c-structure; relevant grammatical statements are found as annotations on c-structure nodes, nonterminal and terminal (i.e. lexical entries), alongside f-structure annotations. These m-structure annotations, like f-structure ones, make direct reference to c-structure nodes. As such, the m-structure is a description-by-analysis of the c-structure, and c-structure elements co-describe the f- and the m-structure. This is also true of the c- to m-structure correspondence of Sadler and Nordlinger (2004).

For Frank and Zaenen (2002), the m-structure is projected from the f-structure; m-structure annotations make direct reference to f-structures and, like f-structure annotations, are found on c-structure nodes. Thus the m-structure is a description-by-analysis of the f-structure, and c-structure elements co-describe the f- and the m-structure.

In the case of Sadler and Nordlinger (2004), there are two more correspondence functions to consider: one from m- to p-structure, and another from m- to f-structure. In the former case, the correspondence function is the paradigm function: it specifies how the m-structure, a set of morphosyntactic features, is to be realized phonologically. In the latter case, transfer rules convert morphosyntactic feature sets to f-descriptions, providing a description-by-analysis of the f-structure. Though both the p- and the f-structure are projected from the m-structure, they are not co-described.

On the basis of evidence to be presented below pertaining to the perfectivity of verb forms in Russian, we take exception to most of the above: there is no need to assume a separate level of m-structure, and thus no need for it to have a formal type. Though our PFM approach to morphology is in line with Sadler and Nordlinger (2004), we have need of only a paradigm function as a correspondence function projecting a p-structure from the c-structure.

3 Russian perfectivity paradigms

Every verb stem in Russian is classified as either imperfective or perfective.[2] Verb roots have an inherent perfectivity value, for example imperfective for *chita-* in (1a), and more complex stems can be augmented with affixes which alter the perfectivity of their base, as with *pro–* in (1b).[3]

(1) a. *Petya chita–l* *knig–u*
Peter read.$\boxed{\text{IMPF}}$–MASC.SG.PST book–ACC
'Peter read/was reading a book.'

b. *Petya pro–chita–l* *knig–u*
Peter $\boxed{\text{PRF}}$–read.~~IMPF~~–MASC.SG.PAST book–ACC
'Peter has read the book.'

Prefixed stems are usually perfective (2a). Imperfective stems are either bare roots (2b), or have a *secondary imperfective* suffix, realized as *–yva* in (2c):

[2]This is true of other Slavic languages as well; however we focus exclusively on Russian.

[3]We notate the perfectivity of verb forms with a box around the relevant value, and strikethroughs for any other perfectivity values. Abbreviations in glosses are as follows: ACC = accusative; CML = cumulative; DAT = dative; DISTR = distributive; GEN = genitive; IMPF = imperfective; INCEP = inceptive; INF = infinitive; MASC = masculine; PL = plural; PRDR = perdurative; PREP = prepositional; PRF = perfective; PST = past; SG = singular.

(2) a. *s–pisa–tj* ‘copy’ (PRF)

b. *pisa–tj* ‘write’ (IMPF)

c. *s–pis–yva–tj* ‘copy’ (IMPF)

Russian verbal prefixes are not a homogeneous class: traditionally they are divided into lexical and superlexical prefixes (Ramchand 2004; Romanova 2004; Svenonius 2004), these two types differ in morphosyntactic properties as outlined below.

Lexical prefixes (LP henceforth) typically have the following properties:

(3) A. Usually attach to telic stems;
B. Allow the secondary imperfective suffix (*–yva/iva*, *–va*, *–a*);
C. Disallow prefix stacking: only one LP can attach to each verb stem;
D. May license argument structure, i.e. turn an intransitive verb into a transitive one;
E. Form an idiosyncratic meaning of the verb.

The following examples from Russian, explained below, illustrate these properties:

(4) *Mne nuzhno s–jezdi–tj v Moskv–u*
I.DAT necessary LP.$\boxed{\text{PRF}}$–go.~~IMPF~~–INF to Moscow–ACC
‘It is necessary for me to go to Moscow.’

(5) a. *rabota–tj*
work.$\boxed{\text{IMPF}}$–INF
‘work’

b. *za–rabota–tj*
LP.$\boxed{\text{PRF}}$–work.~~IMPF~~–INF
‘earn/*work’

c. *za–rabat–yva–tj*
LP.~~PRF~~–work.~~IMPF~~–$\boxed{\text{IMPF}}$–INF
‘earn/*work’

d. ∗*rabat–yva–tj*
work.~~IMPF~~–$\boxed{\text{IMPF}}$–INF
‘work’

(6) a. *lozhy–tj*
put.$\boxed{\text{IMPF}}$–INF
‘put’

b. *po–lozhy–tj*
LP.$\boxed{\text{PRF}}$–put.~~IMPF~~–INF
‘put down’

c. ∗*na–po–lozhy–tj*
LP.$\boxed{\text{PRF}}$–LP.~~PRF~~–put.~~IMPF~~–INF

(7) a. *za–rabota–tj* *piatj dollar–ov*
LP.$\boxed{\text{PRF}}$–work.~~IMPF~~–INF five dollars–GEN.PL
'to earn five dollars'

b. ∗*rabota–tj* *piatj dollar–ov*
work.$\boxed{\text{IMPF}}$–INF five dollars–GEN.PL
'to earn five dollars'

Example (4) shows the LP *s–* perfectivizing the telic imperfective verb of motion *jezdi–tj* 'to go',[4] without otherwise affecting its meaning or argument structure. The examples in (5) show the process of secondary imperfectivization: (5a) shows that the bare verb stem 'work' is imperfective; it can be perfectivized by an LP in (5b), and re-imperfectivized with the secondary imperfective suffix in (5c). However, the secondary imperfective cannot attach to the original imperfective stem: (5d) is ungrammatical. Note also that the LP verbs in (5b,c) obligatorily translate to 'earn', not 'work'. Examples (6a–c) show that attaching more than one LP to the verb stem results in an ungrammatical string. Finally, examples (7a,b) show that the intransitive verb 'work' is not just perfectivized by the LP *za–*, but also transitivized to 'earn': the direct object is not licensed in the absence of the prefix.

Superlexical prefixes (SPs henceforth) are drawn from the same set of forms as LPs but differ in their function. Their typical properties (8A–E) below are directly comparable to (3A–E) above, respectively:

(8) A. Attach to atelic stems;
B. Only allow secondary imperfective if there is also an LP;
C. Allow prefix stacking: SPs can stack on top of each other and on top of LPs;
D. Do not license argument structure;
E. SPs do not change the meaning of lexical root but simply add information about the progress of the event.

These properties are illlustrated by the examples below:

(9) a. *bega–tj*
run.$\boxed{\text{IMPF}}$–INF
'run'

b. *pro–bega–tj*
SP.PRDR.$\boxed{\text{PRF}}$–run.~~IMPF~~–INF
'run' (for some period of time)

c. ∗*pro–beg–iva–tj*
SP.PRDR.~~PRF~~–run.~~IMPF~~–$\boxed{\text{IMPF}}$–INF

[4]Romanova (2004) demonstrates that verbs of directed motion in Russian are telic, whereas verbs of manner of motion are atelic.

(10) a. *kry–tj*
cover.$\boxed{\text{IMPF}}$–INF
'cover'

b. *na–kry–tj*
LP.$\boxed{\text{PRF}}$–cover.~~IMPF~~–INF
'cover'

c. *na–kr–yva–tj*
LP.~~PRF~~–cover.~~IMPF~~–$\boxed{\text{IMPF}}$–INF
'cover'

d. *po–na–kr–yva–tj*
SP.DISTR.$\boxed{\text{PRF}}$–LP.~~PRF~~–cover.~~IMPF~~–~~IMPF~~–INF
'cover many objects'

(11) a. *proda–tj*
sell.$\boxed{\text{IMPF}}$–INF
'sell'

b. *ras–proda–tj*
SP.CML.$\boxed{\text{PRF}}$–sell.~~IMPF~~–INF
'sell out'

c. *po–ras–proda–tj*
SP.DISTR.$\boxed{\text{PRF}}$–SP.CML.~~PRF~~–sell.~~IMPF~~–INF
'sell out piece by piece'

(12) a. *dva mesiatsa pro–lezha–tj v bolnitse*
two months.GEN SP.PRDR.$\boxed{\text{PRF}}$–lie.~~IMPF~~–INF in hospital.PREP
'to spend two months in a hospital'.

b. *dva mesiatsa lezha–tj v bolnitse*
two months.GEN lie.$\boxed{\text{IMPF}}$–INF in hospital.PREP
'to be in a hospital for two months'

(13) a. *za–pe–tj pesnju*
SP.INCEP.$\boxed{\text{PRF}}$–sing.~~IMPF~~–INF song
'to start singing a song'

b. *pe–tj pesnju*
Sing.$\boxed{\text{IMPF}}$–INF song
'sing a song'.

Example (9b) shows that SPs can attach to an atelic stem like 'run' in (9a); however the resulting stem cannot take the secondary imperfective; see (9c). The examples in (10) show that SPs can in fact co-occur with the secondary imperfective, if an LP is also part of the same form: (10a–d) show the incremental assembly of such a verb. Examples (11a–c) illustrate the ability of the SPs to stack: compare the single-SP verb in (11b) to the two-SP verb in (11c). In (12a) the SP *pro–* is attached to the imperfective stem *lezha–tj* 'lie', perfectivizing the verb and adding a perdurative meaning, indicating that the stay in the hospital is over. Compare (12b), which lacks the SP, and means that the person may still be in the hospital.

In (13a) the inceptive SP *za–* is illustrated. There is no way to manipulate the argument structure of any of the verbs in (9–13) by adding or taking away an SP.

Perfectivity and tense Russian distinguishes morphologically between past and nonpast tenses (we use 'preterite' below): there is one set of agreement suffixes for past tense verbs, and another set for nonpast verbs. Morphologically past tense forms and morphologically imperfective forms are relatively well-behaved semantically, in the sense that they are interpreted as past tense and imperfective, respectively. But complications arise with morphological perfectives in the nonpast tenses: though an imperfective root with only a present tense agreement marker is interpreted as imperfective, that same form with a perfectivizing prefix is interpreted as future tense. The upshot is that no verb forms are interpreted as present perfective. Consequently, though semantically future imperfective forms are morphosyntactically periphrastic, future perfective is not: this function is covered and pre-empted by morphologically present perfective forms (on all this, see Smith and Rappaport in Chapter 10 of Smith 2001).

					SP+*build*		SP+*build*+2IMPF
			build	Impf	Perf (Cumul)		Impf (Cumul)
Preterite	Past	1S		*stroi*-la/l/lo	na-*stroi*-la/l/lo		na-*stra*-yva-la/l/lo
		2S		*stroi*-la/l/lo	na-*stroi*-la/l/lo		na-*stra*-yva-la/l/lo
		3S		*stroi*-la/l/lo	na-*stroi*-la/l/lo		na-*stra*-yva-la/l/lo
		1P		*stroi*-li	na-*stroi*-li		na-*stra*-yva-li
		2P		*stroi*-li	na-*stroi*-li		na-*stra*-yva-li
		3P		*stroi*-li	na-*stroi*-li		na-*stra*-yva-li
Nonpret.	Present	1S		*stroi*-u			na-*stra*-yva-yu
		2S		*stroi*-ish			na-*stra*-yva-yesh
		3S		*stroi*-it			na-*stra*-yva-yet
		1P		*stroi*-im			na-*stra*-yva-yem
		2P		*stroi*-ite			na-*stra*-yva-yete
		3P		*stroi*-at			na-*stra*-yva-yut
	Future	1S	bud-u	*stroi*-tj	na-*stroi*-u	bud-u	na-*stra*-yva-tj
		2S	bud-esh	*stroi*-tj	na-*stroi*-ish	bud-esh	na-*stra*-yva-tj
		3S	bud-et	*stroi*-tj	na-*stroi*-it	bud-et	na-*stra*-yva-tj
		1P	bud-em	*stroi*-tj	na-*stroi*-im	bud-em	na-*stra*-yva-tj
		2P	bud-ete	*stroi*-tj	na-*stroi*-ite	bud-ete	na-*stra*-yva-tj
		3P	bud-ut	*stroi*-tj	na-*stroi*-at	bud-ut	na-*stra*-yva-tj

Table 1: Partial paradigm for *stroi–* 'build', including its cumulative forms

This situation is displayed in Table 1, a partial inflectional paradigm for the verb *stroi–* 'build', including in the left column the bare imperfective in all three tenses, and in the remaining columns cumulative forms, meaning 'build (a lot of)', resulting from the prefixation of the *na–* SP — note the stem change with the secondary imperfective.[5] Omitted are forms with a different single SP, and forms with

[5]This paradigm is an exception to property (8B), in that it combines the superlexical prefix *na–* and the secondary imperfective suffix *–yva* with no lexical prefix, yet remains grammatical. We provide it as a cautionary tale about the complexity of the phenomena we discuss.

stacked SPs; we treat LP forms as part of derivational, not inflectional, paradigms.[6] Notably, the center region of the table is empty: the set of forms belonging there morphologically — namely the ones differing from the imperfective present in just the prefix *na–* and from the cumulative imperfective present in just the suffix *–yva* — are semantically future tense forms. As such they are located in the bottom row of the paradigm, where they differ from the other future forms, which are periphrastic. Periphrastic future perfective forms such as $*$*bud–u na–stroi–tj* are ungrammatical.

Pragmatics and the imperfective An additional complication is that the Russian imperfective can be deployed for pragmatic effect. We note in particular the general-factual imperfective and the annulled-result imperfective. In the former usage, a morphologically imperfective verb (for example, 'I already ate.IMPF') can be used to refer to a completed event, with the effect of backgrounding its time reference and focussing its facthood. In the latter usage, an imperfective sentence can be used if the result achieved by the event denoted has been undone — for example 'I closed.IMPF the window' when the window is no longer closed; reference is actually to the result's undoing. See Smith and Rappaport in Chapter 10 of Smith (2001) for details. We assume that these phenomena are to be analyzed at i-structure, but make no attempt to provide such an analysis.

4 Sketch of an m-structure analysis

In this section we sketch one possible analysis of Russian perfectivity paradigms within the morphological framework of Butt et al. (1996), to illustrate its imposition of constraints on the syntax–semantics interface. Recall that in this approach c-structure nodes, including terminal nodes via lexical entries, co-describe both m- and f-structure. We augment the c-structure devices with sublexical rules, which generate sublexical trees that we will display as part of the c-structure. Please note that we will reject this analysis in favor of the one to be presented in the next section.

Our point of departure is the partial *stroi–* paradigm in Table 1. We restrict ourselves to the 3rd person singular present imperfective cumulative form *na–stra–yva–yet*, as this is sufficient for making our point. Our analysis includes the sublexical c-structure rule in (14a) and the partial lexicon in (14b).

(14) a.

$$\begin{array}{ccccc} \text{Vstem} & \rightarrow & \text{Af}_{Pf} & , & \text{Vstem} \\ & & \uparrow{=}\downarrow & & \uparrow{=}\downarrow \\ & & \hat{*}_{\mu} = *_{\mu} & & (\hat{*}_{\mu}\ \text{MARG}) = *_{\mu} \end{array}$$

Furthermore, for the forms listed here as cumulative, there are homophonous non-cumulative forms where *na–* is an LP and which translate to 'tune', as a musical instrument. Some speakers prefer this interpretation to the one we use.

[6]We return to this distinction briefly in Section 5 below.

b.	*stra-*	Vstem	$(\uparrow \text{PRED}) = \text{‘build } \langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle\text{’}$ $(\hat{*}_{\mu}\ \text{VPERF}) = \text{impf}$
	na-	Af_{Pf}	$(\hat{*}_{\mu}\ \text{VPERF}) = \text{perf}$ $(\hat{*}_{\mu}\ \text{MARG VPERF}) = \text{impf}$ $(\uparrow \text{CUMUL}) = +$
	-yva	Af_{Pf}	$(\hat{*}_{\mu}\ \text{VPERF}) = \text{impf}$ $(\hat{*}_{\mu}\ \text{MARG VPERF}) = \text{perf}$
	-yet	Af_{Agr}	$(\uparrow \text{SUBJ PERS}) = 3$ $(\uparrow \text{SUBJ NUM}) = \text{sg}$

The sublexical rule (14a) treats verb stems and perfectivity affixes (Af_{Pf}), in either linear order, as f-structure co-heads (Bresnan 2001). This provides an $\uparrow{=}\downarrow$ path to the arbitrarily deeply embedded PRED of the root. With respect to m-structure, in contrast, perfectivity affixes are heads and stems are their arguments (MARG). This device allows affixes to select for stems with particular morphosyntactic properties. For example, the annotations on the prefix *na–* in (14b) state that its own m-structure has the attribute–value pair [VPERF perf], while its morphological argument MARG is [VPERF impf].

Figure 3 displays the resulting c-, m- and f-structures. Notice that we are assuming an additional sublexical rule that makes f- and m-structure co-heads of verb stems and agreement affixes; furthermore we assume that the top node of this sublexical tree is annotated such that its own f-structure's ASPECT value is equated with its mother's m-structure's VPERF value. This essentially passes the outermost m-structure's perfectivity to the verb's f-structure; perfectivity values more deeply embedded into the m-structure are irrelevant at f-structure.

Perfectivity at the syntax–semantics interface But this cannot be all there is to it: perfectivity is semantically interpreted, and must participate in semantic composition.[7] In the projection architecture this means that perfectivity must be projected via the correspondence function σ from the f-structure to the s-structure. There are two current views of σ: one from the transfer rules tradition (Halvorsen 1983; Crouch and King 2006) and another from the Glue Semantics tradition (Dalrymple 1999).

Transfer rules are a description-by-analysis device whose function is to parse the f-structure and translate its relevant bits into an s-structure. Thus a morphosyntactically imperfective verb in this approach, for example the one in Figure 3, is semantically imperfective because of a transfer rule that converts the f-structure attribute–value pair [ASPECT impf] into the appropriate semantic expression, appropriately composed with the semantics of the clause. There is no need for any individual morpheme to contribute semantic perfectivity to the sentence, since this

[7]The precise semantic contribution of perfectivity is beyond the scope of this paper; we are content to assume it is aspectual definiteness, following Ramchand (2008).

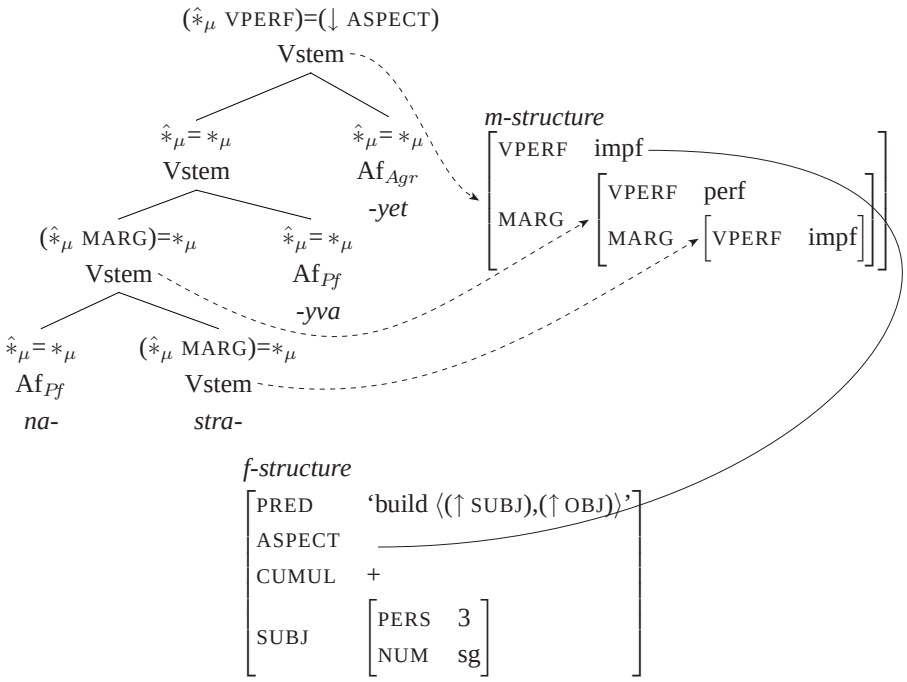


Figure 3: The c-, m- and f-structures of *na–stra–yva–yet*

semantic contribution is made by the transfer rules based on the morpheme's morphosyntactic contribution.

The Glue Semantics approach to σ is co-descriptive: lexical entries include meaning constructors pairing a semantic expression with a formula of linear logic to direct its composition. Using Glue Semantics with our analysis above would require an imperfective verb to carry in its lexical entry a meaning constructor matching (or perhaps replacing) the VPERF value of its m-structure. The complication in Russian is that a single verb form may have multiple morphosyntactic markers of perfectivity; for example the verb in Figure 3 has three such markers, the root and two affixes. These markers contribute perfectivity values which, though co-present at m-structure, do not conflict with each other because of the morphosyntactic embedding implemented by the sublexical rule (14a). But each perfectivity morpheme would have its own meaning constructor, effectively expressing three separate semantic perfectivity values for the single f-structure these morphemes share: a perfective one from *na–* and two imperfective ones from *stra–* and *–yva*. Then either these meaning constructors cannot compose with each other and the structure is ungrammatical, or else they can compose with each other after all but the verb's aspect is re-imperfectivized perfectivized imperfective rather than just imperfective. One apparent way of sidestepping this issue is to make the meaning constructors optional; however this would result in an ambiguous sentence, perfective when the prefix's meaning constructor is chosen, imperfective otherwise. All

of these results are wrong.

The conclusion must then be that given an analysis like that in Figure 3, s-structure is projected via transfer rules, at least where perfectivity is concerned: semantic perfectivity is read off of the f-structure, not contributed by morphemes. This does not necessarily mean that there is no role to play for co-description — a hybrid approach combining transfer rules for symbol values with meaning constructors for grammatical functions and scope (for example) seems possible to envision — just that it is not up to the task of dealing with perfectivity, given this analysis of Russian facts.

We do not adopt this analysis. We presented it to illustrate the point that there can be a dependency between choices that are better kept orthogonal. In this case the dependency is between the approach to m-structure we employed above and the choice of syntax–semantics interface type. The m-structure represents a theoretical commitment to the existence of a set of morphosyntactic generalizations distinct from f-structure generalizations, while the correspondence function σ between f- and s-structure represents a theoretical commitment to the existence of a significant relationship between the contents of these structures. Decisions about these theoretical commitments should be interdependent only if the commitments are found to be related.

5 Paradigm Function Morphology analysis

We follow Sadler and Nordlinger (2004, 2006) in proposing Paradigm Function Morphology as the morphological component in LFG grammars. We introduce our analysis in a manner consistent with their architecture (see the bottom of Figure 2), but end up departing from it significantly in the next section.

In Stump (2001), PFM is presented as a framework primarily for inflectional morphology. Since we believe lexical prefixes to be derivational, our analysis will therefore focus on superlexical prefixes and the secondary imperfective. With regards to lexical prefixes, we take the following to suffice at this point: property (3D), according to which LPs are able to license additional arguments, can be modeled by assuming that these prefixes carry annotations such as ($\hat{*}_{\alpha}$ THEME), effectively projecting an additional argument to the a-structure in its Butt et al. (1997) incarnation. Meanwhile property (3E), according to which LP stems can have idiosyncratic meaning, for example ‘work’ vs. ‘earn’ in (5), are accounted for if such stems are memorized as lexical units.

As to our inflectional affix types, the superlexical prefixes and the secondary imperfective suffix, they combine with lexemes to form paradigms of morphological forms. These are defined in PFM by declaring a paradigm space — a set of morphosyntactic feature sets — and stating rules to realize the forms filling that space. We define the paradigm space for verb lexemes by:

(15) a. declaring a set of possible morphosyntactic features: attributes and their possible values:

Attribute	Values
ASPECT	prf, impf
PRETERITE	$+, -$
TENSE	past, present, future
PERS	1, 2, 3
NUM	sg, pl
GEND	fem, masc, neut
SUBJAGR	{PERS:x, NUM:y, GEND:z}
CUMUL	$+, -$
PRDR	$+, -$
⋮	

b. declaring what constitutes a complete feature set; here we assume this means one of each attribute in the table above along with a valid value for this attribute.

c. declaring feature co-occurrence restrictions; in our case there is at least:

> If a morphosyntactic feature set Σ is an extension of {TENSE:present}, then Σ is not an extension of {ASPECT:prf}.

Formally, the paradigm represented diagrammatically in Figure 1 is a set of complete morphosyntactic feature sets including (16a,b), though not (16c) since it violates (15c):[8]

(16) a. Non-cumulative imperfective present:
{ASPECT:impf, PRETERITE:−, TENSE:pres, CUMUL:−, PRDR:−,
SUBJAGR:{PERS:1|2|3, NUM:sg|pl, GEND:fem|masc|neut} ...}

b. Cumulutative perfective future:
{ASPECT:prf, PRETERITE:−, TENSE:fut, CUMUL:+, PRDR:−,
SUBJAGR:{PERS:1|2|3, NUM:sg|pl, GEND:fem|masc|neut} ...}

c. Cumulative perfective present:
∗{ASPECT:prf, PRETERITE:−, TENSE:pres, CUMUL:−, PRDR:−,
SUBJAGR:{PERS:1|2|3, NUM:sg|pl, GEND:fem|masc|neut} ...}

In the projection architecture of LFG, these complete morphosyntactic feature sets constitute the m-structure of the verb. The lexical entry of a verb lexeme thus contains an equation like $(\hat{*}_{\mu}) = [(16a)|(16b)|\ldots]$ in which the morphosyntactic feature sets constituting the lexeme's paradigm space are enumerated disjunctively.

It is important to notice at this point that each complete feature set contains a single perfectivity value. Each such set also constitutes a unique constellation of features, which can be used to realize a unique form for this constellation.

[8]The conjunctions in the SUBJAGR subsets of features abbreviate the 18 possible combinations of person, number and gender.

Block A (stem choice)

A1. $\mathrm{RR}_{\mathrm{V},\{\textsc{aspect}:\mathrm{impf},\textsc{cumul}:+\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{Y}', \Sigma\rangle$, where Y is X's 2nd stem

A2. $\mathrm{RR}_{\mathrm{V},\{\ \}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{Y}', \Sigma\rangle$, where Y is X's 1st stem

Block B

B1. $\mathrm{RR}_{\mathrm{V},\{\textsc{cumul}:+\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}na', \Sigma\rangle$

Block C

C1. $\mathrm{RR}_{\mathrm{V},\{\textsc{aspect}:\mathrm{impf},\textsc{cumul}:+\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}yva', \Sigma\rangle$

Block D

D1. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:+\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}l', \Sigma\rangle$

D2. $\mathrm{RR}_{\mathrm{V},\{\textsc{aspect}:\mathrm{impf},\textsc{tense}:\mathrm{pres},\textsc{cumul}:+\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}y', \Sigma\rangle$

Block E

E1. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:+,\textsc{subjagr}:\{\textsc{num}:\mathrm{pl}\}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}i', \Sigma\rangle$

E2. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:+,\textsc{subjagr}:\{\textsc{num}:\mathrm{sg},\textsc{gend}:\mathrm{fem}\}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}a', \Sigma\rangle$

E3. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:+,\textsc{subjagr}:\{\textsc{num}:\mathrm{sg},\textsc{gend}:\mathrm{neut}\}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}o', \Sigma\rangle$

E4. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:-,\textsc{subjagr}:\{\textsc{pers}:1,\textsc{num}:\mathrm{sg}\}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}u', \Sigma\rangle$

E5. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:-,\textsc{subjagr}:\{\textsc{pers}:2,\textsc{num}:\mathrm{sg}\}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}esh', \Sigma\rangle$

E6. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:-,\textsc{subjagr}:\{\textsc{pers}:3,\textsc{num}:\mathrm{sg}\}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}et', \Sigma\rangle$

E7. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:-,\textsc{subjagr}:\{\textsc{pers}:1,\textsc{num}:\mathrm{pl}\}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}em', \Sigma\rangle$

E8. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:-,\textsc{subjagr}:\{\textsc{pers}:2,\textsc{num}:\mathrm{pl}\}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}ete', \Sigma\rangle$

E9. $\mathrm{RR}_{\mathrm{V},\{\textsc{preterite}:-,\textsc{subjagr}:\{\textsc{pers}:3,\textsc{num}:\mathrm{pl}\}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}ut', \Sigma\rangle$

E10. $\mathrm{RR}_{\mathrm{V},\{\textsc{aspect}:\mathrm{impf},\textsc{preterite}:-,\textsc{tense}:\mathrm{fut}\}}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \langle \mathrm{X}tj', \Sigma\rangle$

Figure 4: Realization rule blocks for Russian verb lexemes

Projection to p-structure takes place via blocks of realization rules, among each of which the most specific rule applicable, given the lexeme's morphosyntactic feature set, adds phonological material to the p-structure of the lexeme. When no rule within a block is applicable, the identity function applies to add nothing. We assume the rule blocks in Figure 4 (which abstract away from some morphophonological details), and define the paradigm function for Russian verb lexemes as:

(17) $$\mathrm{PF}(\langle \mathrm{X}, \Sigma\rangle) =_{\mathrm{def}} \mathrm{Nar}_{\mathrm{E}}\,(\mathrm{Nar}_{\mathrm{D}}\,(\mathrm{Nar}_{\mathrm{C}}\,(\mathrm{Nar}_{\mathrm{B}}\,(\mathrm{Nar}_{\mathrm{A}}\,(\langle \mathrm{X}, \Sigma\rangle)))))$$

Thus a lexeme X with morphosyntactic feature set Σ has its phonological form defined by the successive application of the narrowest rule from each of rule blocks A through E. A complete example follows.

Assume a lexeme *stroitj* 'build' with first stem *stroi*– and second stem *stra*–, and the morphosyntactic feature set in (18), including GEND:fem to be complete per (15b). The narrowest rule from each block is listed in (19) along with its effect. Rule A1 is narrowest in block A because its conditions of application match more of the lexeme's features than the alternative; thus the lexeme's second stem is selected. Rules B1 and C1 are narrowest in their blocks because their conditions of application match the lexeme's features; if there was no such match, the identity function would apply and nothing would be added to the lexeme's phonological

form. In block D, rule D2 is narrowest because the other rule's conditions of application do not match the lexeme's features. Finally in block E rule E7 applies, as no other rules match the lexeme's features. The output is the form *na–stra–yva–y–em*. The meaning associated with this form is determined by the m-structure's projection to the f-structure and then via σ to the s-structure.

(18) $\langle$*stroitj*,$\{$ASPECT:impf,
PRETERITE:$-$,
TENSE:pres,
CUMUL:$+$,
REPET:$-$,
SUBJAGR:$\{$PERS:1, NUM:pl, GEND:fem$\}$ $\}$ $\rangle$

(19) $\langle$*stroitj*,$\Sigma\rangle$

A1. $\mathrm{RR}_{\mathrm{V},\{\mathrm{ASPECT:impf,CUMUL:+}\}}(\langle \textit{stroitj}, \Sigma\rangle) = \langle stra, \Sigma\rangle$
B1. $\mathrm{RR}_{\mathrm{V},\{\mathrm{CUMUL:+}\}}(\langle stra, \Sigma\rangle) = \langle nastra, \Sigma\rangle$
C1. $\mathrm{RR}_{\mathrm{V},\{\mathrm{ASPECT:impf,CUMUL:+}\}}(\langle nastra, \Sigma\rangle) = \langle nastrayva, \Sigma\rangle$
D2. $\mathrm{RR}_{\mathrm{V},\{\mathrm{ASPECT:impf,TENSE:pres,CUMUL:+}\}}(\langle nastrayva, \Sigma\rangle) = \langle nastrayvay, \Sigma\rangle$
E7. $\mathrm{RR}_{\mathrm{V},\{\mathrm{PRETERITE:-,SUBJAGR:\{PERS:1,NUM:pl\}}\}}(\langle nastrayvay, \Sigma\rangle) = \langle nastrayvayem, \Sigma\rangle$

Projection to the f-structure takes place via transfer rules, which can be summarized as in Table 2, taking f to the f-structure of the lexeme.

Morphosynatctic feature	$\Rightarrow$	f-description
ASPECT:impf		(f ASPECT) = impf
TENSE:pres		(f TENSE) = pres
CUMUL:+		(f CUMUL) = $+$
SUBJAGR:$\{$PERS:1$\}$		(f SUBJ PERS) = 1
SUBJAGR:$\{$NUM:pl$\}$		(f SUBJ NUM) = pl
SUBJAGR:$\{$GEND:fem$\}$		(f SUBJ GEND) = fem
	$\vdots$	

Table 2: Transfer rules projection from m- to f-structure

It follows that the lexeme in (18) corresponds to the form *nastrayvayem* at p-structure, and to the f-structure in (20):

(20)
$$f\begin{bmatrix} \text{PRED} & \text{'build } \langle(\uparrow \text{SUBJ}),(\uparrow \text{OBJ})\rangle\text{'} \\ \text{ASPECT} & \text{impf} \\ \text{TENSE} & \text{pres} \\ \text{CUMUL} & + \\ \text{REPET} & - \\ \text{SUBJ} & \begin{bmatrix} \text{PERS} & 1 \\ \text{NUM} & \text{pl} \\ \text{GEND} & \text{fem} \end{bmatrix} \end{bmatrix}$$

Sadler and Nordlinger (2004) need an additional, intermediary tree representation to complete the f-description: the recursive case stacking phenomenon they analyze allows single forms to specify not just attributes of their own f-structure, but also attributes of other f-structures in which they are embedded, via inside-out function application. We accept the argument in Andrews (2005) that this intermediary representation can be eliminated by complexifying the transfer rules.

Semantic perfectivity One advantage of this analysis over the one presented in Section 4 is that it is compatible with both a co-description and a transfer rule (or some other kind of description-by-analysis) treatment of the syntax–semantics interface. The compatibility of a transfer rule treatment can be straightforwardly observed: the f-structure contains a single perfectivity value, which will be translated to a semantic expression of perfectivity. A co-description treatment would place perfectivity meaning constructors in the lexical entries of lexemes, matching the single morphosyntactic perfectivity value they project to m-structure. There is no issue of conflicting perfectivity values arising from morpheme concatenation, because the verb's features are not calculated from the sum of its morphemes' features; rather the verb's form is calculated based on the lexeme's features.

6 PFM without m-structure

While the Paradigm Function Morphology presented in the previous section solves the issue of morphology constraining the choice of syntax–semantics interface type, its architecture seems unnecessarily complicated to us. In particular, we believe that the m-structure can be completely eliminated with no detriment to the analysis.

Consider again Table 2, and notice what little value is added by having both a set of morphosyntactic features and a set of f-structure features when they are essentially identical. Suppose that we replace (15), where paradigm space is defined with respect to morphosyntactic features, with the following:

(21)
a. a declaration of a set of possible f-structure features: attributes and their possible values — note that this is a required feature of XLE grammars (Butt et al. 1999; Crouch et al. 2008);
b. a declaration of what constitutes a complete f-description for lexemes;
c. a declaration of feature co-occurrence restrictions, at potentially arbitrary outside-in and inside-out depths, to cover phenomena like case stacking.

Paradigm functions, projecting to p-structure, can then be defined to apply to the f-descriptions of lexemes. The recursive morphosyntactic feature sets that Sadler and Nordlinger (2006) use to account for case stacking find their analogue in the

recursiveness of f-structure descriptions. Since in this view lexemes already have f-descriptions and already project to the f-structure, there is no need for transfer rules to do this. We therefore argue that the architecture in Figure 5 is appropriate for LFGs with a PFM morphological component:

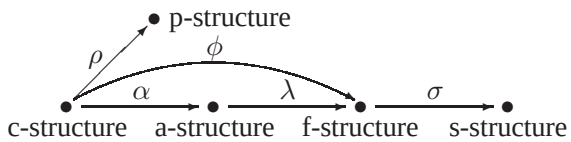


Figure 5: The projection architecture without m-structure

There is an implied theoretical innovation in this approach: though a lexeme is standardly understood as having a lexical entry containing its own f-description, in a c-structure that lexeme merely contributes to the f-description of the sentence, within which the provenance of functional equations is effaced. We assume that it is possible to revise this conception such that the f-descriptions of different terminal nodes can be differentiated. Paradigm functions then apply to terminal nodes based on their f-descriptions.

One possible objection to the elimination of morphosyntactic features in favor of f-structure features, which was raised in the question period of our presentation, is that there is a need for such things as conjugation classes, and possibly other kinds of purely morphological categories that have no business in f-structure. But the conjugation class issue is already taken care of in the PFM framework: the application conditions of realization rules include not just morphosyntactic features, but also category features (Stump 2001). In Figure 4, our rules apply to lexemes of category V, but we could decompose V into features and have rules apply to Vs with only some of these features.[9] A formalization of this might recruit some of the same tools that Bresnan (2001) uses in her theory of structure–function mapping.

As for the issue of other features that do not belong in the f-structure, we suggest that PFM work should proceed by eliminating them from analyses: by hypothesis, only meaningful features will be used to cross-classify the forms in a well-specified paradigm. As long as the cross-classification is total, the end result will be that each cell in the paradigm formally constitutes a unique constellation of functional schemata: an f-description unique within the paradigm, which a well-designed paradigm function can spell out as a unique form. We only appear to violate this methodological principle with our use of PRETERITE among morphosyntactic features, but not in f-structures, to differentiate past from nonpast tenses — compare (18) with its f-structure in (20). If paradigm functions apply based on the f-descriptions of lexemes, then this feature can be eliminated because the f-description language allows negative expressions like $(f\ \text{TENSE}) \neq \text{past}$.

Finally, we note that Saléschus and Hautli (2008) generate correct verb forms by using finite-state tools to state long-distance dependencies between morphemes.

[9] In fact we do need to do this since Russian has conjugation classes.

While we do not believe that their analysis has the scope of ours, it does raise the question of whether assuming Paradigm Function Morphology is necessary.

7 Conclusion

The expression of perfectivity in Russian lies at a busy intersection: morphosyntactically it arises from nontrivial patterns of affixation, which do not interact straightforwardly with tense interpretations and are sensitive to telicity; it seems to have a foot in both inflectional and derivational paradigms; it affects argument structure and interacts with information structure. This makes it a privileged vantage point from which to view issues concerning the projection architecture.

We focussed on a particular problem arising from some of these Russian facts: the analysis type indicated by an m-structure approach like that of Butt et al. (1996) in not compatible with a purely co-descriptive approach to the syntax–semantics interface, and requires at least some admixture of description-by-analysis. In contrast, a Paradigm Function Morphology treatment of the Russian data is trivially compatible with either form of syntax–semantics interface.

We do not purport to have resolved the issue of morphology in the projection architecture, nor to have a complete analysis of Russian perfectivity paradigms. However we do believe that we have demonstrated the practicality of the general course we advocate, of freeing theoretical choices from orthogonal constraints. In frameworks like LFG where grammars are highly modular by design, and deciding on the architecture of the grammar is part of object of research, compartmentalizing theoretical choices is a sound methodological strategy that need not interfere with detailed grammatical description.

References

Andrews, Avery D. 2005. F-structural spellout in LFG morphology, MS, Australian National University.

Bresnan, Joan. 2001. *Lexical-functional syntax*. Oxford: Blackwell Publishers.

Butt, Miriam, Dalrymple, Mary and Frank, Anette. 1997. An architecture for linking theory in LFG. In Miriam Butt and Tracy Holloway King (eds.), *The proceedings of the LFG '97 conference*, University of California, San Diego.

Butt, Miriam and King, Tracy Holloway. 1998. Interfacing phonology with LFG. In Miriam Butt and Tracy Holloway King (eds.), *The proceedings of the LFG '98 conference*, University of Queensland, Brisbane.

Butt, Miriam and King, Tracy Holloway. 2000. Null elements in discourse structure. In K.V. Subbarao (ed.), *Papers from the NULLS seminar*, Delhi: Motilal Banarasidas.

Butt, Miriam, King, Tracy Holloway, Niño, María-Eugenia and Segond, Frédérique. 1999. *A grammar writer's cookbook*. Stanford, CA: CSLI Publications.

Butt, Miriam, Niño, María-Eugenia and Segond, Frédérique. 1996. Multilingual processing of auxiliaries in LFG. In D. Gibbon (ed.), *Natural language processing and speech technology: results of the 3rd KONVENS conference*, pages 111–122, Mouton de Gruyter, universität Bielefeld, 7–9 October 1996.

Crouch, Dick, Dalrymple, Mary, Kaplan, Ron, King, Tracy Holloway, Maxwell, John and Newman, Paula. 2008. XLE documentation, Palo Alto Research Center.

Crouch, Dick and King, Tracy Holloway. 2006. Semantics via f-structure rewriting. In Miriam Butt and Tracy Holloway King (eds.), *The proceedings of the LFG '06 conference*, University of Konstanz, Germany.

Dalrymple, Mary (ed.). 1999. *Semantics and syntax in Lexical Functional Grammar: the resource logic approach*. Cambridge, MA: The MIT Press.

Frank, Anette and Zaenen, Annie. 2002. Tense in LFG: syntax and morphology. In Hans Kamp and Uwe Reyle (eds.), *How we say WHEN it happens: contributions to the theory of temporal reference in natural language*, Tübingen: Niemeyer, to appear.

Halvorsen, Per-Kristian. 1983. Semantics for Lexical Functional Grammar. *Linguistic Inquiry* 14(4), 567–615.

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 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, Cambridge, MA: The MIT Press.

Mycock, Louise. 2006. *The typology of constituent questions: a Lexical-Functional Grammar analysis of 'wh'-questions*. Ph. D.thesis, University of Manchester.

O'Connor, Rob. 2006. *Information structure in Lexical-Functional Grammar*. Ph. D.thesis, University of Manchester.

Ramchand, Gillian Catriona. 2004. Time and the event: the semantics of Russian prefixes. *University of Tromsø Center for the Advanced Study of Theoretical Linguistics working papers, Nordlyd* 32, 323–361.

Ramchand, Gillian Catriona. 2008. Perfectivity as aspectual definiteness: time and the event in Russian. *Lingua* 118, 1690–1715.

Romanova, Eugenia. 2004. Superlexical vs. lexical prefixes. *University of Tromsø Center for the Advanced Study of Theoretical Linguistics working papers, Nordlyd* 32, 255–278.

Sadler, Louisa and Nordlinger, Rachel. 2004. Relating morphology to syntax. In Louisa Sadler and Andrew Spencer (ed.), *Projecting morphology*, pages 159–186, Stanford, CA: CSLI Publications.

Sadler, Louisa and Nordlinger, Rachel. 2006. Case stacking in realizational morphology. *Linguistics* 44, 459–487.

Saléschus, Dirk and Hautli, Annette. 2008. On dissolving morphological long distance dependencies in Russian verbs. *Linguistic Issues in Language Technology* 1, 1–27.

Smith, Carlotta S. 2001. *The parameter of aspect*. Dordrecht: Kluwer Academic Publishers.

Stump, Gregory T. 2001. *Inflectional morphology: a theory of paradigm structure*. Cambridge, UK: Cambridge University Press.

Svenonius, Peter. 2004. Slavic prefixes inside and outside the VP. *University of Tromsø Center for the Advanced Study of Theoretical Linguistics working papers, Nordlyd* 32, 205–253.

Send correspondence to:
Jean-Philippe Marcotte
University of Minnesota
Institute of Linguistics
315 Pillsbury Dr. SE
Nolte Center 214
Minneapolis, MN 55455
United States
email: marco043@umn.edu

INCORPORATION AND COMPLEX PREDICATION IN PERSIAN

Fatemeh Nemati

University of Tehran, University of Konstanz

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

This study deals with the nature of N+V sequences in Persian and suggests a sub-classification of these sequences into Noun Incorporation and Complex Predication. This classification is grounded in the lexical and phrasal properties of the nouns involved in these sequences. Noun Incorporation cases are analyzed in terms of head-adjunction of the non-projecting Noun at the level of c-structure (Toivonen 2001). Complex Predication is dealt with in terms of the predicate composition proposed in Butt (1995, 1997) and Alsina (1997), along with some adaptations from Pustejovsky's (1995) theory of the generative lexicon.

1 Introduction

Persian shows a strong preference for using multiword verbal expressions over simple verbs[1]. Sadeghi (1993) has stated the number of simple verbs used in both spoken and written Persian do not exceed 150. He also claims that this is not a new tendency and that the formation of multiword verbal expressions has been used extensively even before the enormous borrowings from Arabic, and later from other foreign languages. Therefore, Persian must have been using productive processes to conceptualize new ideas and add new verbs to its repository of verbal expressions. Since these processes have been used over centuries, it is no surprise if a once-productive process of verb formation is not accessible to the Persian speakers anymore (locative incorporation as in *piS raftan* ‘front going’ (to move forward)); and if the idiomaticity of some of the verbal expressions obscures their internal structures (consider *zamin xordan* ‘ground hitting’ (to fall) where object of proposition is incorporated by the verb). This short article would surely do not do justice to reflect upon the whole process and, thus, the focus of the present study will be limited to a synchronic investigation of frequent multiword-verbal expressions, mostly labeled as complex verbs, composite verbs or compound verbs in the state-of-the-art.

A close look at the constructions shows that these verbal expressions vary systematically from each other in terms of the semantic relation of the non-verbal element to the verbal element. They also show different syntactic behaviors. This study aims at investigating the nature of these multiword

1 I am grateful to Miriam Butt for her support and long discussions during my research period at the University of Konstanz. My thanks are also due to Ali Ashraf Sadeghi for his insightful seminars on Persian morphology. I would like to show my gratitude towards Yehuda Falk, George Aoran Broadwell, Lachlan Duncan, Louisa Sadler and Rachel Nordlinger for their helpful comments. I am also very grateful to my colleagues at the University of Konstanz, Thomas Mayer, Michael Spagnol, Maialen Iraola Azpiroz and my colleagues in Iran, Fatemeh Alavi and Sahar Bahrami for helping me at different stages of the work.

verbal expressions based on these two factors and will show that there are two different processes involved in the formation of verbal expressions, namely incorporation vs. complex predication. In the next section, I will touch on the differences of these two subclasses of multiword verbal expressions, limiting the scope of it to N+V sequences. In section 3 and 4, an attempt will be made to analyze these two types of multiword verbal expressions in the light of LFG findings and to show how they can be represented linguistically.

2 Differences between Multiword Verbal Expressions: Motivations for a sub-classification

2.1 The Basic Data

To start with the differences between multiword verbs in Persian, a very brief introduction to the sentence structure in Persian seems to be due. Persian is an SOV, pro-drop language with a partially free word order, to the extent that Sadeghi (pc) claims that it is a non-configurational language. The canonical word order is illustrated in (1a). Other possible word orders are given in (1b, c, d, and e).

(1) a. [āryā] [ketāb rā] [be man] [dād]
Arya book OM to me give.Past.3.Sg.
Subj OBJ OBL V
'Arya gave me the book.'
b. [ketāb rā] [āryā] [be man] [dād]
c. [be man] [āryā] [ketāb rā] [dād]
d. [āryā] [be man] [ketāb rā] [dād]
e. [ketāb rā] [be man] [āryā] [dād]

As can be seen from the above sentences, the canonical position of the verb is sentence final. The verb rarely undergoes scrambling. The object appears obligatory followed by 'rā' as the accusative case marker, if definite, as in the above sentences. Otherwise, it can be followed by an indefinite clitic (2), having the same ordering possibilities as a definite object.

(2) [āryā] [ketāb-i] [be man] [dād]
Arya book-INDF to me give.Past.3.Sg.
Subj OBJ OBL V
'Arya gave me a book.'

The OBL(ique) or indirect object always appears as the complement of a preposition and receives case as the object of preposition. The only noun in the sentence with a covert case is the subject, receiving a nominative case.

Keeping these characteristics of Persian sentence grammar in mind, we proceed to our topic: multiword verbal expressions. As an example of such constructions, consider the example sentences (3b) and (3c) compared to (3a).

(3) a. minā **Gazā** rā be bače **dād**
Mina food OM to child give.Past.3.sg
'Mina gave the food to the child.'

b. minā be bače **Gazā** **dād** (Noun Incorporation)
Mina to child food give.Past.3.sg
Lit. 'Mina food-gave to the child.'

c. minā āryā rā **šekast** **dād** (Complex Predication)
Mina Arya OM defeat give.Past.3.sg
'Mina defeated Arya.'

The verbal in all the three sentences is *dād*. (3a) and (3b) have an NP with the same semantic content (*Gazā*). They differ in that the noun in (3a) is followed by a case marker, giving it a specific interpretation, whereas the noun in (3b) appears caseless and adjacent to the verb. There are strict constraints on the interpretation and the order of this caseless noun and it also varies from the noun in the (3a) in having a generic interpretation and joining the verb to denote a unitary activity. I argue that this N+V sequence in (3b) is a case of Noun Incorporation (NI).

The N+V sequence in (3c) will, in contrast to (3b), be discussed as a Complex Predicate (CPr). The noun here appears caseless as well, but it has no argument relation to the verb, as can be inferred from the semantic content of the noun. The noun is part of the predication and contributes to the argument structure of the complex. What is predicated in (3c) does not refer to the main semantic content of the verb 'the transference of something from Arya to Mina', but that Arya has brought something on Mina that is 'defeat'. Besides this semantic difference, I point to other differences between these two N+V sequences in the remainder of the paper, arguing that they should receive different syntactic analyses.

2.2 Syntactic Behavior of CPrs vs. NI

2.2.1 Modification

The noun in incorporation sentences resists modification by adjectives and quantifiers (4a). Otherwise, modification changes the semantic interpretation of the complex as a conceptual whole to its non-incorporated counterpart, as it is can be seen in (4b). On the other hand, when the noun in CPr is modified (as in (4c)), the scope of the modification is extended over the whole event denoted by the N+V sequence and the semantics of the noun, in terms of

definiteness, does not change. This result is predicted because the noun in (4c) is an essential part of the predication.

(4) a. *minā koll-e **nāme** nevešt
Mina all-of letter write.Past.3.Sg
Lit. '*Mina the whole letter-wrote.'

b. minā **nāme**-'i *tulāni* nevešt
Mina letter-INDF long write. Past.3.Sg
'Mina wrote a long letter.'

c. (ānhā) dar gilān šekast-e *sangin*-i xord-and
(they) in Gilan defeat-Ez hard-INDF eat- Past.3.Pl
'They were defeated severly in Gilan'.

2.2.2 Relativization

The noun in the syntactic paraphrase of *Gazā xord-im* (4a) can be relativized and thus it becomes specific, referring to a certain instance of food (consider 4b). Relativization of the nominal in CPrs does not result in specific reading of the noun: the noun fills the subject position of the main clause, but the whole event predicated jointly by the noun and the verb is inferred to be functioning as the subject.

(5) a. diruz tu resturān Gazā xord-im.
Yesterday in restaurant food eat-Past.1.Pl.
Lit. 'We food-ate in the restaurant yesterday.'

b. Gazā-'i ke diruz xord-im xošmaze bud
food-INDF that yesterday eat-Past.1.Pl delicious be.Past.3.Sg.
'The food that we ate yesterday, was delicious.'

c. šekast-i ke mā xord-im be dalil-e
defeat-INDF REL we eat- Past.1.Pl to reason-EZ
na-dāštan-e barnāme bud
not-having-Ez plan be- Past.3.Sg
'We were defeated because of not having a plan.'

2.2.3 Scrambling

As Persian is a partially free word order language, the constituents might appear in almost any order before the verb (in the spoken register some constituents might appear after the verb as well). A bare noun with a generic interpretation as observed in (6a) cannot scramble. When the noun scrambles, it is modified and obtains a specific reading (6b). Scrambling of the nominal element of the CPr does not result in the specific interpretation of the noun.

Rather, it gives the event a pragmatic prominence, putting it in a focus position (6c).

(6) a. minā tamām-e ruz rā xub dars xānd
Mina all-Ez day OM well lesson read.Past.3.sg.
'Mina studied well all day long.'

b. minā dars-hā-yaš rā xub xānd
Mina lesson-Pl.-her OM well read.Past.3.sg.
'Mina studied her lessons well.'

c. diruz kotak-e šadidi-i āryā az bābā-š
yesterday beating-EZ harsh-a Arya from father-PossC
xord
eat.past.3.sg.
'Arya was beaten harshly by his father yesterday.'

2.2.4 Pronominal Cliticization

In Persian, clitics attach to the outer edge of phrasal constituents (7a) and (7b); they do not have access to the internal structure of words (7c).

(7) a. ketāb-aš
book-PosCl3.sg.
' her/his book'

b. ketāb-xāne-aš
book-house-PosCl3.sg.
'her/his/its library'

c. *ketāb-aš-xāne
book-PosCl3.sg.-house
'her/his/its library'

NI disallows cliticization which follows from the lexical status of the Noun. When a clitic attaches a noun in its corresponding non-incorporated paraphrase, the noun becomes specific (compare (8a) and (8b)). The nominal in CPr, however, allows cliticization, the clitic has no effect on the nonspecific interpretation of the noun (consider (8c) and (8d)). It is also worth mentioning that the type of clitics attaching the noun in the syntactic counterpart of incorporation is different from the clitic attaching the nominal in CPrs: in the former, it is a possessive clitic; while in the latter, it is a pronominal clitic satisfying one of the grammatical functions in the sentence (compare (8b) with (8d) and (8f) for the difference).

(8) a. az bānk **vām** **gereft** (NI)
from bank loan take.Past.3.sg
'S/He got a loan from the bank.'

b. **vam**-aš rā az bānk **gereft**
loan-POSCl.3.sg. OM from bank take.Past.3.sg
Lit. 'S/He got her/his loan from the bank.'

c. minā nasrin rā be mehmāni **da'vat** **kard** (CPr)
Mina Nasrin OM to party invitation do-Past.3.sg
'Mina inviter her/him to the party.'

d. minā be mehmuni **da'vat**-eš **kard** (spoken register, CPr)
Mina to party invitation-PCl3.sg. do-Past.3.sg
'Mina inviter her/him to the party.'

e. minā be nasrin **komak** **kard** (CPr)
Mina to Nasrin help do- Past.3.sg
'Mina helped Nasrin.'

f. minā **komak**-eš **kard** (spoken register, CPr)
Mina help-PCl3.sg. do-Past.3.sg
'Mina helped her.'

The fact that the noun in NI is invisible to syntactic processes shows that it has a lexical status. CPrs, in contrast, are syntactic and the noun has a phrasal status in that it can function as a host to clitics and it can be modified, relativized and scrambled.

2.3 An Overview of N+V Sequences in the State-of-the art

The investigation of Persian N+V sequences in the state-of-the-art has opted for either a lexical or a syntactic approach. Some researchers claim that all multiword verbal expressions are lexical and that they are the result of the morphological processes of incorporation and combination (Dabir-Moghaddam 1997, Vahedi-Langrudi 1996). However, a lexicalist approach falls short of explaining the syntactic behavior of CPrs, as discussed above.

Other researchers discuss multiword verbal expressions as syntactic constructions, but they fail to observe the distinction between the two types of the nouns in N+V sequences (Karimi 1997, Karimi-Doostan 1997, Folli et al. 2005, Pantcheva 2010). Mostly they have ignored the possibility of an incorporation account for some of the N+V sequences, as well as ADV+V

sequences whose resistance to separation can be explained better in the light of an incorporation account. Megerdoomian (2006), in contrast, deals with the syntactic and semantic differences between these two N+V sequences, treating the nominal part under the term bare nominal as opposed to preverbal nominal (in my analysis the incorporated noun in NI and the nominal part of the CPr, respectively). She, however, does not deal with scrambling, topicalization, and relativization of the preverbal noun in CPrs, and the separability of the CPr elements by these processes pose a challenge to the derivational framework she has adopted for analyzing these sequences. The incorporation analysis I put forward not only accounts for the bare nominals in Megerdoomian's analysis, but it can also be extended to include another type of multiword verbal expression in Persian, ADV+V sequences which are incorrectly treated in the literature as CPrs (Foli et al. 2005; Megerdoomian 2006; Pantcheva 2010).

In the next two sections, I will try to give an analysis of these two N+V sequences from the perspective of LFG. As my point of departure for distinguishing these two N+V sequences, I draw on the definition of CPrs given in Mohanan (1997) and Butt (1997) as a construction in which two semantically predicative elements jointly determine the argument structure of a single syntactic clause. The co-predication results in a complex argument structure, but a flat grammatical function structure, like that of a simple predicate. Based on this definition, cases of incorporation where an explicit or implicit argument of the verb and the verb make a complex are excluded, because incorporation does not give rise to a complex argument structure.

In order to account for cases of incorporation, I avail myself of the non-projecting words analysis put forward in Toivonen (2001), to argue that nouns as well as some adverbs in some ADV+V sequences are all non-projecting nodes, head adjoined to the V. As for the CPrs, I follow Butt (1995) and Alsina (1997) to account for the co-predication of their constitutive elements in terms of argument fusion.

3 Incorporation

That some of Persian multiword verbal expressions are the results of incorporation has already been discussed in Dabir-Moghadam (1997) and Vahedi Langrudi (1996). While I include their findings, in contrast to their conclusion, I claim that the results of incorporation are not CPrs. As observed in (2.2), these two N+V sequences contrast in their syntactic behaviour. NI resists separability by modification, relativization, scrambling and cliticization. These facts point to the lexical status of incorporation cases.

There are other facts discussed in Dabir-Mogaddam (1997), also in line with the list of defining characteristics in Mithun and Corbett (1999) that points towards the lexical status of Noun in the NI. Phonologically, they are pronounced as a whole, with no pause in between. They are treated as one

phonological word, with the noun bearing the primary stress. The pause after bare noun shows focus, and gives the noun prominence and a specific interpretation. They are highly productive and it can apply to most of the Object + Verb sequences to have an N+V incorporated sequence, as long as the noun is inanimate. Pragmatic factors, however, are at work to give a sequence the status of a well-established incorporation construct. *Gazā xordan* 'food-eating' is recognized by native speakers to function as a unitary activity and is packaged as conceptual whole by the native speaker. *havij xordan* 'carrot-eating' is not considered as such, although it has the potential to acquire the activity reading discerned in *Gazā xordan*. Semantically, they are transparent and the meaning of it corresponds to its parts. Even in the idiomatic ones, such transparency can be detected. Compare the two meanings of *dars xāndan* (Lit. ‘lesson-reading’) ‘reading a lesson’ vs. ‘studying in an institute such as a university’.

Some other syntactic tests for constituency also illustrate that the Noun in NI is not a phrasal constituent. Due to the unavailability of the noun as a constituent on the c-structure, the following syntactic operations are not allowed: Gapping and the coordination of the incorporated noun with a specific noun (9), binding the pronominal (10), and nominal ellipsis (11).

(9) *man ham Gazā xord-am va ham mive rā
I also food eat-past.1.sg. and both fruit OM
Lit. '*I both food-ate and the fruit.'

(10) *man Gazā xord-am va kami az ān rā
I food eat-past.1.sg. and some from it OM
be gorbe dād-am
to cat give-past.1.sg.
Lit. '*I food-ate and gave some of it to the cat.'

(11) *ali Gazā xord va be bače-hā ham dād
Ali food eat.past.3.sg. and to child-Pl also give.past.3.sg.
Lit. '*Ali food-ate and gave (it) to the children.'

3.1 Is NI Lexical?

Mohanan (1995) introduces two different conceptions of lexicality: (a) 'lexical' as belonging to the lexicon as a module where items are formed; (b) lexical as the category of the unit formed. The data put forward so far illustrates that Persian NI can be lexical in both senses. There are, however, postlexical morphological facts that run counter to such an analysis. In the face of these facts, I argue that Persian NI is not created in the lexicon, but that it consists of a V^0 through head adjunction of the noun as a non-projecting lexical item with V^0 in the c-structure, as proposed in Toivonen (2001).

In many languages in the state-of-the-art on incorporation (Mohanan 1995, Mithun 1984, to name just two), the inflectional morphology appears on the edges of the N+V sequence as prefixes or suffixes, treating the whole sequence as a lexical item. In Persian, in contrast, the inflectional morphology appears on the verb as the host, thus intervening between the noun and the verb. These inflectional elements include the present tense (*mi-*), the subjunctive (*be-*), and the negative (*na-*) prefixes. If we assume that NI belongs to the lexicon, these facts are at odds with the assumptions of lexical morphology, based on which compounding (and incorporation as an instance of it) occurs before the word receives inflectional morphology. The prediction is that morpho-syntactic elements do not intervene between the elements, and they appear at the edges. This prediction is not attested in Persian NI as is illustrated by the following data.

(12) āryā **Gazā** *ne-mi-***xor-ad**
Arya food neg-IMP-eat-pres.3.sg.
'Arya doesn't eat food.'

It should be noted that *na-* in this construct has scope over the whole N+V and is not limited to the verb. To further clarify the facts about the scope of the negative marker, consider the following sentences. (The data is from spoken register.)

(13) āryā emruz hič **dars** **na-xund** (NI)
Arya today nothing lesson Neg-read.past.3.sg
'Arya didn't study at all today.'
(14) āryā emruz hič **dars**-iš ro **na-xund**
Arya today nothing lesson-3PossC OM Neg-read.past.3.sg
'Arya didn't study any of his lessons today.'

What (13) conveys is that Arya did not do the activity of reading (lit. lesson-reading) and the negative marker *na-* and the intensifying negative quantifier *hič* have scope over the whole activity denoted by N+V *dars xāndan* and consequently they have the whole sentence in their scope. In (14), the scope of *hič* is limited to the noun and *na-* has scope over the verb *xāndan* and as a consequence over the whole sentence. These facts about the scope of negative, then, point at the lexical status of the noun, while the separation by these prefixes questions the lexical status of the whole N+V sequence.

The future auxiliary *xāhad* also intervenes between the Noun and the Verb.

(15) āryā az aval-e mehr dar dāneSgāh-e tehrān
Arya from first-EZ Mehr in university-EZ Tehran
dars xāhad xānd
lesson FUT.3.sg. read
'Arya will begin his studies in the University of Tehran on 1st Mehr.'

Except for this functional word for future tense and affixes, Persian NI resists separation by content words and the noun remains adjacent to the verb.

3.2 Incorporated Nouns as Non-Projecting Words

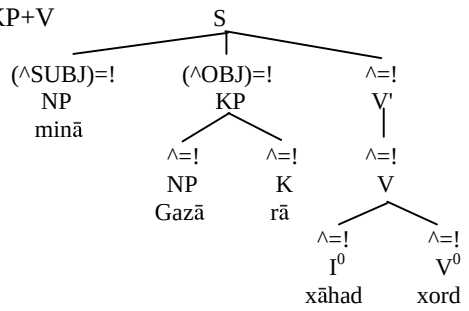
Considering the data in favor of the lexical status of the noun in NI and the data in (12) and (15), the analysis should consider the option of inflecting the verb for tense, aspect, person and number before they make a V^0. Persian verbs never appear in the root form at c-structure. They only appear as word forms, which is in line with the constraint of wordhood on the leafs in the c-structure, proposed by Bresnan (2001). The analysis that best captures the behavior of Persian NI is head-adjunction. In this analysis, the noun is treated as a non-projecting word that makes a V^0 when adjoined to a V^0 inflected for tense and aspect. Compare the following representations (18) and (19) for the NI (17) and its corresponding non-incorporated sentence (16).

To account for the close affinity of the future auxiliary and the other auxiliaries with the verb, head-adjunction of I^0 and V^0 is also posited, which is line with the recursive head adjunction discussed in Sadler (1998). Another explanation is due regarding the c-structure of Persian sentences, given its partial free word order. The Object (KP) and the Oblique (PP) receive case from the accusative marker *rā* and the preposition, respectively. Persian is also a pro-drop language, therefore the semantics of the Subject when dropped, is retrieved from the person/number agreement on the verb. Given that the grammatical functions in Persian are not dependent on their configurational positions in c-structure trees, a flat c-structure representation is posited for this language.

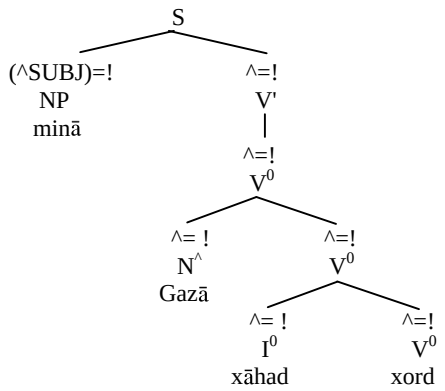
(16) minā Gazā rā xāhad xord
Mina food OM FUT.3.sg. eat-Past
'Mina will eat the food.'

(17) minā Gazā xāhad xord
Mina food FUT.3.sg. eat-Past
Lit. 'Mina food-ate.'

(18) c-structure for non-incorporated KP+V



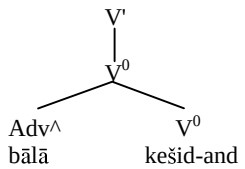
(19) c-structure for NI



As I have said above, the incorporation analysis can be extended to cover cases of the Adv+V sequences as well. As Persian is a free word order language with a flat c-structure, the verb can incorporate an adjacent argument or an adjunct, giving rise to new verbal complexes. These incorporated adverbs denote location and have the semantic role of Goal. As such they have the role predicted by Mithun to be among the roles that can be incorporated by a verb (Mithun 1984). Folli et al. (2005) have neglected this argument relation and classify these constructs as cases of CPr; but they correctly predict that Adv+V verbs have telic interpretations. These adverbs show lexical properties in that they resist modification and other constituency tests. In these cases, the Adv as a non-projecting word is head-adjoined to the V^0. As an instance of one such construction, consider (20) and the c-structure representation of the verbal complex in (21).

(20) xodešān u rā **bālā** **kešid-and**
themselves he OM up pull-3.Pl.Past
'They have promoted him themselves.'

(21)



The idiomaticity of the constructs obscures the argument relationship between the Adv and the V. Therefore, they have been treated as CPrs by most researchers (Karimi 1997; Folli et al 2005; Megerdoomian 2006; Pantcheva 2010). However, if one looks at the semantics of the verbal element in these sequences and the argument and adjunct positions it allows, and decomposes the idiomatic whole to its constitutive semantic components reflected in the literal meaning of each element of these sequences, the incorporation analysis seems plausible.

4 Complex Predication

CPrs in Persian consist of a nonverbal element and a light verb. The nonverbal element are claimed to be adjective, prepositional phrase, adverb or noun. As mentioned above, the scope of this study is limited to the noun and light verb combinations. In the section on the differences between NI and CPrs, we observed that the noun in CPrs is phrasal (it can be modified and it allows cliticization). It also allow scrambling and topicalization which are strong constituency tests. This fact about the phrasal status of the constitutive elements of CPrs is widely acknowledged in the state-of-the-art on Persian Cprs (Karimi 1997, Karmi-Doostan 1997, Megerdoomian 2006, Muller 2009). Karimi (1997) illustrates cases where the noun is modified by the quantifiers and adjectives. She also points at cases where the noun is separated by an intervening propositional phrase subcategorized by the noun in some of the CPrs (22).

(22) kimea unā ro **da'vat** be mehmuni **kard**
Kimea them rā invitation to party did
'Kimea invited them to a party.' (Karimi 1997: 281)

These cases cannot be straightforwardly accounted for in derivational approaches favored by most scholars in their studies of Persian CPrs: the noun must be accessible for both types of movements, that is the movements giving rise to CPr formation and also to undergo the syntactic operations of scrambling and modification (Vahedi-Langrudi 1996, Folli et al. 2005, Megerdoomian 2006, Pantcheva 2010). This has resulted in the fact that in the analyses, either only one type of movement has been the focus of the study or they have to assume many successive movements in the derivation.

LFG, as a theory which allows different independent interacting levels of representations, explains CPr formations by appealing to the predicate composition at c-structure and argument fusion at a-structure independently of each other. To account for complex predication in Persian, I draw on the theory of predicate composition proposed in Butt (1995) and Alsina (1997). In this theory, CPr formation occurs at the level of a-structure, which is independent of c-structure representations. The independence of c-structure allows for the possibility of discontinuous constituents mapping onto one single PRED value in the f-structure, a level of representation that is linked to both c-structure and a-structure.

In the remainder of this section, I deal first with CPr formation at the a-structure based on the argument fusion analysis developed in Butt (1995) and Alsina (1997). I avail myself of Pustejovsky's (1995) theory of Generative Lexicon towards a more fine-grained analysis of argument fusion and event fusion and also to be able to account for the cases where the noun combined with the light verb is not eventive. Then I attempt at c-structure analysis of Persian CPrs in line with Alsina (1997) in terms of a PRED sharing constraint.

4.1 Argument Structure of Persian CPrs

According to the theory of predicate composition (Alsina 1997; Butt 1995), the argument structure of two semantic heads that carry the PRED value contribute to the overall argument structure or event structure of the CPr. What triggers argument fusion is an incomplete predicate which has a complete predicate in its argument structure. This complete predicate is represented as P* in Alsina (1997) to stand for any predicate that "will fully specify the underspecified argument structure of the incomplete predicate" (234). Butt (1995, 1997) states the same idea by integrating a Transparent Event ({E_T}) in the argument structure of the light verb. This {E_T}, which stands for an argument taking predicate, triggers CPr formation and the argument fusion of the highest argument of the embedded predicate with the lowest argument in the embedding incomplete predicate. I follow Butt's (1995) analysis of complex predicate formation and argument fusion with minor changes and also posit in the same line that the light verb is an incomplete predicate which selects for a transparent event, e_t. The eventive predicate combined with the light verb, then, maps onto this event, e_t. In the constructions under study, N+V sequences, the predicate that combines with the light verb is a nominal. Therefore this analysis should be modified to integrate mechanisms for mapping a noun onto an event regardless of the semantics of the noun as a predicative one as in *da'vat kardan* 'invitation doing' (to invite), or a non-predicative and non-eventive one as in *guš kardan* 'ear doing' (to listen). This calls for a deep lexical semantic theory of nominals as developed in Pustejovsky (1995).

Pustejovsky (1995) assumes a matrix for lexical semantic representation of words in terms of four different levels of argument structure (ARGSTR), event structure (EVENTSTR) and qualia structure (QUALIA) and lexical inheritance structure. Of these, the first three which are relevant for this study, will be posited in the representations. ARGSTR provides information about the number and the types of the arguments. EVENTSTR gives a description of an event in terms of its type (state, process, and transition), its internal structure and the subevents involved. QUALIA includes (a) the information of how an object and its constituents are related (CONSTITUTIVE role); (b) what distinguishes the object within a larger domain such as size, color (FORMAL role); (b) the purpose and the function of an object (TELIC role); and (d) factors involved in its origin or bringing it about (AGENTIVE role). Every word based on its type will be specified for the relevant kind of semantic information. What is activated is determined by the context in which the word is used. To clarify how this works in a

representation of light verbs, let us consider the following representation for *kardan* 'to do'[2].

(23)

$$\begin{bmatrix} \text{kardan 'to do'} & \\ \text{ARGSTR} & \begin{bmatrix} \text{ARG}_1\text{= x: ag} \\ \text{ARG2=e}_t \end{bmatrix} \\ \text{EVENTSTR} & \begin{bmatrix} \text{E}_1\text{=e1: process} \\ \text{E}_2\text{=}E \end{bmatrix} \\ \text{QUALIA} & \begin{bmatrix} \text{AGENTIVE= act (e}_1\text{, x, e}_t\text{)} \end{bmatrix} \end{bmatrix}$$

What this representation says is that the light verb *kardan* has two arguments: an x which is specified as the agent (I have departed from Pustejovsky's analysis in annotating the arguments for their thematic roles in order to serve the linking theory in mapping from arguments to grammatical functions), and an e_t which is the event which combines with the light verb to make it complete. In the event structure, it comprises of two events: e_1, a process which is the inherent event property of *kardan*, and the second one e_2 is the event contributed by the predicative noun and yet unspecified. The AGENTIVE role states that the agent (x) does the event denoted by e_2. When light verb combines with the eventive noun, the nominal's argument, event and qualia structure merge to give rise to a CPr. Following are the representations for *da'vat* 'invitation' (24) and the CPr formed as a result of the semantic information merging of these two predicates at different levels. (24) says that the eventive noun *da'vat* has three arguemnts, involved in an event made up of two subevents, process (of inviting) and state (of the invited being at z). RESTR in the EVENTSTR puts a restriction on the precedence of the events, here saying that the process subevent comes first. The two events are embodied in terms of TELIC role and AGENTIVE role in the QUALIA. The headedness principle says which subevent is the head in the internal structure of the event. The head subevent is the event that is linked to the syntax to be realized, which corresponds to TELIC role or AGENTIVE role. Here the head is process, which will be the one selected to fill in the e_t slot in the AGENTIVE role of *kardan*, as illustrated in (25).

2 Since there is not enough space for discussing the semantics of all the light verbs, the scope of the analysis will be further limited to the productive light verb *kardan* 'to do'. This light verb contributes agentivity to the complex.

(24)

da'vat 'invitation'

ARGSTR
- ARG_1= x: ag
- ARG_2=y: th
- D-ARG=z: loc

EVENTSTR
- $E_1=e_1$: process
- $E_2=e_2$: state
- RESTR=<
- Head=e_1

QUALIA
- TELIC= AT(e2, y, z)
- AGENTIVE= act (e_1, x, y, z)

(25)

da'vat kardan 'invitation doing'

ARGSTR
- ARG_1= x: ag
- ARG_2= ARGSTR-da'vat
  - ARG_1= x: ag
  - ARG_2=y: th
  - D-ARG=z: loc

EVENTSTR
- $E_1=e_1$: process
- E_2=EVENTSTR-da'vat:
  - $E_1=e_1$: process
  - $E_2=e_2$: state
  - RESTR=<
  - Head=e_1
- RESTR= o

QUALIA
- TELIC= AT(e2, y, z)
- AGENTIVE= act (e_1, x_i, (x_i, y, z))

The above representation illustrates how the different structures of these predicate merge to predicate jointly. The CPr inherits all the arguments of the constitutive predicates and in the level of the AGENTIVE role, the argument structure of the incomplete predicate is completed as the AGENTIVE role of the complete predicate is added. The presence of e_t triggers the merging of the two AGENTIVE roles. Accordingly, the higher argument of the embedded predicate is co-indexed with the only argument available for binding in the matrix predicate, since their thematic roles are compatible, both agents here. The composition of the two event structures into one follows from the event headedness principle: the event that is the head of the embedded predicate (process in *da'vat*) will be the event that is contributed at the level of event structure and maps to e_t in the AGENTIVE role of *kardan*. The composition ends in two simultaneous process events which leaves the question of headedness of the event at the matrix level irrelevant. The state event in the embedded event contributes to the *aktionsart* of the complex predicate, as reflected in the TELIC role of the CPr. Any event which has

process and state subevents in its internal structure, with process preceding the state and the process being the head subevent, is an accomplishment. *da'vat kardan* with this internal event structure is, therefore, an accomplishment.

The predicative nouns have an event structure that satisfies the type required by the light verb to combine with, but the non-predicative nouns lack an event structure and do not have the event type required by the light verb. The combination of the light verb with a non-predicative noun results in a type clash, hence such a combination is predicted to be ruled out. The data, however, shows that it is possible for a non-predicative noun to combine with a light verb. *guš kardan* 'ear doing' (to listen) is one such construction, as used in (27). The assumption is that the QUALIA structures of the nouns have AGENTIVE and TELIC roles that make it possible for the light verb to select the event from there through the application of Selective Binding. Selective Binding is defined as follows (26).

(26) "SELECTIVE BINDING:
If α is of type <a,a>, β is of type b, and the qualia structure of β, QS_β, has quale, q of type a, the $\alpha\beta$ is of type b, where $[[\alpha\beta]]=\beta\cap\alpha(q_\beta)$." (Pustejovsky 1995: 129)

The application of the selective binding provides the light verb with the event type it requires to become complete. *kardan* looks into the representation for the non-predicative noun and it finds an event encoded in the TELIC role of the QUALIA of the noun, it selects that event and binds it into the ARGSTR. This event will be value of the e_t in the AGENTIVE role of *kardan*. The two representations below, for *guš* 'ear' (28) and the CPr *guš kardan* 'ear doing' (to listen) (29) are meant to clarify how selective binding works in the predicate composition of non-predicative *guš* with *kardan*. Notice that arguments represented for *guš* are those of the event denoted in its TELIC quale. Since the non-predicative noun does not have an event structure, it contributes nothing to the EVENTSTR. Thus, the event structure of this CPr has only a process in its internal event structure, giving rise to an activity reading. In the AGENTIVE role, the two events merge and the identical arguments of them are co-indexed and unified.

(27) man diruz be rādiyo guš kardam
I yesterday to radio ear do-Past.1.sg.
'Yesterday, I listened to the radio.'

(28)

guš 'ear'
ARGSTR [ARG_1= x: instrument]
QUALIA [FORMAL= x
TELIC= listen (e, y:ag, z:go)]

(29)

guš kardan 'ear doing'
ARGSTR [ARG_1= x: ag
ARG_2= TELIC-guš [ARG_1=x: ag
ARG_2=y: go]]
EVENTSTR [E_1=e_1: process
Head=e_1]
GUALIA [AGENTIVE= act (e_1, x_i, (x_i, y))]

In order to map the arguments onto the relevant grammatical functions, lexical mapping theory (LMT), as discussed in Butt (1995) and Bresnan (2001), applies. Before applying LMT, the headedness principle (Pustejovsky 1995) determines which arguments are mapped obligatory and which arguments are optional. The arguments in the head event are obligatorily mapped onto the grammatical functions. The f-structure representations as a result of mapping from arguments onto grammatical functions for the verbs *da'vat kardan* (30) and *guš kardan* (31) are given below.

(30) da'vat kardan: AGENTIVE= act (e_1, x_i, (x_i, y, z))

ag	th	loc
-o	-r	-o
-r		+r
SUBJ	OBJ	OBL

(31) guš kardan: AGENTIVE= act (e_1, x_i, (x_i, y))

ag	go
-o	-o
-r	+r
SUBJ	OBL

As it can be read from the above representation, the two predicates are combined to make one predicate on the a-structure and f-structure levels and they share one argument structure.

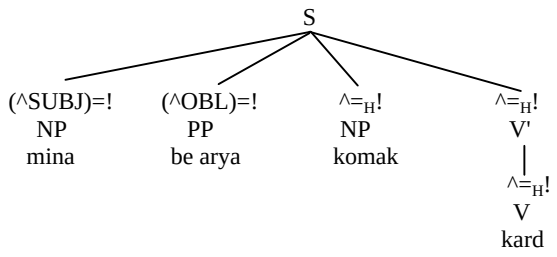
4.2 Tree Structure of Persian CPrs

As concluded above, Persian CPrs map on one PRED value at the level of f-structure. The constitutive elements of CPrs, however, are not adjacent in the c-structure and the noun shows properties typically associated with phrasal categories, such as modification, relativization and scrambling. In order to allow for two constituents specified with the PRED value to compose on the c-structure and at the same time to account for discontinuous constituents of CPrs on the phrasal structure level, Alsina (1997) suggests annotating the relevant nodes of the constitutive elements with $\uparrow=_{H}!$. This constraint says that f-structure values of the mother except for the PRED value are unified and the PRED value of the mother node is the result of the composition of the PRED value of that node with that of its head sister constituents. To illustrate how this works, c-structure representations for sentences (32) and (33) in are given in (34) and (35), respectively.

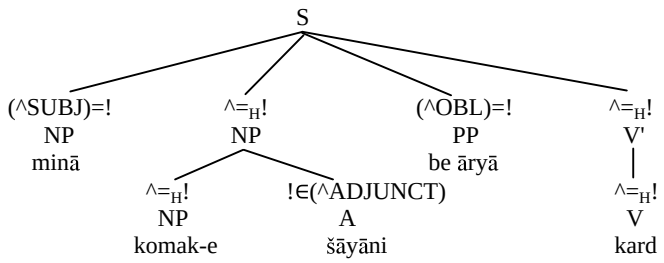
(32) minā be āryā komak kard
Mina to Arya help do.past.3.sg
'Mina helped Arya.'

(33) minā komak-e šāyāni be āryā kard
Mina help -EZ considerable to Arya do.past.3.sg
'Mina helped Arya a lot.'

(34)



(35)



The c-structure (34) shows that both the light verb and the noun are annotated by a PRED composition constraint, $\uparrow=_{H}!$, requiring the PRED value of the annotated nodes to be composed to the PRED value of the

mother node. The tree structure in (35) also illustrates that the scope of modification is not limited to the noun *komak* 'help' and the PRED value is shared by the whole CPr.

5 Conclusion

N+V sequences show different syntactic behaviors. Some of these N+V sequences are classified as NI because there is an argument relationship between the noun and the verb and there is a strict constraint on the adjacency of the noun and the verb. NI cases are analyzed based on head-adjunction of non-projecting words proposed by Toivonen (2001). Based on this analysis, the noun has a lexical status and is not able to project as a phrasal category. The other N+V sequences are discussed as cases of complex predication, where the noun combines with an incomplete predicate to project on a single PRED value with a shared argument structure. To analyze CPrs, I used the CPr formation analysis proposed in Butt (1995, 1997) and Alsina (1997) along with adaptations from Pustejovsky's multi-layered semantic representations. To extend the analysis of the Noun+V CPrs from eventive nouns to non-eventive nouns, selective binding (Pustejovsky 1995) of semantic information was considered. Finally, an attempt is made to account for the discontinuous constituent structure of CPrs by employing the PRED composition constraint developed in Alsina(1997) to apply on the c-structure representations. The scope of this research was limited to N+V sequences and in the analysis of CPrs I limited the scope to the light verb *kardan* 'to do'.

References

Alsina, Alex. 1997. A Theory of Complex Predicates: Evidence from Causatives in Bantu and Romance. In Alex Alsina, et al. (eds.), *Complex Predicates*, pages 203-246 . Stanford: CSLI Publications.

Bresnan, Joan. 2001. *Lexical Functional Syntax*. Oxford: Blackwell Publishing.

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

Butt, Miriam. 1997. Complex Predicates in Urdu. In Alex Alsina, et al. (eds.), *Complex Predicates*, pages 107-149. Stanford: CSLI Publications.

Dabir-Moghaddam, Mohammad. 1997. Compound Verbs in Persian. *Studies in Linguistic Sciences* 27, 25-59.

Folli, Raffaella, Heidi Harley, and Simin Karimi. 2005. Determinants of Event Type in Persian Complex Predicates. *Lingua* 115(10),1365–1401.

Karimi, Simin (1997) Persian Complex Verbs: Idiomatic or Compositional? *Lexicology* 3.1, 273-318.

Karimi-Doostan, Gholamhossein. 1997. *Light Verbs Constructions in Persian*. Ph.D. thesis, University of Essex.

Megerdoomian, Karine. 2006. The Status of the Nominal in Persian Complex Predicates. available at http://www.zoorna.org/papers/preverbs.pdf

Mithun, Marianne. 1984. The evolution of noun incorporation. *Language* 60, 847–894.

Mithun, Marriane and Gerville Corbett. 1999. The effect of noun incorporation on argument structure. In Lunella Mereu (ed) *The Boundaries of Morphology and Syntax,* pages 49-71. Amsterdam: John Benjamins.

Mohanan, Tara. 1995. Wordhood and Lexicality: Noun Incorporation in Hindi. *Natural Language and Linguistic Theory* 13, 75-134.

Mohanan, Tara. 1997. Multidimensionality of Representation: NV Complex Predicates in Hindi. In Alex Alsina, et al. (eds.), *Complex Predicates,* pages 431-471. Stanford: CSLI Publications.

Pantcheva, Marina. 2010. First Phase Syntax of Persian Complex Predicates: Argument Structure and Telicity. *Journal of South Asian Linguistics* 2, 53-72.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge: MIT Press.

Sadler, Louisa. 1998. Welsh NPs without Head Movement. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG98 Conference,* CSLI On-line Publications.

Toivonen, Ida. 2001. *The Phrase Structure of Non-Projecting Words*. Ph.D. Thesis, Standford University.

Vahedi-Langrudi, Mohammad-Mehdi. 1996. *The Syntax, Semantics and Argument Structure of Complex Predicates in Modern Farsi*. Ph.D. thesis, University of Ottawa.

ON SNAKES AND LOCATIVE BINDING IN HUNGARIAN

György Rákosi

University of Debrecen

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King
(Editors)

2010

CSLI Publications

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

Abstract

The paper investigates the peculiar pattern of reflexivity marking in Hungarian locative and directional PPs. Unlike in English, where both a pronoun and a reflexive can serve this purpose, only the reflexive is grammatical in standard Hungarian in these contexts. For many speakers, however, pronominal coding of reflexivity is also an option in first and second persons in locative PPs. The paper presents an LFG-theoretic analysis that rests on the assumption that this option is only available for speakers who treat the PP as a possessive structure, hence licensing the pronoun in what is essentially a non-local binding dependency.

1. Introduction

Locative and directional PPs represent a well-known context in English where the usual complementarity between reflexives and pronouns breaks down. (1) below is a much-cited example in the pertinent literature.

(1) *John*$_i$ *saw a snake near him*$_i$ / *himself*$_i$.

In general, the pronoun counts as the unmarked choice in these contexts, and there is some disagreement over the status of the reflexive.[1] Nevertheless, it is a fact that both the pronoun and the reflexive are acceptable in (1) and generally in sentences of this kind, which I will be referring to as "snake-sentences" for ease of exposition.

It is also well-known that there is significant variation across languages in how such reflexive dependencies are encoded. This article discusses snake-sentences in Hungarian, a language which shows a particularly interesting distribution of reflexive markers in this construction. In standard Hungarian,

[1] Here are some illustrative quotes on judgements concerning reflexivity marking in locative PPs:

- Faltz (1985: 100) : "With normal intonation, reflexive pronouns are at best odd in these positions, ..."
- Reinhart & Reuland (1993: 687): "The use of the anaphor ... is much more marked ..."
- Huddleston & Pullum (2002: 1489): "There is variation across speakers and particular examples, but for many, the non-reflexive form is preferred ..."

only the reflexive is grammatical in third person and speakers consistently reject the bound reading of the pronominal PP.[2]

(2) *János*$_i$ *látott egy kígyó-t* *maga* *mellett*$_i$ / **mellett-e*$_i$.
John saw a snake-ACC himself beside beside-3SG
'John saw a snake beside him.'

In first and second persons, however, many speakers also accept pronominal coding of reflexivity. Examples (3a) and (3b) represent the judgements of these speakers.

(3) a. *Lát-t-am* *egy* *kígyó-t* *magam mellett* / *mellett-em*
see-PAST-1SG a snake-ACC myself beside beside-1SG
'I saw a snake beside myself/me.'

b. *Lát-t-ál* *egy* *kígyó-t* *magad mellett* / *mellett-ed*
see-PAST-2SG a snake-ACC yourself beside beside-2SG
'You saw a snake beside yourself/you.'

Where snake-sentences are mentioned in the literature on Hungarian, only the reflexive is claimed to be grammatical (cf., for example, É. Kiss 1987 and Marácz 1989). The fact that many speakers also accept the pronominal is only mentioned in passing in den Dikken et al. (2001: 147-48, ft. 9).

My main objective here is to provide an explanation for the Hungarian facts. In Hungarian, pronominal encoding of reflexivity is a marked option in the sense that (i) some speakers accept only reflexives, but not pronouns in these contexts, (ii) even the more permissive speakers tend to experience the reflexive as more natural, and (iii), as we will see, there are no instances of obligatory pronoun marking in snake-sentences, but there are many constructions where even the more permissive speakers accept only the reflexive.

In this article, I seek an answer to the following two questions to explain this distribution. First, when exactly is it possible to use pronominals instead of reflexives to encode reflexivity in snake-sentences in Hungarian? Second, why is it that, unlike in English, the pronominal is a marked choice in locative PPs? I show that in fact the reflexive *must* consistently be used in local contexts, and locative PPs count as the local domain for purposes of binding in Hungarian. Pronominal marking is only licensed when the inflecting postposition acts as the head of a possessive structure, with its (possibly *pro*-dropped) pronominal complement being concomitantly analyzed as a structural possessor. Since now the dependency between the

[2] The grammar of agreement-marked Hungarian postpositions is briefly discussed in Section 3.2.

"possessor" and its antecedent is non-local, pronominal coding of reflexivity becomes an option in snake-sentences − at least for those speakers who allow for the possessive analysis.

The structure of the paper is as follows. In Section 2, I briefly overview the standard LFG approach to locative binding, and comment on the typological picture. In Section 3, I discuss the Hungarian data and pair them up with the corresponding English structures. In Section 4, I provide an analyses of the Hungarian facts paying special attention to why the pronoun is the marked choice. In addition, I also make some tentative comments on the origin of the third person/non-third person difference manifested in (2) and (3) above. Finally, I conclude this paper in Section 5.

2. The standard LFG approach and a typological outlook

The standard LFG account of reflexivity marking in English snake-sentences builds on the core assumption that there is a binding domain asymmetry between reflexives and pronouns (see especially Dalrymple 1993, and Büring 2005). In standard LFG, Reflexives are +NUCLEAR in the sense that they are constrained to find an antecedent within the *minimal complete nucleus*, i.e. the smallest f-structure that contains the f-structure of the reflexive and a SUBJ.[3] Pronouns are −NUCLEAR in the sense that they are constrained to be disjoint from their coarguments. The coargument domain is defined by the PRED feature, and this domain need not include a SUBJ. Therefore locative PPs, being predicative, will constitute a negative binding domain for pronouns, but, lacking a SUBJ, they will not constitute a (positive) binding domain for reflexives.

---

[3] I quote here the formal definition of the *Minimal Nucleus Condition* of Bresnan (2001: 271):

(i) **Minimal Nucleus Condition**

A binding constraint designator ((GFα↑)GF') in a nuclear (respectively, nonnuclear) binding constraint is subject to the minimal nucleus condition if
(i) GF and GF' are argument functions and
(ii) if the attribute string α is nonempty, then setting α=*xa* for some attribute *a* and possibly empty string of attributes *x*, the off-path constraint ¬(→SUBJ) (respectively ¬(→PRED)) holds for every attribute in GF *x*.

Notice that this condition requires the f-structure that contains the bound element (GF) to be an argument function. This may be too restrictive for the purposes of describing locative binding, since the reflexive is licensed both in argument and adjunct PPs (see example 4).

As has been repeatedly observed, locative binding per se is not sensitive to whether the PP is an argument or an adjunct of the main predicate (see Dalrymple 1993 and Lødrup 2005 in the LFG literature, as well as, a.o., Reinhart & Reuland 1993 and Büring 2005). So both (1), repeated here as (4a), and (4b) show the same pattern even though the PP is an adjunct in the former but an argument in the latter.

(4) a. *John*$_i$ *saw a snake near him*$_i$ / *himself*$_i$.

b. *John*$_i$ *placed the snake near him*$_i$ / *himself*$_i$.

The f-structure in (5) represents (4b), and it serves to illustrate the binding proposal sketched above.[4]

(5)

$$f1: \begin{bmatrix} \text{SUBJ} & [\text{PRED } \text{‘John’}]_i \\ \text{PRED} & \text{‘place} <(f_1\ \text{SUBJ})\ (f_1\ \text{OBJ})\ (f_1\ \text{OBL}_{loc})>\text{’} \\ \text{TENSE} & \text{past} \\ \text{OBJ} & [\text{PRED } \text{‘snake’}] \\ \text{OBL}_{loc} & f_2: \begin{bmatrix} \text{PRED} & \text{‘near} <(f_2\ \text{OBJ})>\text{’} \\ \text{OBJ} & [\text{PRED } \text{‘him | self’}]_i \end{bmatrix} \end{bmatrix}$$

The positive binding domain for *himself* is f_1, since this is the smallest f-structure that contains a SUBJ. The f-structure of the oblique argument, f_2, is a negative binding domain for the pronoun *him*. This is so because f_2 has a PRED feature and that defines a binding theoretically relevant domain for the pronoun within which it cannot be bound. As a result, the relation between the pronoun and the antecedent is non-local at the level of f-structure, and therefore the dependency is licensed.

In many languages, however, reflexives and pronouns *are* in complementary distribution even in snake-sentences. German is one such language, consider the following examples:

(6) a. *Hans*$_i$ *sah eine Schlange neben *ihm*$_i$ / *sich*$_i$.
Hans.NOM saw a snake near him.DAT self.DAT

b. *Ich sah eine Schlange neben mir / *sich.*
I. NOM saw a snake near me. DAT self. DAT

[4] I use the notation “self” simply as the PRED value of the reflexive. Since nothing crucial hinges on a more elaborate semantic representation, here and throughout I use indices on the respective f-structures to represent binding dependencies.

In the third person, the morphologically simple reflexive *sich* must be used, and the pronoun is ungrammatical (6a). In first and second persons only the pronoun is grammatical, and *sich* is out (6b). Notice, however, that the primary reflexive strategy is consistently *sich* in the third person, but in non-third persons the pronoun is used for the same purpose.[5] The following two transitive structures represent this point.

(7) a. *Hans* *sah* **ihn*$_i$ / *sich*$_i$.
Hans. NOM saw him. ACC self. DAT

b. *Ich* *sah* *mich* / **sich*.
I. NOM saw me. ACC sich. ACC

Thus in German snake-sentences represent no special binding context: they show the same complementary distribution of pronominal and anaphoric markers of reflexivity as transitive structures do.

The deeper question is why German differs from English in this respect. Not being aware of a solution in the syntactic literature that predicts this difference in the two languages, I simply point out that it nevertheless complies with the typology that Faltz (1985) makes. Faltz observes that there is a correspondence between the morphological make-up of the primary reflexive of a given language and the coding of reflexivity in snake-sentences. In particular, complex reflexives, like the English *himself*, compete as reflexive markers with pronominals in locative PPs. Morphologically simple reflexives, like the German *sich*, tend to be obligatory markers of reflexivity in snake-sentences.

The German facts can be analyzed in the LFG binding theory by assuming that there is no domain asymmetry between pronouns and reflexives in German. Crucially, this means that predication per se will play no role in determining the (negative or positive) binding domain. Rather, the binding domain is defined with respect to presence of a SUBJ function.[6] Thus, for example, *ihn* 'him' cannot be bound within an f-structure that contains a SUBJ, but the first person pronoun *mich* 'me' can. In the same vein, *sich* must be bound within an f-structure that contains a SUBJ.

In essence, I will argue here that Hungarian is like German as far as the definition of the binding domain is concerned. The relevant binding domain is defined by the presence of a SUBJ. Reflexives must consistently be bound in this domain, but pronominals cannot be. The reason why we can still get

---

[5] The primary reflexive strategy is the strategy that is used in transitive sentences.

[6] This is the more traditional approach to defining binding domains. In the recent LFG literature, Lødrup (2007) makes detailed arguments to show that we need this approach - instead of the predicate-based binding approach of, for example, Dalrymple (1993) and Reinhart & Reuland (1993) - to properly describe the distribution of the Norwegian anaphors *seg* and *seg selv*.

pronominal marking of reflexivity in Hungarian snake-sentences is that in every such case the P-element is reanalyzed and a more complex structure is created than what first meets the eye. I proceed now to give a detailed description of how this happens in the following two sections.

3. Locative binding in Hungarian: the empirical background

3.1. Introduction

The main purpose of this section is to establish the empirical facts that the analysis in Section 4 is intended to cover. I first overview the grammar of P-elements in Hungarian, with special respect to inflecting postpositions, the P-elements that most prominently figure in snake-sentences (3.2). Inflecting Ps developed out of possessive structures, and this history plays a key role in understanding the grammar of locative binding Hungarian. After the descriptive overview of P-grammar, I give a survey of the basic locative binding facts of Hungarian and link them up with the English data (3.3).

3.2. On P-markers in Hungarian

Hungarian has three types of P-elements. The core set, which we typically find in snake-sentences, contains postpositions that Marácz (1989) calls "dressed Ps". These postpositions take nominative complements, and agree with this complement if it is pronominal. Hungarian is a highly inflecting language and agreement markers license *pro*-drop. This typically happens in neutral discourse conditions in PPs headed by dressed Ps:

(8) (*én*) *mellett-em*
(I.NOM) beside-1SG
'beside me'

The construction shows some surface similarities with possessive constructions, compare (8) with (9):

(9) *a*(*z* *én*) *fej-em*
the I.NOM head-1SG
'my head'

The same agreement morphology is utilized in both constructions, but it licences a postpositional complement in one case and a possessor in the other.

The analogy between possessive noun phrases and PPs is not a forced one, nor are the similarities accidental. All Hungarian P-elements come historically from nominal sources. *Mellett* 'beside', for example, is the complex of the noun *mell* 'chest' and the archaic locative marker *-tt*. This origin has become obscure for native speakers, and there are good reasons

not to collapse the grammar of possessive structures and dressed PPs (see Bartos 1999 and Asbury 2008).

I discuss here two arguments supporting a differential analysis. In the case of pronominal possessors, as in (9), the possessive structure must take the definite article, which can never be dropped. A definite article, however, is not licensed in PPs:

(10) (*az) (*én*) *mellett-em*
the (I.NOM) beside-1SG
'beside me'

Second, there is a gap in the agreement paradigm of postpositions: non-pronominal complements do not trigger agreement (11a). This is not so in possessive structures, where the agreement paradigm is full (11b):

(11) a. *János* *mellett*(*-*e*)
John beside-3SG
'beside John'

b. *János fej*-*(*e*)
John head-3SG
'John's head'

There are further differences between the two constructions, but (10) and (11) suffice to prove the point.

Accordingly, the agreement markers themselves must have distinct lexical representations, cf.:

(12) a. (*én*) *mellett-em* -(*e*)*m*: [P_]P, ((↑OBJ PRED) = 'pro')
I.NOM beside-1SG (↑OBJ PERS) = 1
'beside me' (↑OBJ NUM) = sg

b. *a*(*z én*) *fej-em* -(*e*)*m*: [N_]N, ((↑POSS PRED) = 'pro')
the I.NOM head-1SG (↑POSS PERS) = 1
'my head' (↑POSS NUM) = sg

The only relevant difference is that possessive morphology combines with nouns and licenses a POSS argument, whereas postpositional agreement markers combine with P-elements and license pronominal OBJ complements.

I will argue below that ultimately it is their historical origin that allows inflecting Ps to be analyzed as possessive structures, at least for some speakers and only as a marked option. What exactly should trigger this analysis is what I discuss in the next subsection.

Before concluding this section, I briefly want to point out that Hungarian has two other types of P-elements. Some locative Ps are in fact suffixal markers which cannot be separated from the noun head (13a). Some others are true postposition that take case-marked complements, and unlike dressed Ps, never agree with them (13b).

(13) a. *Kati-hoz*
John-to
'to John'

b. *Kati-val szemben*
Kate-with in.front.of
'in front of Kate'

I do not discuss these two P-markers in this article, and refer the reader to the extensive literature available for details of their grammar.[7]

3.3. A parallel presentation of English and Hungarian data

It is well-known that there are contexts in which the reflexive is obligatory in locative and directional phrases in English, and even if both the reflexive and the pronoun are allowed as reflexive markers, they need not convey the same meaning. Interestingly, a very similar distributional variation is attested in Hungarian locative and directional PPs.

I have already pointed it out that it does not much matter whether the PP is an adjunct or an oblique argument, the binding facts are still the same in English. This is true as long as the P-element is *semantic* in the sense of Butt et al. (1999).[8] If it is selected by the verb and is void semantically, either as a non-semantic case marker or as part of a possibly larger idiom, then normally only the reflexive is grammatical (see, a.o., Reinhart & Reuland 1993 and Büring 2005). Consider the following set of examples:

(14) a. *John*$_i$ *gave the snake to *him*$_i$ */himself*$_i$.

b. *John*$_i$ *looks after *him*$_i$ */himself*$_i$.

c. *John*$_i$ *was beside *him*$_i$ */himself*$_i$ *with rage.*

---

[7] See, among others, Ackerman (1990), Ackerman & Webelhuth (1998), Asbury (2008), Bartos (1999), É. Kiss (1998, 2002), Marácz (1989), and Surányi (2009a,b).

[8] A semantic P is one which has a PRED feature and introduces a subcat frame of its own. For more detailed LFG-theoretic discussions of differences between semantic and non-semantic Ps, I refer the reader to Bresnan (1982) and Dalrymple (2001). I thank Miriam Butt for useful suggestions to improve the presentation here.

Similarly, the pronominal PPs are consistently out in Hungarian even for the less restrictive speakers if the PP has an idiomatic contribution:

(15) *Ki-tart-ok* **mellett-em / magam* *mellett.*
out-hold-1SG beside-1SG myself beside
'I stand by myself.'

(16) a. *Vág-om a fá-t* *alatt-am.*
cut-1SG the tree-ACC under-1SG
(i) 'I am cutting the tree under me.'
(ii)*'I am cutting my own throat.'

b. *Vág-om a fá-t* *magam alatt.*
cut-1SG the tree-ACC myself under
(i) 'I am cutting the tree under me.',
(ii)'I am cutting my own throat.'

(15) includes a postposition that is selected by the verb. The sentence in (16) can be interpreted literally or figuratively - this latter option is only licensed, however, by the reflexive PP. As far as I am aware, this is always the case in Hungarian. If the postposition is not interpreted literally, then reflexivity can only be coded via the reflexive even for speakers that otherwise allow for pronominal coding in snake-sentences.

There are certain contexts where even though both pronominal and reflexive PPs are acceptable, they seem to have a distinct meaning contribution. It has been pointed out repeatedly that the reflexive, unlike the pronoun, often triggers body-oriented readings (see Bresnan 2001 and Lødrup 2007 in the LFG literature). Examples (17) and (18) are from Rooryck & Vanden Wyngaerd (2007: 54, 59).[9]

(17) a. *When he*$_i$ *woke up, John*$_i$ *found a rope around him*$_i$.
√*It described a neat circle 4 meters in diameter.*

b. *When he*$_i$ *woke up, John*$_i$ *found a rope around himself*$_i$.
**It described a neat circle 4 meters in diameter.*

(18) a.**The earth*$_i$ *revolves around it*$_i$.
b. *The earth*$_i$ *revolves around itself*$_i$.

[9] Examples of this sort lead Rooryck and Vanden Wyngaerd (2007) to argue against predicate-based binding proposals (see also Lødrup 2007 for similar arguments). While the predicate-based approach (described here in Section 2) does not predict in itself the difference between the (a) and (b) examples above, such data do not necessarily refute it. My analysis of the Hungarian data will not force me to decide on the right analysis of the English facts, and therefore I leave this issue open.

As opposed to the reflexive, the pronoun is used in (17a) and (18a) to refer not to the body of the antecedent's referent, but to a possibly more abstract/extended location identified in relation to him or it.

Such a difference is manifest also in Hungarian, and in fact it seems to me that this effect is stronger in Hungarian than in English. Consider the following two examples, which are essentially analogous to the English ones discussed in the previous paragraph:

(19) a. *Érez-t-em* *a* *kígyó-k-at* *körülött-em.*
feel-PAST-1SG the snake-PL-ACC around-1SG
'I felt the snakes around me.' <body reading is out>

b. *Éreztem* *a* *kígyó-k-at* *magam* *körül.*
feel-PAST-1SG the snake-PL-ACC myself around
'I felt the snakes around myself.' <body reading is ok>

(20) a.**Lassan* *forg-ok* *körülött-em.*
slowly revolve-1SG around-1SG
'*I am slowly revolving around me.'

b. *Lassan* *forg-ok* *magam* *körül.*
slowly revolve-1SG myself around
'I am slowly revolving around myself.'

(20), just like the English (18), is nonsensical when the pronominal PP is used. (19a) is acceptable when the snakes do not touch the speaker's body. When the snakes are directly swirling around the speaker's body, the reflexive must be used (19b).

A further interpretative difference between reflexives and pronouns in snake-sentences is triggered by variation in point-of-view. In particular, it has been argued that reflexives always have a *logophoric* use in an extended sense in these contexts (see especially Reinhart & Reuland 1993 and Rooryck & Vanden Wyngaerd 2007). The contrasting pair in (21) is from Rooryck and Vanden Wyngaerd (2007: 35).

(21) a. *They*$_i$ *placed their guns, as they looked at it,*
*in front of *them*$_i$*/themselves*$_i$.

b. *They*$_i$ *placed their guns, as I looked at it,*
in front of them$_i$*/*themselves*$_i$.

Whether such effects indeed manifest themselves across the board in English or not, contrastive pairs of the sort that (21) represents are difficult to reproduce in Hungarian. This is probably so because the reflexive is overwhelmingly the default choice in snake-sentences even for the more

restrictive speakers, and therefore perspective effects do not seem to have been grammaticalized in a reflexive/pronominal PP opposition.

Discussing the Hungarian data, two facts have arisen here that are in need of explanation. First, pronominal marking of reflexivity is only licensed if the P-element is semantic and has its own subcategorization frame. Where its contribution is non-semantic or idiomatic, speakers only find the reflexive grammatical. Second, there seems to be a distribution of labour between pronominal and reflexive PPs when both are acceptable: the former are used primarily for body-oriented readings, whereas the latter necessarily seem to trigger a reading where the relevant location is an area vaguely construed in relation to the position of the referent of the antecedent. In the next section, I make an attempt at explaining these two facts.

4. Possessive analysis of postpositions and its relevance in binding

4.1. Independent evidence for the possessive analysis

To explain the facts observed in the previous section, I propose here that many speakers can still analyze inflecting postpositions in Hungarian as heads of possessive structures. How that explains the binding data is an issue that I pick up in the next subsection. In this subsection, I present independent evidence that such a possessive analysis is indeed an existing phenomenon in Hungarian and is therefore not an artefact of the current analysis.

As noted in 3.2, inflecting postpositions all go back historically to possessive sources. It is actually the case that some postpositions still have not fully undergone this diachronic change and have retained some possessive traits that are synchronically available. Such a postposition is *számára* ‘for’. Literally the opaque complex of the noun *szám* ‘number’ and the suffix *-ra* ‘onto’, this postposition marks benefactives and certain types of experiencers. (22a) below is an example for this postposition, (22b) is a possessive construction, (22c) is a regular inflecting postposition (with its opaque internal complexity made explicit).

(22) a. (*a*) *szám-om-ra*
the number-1SG-onto
‘for me’

b. *(*a*) *fej-em-re*
the head-1SG-onto
‘onto my head’

c. (**a*) *mell-ett-em*
the chest-LOC-1SG
‘beside me’

There are two important ways in which *számára* ‘for him’ differs from regular postpositions. As we have seen already in 3.2, regular Ps do not take the definite article (22c), whereas possessive structures with a pro-dropped or overt pronominal possessor must take it (22b). The postposition *számára* ‘for him’ can still optionally co-occur with the definite article (22a). Second, in the case of regular postpositions, the agreement morphology always comes last (22c), following any possible (obscure) locative morphology. In possessive constructions and in the case of *számára*, agreement morphology precedes locative/directional morphology.

The upshot of this discussion is that there exist inflecting postpositions in the synchronic system which have more possessive-like behaviour than usual and which may provide analogical grounds for a possessive analysis of other, more grammaticalized postpositions. In fact, Surányi (2009b) argues that the assumption of such an analysis may be necessary to explain the data in (23).

(23) a. *Kati-nak* *ellopt-ák* *a* *bicikli-jé-t.*
Kate-DAT stole-3PL the bike-POSS.3SG-ACC
‘They stole Kate’s bike.’

b. *Mögé-dobt-am* *a* *kígyó-t* *Kati-nak.*
to.behind.3SG-threw-1SG the snake.ACC Kate-DAT
‘I threw the snake behind Kate.’

(23a) is possessive construction with the possessor having been extracted. When that happens, the possessor receives dative case. In (23b), the directional inflecting postposition forms a complex predicate with the verb (see Forst, King & Laczkó (this volume) on particle-verb complex predicate formation in Hungarian). The complement of the incorporated postposition may appear further away from the complex, and just like extracted possessors, it receives dative case. Surányi takes this to be an instance of possessor extraction, and then (23a) and (23b) are analogous in this respect. It is important to note that not every speaker finds the construction in (23b) perfectly acceptable. Crucially, it implies that the possessive analysis of postpositions is a marked option, and this is in fact what I claim here.

4.2. Accounting for the binding facts

Recall that the primary target of this article is to explain the contrast between (24a) and (24b). Many speakers do not like (24a), and even those who do, prefer (24b) if the constraints discussed in 3.3 do not intervene.

(24) a. *Lát-t-am* *egy* *kígyó-t* *mellett-em.*
see-PAST-1SG a snake-ACC beside-1SG
‘I saw a snake beside me.’

b *Át-t-am* *egy* *kígyó-t* *magam* *mellett*.
see-PAST-1SG a snake-ACC myself beside
'I saw a snake beside myself.'

Pronominal coding of reflexivity is thus a marked option in Hungarian snake-sentences. Why is this so?

What I propose here is that inflecting postpositions have two lexical entries, and that creates an important, binding-theoretically relevant structural difference between (24a) and (24b). The two lexical entries are as follows:

(25) a. *mellett-em*$_1$: P, (↑PRED) = 'BESIDE <(OBJ)>'
((↑OBJ PRED) = 'pro')
(↑OBJ PERS) = 1
(↑OBJ NUM) = sg

b. *mellett-em*$_2$: P, (↑PRED) = 'BESIDE <(OBJ)>'
(↑OBJ PRED) ='PLACE<(POSS)>'
((↑OBJ POSS PRED) = 'pro')
(↑OBJ POSS PERS) = 1
(↑OBJ POSS NUM) = sg

Mellettem$_1$ is a run-of-the-mill inflecting postposition, and this is the lexical entry that occurs in (24b). *Mellettem*$_2$ is a possessively analyzed variety of *mellettem*$_1$.[10] Not every speaker possess this lexical entry, and those who do are the ones that find (24a) acceptable, for (24a) contains this entry.

Essentially, (25a) means 'beside me', whereas (25b) means something like 'beside my place'. In other words, a silent PLACE predicate is introduced into the lexical representation of the possessively analyzed postposition. That may sound a marked feature of the analysis in a theory that generally prefers not employ silent elements. Note nevertheless that Bresnan (1994: 110) makes use of the same device to explain why locative PPs may occupy positions typically reserved for noun phrases in English (26). Her analysis is in (27).

(26) a. *Is* [*under the bed*] *a good place to hide*?

b. [*Under the bed*] *is a good place to hide, isn't it*?

(27) [NP (*a place*) [PP under the bed]]

[10] I refer the reader to Laczkó (1995, 1997, and 2001) for a more detailed LFG-theoretic analysis of possessive constructions in Hungarian.

One difference is that whereas I assume that the silent predicate is introduced in the lexicon, Bresnan argues that it is a missing nominal head that is "contextually interpreted as an instance of ellipsis".

Let me now illustrate how this analysis explains our binding data. I start with (24b) repeated as (28), whose f-structure is in (29). This f-structure contains the lexical entry *mellett-em*$_1$.

(28) *Lát-t-am* *egy* *kígyó-t* *magam* *mellett*.
see-PAST-1SG a snake-ACC myself beside
'I saw a snake beside myself.'

(29)
$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 1 \end{bmatrix}_i \\ \text{PRED} & \text{'see <(SUBJ) (OBJ)>'} \\ \text{TENSE} & \text{PAST} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'snake'} \\ \text{DEF} & - \end{bmatrix} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'beside <(OBJ)>'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'self'} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 1 \end{bmatrix}_i \end{bmatrix} \right\} \end{bmatrix}$$

Nothing special needs to be said about (29): it includes the regular inflecting postposition, and its complement, the reflexive finds its antecedent in the smallest f-structure that includes a SUBJ. That f-structure is the f-structure of the clause.

Sentence (24a) is repeated below as (30), and its f-structure is in (31) on the next page.

(30) *Lát-t-am* *egy* *kígyó-t* *mellett-em*.
see-PAST-1SG a snake-ACC beside-1SG
'I saw a snake beside me.'

(30) includes the possessively analyzed entry *mellett-em*$_2$. I claim that only those speakers accept (30) who have this entry. The result is an f-structure that is richer than (28), because it includes an extra possessive layer.

(31)

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{‘pro’} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 1 \end{bmatrix}_i \\ \text{PRED} & \text{‘see <(SUBJ) (OBJ)>’} \\ \text{TENSE} & \text{PAST} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{‘snake’} \\ \text{DEF} & - \end{bmatrix} \\ \text{ADJ}_{\text{loc}} & \left\{ \begin{bmatrix} \text{PRED} & \text{‘beside <(OBJ)>’} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{‘PLACE<(POSS)>’} \\ \text{POSS} & \begin{bmatrix} \text{PRED} & \text{‘pro’} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 1 \end{bmatrix}_i \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix}$$

The object of the postposition, which includes the extra possessive layer, is marked by grey shading in the f-structure in (31).

This analysis explains the Hungarian facts in the following way. Pronominal marking of locative coreference is a marked option in Hungarian, because it includes an extra structural layer otherwise not present in snake-sentences by default. Furthermore, this extra layer is projected by a possessively analyzed lexical entry of the postposition, and since not every speaker stores this entry in his/her lexicon, not every speaker will accept pronouns in snake-contexts. [11]

What triggers the use of the possessive postpositional structure is the distribution of labour between pronominals and reflexives that we observed to exist in snake-sentences (section 3.3). The reflexive is preferred in body-

[11] Following Laczkó (1995, 1997) and Bresnan (2001: 254), I assume that POSS is a SUBJ-like function and thus it defines a negative binding domain for pronouns. That this is so is evident in the case of regular possessives:

(i) *Lát-t-am* *egy* *kígyó-t* *a* *hely-em-en*.
see-PAST-1SG a snake-ACC the place-1SG-at
‘I saw a snake at my place.’

oriented readings. The pronominal PP is used consistently when we make reference to the extended location around the referent of the P-object. This semantics follows now from the possessive semantics of the P-element itself and the silent PLACE predicate that is introduced. Finally, pronominal PPs have been found to be grammatical markers of reflexivity only if the P-element is literally interpreted. If the current analysis is on the right track, this requirement is in fact a prerequisite for the possessive analysis. Only semantic Ps can be analyzed meaningfully as possessive structures along the lines of (31).

4.3. On the third person constraint

Remember that every speaker rejects pronominal marking of reflexivity in snake-contexts if the antecedent is third person. This judgement is very strong when the antecedent is a lexical noun. I repeat (2) as (32) to illustrate.

(32) *János*$_i$ *látott egy kígyó-t maga mellett*$_i$ / **mellett-e*$_i$.
John saw a snake-ACC himself beside beside-3SG
'John saw a snake beside him.'

The sentence actually gets better if the antecedent is a pronoun (33a), and it is best, but still not fully grammatical, if the pronoun is dropped (33b). The question marks represent the judgements of the speakers of the less restrictive dialect. Other speakers may find these sentences completely ungrammatical.

(33) a. ??*Ő*$_i$ *lát-ott egy kígyó-t mellett-e*$_i$.
he see-PAST.3SG a snake-ACC beside-3SG
'He saw a snake beside him.'

b. ?*Lát-ott*$_i$ *egy kígyó-t mellett-e*$_i$.
see-PAST.3SG a snake-ACC beside-3SG
'(He) saw a snake beside him.'

I do not have a full account of why we have this particular difference between third person and non-third person. Notice nevertheless that under the current analysis, the non-acceptability of the pronominal form implies that the postposition cannot be analyzed here as a possessive head. Why should this be so?

It is interesting to note once again that inflecting postpositions only agree with pronominal complements, and not with lexical ones. I repeat (11a) as (34a), and compare it to (34b)

(34) a. *János mellett*(**-e*)
John beside-3SG
'beside John'

b. *ő-* *mellett-*(*e*)
he beside-3SG
'beside him'

Possibly, the fact that the possessive analysis of inflecting postpositions is blocked in the presence of lexical antecedents (32) is related to the fact that such non-pronominal noun phrases do not otherwise trigger agreement with the postposition (34a). Similarly, the fact that pronominal antecedents fare a bit better (33) might be related to the fact that pronominal complements must agree with the inflecting postposition (34b). In other words, it might be that the facts of (34) influence the pattern in (32-33) on analogical grounds.[12]

There also exists a more general functionalist account for the binding theoretically relevant divide between third and non-third persons (see, for example, Faltz 1985 and Reuland 2008). Unlike the interpretation of third person pronominals, the interpretation of non-third person pronominals is kept constant per reportive domain. First and second person pronouns are normally not ambiguous referentially in a given context of use. The reference of third person pronominals may switch from one individual to another even in the same discourse. For this reason, languages often employ special reflexive forms in third person which cannot be interpreted ambiguously. In first and second persons, there is no necessary drive to employ reflexive markers. Thus in German first and second person pronouns are used to encode local reflexive dependencies (Section 2). Standard Hungarian uses a reflexive in snake-sentences in first and second persons, but pronominals can also be used for the same purpose since their use does not trigger referential ambiguities anyway.

5. Conclusion

I have shown here that locative and directional PPs are a special context in Hungarian as well as in English as far as the encoding of reflexive dependencies is concerned. The similarities between the two languages are the following:

- Pronominal marking of reflexivity is only an option in snake-sentences if the P-element is semantic and is literally interpreted.
- Reflexives in these positions often trigger body-oriented readings, whereas the pronoun is used when the relevant location is an area vaguely construed in relation to the position of the referent of the antecedent.

[12] This possibility has been suggested to me by Miriam Butt.

I have also shown that the Hungarian pattern differs from the English one in the following two respects:

- The default coreference marker is always the reflexive in Hungarian in snake-sentences, and pronominal coding is a marked option, available only for a subset of speakers.
- There is a divide between third and non-third persons: pronominal marking of coreference is only acceptable in the latter case.

I presented an account of these facts which relies on the independently motivated assumption that inflecting Ps can be analyzed as heads of possessive constructions. The possessive construction creates an extra layer of embedding, which can license a non-local coreference relation between the "possessor" pronoun and the subject antecedent. In this account, there is no domain asymmetry between pronouns and reflexives in Hungarian: the local context that is a negative binding domain for the pronoun and a positive binding domain for the reflexive is the smallest domain that includes a SUBJ or a POSS function.

Acknowledgements

I gratefully acknowledge that the research reported here has been supported in part by OTKA (*Hungarian Scientific Research Fund*), grant number: K 72983. I also acknowledge the support of the TÁMOP 4.2.1-08/1-2008-003 project, which is implemented through the New Hungary Development Plan co-financed by the European Social Fund, and the European Regional Development Fund.

I thank the LFG10 reviewers and the audience of the Ottawa conference for useful comments and discussions. I also thank the editors of this volume for their assistance, and I am especially grateful to Miriam Butt for extensive comments on the first draft. Any errors that remain are solely mine.

Bibliography

Ackerman, Farrell. 1990. The morphological blocking principle and oblique pronominal incorporation in Hungarian. In K. Dziwirek et. al. eds. *Grammatical Relations: A Cross-Theoretical Perspective.* Stanford: CSLI Publications. 1-19.

Ackerman, Farrell & Gert Webelhuth. 1998. The composition of (dis)continuous predicates: lexical or syntactic? *Acta Linguistica Hungarica* 44. 317-340.

Asbury, Anna. 2008. *The morphosyntax of case and adpositions*. PhD dissertation. University of Utrecht.

Bartos, Huba. 1999. *Morfoszintaxis és interpretáció: A magyar inflexiós jelenségek szintaktikai háttere*. [Morphosyntax and interpretation: The syntactic background of inflectional phenomena in Hungarian.] PhD dissertation. Budapest: ELTE.

Bresnan, Joan ed. 1982. *The Mental Representation of Grammar*. Cambridge, MA: The MIT Press.

Bresnan, Joan. 1994. Locative inversion and universal grammar. *Language* 70 (1). 72-131.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Büring, Daniel. 2005. *Binding Theory*. Cambridge: Cambridge University Press.

Dalrymple, Mary. 1993. *The Syntax of Anaphoric Binding*. Stanford: CSLI Publications.

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

den Dikken, Marcel; Lipták, Anikó & Zvolenszky, Zsófia. 2001. On inclusive reference anaphora: new perspectives from Hungarian. In K. Megerdoomian & L. Bar-el eds. *WCCFL 20 Proceedings.* Somerville, MA: Cascadilla Press. 137-149.

Faltz, Leonard. 1985. *Reflexivisation: A study in universal syntax*. New York: Garland.

Forst, Martin; King, Tracy Holloway & Laczkó, Tibor. 2010. Particle verbs in computational LFGs: Issues from English, German and Hungarian. *This volume*.

Huddleston, Rodney & Pullum, Geoffrey K. eds. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

É. Kiss, Katalin. 1987. *Configurationality in Hungarian*. Dordrecht: Reidel.

É. Kiss, Katalin. 1998. Verbal prefixes or postpositions? Postpositional aspectualizers in Hungarian. In C. de Groot & I. Kenesei eds. *Approaches to Hungarian 6*. Szeged: JATE.

É. Kiss, Katalin 2002. *The Syntax of Hungarian*. Cambridge: Cambridge University Press.

Laczkó, Tibor. 1995. *The syntax of Hungarian noun phrases. A Lexical-Functional Approach*. Frankfurt am Main: Peter Lang.

Laczkó, Tibor. 1997. Action nominalization and the possessor function within Hungarian and English noun phrases. *Acta Linguistica Hungarica* 44. 413-475.

Laczkó, Tibor. 2001. A magyar birtokos DP inflexiós morfológiájáról lexikai-funkcionális megközelítésben. [On the inflectional morphology of the Hungarian possessive DP in a Lexical-Functional approach.] In: M. Bakró-Nagy et al. eds. *Újabb tanulmányok a strukturális magyar nyelvtan és a nyelvtörténet köréből. Kiefer Ferenc tiszteletére barátai és tanítványai*. Budapest: Osiris Kiadó. 59–77.

Lødrup, Helge. 2007. A new account of simple and complex reflexives in Norwegian. *The Journal of Comparative Germanic Linguistics* 10 (3). 183-201.

Marácz, László. 1989. *Asymmetries in Hungarian*. PhD dissertation. Rijskuniversiteit Groningen.

Reuland, Eric. 2008. Anaphoric dependencies: How are they encoded? Towards a derivation-based typology. In E. König & V. Gast eds. *Reciprocals and Reflexives. Theoretical and Typological Explorations. Trends in Lingusitics 192*. Berlin/New York: Mouton de Gruyter. 499-556.

Rooryck, Johan & Vanden Wyngaerd, Guido. 2007. The syntax of spatial anaphora. In M. Bašić et al. eds. *Tromsø Working Papers on Language and Linguistics: Nordlyd 34.2*. Tromsø: CASTL. 33-85.

Surányi, Balázs. 2009a. Locative particle and adverbial interpretation at the interfaces. In K. É. Kiss ed. *Adverbs and adverbial adjuncts at the interfaces*. Berlin: Mouton de Gruyter. 39-74.

Surányi, Balázs. 2009b. Adpositional preverbs, chain reduction and phases. In M. den Dikken & R. Vago eds. A*pproaches to Hungarian 9. Papers from the 2007 New York Conference*. Amsterdam/Philadelphia: John Benjamins. 217-250.

APPLICATIVIZING COMPLEX PREDICATES: A CASE STUDY FROM MURRINH-PATHA

Melanie Seiss and Rachel Nordlinger
University of Konstanz University of Melbourne

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

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

Abstract

In this paper we give an analysis of Murrinh-Patha verbs as morphological complex predicates. We argue that the different parts of the complex predicate provide information for different layers of the argument structure; more precisely, that classifier stems determine the number of arguments a verbal complex takes while lexical stems contribute thematic information. We further show how the argument structure composition interacts with valency-changing processes such as applicativization and reflexivization/ reciprocalization and that the argument structure has to be built up from right to left.

1 Introduction

Complex predicates have been discussed widely in the LFG literature, e.g. Mohanan (1994), Butt (1995), Alsina (1996), Alsina et al. (1997), Andrews and Manning (1999), Wilson (1999).[1] However, little work has dealt with the interaction of complex predicates with other valency-changing morphological processes.

This article focuses on the formation of complex predicates and its interaction with applicativization and reflexivization/reciprocalization. We are concerned with complex predicates in Murrinh-Patha, a polysynthetic language from Northern Australia. Murrinh-Patha has a bipartite verb system in which an inflecting element with a rather generic meaning combines with an invariable element, which carries more specific meaning, into a single morphological word. In this paper we use the term 'classifier stem' for this generic element (following previous work on Murrinh-Patha and related languages and McGregor 2002), and 'lexical stem' for the other element; the two combine to form a complex predicate.

However, these complex predicates seem to work differently from what has been previously established in the literature. We show that these complex predicates can be accounted for by assuming two different levels on which the different parts of the complex predicate operate: the classifier stem provides the argument structure (slots) while the lexical stem provides thematic information. A distinction between these different levels has been used in previous analyses of complex predicates, for example, by Alsina (1997) or Mohanan (1997). However, in those analyses both parts of the complex predicate jointly contribute information for these levels. The analysis we propose thus adds a new dimension to the discussion around

[1] Unless otherwise specified, the data in this paper is taken from Rachel Nordlinger's fieldnotes collected in Wadeye between 2005-2009. Rachel would like to thank the Murrinh-Patha speakers who have so patiently taught her their language, especially Carmelita Perdjert, Norma Kulumboort, Bonaventure Ngarri and Theodora Narndu. She is also grateful to the Arts Faculty at the University of Melbourne, and the Australian Research Council (DP0343354 and DP0984419) for funding fieldtrips. For helpful input into the analysis presented here, we would like to thank the participants of the LFG Conference 2010 in Ottawa. Special thanks go to Miriam Butt who provided very valuable ideas and commented on various versions of the poster and this article.

complex predicates cross-linguistically.

To help the reader follow the complex examples in the later sections, we first provide a short overview of the Murrinh-Patha language and a discussion of its verbal structure in section 2.

In section 3 we discuss the composition of the argument structure for the basic verbal complex. We show that the number of arguments of the complex predicate mainly follows the number of arguments provided by the classifier stem. In contrast, the thematic information is provided by the lexical stem. Argument structure alternations as known in other languages, e.g. the locative or causative alternation, are thus realized by combining the same lexical stem with different classifier stems.

Our approach is also compatible with the interaction of complex predicates with morphological valency-changing processes. Murrinh-Patha has a morphological applicative marker *–ma* which promotes a source to the function of a direct object (Nordlinger 2009). We discuss this applicative marker and its interaction with the argument structure of complex predicates in section 4.

Finally, section 5 deals with the interplay of the applicative and the various reflexive/reciprocal constructions in Murrinh-Patha. Different coindexation properties exist when applicatives and reflexive/reciprocal processes interact. We account for these differences by assuming that reflexivization/reciprocalization can either coindex two arguments or may result in an intransitive verbal complex. The interaction of the applicative and reflexive/reciprocal processes further shows that the argument structure is built up in the inverse order of the morphological markers in the verbal template, i.e. from right to left.

2 Murrinh-Patha overview

Murrinh-Patha is spoken in and around Wadeye (Port Keats) in the Daly River region of the Northern Territory of Australia. It is one of a small number of Australian languages which is still acquired as a first language by children in the community. Currently, there are approximately 2500 speakers – probably more than pre-contact because it has developed into the lingua franca of the region. It is a head-marking, polysynthetic language which incorporates body parts, adverbials etc. in the verbal complex and has minimal case morphology.

The nominal system comprises nominal classifiers, nouns, adjectives, demonstratives and numerals (Blythe 2009). A single noun may be used with one or several of the 10 semantically motivated noun classes (Walsh 1997). As the focus of this article is on complex predicates, we will not deal with the nominal system in any detail here.

As has been stated before, verbs are bipartite, consisting of a classifier stem (traditionally glossed with a number) combined with a lexical stem into a single morphological word. The classifier stems are inflected for subject person, number and tense/aspect/mood. This information is encoded in portmanteau forms. There are approximately 38 classifier stem paradigms.

Lexical stems on the other hand are invariable, and their class is

considerably larger. The combination of classifier stem and lexical stem determines the verbal predicate. Some straightforward examples are given in (1). (1a,b) show examples of the same classifier stem combining with different lexical stems. In contrast, (1c,d) show examples of the same lexical stem combining with different classifier stems.[2]

(1a) *bangarntal*
bangam-rtal
3sS.BASH(14).nFut-chop
'He chopped it (with an axe).'

(1b) *bangamelmel*
bangam-melmel
3sS.BASH(14).nFut-flatten
'He flattened it (with a hammer).'

(1c) *nungarntirda*
nungam-rirda
3sS.FEET(7).nFut-push
'He kicked him.'

(1d) *marntirda*
mam-rirda
3sS.HANDS(8).nFut-push
'He pushed him (with his hands).'

Some classifier stems, primarily classifier stems 1-8, can also occur without a lexical stem. This is exemplified in (2) which shows two portmanteau forms of the classifier stem 'SIT(1)'. In contrast to lexical stems in other Australian languages, e.g. in Wagiman (Wilson 1999), lexical stems in Murrinh-Patha can never function as the sole verbal stem.

(2a) *dim*
3sS.SIT(1).nFut
'He's sitting.'

(2b) *pirrimka*
3dS.SIT(1).nFut
'They two are sitting.'

As mentioned above, classifier stems are inflected for subject person and number and tense/aspect/mood (TAM). There are five major TAM categories: non-Future (nFut), Past Imperfective (PImp), Future (Fut), Future Irrealis (FutIrr) and Past Irrealis (PstIrr). Apart from non-Future, the TAM categories require an additional tense marker on the combination of classifier and lexical stem. Additionally, the classifier stems 1 to 7 carry aspectual information when they are combined with a lexical stem or cliticize onto another classifier and lexical stem combination. We will not be concerned with this phenomenon here, for more information see Street (1996) and Nordlinger (2010a).

Subject number is also jointly determined by the portmanteau forms of the classifier stem and number markers which morphologically attach to the classifier and lexical stem combination. The classifier has three number categories. The special number markers distinguish between dual and plural in two different genders and sibling classes. Different combinations of these classifier forms and number markers then yield a five-way number and

[2] The following non-obvious abbreviations are used: *NC* noun classifier (NC:anim 'animate class', NC:hum 'human class', NC:water 'water class'), *DO* direct object marker, *DM* discourse marker (function unknown), *IO* indirect object marker (both object markers may be inflected for gender and number), *sS/dS/pS* singular/dual/plural subject form of classifier stem, *du.m/du.f* dual masculine/feminine subject or object marker, *pauc.m/pauc.f* paucal (3-10 people) masculine/feminine subject or object number marker, *FOC* focus marker, *RR* reflexive/reciprocal marker.

Where data is taken from Street (1989), we provide our own glosses.

sibling contrast, which is schematicized in table 1 (see Nordlinger 2010a for further discussion).

Classifier form	**Number Marker**	**Subject Properties**
SING	unmarked	singular
SING	dual non-sibling	dual non-sibling
DU	unmarked	dual sibling
DU	paucal non-sibling	paucal non-sibling
PL	unmarked	paucal sibling/plural

Table 1: Subject number and sibling marking

Object agreement is also marked on the verbal complex. Example (3a) shows 1st person singular direct object marking which attaches directly to the classifier stem. 3rd person singular direct object marking, however, is unmarked, which can be seen in (3b).

(3a) *dirra-**ngi**-wintharrarr-dha*
3sS.WATCH(28).PImp-**1sDO**-seek-PImp
'He was looking for me.'

(3b) *dirra-wintharrarr-dha*
3sS.WATCH(28).PImp-seek-PImp
'He was looking for him/her.'

Similarly to direct object marking, indirect objects are also marked in the same slots of the verbal template. An example is given in (4).

(4) *nhinhi-re thim-**na**-ku*
2s-Erg 2sS.SIT(1).nFut-**3smIO**-hit
'You punched him.'

Beside classifier and lexical stem, markers for TAM, subject number and object, the verbal complex may include incorporated body parts and an applicative marker, adverbials and a reflexive/reciprocal marker as can be seen in the simplified verbal template in table 2. We will come back to the applicative and reflexive/reciprocal markers in later sections when we deal with the phenomena in more detail.

This section gave the description of the verbal system of Murrinh-Patha required to follow the examples in the following sections. For a more detailed view on the various template slots and the interaction between the different markers see Blythe (2009) and Nordlinger (2010b). For further references on other phenomena of the language see e.g. Walsh (1976, 1996), Street (1987, 1989) and Nordlinger (2009, 2010a).

1	2	3	4	5	6	7	8	9
CS.SUBJ.TNS	SUBJ.NUM / OBJ	RR	IBP / APPL	LEXS	TNS	ADV	SUBJ.NUM / OBJ.NUM	ADV

Key:		
	CS.SUBJ.TNS:	classifier stem, marked for tense/aspect/mood & subject person/number
	SUBJ.NUM:	subject number markers for dual & paucal
	OBJ:	object agreement marker
	OBJ.NUM:	object number marker for dual & paucal
	RR:	reflexive/reciprocal marker
	IBP:	incorporated body part
	APPL:	applicative marker
	LEXS:	lexical stem
	TNS:	tense marker
	ADV:	adverbial

Table 2: Simplified verbal template

3 Argument structure of classifier and lexical stem

In this section we provide an analysis of the composition of argument structure in the verbal complex. Following work on similar constructions in other Australian languages (e.g. Wilson (1999) for Wagiman, Schultze-Berndt (2000) for Jaminjung, and Bowern (2004) for Bardi), we treat the two parts of the verb – the classifier stem and the lexical stem – as providing aspects of the argument structure, which combine to form the argument structure of the whole.

McGregor (2002) emphasizes the (semantic) classificational role of classifier stems, i.e. he treats the classifier stem as a verbal classifier similar to semantically motivated nominal classifier systems. He proposes that three parameters may be relevant for the kind of classifier system that can be found in Murrinh-Patha: aktionsart, valency and vectorial configuration. In this paper we focus only on valency, as mediated through argument structure. Aktionsart and other aspects of lexical semantics play a substantial role in determining permissible combinations of classifier and lexical stems, the resulting semantics of the complex predicate and the semantic effects of alternative possible combinations. A complete analysis of the Murrinh-Patha system therefore needs to incorporate an analysis of these issues as well. Such an analysis is well beyond the scope of this study, however, and so we will restrict our discussion to argument structure issues alone.[3]

The main semantic load of the verbal complex is carried by the lexical stem. Therefore, the thematic information of the verbal complex is provided by the lexical stem. We will show that the classifier stem does not provide thematic information when it combines with a lexical stem. For this reason,

[3] Recall from the discussion in section 2 that lexical stems can never occur alone, and neither can the vast majority of the classifier stems. This means that an analysis of the contributions of each element of the complex predicate can only be arrived at through careful analysis of all possible combinations that each element is found in. Such an analysis requires substantial expansion of the present corpus, and we leave it for future research.

we use variables to represent the argument slots in the argument structure of the classifier stem.

This analysis thus calls on a distinction between semantic/thematic structure and argument structure. A distinction between these levels has been used in the analysis of complex predicates for example by Alsina (1997) and Mohanan (1997). However, in these analyses the different parts of the complex predicates determine both levels jointly, e.g. the complex predicate itself determines the thematic roles and the number of argument slots in the two different levels. In contrast, we argue that the Murrinh-Patha classifier stems alone determine the number of argument slots, and the lexical stems provide thematic information. These two parts of the argument structure then have to combine before linking to grammatical functions can commence, accounting for why neither part can constitute a verbal predicate on its own.

We provide some initial straightforward examples here. Following Reid (2000) in his treatment of the verbal structure of Ngan'gitymerri, a Daly river language related to Murrinh-Patha, we assume that classifier stems can be divided into intransitive, transitive and reflexive/reciprocal (RR) stems. Intransitive classifier stems provide one argument <x> while transitive classifier stems provide two arguments <x y>. Thus, in (5), the classifier stem provides one argument slot and the lexical stem provides the thematic information 'agent'. The argument structure of the verbal complex as a whole thus requires one agent as its argument.

A transitive classifier stem as in (6) provides two argument slots. The lexical stem provides an agent and a patient: thus, the verbal complex together requires an agent and a patient.

(5a) *kanam-kaykay*
3sS.BE(4).nFut-call.out
'He continually calls out.'(Street 1989)

(5b) BE(4) < x >
kaykay, 'call out' < ag >
=> BE(4)-*kaykay* < ag >

(6a) *mam-kurrk*
1sS.HANDS(8).nFut-scratch
'I scratched something.'

(6b) HANDS(8) <x y>
kurrk, 'scratch' <ag th>
=> HANDS(8)-*kurrk* <ag th>

Having laid out the basic analysis here, we provide evidence for our claims in the remainder of the section. In 3.1 we discuss evidence for the claim that the lexical stems provide the thematic information while classifier stems only provide argument structure slots. In 3.2 we present three argument structure operations which show that the number of arguments mainly follows the number of slots provided by the classifier stem.

3.1 Evidence: Lexical stems determine theta roles

In this subsection we show that lexical stems provide more fine grained meaning and thematic information while classifier stems may not place strict requirements on the thematic information provided by the lexical stem. Evidence for this claim comes from minimal pairs such as (7) in which different lexical stems combine with the same classifier stem. In (7a) the lexical stem *ngkamumur,* 'be blind' provides a theme while in (7b) the lexical stem *kaykay,* 'call out' provides an agent. These theta roles are also the theta roles of the whole complex predicate.

(7a) *ngani-ngkamumuy-nu*
1sS.BE(4).Fut-be.blind-Fut
'I'll be blind.' (Street 1989)

(7b) *kanam-kaykay*
3sS.BE(4).nFut-call.out
'He continually calls out.' (Street 1989)

Examples of such alternations can also be found for transitive classifier stems. In (8) the same classifier stem combines with two different lexical stems and these determine the thematic roles, i.e. an experiencer and theme in (8a), and an agent and theme in (8b).

(8a) *nakurl ba-nhi-ngkardu-nu*
later 1sS.13.Fut-2sDO-see-Fut
'I'll see you later.'

(8b) *kura patha ba-gurduk-nu*
NC:water good 1sS.13.Fut-drink-Fut
'I will drink water.'[4] (Street 1989)

Thus, classifier stems can vary in terms of the thematic roles depending on the lexical stem involved. The classifier stems are not, however, semantically empty. This can be seen by the fact that not every classifier stem can combine with every lexical stem that provides the right number of thematic roles. Furthermore, different classifier stems lead to different predicate meanings when combined with the same lexical stem, as shown in the examples in (1c) and (1d). Nevertheless, the semantic information provided by the classifier stem can be so opaque in some cases that lexical stems with different thematic roles can combine with the same classifier stem. We leave it to future research to work out a thorough semantic analysis of the combination of classifier and lexical stems.

[4] *Kura patha* is an idiom meaning 'drinking water'.

3.2 Argument structure operations

In the previous subsection we laid out our basic assumptions about the argument structure of Murrinh-Patha verbal complexes and presented evidence that the thematic information is provided by the lexical stem. In this subsection we show that it is the classifier stem which accounts for the number of arguments of the complex predicate.

Determining the number of arguments of the complex predicate is quite straightforward, e.g. direct object marking and/or the presence of an unmarked NP in a patient/theme role indicates that the complex predicate is transitive. However, it is more difficult to determine whether classifier and lexical stems are transitive or intransitive. This can only be determined by looking at the multitude of possible combinations of lexical and classifier stems and working out the semantic meaning of the different parts of the complex predicate.

We consider three different combinatorial types: firstly, we consider cases in which the number of argument slots contributed by the classifier stem and the number of thematic roles contributed by the lexical stem match. Secondly, we look at cases in which the lexical stem provides more thematic roles than argument slots provided by the classifier stem. These cases can be divided into two subgroups: some examples seem to be clearly intransitive in that only one of the thematic roles provided by the lexical stem is realized. These examples will be discussed in 3.2.2. The second subgroup comprises examples with an openly expressed NP which seem to be a direct object. Thus, at first glance these examples seem to contradict our claims. However, as will be discussed in 3.2.3, the NPs in these cases are generic 3rd person objects which seems to be non-referential and have to be treated differently from "normal" objects.

3.2.1 Transitivity Matching

In most cases, the number of argument slots provided by the classifier stem and number of arguments provided by the lexical stem match, i.e. intransitives combine in (9) while transitives combine in (10).

(9a)	*dim-karrk*	(9b)	SIT(1)	< x >
	3sS.SIT(1).nFut-cry		*karrk*, 'cry'	< ag >
	'He's crying.' (Street 1989)		=> SIT(1)-*karrk*	< ag >

(10a)	*mam-lerrkperrk*	(10b)	HANDS(8)	< x y >
	1sS.HANDS(8).nFut-crush		*lerrkperrk,* 'crush'	<ag th>
	'I crushed it.'		=>HANDS(8)-*lerrkperrk*	<ag th>

As stated before, lexical semantic restrictions might further restrict which lexical stems might combine with which classifier stems. We leave the detailed working out of the semantics to future research.

3.2.2 Reduction of Thematic Roles

If the classifier stem only provides one argument, not all thematic roles licensed by the lexical stem may be realized**.** Depending on the lexical semantics of the verb, either the theme or agent might be suppressed. Thus, similar to the distinction between 'eat' and 'eat something' in English, certain lexical stems may omit their theme when combined with an intransitive classifier stem. For example in (11), *dhegdhek,* 'play' may combine with the transitive classifier stem HANDS(8) in (12a) which then yields a transitive verbal complex. It can, however, also combine with the intransitive classifiers stem BE(4) which then yields an intransitive activity reading.

(11a) *pumamka-dhegdhek-ngime*
3duS.HANDS(8).nFut-play-pauc.f
'They're playing around with that girl/boy.'

(11b) *parnamka-dhegdhek-ngime*
3duS.BE(4).nFut-play-pauc.f
'They're playing.'

(12a)	HANDS(8)	< x y>
	dhegdhek, 'play with'	<ag th>
	=> HANDS(8)-*dhegdhek*	<ag th>
(12b)	BE(4)	< x >
	dhegdhek, 'play with'	<ag th>
	=> BE(4)-*dhegdhek*	< ag >

The argument structure of (11a) is presented in (12a) which is a simple transitive combination in which the number of arguments and the number of thematic roles match. In contrast, in (12b), which provides the argument structure for (11b), the classifier stem only provides one argument slot but the lexical stem provides two thematic roles. In the combination of classifier and lexical stem, only the agent is realized.

In contrast, transitive lexical stems with a causative meaning can combine with SIT(1), which deletes the agent and thus triggers an anticausative/resultative reading as in the examples in (13). The lexical stems *lerrkperrk,* 'crush' and *warnta*, 'split open' normally combine with transitive classifier stems which denote the kind of action that leads to the state of being smashed or split open. An example of *lerrkperrk,* 'crush' with a transitive classifier stem is given in (14).

These lexical stems can thus be considered to have an agent and a theme as their thematic roles. When combining with SIT(1), only the theme is realized and the combination has an anticausative reading. The argument structure of (13a) is given in (15).

(13a) *dim-lerrkperrk*
3sS.SIT(1).nFut-crush
'It's smashed.'

(13b) *dim-warnta*
3sS.SIT(1).nFut-split.open
'It's cracked.'

(14) *ku* *tumtum* *mam-lerrkperrk*
NC:anim egg 1sS.HANDS(8).nFut-crush
'I crushed the egg with my hand.'

(15) SIT(1) < x >
lerrkperrk, 'smash' <ag th>
=> SIT(1)-*lerrkperrk* < th >

We suspect that the lexical semantics of the lexical stem determines which thematic role is realized if the classifier stem does not provide enough argument slots. We leave the testing of this for future field work.

To sum up, this subsection has dealt with examples of intransitive classifier stems combining with transitive lexical stems which result in an intransitive verbal complex. The next subsection looks at instances in which the lexical stem also provides more thematic roles than the classifier stem provides argument slots. However, in these examples the thematic roles seem to be realized in the verbal complex.

3.2.3 'Adding' Arguments

We claimed above that the classifier stem determines the number of arguments which are realized in the verbal complex. However, there are a couple of examples in which an intransitive classifier stem seems to result in a transitive verbal complex. In the examples in (16), an intransitive classifier stem combines with a transitive lexical stem and an overt NP.

(16a) *thamul* *pirrim-nga-batbat*
spear 3sS.STAND(3).nFut-1sIO-throw(RDP)
'He always throws the spear at me.' (Street 1989)

(16b) *kura patha kanamkurdugurduk*
kura patha kanam-gurdugurduk
NC:water good 3sS.BE(4).nFut-drink(RDP)
'He continually drinks water.' (Street 1989)

(16c) *ku* *ngurlmirl* *wurran-ku*
NC:anim fish 3sS.GO(6).nFut-fish
'He continually catches fish.' (Street 1989)

At first glance, these examples seem to contradict our claim. However, when looking at the examples more closely, all take a narrow semantic class of generic objects. These objects seem to be non-referential and serve to characterize the action rather than pick out a participant (cf. Van Valin & LaPolla 1997:148ff). Reid (2000), reporting on similar constructions in Ngan'gityemerri, claims that these constructions focus on the subject's posture or activity and are thus less transitive (in a Hopper and Thompson (1980) sense) than regular transitive combinations. Thus, these examples seem to resemble (pseudo) noun incorporation as discussed by Mohanan

(1995), Ball (2004) and Duncan (2007).

However, it is possible to find other (albeit infrequent) examples in which an intransitive classifier stem results in a transitive complex predicate with a clearly referential direct object, as in the following examples:

(17a) *kumparra warra punni-dha* *berematha gathu*
first first 3plS.GO(7).PImp-PImp that's all towards

warda ***pirrim-pun****-mardaputh*
after that **3sS.STAND(3).nFut-3plDO**-load_up_a_truck

'A big mob went in front, after that he picked them all up on the truck.'

(17b) ***ngani-nan****-part-nu-warda* *ngurru-warda*
1sS.BE(4).Fut-2plDO-leave-Fut-now 1sS.GO(6).Fut-now
'I've got to leave you behind, I'm going.'

The empirical facts remain to be fully determined in terms of which intransitive classifiers can combine with direct objects in these ways, and under what conditions. It may be that some intransitive classifiers in fact contain an optional second argument position that enables them to combine with certain transitive lexical stems as well as intransitive lexical stems. This does not invalidate the overall analysis, however: we have no examples in which transitive classifier stems are found in intransitive complex predicates, and in the vast majority of examples the valency of the classifier stem correlates directly with that of the complex predicate.

In this section we have presented our basic assumptions about the argument structure of the verbal complex in Murrinh-Patha and provided evidence for our claim that the number of arguments is generally determined by the classifier stem while the thematic information is provided by the lexical stem. The analysis suggests that the classifier stem and the lexical stem function as co-heads in the verbal complex as they account for different parts of the combined argument structure.

4 Applicatives in Murrinh-Patha

In this section we show how the argument structure of the verbal complex as laid out in the previous section interacts with the valency-changing process of applicativization. The applicative marker *-ma* attaches to the verbal complex in the slot for incorporated body parts (see Nordlinger (2009) for justification of the applicative analysis, and discussion of the relationship between the applicative marker and the incorporated body part *–ma* 'hand').

The applicative promotes a source to the function of a direct object as can be seen in examples (18)-(20). (18a) shows a simple verbal complex with an oblique argument specifying the source. In contrast, (18b) shows the applicativized version in which the source has been promoted to the function of a direct object which is now marked on the verbal complex. (19) is a

similar example with a 2nd person direct object marking the source, while (20) shows that the original object can still be expressed as an overtly expressed object$_{\theta}$.

(18a) *Truck darrarart pumangan-art ngarra ngay.*
truck stolen 3plS.SNATCH(9).nFut-get LOC 1sg
'They stole a truck from me.'

(18b) *pumanganngimart*
pumangan-ngi-ma-art
3plS.SNATCH(9).nFut-1sDO-APPL-get
'They took it from me.'

(19) *nganam-nhi-ma-kut*
1sS.BE(4).nFut-2sDO-APPL-collect
'I collected (the money) from you.'

(20) *mangan-nhi-ma-art kura*
1sS.SNATCH(9).nFut-2sDO-APPL-get NC:water
'I got (some) water from you.'

Based on work by Alsina & Mchombo (1993) on Chicheŵa applicatives, we assume the basic argument structure in (21): the applicative adds an argument to the already existing thematic roles. However, while the Chicheŵa applicative operates on the verb, the Murrinh-Patha applicative adds a source argument to the thematic roles provided by the lexical stem (LS) and adds an argument variable in addition to those contributed by the classifier stem.

(21) **APPL < source >** (**< z >**) **LS < $\theta_{1,\ldots}$ θ_n >** (**< z >**)
APPL – LS < θ_1 … source … θ_n >

(22) displays the argument structure for (18b) in which the source argument is added to the argument structure of the lexical stem and another argument slot is made available for the argument structure of the classifier stem.

(22) SNATCH(9)	<x, y>
art, 'get'	<ag th>
	< z >
=>APPL – *art*	<ag source th>
=>SNATCH(9) – APPL – *art*	<ag source th>

Once the argument structure of the verbal complex is in place, the linking follows the linking principles as put forth in Alsina & Mchombo (1993). Thus, we assume that intrinsic arguments such as themes and applied

arguments receive [-r] and theta roles inherently lower than goal may receive [+o]. We assume that Murrinh-Patha, like Chichewa, carries the constraint that only one [-r] and only one [+o] can be assigned intrinsically. With standard Lexical Mapping Theory (Bresnan & Zaenen 1990) we assume that other theta roles receive [-o] intrinsically, that the highest theta role receives [-r] as default and all others receive [+r] as default.

Applying these principles to the sentence in (18b) renders the linking in (23), when starting with the argument structure of the verbal complex in (22).

(23) SNATCH(9) – APPL – *art*

	<ag	source	th>
IC:	[-o]	[-r]	[+o]
Default:	[-r]		[+r]
	S	O	O_θ

Source is usually not part of the thematic hierarchies in LFG. In Kiparsky's (1985) hierarchy, however, source is higher than goal. If we follow this hierarchy, source cannot receive [+o] as its intrinsic classification and (23) is the only possible mapping for (18b).

Summing up, this section introduced the applicative construction in Murrinh-Patha by which a source is promoted to the function of a direct object. We presented an analysis in which the applicative adds a source argument to the thematic information of the lexical stem. We will justify this analysis in the following section, which deals with the interaction of the applicativization process with reflexivization/reciprocalization. This interaction provides evidence that the applicative first combines with the lexical stem as has been suggested by our analysis. In fact, taking together applicativization and reflexivization/reciprocalization shows that the argument structure has to be built up from right to left.

5 Reflexivization/ Reciprocalization and its interplay with Applicativization

This section introduces two ways to express reflexivity/reciprocality, discusses the interplay of these processes with applicativization and presents an argument structure analysis of this interaction. To account for the different binding relations, the argument structure has to be composed from right to left in productive cases. This provides evidence for the claim of the last section, namely that applicativization has to operate on the lexical stem first.

5.1. Reflexive/reciprocal classifier stems

One way to encode reflexivity/reciprocality in Murrinh-Patha is using a reflexive/reciprocal (RR) classifier stem. Some transitive classifier stems have corresponding RR classifier stems, i.e. transitive lexical stems which combine with a transitive classifier stem may also combine with the

corresponding RR classifier stem.[5] For example, (24a) shows a transitive combination of classifier and lexical stems. In (24b), the corresponding RR classifier stem is used which triggers a reflexive meaning.

(24a) *mam-kurrk*
1sS.HANDS(8).nFut-scratch
'I scratched something.'

(24b) *mem-kurrk*
1sS.HANDS:RR(10).nFut-scratch
'I scratched myself.'

(25) provides the argument structure of the complex predicate in (24b). We assume that the RR classifier stem provides two coindexed argument slots $<x_i, y_i>$. When combining with a lexical stem, the thematic roles of the lexical stem will be coindexed. These thematic roles may then be linked to one grammatical function (Alsina 1996:116ff).

(25)	HANDS:RR(10)	$<x_i \;\; y_i>$
	kurrk, 'scratch'	<ag th>
	=>HANDS:RR(10)-*kurrk*	$<ag_i \;\; th_i>$

RR classifier stems may also trigger a reciprocal meaning when the subject is in non-singular form. An example is given in (26b). We will not be concerned with the semantic difference between reflexives and reciprocals here and treat reflexives and reciprocals alike.

(26a) *ngu-nhi-bat-nu*
1sS.SLASH(23).Fut-2sDO-hit-Fut
'I'm going to hit you.'

(26b) *puy-bat-nu*
1incS.SLASH:RR(24).Fut-hit-Fut
'We are going to hit each other.'

RR classifier stems plus lexical stem combinations may also result in non-reflexive or non-reciprocal meanings. In examples like (27), transitive lexical stems combine with RR classifier stems. However, the combination does not result in a coindexation of the thematic roles involved. Rather, the combination denotes a resultant state, e.g. in (27a), the speaker reports of a state of 'being confused', in which the speaker is not necessarily the source of the confusion himself. Similarly, in (27b), the source of the amazement are fish, not the men themselves, which would be the case if we treat the RR classifier as coindexing the thematic roles of the lexical stem.

(27a) *ngurdampengkawuy*
ngurdam-wengkawuy
1sS.30:RR.nFut-confuse
'I'm confused' (Street 1989)

[5] For a detailed discussion of RR classifier stems and the RR marker, especially which combinations are possible, see Nordlinger (2008).

(27b) *kardu* *ngamere-ka* *pumem-mardat*
NC:hum few-FOC 3plS.HANDS:RR(10).nFut-amaze

ku *ngurlmirl nhini-nu-yu*
NC:anim fish Dem-DAT-DM
'and the few men were amazed at all those fish.' (Street 1989)

For the examples in (27), corresponding non-RR classifier stem combinations exist which help us to determine the thematic roles of the lexical stem. Thus, in (28a) *wengkawuy*, 'confuse' is used with the non-RR classifier stem 23 and results in a combination with an agent and a theme. Similarly, (28b) involves *mardat*, 'amaze' in combination with HANDS(8), a transitive classifier prototypically used to describe actions carried out with one's hands.

(28a) *pan-ngi-wengkawuy*
3sS.SLASH(23).nFut-1sDO-confuse
'He confused me.' (Street 1989)

(28b) *ma-nhi-mardat-nu*
1sS.HANDS(8).Fut-2sDO-amaze-Fut
'I'll amaze you.' (Street 1989)

More difficult cases exist in which the corpus contains no examples with a corresponding non-RR classifier stem combining with the relevant lexical stems. This is the case with the examples in (29). The lexical stems *nham*, 'fear' and *ngkabat,* 'surprise' cannot combine with other classifier stems (as far as we are aware).

(29a) *nhem-nham*
1sS.POKE:RR(21).nFut-fear
'I'm afraid.'

(29b) *be-ngkabat-nu*
1sS.15:RR.Fut-surprise-Fut
'I'll be surprised.' (Street 1989)

We thus treat cases like (29), and by analogy also (27), as providing evidence for lexicalized combinations with non-compositional argument structures. Thus, we consider these combinations to have a monovalent argument structure despite the presence of the RR classifier stem. These combinations are not formed according to the principles outlined in section 3, but rather are lexicalised complex predicates stored as a whole and requiring only an experiencer argument.

That RR markers are used in non-RR constructions like in the examples above is quite common cross-linguistically. The Murrinh-Patha examples resemble what Steinbach (2002) calls German "inherent reflexive constructions" (30) in which the reflexive pronoun is not a semantic argument of the verb, but is needed syntactically.

(30) Er fürchtet sich.
He be.afraid reflexive-pronoun-ACC
'He is afraid.'

Similarly, Kemmer (1993) treats reflexive markers in constructions like the

German and Murrinh-Patha examples as middle markers which should be given a semantic analysis distinct from reflexives. She considers several subclasses of verbs in which reflexive markers are often used as middle markers cross-linguistically. Besides body care verbs and different verbs of motion events, emotion verbs like the Murrinh-Patha examples above are named as prototypical verb classes in which reflexive markers are used as middle markers.

Thus, while we treat the Murrinh-Patha examples in (27) and (29) as lexicalized combinations of classifier and lexical stems, we acknowledge the fact that there may be crosslinguistic tendencies which allow reflexive markers to be used in intransitive, stative events.

To sum up the discussion on RR classifier stems, in productive uses these stems coindex the two thematic roles provided by the lexical stem. Lexicalized combinations of RR classifiers and lexical stems, however, are intransitive and usually denote a resultant state.

5.2 Reflexive/reciprocal marker *–nu*

Reflexivization/reciprocalization can also be achieved by a special reflexive/reciprocal morpheme *–nu* which is positioned after the optional marker for subject number in the verbal template. The marker *–nu* (or *–nunggu* for paucal subjects) can combine with RR classifier stems with little change in meaning (as in 32). It is also used, however, to encode a reflexive/reciprocal meaning for classifier plus lexical stem combinations for which no corresponding RR classifiers exists. An example of the latter type is given in (31b) for the non-reflexive/reciprocal combination (31a).

(31a) *nungarntirda*
nungam-rirda
3sS.FEET(7).nFut-push
'He kicked him.'

(31b) *nungam-ngintha-**nu**-rirda*
3sS.FEET(7).nFut-du.f-**RR**-push
'They kicked each other.'

The RR morphological marker only has the RR function. In terms of argument structure this means that *-nu* always coindexes the two arguments of the lexical stem.

5.3. Interaction with the Applicative

The different behaviours of the RR classifier stem and the RR marker can explain the contrast in (32) and (33). In the productive RR combination in (32), the source is coindexed with the agent in both sentences, independent of whether *–nu* is present; in other words, in both cases the reciprocal relation holds between the 'tearers' and those the clothes were torn from (cf. 'they tore each other from the tree.'). In contrast, in the lexicalized (33b), without

-nu, there is no reflexive/reciprocal relation holding between the subject and applied object arguments. In order to express such a relation, the RR marker *–nu* needs to be added, as in (33c).

(32a) *pam-ngintha-**ma**-rartal*
3sS.SLASH:RR(24).nFut-duf-**APPL**-cut.off(RDP)
'They tore the clothes from each other.'

(32b) *pam-ngintha-**nu-ma**-rartal*
3sS.SLASH:RR(24).nFut-duf-**RR-APPL**-cut.off(RDP)
'They tore the clothes from each other.'

(33a) *nhem-nham*
1sS.POKE:RR(21).nFut-fear
'I'm afraid.'

(33b) *nhem-**nhi**-**ma**-nham*
1sS.POKE:RR(21).nFut-**2sO**-**Appl**-fear
'I'm afraid of you.'

(33c) *them-**nu-ma**-nham*
1incS.POKE:**RR**(21).nFut-**RR-APPL**-fear.
'We're (inclusive) frightened of each other.'

The argument structure of (32) is presented in (34). The RR classifier coindexes two arguments of the stem it combines with and thus there is no difference between (32a) and (32b). We get the same result irrespective of whether it is the classifier stem alone that binds the two arguments (as in 32a, shown in 34a), or whether the RR marker *-nu* binds them first (as in 32b, shown in 34b). To receive the correct binding relations, namely that it is the source (rather than the theme) which is coindexed with the agent, argument composition has to proceed strictly from right to left. If the classifier and lexical stem combined first (before applicativization), the agent and theme would necessarily be coindexed. However, this is not the case in the examples in (32). Thus, the applicative has to combine with the lexical stem first before the reflexivization/reciprocalization process coindexes the two arguments. Lexical mapping theory will ensure that it is the source which is coindexed with the agent in the combined structure, and not the theme, as we show in (35) below.

(34a) SLASH:RR(24)	$<x_i \;\; y_i>$
rartal, 'cut off'	<ag th>
=> APPL – *rartal*	<ag source th>
=> SLASH:RR(24)-APPL-*rartal*	$<ag_i \; source_i \; th>$

(34b) SLASH:RR(24)	$<x_i \;\; y_i>$
rartal, 'cut off'	<ag th>
=> APPL – *rartal*	<ag source th>
=> *nu* – APPL – *rartal*	$<ag_i \; source_i \; th>$
=>SLASH:RR(24)-*nu*-APPL-*rartal*	$<ag_{ij} \; source_{ij} \; th>$

The coindexed arguments are mapped onto one grammatical relation as outlined in Alsina (1996: 115ff). In principle, two different coindexations would be possible, one which coindexes the agent with the source and another which coindexes it with the theme, as illustrated in (35).

(35) SLASH:RR(24)-*(nu)*-APPL-*rartal*

	$<ag_i$	$source_i$	th>	or	$<ag_i$	source	$th_i>$
IC:	[-o]	[-r]	[+o]		[-o]	[-r]	[+o]
Default:			[+r]				
		S	O_θ				

Following the linking principles as put forth in the previous section, the source must be assigned [-r], which therefore requires the theme to be linked to [+o]. This then makes it impossible for the agent and theme to be coindexed and thus it follows that it is the source that enters into the reflexive/reciprocal relationship in (32).

In contrast to the productive use of the RR classifier stem in (32), (33) involves a lexicalized combination of classifier and lexical stem in which there is only a single argument slot. (36a) shows the argument structure of (33a). (36b) is the argument structure of (33b) which is applicativized and thus a source is added. Adding *–nu* then coindexes the source and the experiencer (36c).

(36) a. POKE:RR(21)-*nham*, 'fear' < ex >
b. POKE:RR(21)-APPL-*nham* < ex source >
c. POKE:RR(21)-*nu*-APPL-*nham* $< ex_i$ $source_i >$

To sum up, the RR classifier stems either coindex two thematic roles of the (possibly complex) stem they combine with or form an intransitive lexicalized combination with a lexical stem. The RR marker *–nu* on the other hand always coindexes two thematic roles. This difference explains the different behaviour of classifier plus lexical stem combinations when they are combined with an applicative, presuming the morphological processes proceed from right to left.

6 Conclusion

In this paper we have provided an analysis of Murrinh-Patha verbs as morphological complex predicates, in which the argument structure of the whole is composed of different types of information coming from each of the component parts. We have argued that the classifier stems determine the *number* of arguments a verbal complex takes while the lexical stems contribute the *thematic* information for those arguments. We further extended our analysis to the interaction of argument structure composition with valency-changing processes such as applicativization and reflexivization/ reciprocalization, thus bringing new data into the discussion of complex

predicate formation. This work thus adds a new perspective to the different levels of argument structure and thematic structure and their interplay, and extends the discussion of complex predicate formation into a new typological domain.

References

Alsina, Alex. 1996. The Role of Argument Structure in Grammar. Stanford: CSLI Publications.

Alsina, Alex. 1997. Causatives in Bantu and Romance. In Alex Alsina, Joan Bresnan and Peter Sells (eds), *Complex Predicates*, pages 203-246, Stanford: CSLI Publications.

Alsina, Alex and Sam Mchombo. 1993. Object Asymmetries and the Chicheŵa Applicative Construction. In Mchombo, Sam (ed), *Theoretical Aspects of Bantu Grammar,* pages 1-44, Stanford: CSLI Publications.

Alsina, Alex, Joan Bresnan and Peter Sells (eds). 1997. *Complex Predicates*. Stanford: CSLI Publications.

Andrews, Avery D. and Christopher D. Manning. 1999. *Complex Predicates and Information Spreading in LFG*. Stanford: CSLI Publications.

Ball, Douglas. 2004. Pseudo-Noun Incorporation and Argument Structure in Niuean. In M. Butt and T.H. King (eds), *LFG Online-Proceedings of the 2004 Lexical-Functional Grammar Conference, Christchurch,* Stanford: CLSI Publications.

Blythe, Joe. 2009. *Doing Referring in Murriny Patha conversation*. PhD thesis, University of Sydney.

Bowern, Claire. 2004. *Bardi Verbal Morphology in Historical Perspective*. Ph.D. Thesis, Harvard University.

Bresnan, Joan and Annie Zaenen. 1990. Deep Unaccusativity in LFG. In Katazyna Dziwirek, Patrick Farrell and Errapel Mejias-Bikandi (eds), *Grammatical Relations: A Cross-Theoretical Perspective*, pages 45-57, Stanford: CSLI Publications.

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

Duncan, Lachlan. 2007. Analytic Noun Incorporation in Chuj and K'ichee' Mayan. In M. Butt and T.H. King (eds), *LFG Online-Proceedings of the 2007 Lexical-Functional Grammar Conference*, Stanford: CSLI Publications.

Hopper, Paul J. and Sandra Thompson. 1980. Transitivity in Grammar and Discourse. *Language* 65: 251-299.

Kemmer, Suzanne. 1993. *The Middle Voice*. Amsterdam: John Benjamins.

Kiparsky, Paul. 1985. *Morpohology and Grammatical Relations*, unpublished ms., Stanford University, Stanford.

McGregor, William B. 2002. *Verb Classification in Australian Languages*. Berlin: Mouton de Gruyter.

Mohanan, Tara. 1994. *Argument Structure in Hindi*. Stanford: CSLI Publications.

Mohanan, Tara. 1995. Wordhood and Lexicality: Noun Incorporation in

Hindi. *Natural Language and Linguistic Theory* 13: 75-134.
Mohanan, Tara. 1997. Multidimensionality of Representation: NV Complex Predicates in Hindi. In Alex Alsina, Joan Bresnan and Peter Sells (eds), *Complex Predicates*, pages 431-472, Stanford: CSLI Publications.
Nordlinger, Rachel. 2008. Reciprocals in Murrinh-Patha. Under review.
Nordlinger, Rachel. 2009. Body Part Applicatives in Murrinh-Patha (Australia). Paper presented at SOAS, 17th Novermber 2009.
Nordlinger, Rachel. 2010a. Agreement Mismatches 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. Verbal Morphology in Murrinh-Patha: Evidence for Templates. *Morphology* 20(2). To appear.
Reid, Nick. 2000. Complex Verb Collocations in Ngan'gityemerri: A Non-derivational Strategy for Encoding Valency Alternations. In R.M.W. Dixon and Alexandra Y. Aikhenvald (eds.), *Changing Valency: Case Studies in Transitivity*, pages 333–359, Cambridge/Melbourne: Cambridge University Press.
Schultze-Berndt, Eva. 2000. *Simple and Complex Predicates in Jaminjung*. Nijmegen: University of Nijmegen.
Steinbach, Markus. 2002. *Middle Voice: A Comparative Study in the Syntax-Semantics Interface of German*. Amsterdam: John Benjamins.
Street, Chester. 1987. *An Introduction to the Language and Culture of the Murrinh-Patha*. Darwin: SIL.
Street, Chester. 1996. Tense, Aspect and Mood in Murrinh-Patha. In William McGregor (ed), *Studies in Kimberley Languages in Honour of Howard Coate*. München: Lincolm Europe.
Street, Chester and Lyn Chester. 1989. Murrinh-Patha vocabulary. Electronic MS Word file.
Van Valin, Robert D. and Randy R. LaPolla. 1997. *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.
Walsh, Michael. 1976. *The Murinypata Language of North-west Australia*. PhD thesis, ANU.
Walsh, Michael. 1996. Body Parts in Murrinh-Patha: Incorporation, Grammar and Metaphor. In Hilary Chappell and William McGregor (eds), *The Grammar of Inalienability*, pages 327-380, Berlin: Mouton de Gruyter.
Walsh, Michael. 1997. Noun Classes, Nominal Classification and Generics in Murrinhpatha. In Harvey, Mark and Nicholas Reid (eds), *Nominal Classification in Aboriginal Australia,* pages 255-292, Amsterdam: John Benjamins.
Wilson, Stephen. 1999. *Coverbs and Complex Predicates in Wagiman*. Stanford: CSLI Publications.

RELATIONAL-REALIZATIONAL SYNTAX: AN ARCHITECTURE FOR SPECIFYING AND LEARNING MORPHOSYNTACTIC DESCRIPTIONS

Reut Tsarfaty
Uppsala University

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

This paper presents a novel architecture for specifying rich morphosyntactic representations and learning the associated grammars from annotated data. The key idea underlying the architecture is the application of the traditional notion of a "paradigm" to the syntactic domain. N-place predicates associated with paradigm cells are viewed as relational networks that are realized recursively by combining and ordering cells from other paradigms. The complete morphosyntactic representation of a sentence is then viewed as a nested integrated structure interleaving function and form by means of realization rules. This architecture, called *Relational-Realizational*, has a simple instantiation as a generative probabilistic model of which parameters can be statistically learned from treebank data. An application of this model to Hebrew allows for accurate description of word-order and argument marking patterns familiar from Semitic traditional grammars. The associated treebank grammar can be used for statistical parsing and is shown to improve state-of-the-art parsing results for Hebrew. The availability of a simple, formal, robust, implementable and statistically interpretable working model opens new horizons in computational linguistics — at least in principle, we should now be able to quantify typological trends which have so far been stated informally or only tacitly reflected in corpus statistics.

1 Introduction

Precision grammars and treebank grammars present two alternatives for obtaining an accurate, consistent and maximally complete syntactic analysis of natural language sentences. Precision grammars are developed by trained linguists who seek to encode their observations and generalizations as formal statements and constraints, in order to objectively describe natural language phenomena and formulate predictions in terms of inductive hypotheses. Treebank grammars are developed in engineering-oriented natural language processing environments to satisfy the need of technological applications to access an abstract representation of sentences and pass it on to downstream modules for further (e.g., semantic) processing.

For a long time these two research endeavors have been conducted in separate communities and optimized for disparate goals. The development of precision grammars emphasizes stating explicitly how surface expressions map to abstract grammatical functions. A formal framework such as Lexical Functional Grammar (LFG), for instance, has been used to articulate theories on how surface forms are mapped to feature structures via an imperfect correspondence function in different languages (Bresnan, 2000). This is intended to reveal rules and regularities and to provide a realistic way to approach the study of linguistic *universals*. Treebank

[†]The research reported in this article has been conducted at the Institute for Logic, Language and Computation at the University of Amsterdam, and funded by the Dutch Science Foundation (NWO), grant number 017.001.271. The author is currently a researcher at Uppsala University, funded by the project 'Syntactic Analysis for Synthetic Languages'. The author is indebted to the participants of LFG 2010 for stimulating discussion and to Joakim Nivre for insightful comments on earlier drafts.

grammars (Charniak, 1996) are developed so that they are easy to acquire and robust in the face of noise. The development of treebank grammars is often sensitive to annotation idiosyncrasies and more often than not it does not reflect an articulated theory. The resulting analysis is useful in as much as it helps to recover predicate argument relations. From a downstream application point of view, no attention is required to the kind of formal means used to realize these relations.

Recently, these somewhat disparate research efforts started to acknowledge their usefulness for one another. On the one hand, augmenting simple treebank grammars with surfacy linguistic notions such as the position of the head (Collins, 2003) or topological fields (Kübler, 2008) was shown to improve the precision of these treebank grammars. On the other hand, acquiring deep, precision grammars that represent multiple layers of linguistic description from treebank data, e.g., by Hockenmaier and Steedman (2002), Miyao and Tsujii (2008), and Cahill et al. (2008), helped to improve the coverage of such deep grammars and increase their robustness in the face of noise. Notwithstanding, through these and other research efforts it has also become apparent that borrowing terms, constructs and techniques from one research vein and applying them to another may be too simplistic an approach for specifying and statistically learning complex linguistic phenomena.

A case in point is the use of *morphology* in parsing. Treebank grammars developed for configurational languages such as English do not always carry over to less configurational languages effectively because existing models assign probabilities based on configurational positions. Rich morphosyntactic interactions which are orthogonal to syntactic configurations do not get assigned their own probability mass. This entails that the competition between morphology and syntax cannot be fully materialized when such grammars are used, e.g., for parsing (Tsarfaty et al., 2010). On the other hand, a precision grammar such as LFG exploits a wide-range of dependencies in a parallel architecture which makes it challenging to assign a probabilistic interpretation to them. A probabilistic interpretation requires the explicit specification of correlations in the form of conditional independence. Because of the huge space of possible morphosyntactic combinations, trying to learn all the possible options for the integration of the different levels often leads to an explosion of the space of parameters, which in turns leads to extreme sparseness.

This paper takes a step back to consider the task from first principles and develops a novel architecture which remains faithful to both kinds of goals. The proposed architecture, called *Relational-Realizational*, is adequate for economically describing rich morphosyntactic interactions, and, at the same time, it can be used to define an interpretable grammar that can be read off of treebank data and may be used for efficient parsing. The key idea underlying the architecture is the proposal to apply the traditional notion of a "paradigm" to the syntactic domain. Syntactic constituents and their associated features are related to one another in the same way that the set of inflected word forms of a lexeme defines a paradigm. The feature combinations define different cells in the inflectional class of the paradigms, and the N-place predicate associated with paradigms are viewed as relational networks that are realized recursively by combining and ordering cells from other paradigms.

In this paper we show that viewing complex morphosyntactic representation of a sentence as a nested paradigmatic structure allows us to describe profound linguistic facts concerning Modern Hebrew morphosyntax and at the same time instantiates a generative probabilistic model that improves parsing results for Hebrew. The parameters that are learned from the data can be interpreted as capturing different dimensions of realization, and the parameter tables read off of the treebank can be shown to quantify observations about argument marking which have traditionally been stated qualitatively. While this modeling strategy shares a lot of underlying assumptions with LFG, its integrative and realizational nature opens new horizons in the attempt to marry the theoretical and the statistical approaches.

The remainder of this paper is organized as follows. Section 2 outlines the proposal to extend the paradigmatic view from the morphological to the syntactic domain, and shows how appropriate independence assumptions turn it into a generative probabilistic model that can be effectively used for statistical parsing. Section 3 demonstrates how this architecture can be used to describe argument marking patterns in Modern Hebrew. Here we focus on word-order and differential object-marking, but the same methodology carries over to other morphosyntactic phenomena such as agreement. Section 4 summarizes the results of a quantitative evaluation of the resulting treebank grammars and in section 5 we summarize the contribution and conclude.

2 The Model: Relational-Realizational

For a statistical model to meet the challenge of linguistic adequacy it is ultimately required to learn how abstract grammatical functions, such as *subject, object, past tense* or *grammatical gender*, are manifested through a range of language-specific forms, such as *word position*, *affixes*, *phrase-level agreement*, and so on. The view that *form* and *function* in natural language are independent from one another has been wide spread in typological studies (cf. Sapir (1921)), and is a fundamental principle motivating the projection architecture in LFG, where c-structures are mapped to f-structures through an *imperfect* correspondence function. On top of that, the idea that form and function are independent has been mastered and extensively utilized in theoretical morphology, where form-function separation guides the descriptions of the ways morphosyntactic representations map to words (Anderson (1992); Aronoff (1994); Blevins (2010); Matthews (1974); Stump (2001)).

Let us illustrate form-function independence in morphology. Consider, for instance, the realization of the grammatical property [+plural] in English. The property [+plural] in English is expressed in a variety of forms, such as ‘kids’, ‘children’, ‘men’, ‘sheep’, ‘oxen’, and so on. It falls out of this variation that the morphological exponent ‘s’ is not a necessary condition for the realization of [+plural]. At the same time, the exponent ‘s’ associated with English [+plural] expresses also the present-tense third-person singular property-bundle in the morphology of verbs (as in ‘eats’), so ‘s’ is not even sufficient for determining [+plural] in English.

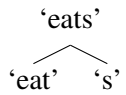


Figure 1: Morpheme-based morphology

Different morphological models approach differently the fundamental question how form-function association are stored, and how form-function relations between property-bundles and surface word forms are being computed. Stump (2001) singles out paradigmatic, realizational models as adequate for modeling complex form-function correspondence patterns in morphology. This paper articulates a proposal to extend this modeling strategy from the morphological to the syntactic domain. Here, instead of looking at grammatical *properties* such as *gender, number* etc., we consider grammatical *relations* such as *subject of*, *object of* and so on. Instead of analyzing property-bundles, we look at sets of relations that define complete argument structures. The result of this exercise leads to a paradigmatic and realizational model which is adequate for describing many-to-many, form-function correspondence patterns between sets of grammatical relations and surface syntactic structures which may be intertwined with complex morphology.

2.1 Models for Morphology

Morphologists seek to describe exponence relations — associations between properties of words and surface formatives — of several kinds. A **Simple Exponence** relation is a one-to-one relation between a property and a formative (such as [z] in ‘seas’ or [d] in ‘sailed’). This clear association between formatives and properties is often the case in radically agglutinating languages, such as Turkish. **Cumulative Exponence** is a one-to-many relation common in fusional languages such as Latin, where a single ending may realize, e.g., number/case feature combinations. Many-to-one relations are called **Extended Exponence**, where the joint contribution of different formatives is required for expressing a single property. An example is the Greek verb *e-le-ly-k-e-te* where the perfective is marked by at least three forms: ‘le’, ‘y’ , and ‘-te’ interleaved with other exponents (Matthews, 1974, p. 180).

In theoretical morphology, there exist at least two different ways to approach such descriptions. In the American structuralist tradition (Bloomfield (1933) and followers), a word form like ‘eats’ is seen as a combination of two different forms, a root ‘eat’ and a suffix ‘s’. These forms are defined to be *morphemes* — minimal meaningful units in the language linking sound to meaning in the Saussurian sense. The functional characterization of the word ‘eats’ is then derived from combining the functions of the parts, and its form is the result of their concatenation. This is the *Morpheme-Based (MB)* view of morphology, illustrated in figure 1.

A different way to analyze the word form ‘eats’ is to view the set {‘eat’, ‘eats’, ‘ate’, ‘eaten’} as associated with a single lexical entry, a *lexeme* EAT. The word

/EAT/	1Sing	2Sing	3Sing	1Pl	2Pl	3Pl
Past	1SingPast	2SingPast	3SingPast	1PlPast	2PlPast	3PlPast
Present	1SingPres	2SingPres	3SingPres	1PlPres	2PlPres	3PlPres
Perfect	1SingPerf	2SingPerf	3SingPerf	1PlPerf	2PlPerf	3PlPerf

/EAT/	,	/EAT/	,	/EAT/	,	/EAT/
+1SingPast		+3SingPast		+1SingPres		+3SingPres
|		|		|		|
'ate'		'ate'		'eats'		'eat'

Figure 2: Word and Paradigm Morphology

forms {'eat', 'eats', 'ate', 'eaten'} then realize different abstract descriptions associated with the lexeme; 'the past form of EAT', 'the present tense third-person singular form of EAT', etc. Under this view, this set is a concrete realization of a *paradigm* that has cells that associate a lexeme with different combinations of morphosyntactic representations of the feature values of *tense, gender* and *number*. This is the *Word-Based (WB)* view of morphology.

The graphical depiction of this alternative strategy is shown in figure 2. The paradigm defines different content cells, i.e., the different functions that a word-form associated with this lexeme may fill up in the syntactic structure. The mechanism used to manipulate the set of properties (e.g., syntactic requirements) is assumed to be distinct of the mechanisms that construct forms (e.g., a finite-state machine). *Morphological paradigms* define how the different word-forms are related, and *realization rules* define the mapping between well-formed property-bundles defining paradigm cells and the appropriate word-forms that make up the paradigm.

Stump (2001) isolates two modeling assumptions that jointly characterize the differences between these two opposing views. Distinguishing *lexical* and *inferential* models is concerned with how the association of properties to exponents is stored in the grammar, and the distinction between *incremental* and *realizational* approaches is about the ways multiple properties get associated with word-forms.

- ***Lexical* vs. *Inferential* Approaches** In *lexical* approaches, form is primary to function. Forms are listed in a lexicon where they are associated directly with discrete functions. In *inferential* approaches, functions are primary, and the model explicitly computes the associated forms in the course of analysis.

- ***Incremental* vs. *Realizational* Approaches** In *incremental* models, properties are accumulated incrementally. In *realizational* ones, complete property-bundles are the precondition for, rather than the outcome of, the application of spell-out rules. In *incremental* models, words are artifacts. In *realizational* ones, words have an independent formal status beyond the combination of parts, and they are related to one another through the notion of a paradigm.

Morpheme-based (MB) approaches, as alluded to above, are *lexical* and *incremental.* They constitute the simplest, most intuitive way to view morphology. Describing morphological patterns involving formatives that go beyond simple exponence, however, may become cumbersome with MB approaches. An example comes from null realization. In the case of "sheep", for instance, it is necessary for *lexical-incremental* approaches to stipulate an empty formative and associate it with a [+plural] property. Empirical evidence for such 'empty morphemes' is hard to establish, but without it the description collapses. Another challenge faced by *lexical-incremental* models is the 'selection problem', that is, the challenge of incrementally choosing morphemes that 'go together' when generating feature combinations in the lexicon. In actuality, feature occurrences may be interdependent.

Old prescriptive grammars never face such challenges because they invoke *Word and Paradigm (W&P)* approaches which are *inferential* and *realizational*. Their modern conception, *Extended Word and Paradigm (EW&P)* approaches (Anderson, 1992; Stump, 2001) define a paradigm by means of an abstract lexeme and a set of well-formed feature-bundles a priori associated with the lexeme. The morphosyntactic representation of cells in the paradigm is primary to word forms. Well-formed property-bundles representing these cells are delivered to the morphological model (say, by the syntax), and the morphological component consists of a set of realization rules which *interpret* these property-bundles. This interpretation may specify nothing (as is the case in 'sheep' plural) or it may articulate mapping of a subset of the property to noncontiguous parts of a form, as in the Greek case.

Lexical-incremental theories then work to perfection in cases of radical agglutination, but face challenges with more complex morphology. The flexibility of combining paradigmatic associations with realization rules in *inferential-realizational* approaches, however, makes them well suited for describing exponence relations of all kinds, as seen in Anderson (1992), Aronoff (1994), Blevins (2010) and more.

2.2 Models for Syntax

It should be clear from the outset that syntactic analysis of natural language sentences needs to cope with a range of exponence relations which is as diverse as morphological exponence. By Syntactic Exponence I refer to the relationship between abstract grammatical relations and their surface manifestation. **Simple Exponence** is a one-to-one relation between abstract entities and configurations, such as the relation between an NP dominated by an S and the *subject* function, as is articulated in early versions of X-bar theory. **Cumulative Exponence** is the realization of multiple syntactic functions by means of a single syntactic formative; this happens, for instance, in structures involving *clitics*, phonologically reduced elements that indicate an independent grammatical entity in addition to the one associated with their host. **Extended Exponence** is a many-to-one relation between formatives and functions, manifested through, e.g., periphrasis, functional co-heads (such as AUX in Warlpiri) or referring to the morphology of multiple forms, e.g., in differential marking or agreement. Given the diversity of exponence relations it may be fruit-

ful to identify general principles according to which one could derive modeling strategies to describe syntactic exponence relations. This paper asks two orthogonal questions, parallel to the ones that concerned models for morphology: firstly, how the model stores form-function associations, and secondly, how complete sets of relations and properties get associated with the morphosyntactic structures that realize them.

- ***Configurational* vs. *Relational* Approaches**
  In *configurational* approaches configurations are primary to functions. Configurational pieces are used to define grammatical functions, and grammatical relations are derived notions. *Relational* approaches take grammatical relations as primary and primitive and separate them from their surface manifestation. The model then calculates form in the course of the analysis.

- ***Incremental* vs. *Realizational* Approaches**
  In *incremental* approaches, the theoretical primitives (configurational or relational) are accumulated incrementally in the course of the syntactic analysis. Argument-structure is an artifact of the combination of syntactic pieces. In *realizational* approaches, complete sets of primitives are a precondition for, rather than the outcome of, the application of syntactic realization rules. Argument-structure then has a formal status beyond the sum of its parts.

The *configurational-incremental* view is compatible with configurational languages, where associations of configurational positions and grammatical relations are very tight. In nonconfigurational languages, configurational positions do not always stand in one-to-one correspondence to grammatical relations, and morphological exponents in the syntactic structure may alter or supplement form-function associations. To effectively capture argument marking patterns that have to do with the interactions of configurations with complex morphology, we propose to extend the architectural design of the W&P approach to the syntactic domain. The model is thus required to be (i) *relational*, i.e., function is primary to form, grammatical relations are primary to complete surface structures, and (ii) *relatizational*, i.e., a complete set of functions is a pre-condition for the application of realization rules which interpret them as morphosyntactic forms. The key idea underlying the proposal is that forms of higher-level constituents in a syntactic structure define the function of lower-level constituents, and the recursive realization process goes on to unfold the surface structure. As the structure unfolds, feature-bundles become increasingly specific until they can be fed into a model of W&P morphology.

2.3 Relational-Realizational Syntax

The design of the relational-realizational architecture starts off from an assumption that has been pertinent in syntactic theory since Relational Grammars (Postal and Perlmutter, 1977), and which has also inspired the architectural design of LFG — that grammatical relations are theoretical primitives, and that the description of

S⟨PRED⟩ FEATS / ARG-ST	Affirmative	Interrogative	Imperative
intransitive	S_{affirm}+{SBJ,PRD}	S_{inter}+{SBJ,PRD}	S_{imper}+{SBJ,PRD}
transitive	S_{affirm}+{SBJ,PRD,OBJ}	S_{inter}+{SBJ,PRD,OBJ}	S_{imper}+{SBJ,PRD,OBJ}
ditransitive	$\boxed{S_{affirm}+\{SBJ,PRD,OBJ,COM\}}$	S_{inter}+{SBJ,PRD,OBJ,COM}	S_{imper}+{SBJ,PRD,OBJ,COM}

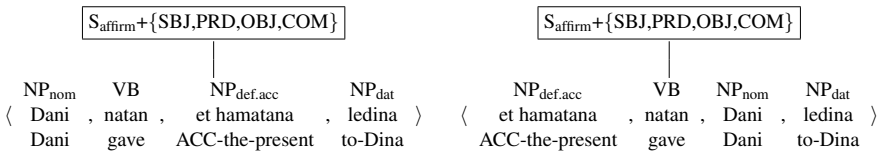


Figure 3: The Relational-Realizational Architecture

how they are realized can vary from one language to another. This is the *relational* assumption. In RGs, Relational Networks (in LFG, N-place predicates) group together sets of relations which are associated with the complete clauses or sentences that realize them. This is the *realizational* assumption. Here we aim to define a generative device that allows us to generate such function-to-form associations in an integrated representation. This generative-integrated view of syntax will then be instrumental in turning the formal description into a proper probabilistic model.

2.3.1 Paradigmatic Organization

The intuitive idea of viewing syntax in terms of paradigms goes a long way back. Pike (1963) has shown that describing the syntactic constructions in a language by means of a feature-based paradigm, as was used for phonological descriptions at that time, provides for an economic and intuitive way to compare the grammar of different languages. Matthews (1981) suggested the notion of a paradigm to capture a set of syntactic alternations that are transformationally related. Here we extend this intuitive idea, that the category S represents a syntactic paradigm, to all syntactic categories. We abstract away from transformations as a mechanism for realizing the paradigmatic alternations, later to be replaced by *realization rules*.

We propose to associate syntactic categories with a feature-geometry that defines the functions that phrases of different types may fill. A syntactic inflectional class associated with a lexical predicate (PRED) and a feature geometry defines a syntactic paradigm. The lexical predicate designates a set of grammatical relations, here conceived as the *initial* level of relational networks (RNs) in the sense of RG. The features defined by the inflectional class may refine the network of arguments to be realized, here conceived as the *final* level of an RN in RG. Syntactic constituents are instantiations of particular cells in syntactic paradigms, each of which realizes a set of abstract relations and properties that defines the function of this constituent. Figure 3 illusrates a paradigm associated with constituents of type S.

2.3.2 Realization Rules

A syntactic category, a property-bundle and a lexical predicate designate a cell in a syntactic paradigm. For each cell in the paradigm, we would like to specify how this overall function is realized in the syntactic structure. Focusing on the realm of morphosyntax,[1] there are at least three ways in which a grammatical function may be locally realized: by designating certain positions for the realization of a relation (e.g., in SVO languages), by delegating a property to a dominated constituent (e.g., by simple case marking), and/or by distributing properties to a set of dominated constituents that stand in a certain relation (e.g., agreement).

Realization rules are formal rules that map the abstract representation of constituents as cells in paradigms onto a sequence of partially specified cells in other paradigms. Figure 3 shows an exemplar S paradigm and realization rules that implement two possible ways in which a particular content cell in the paradigm may be realized. The sequence of labels and features specify regions in other syntactic paradigms, which in turn may be associated with their own lexical predicate and relational network, and this realization process proceeds until the paradigm cells are fully specified and may be handed over to a model of W&P morphology.

While the organization principles of categories as paradigms is assumed to be universal, the ways in which the high-level category can group, order or distribute features to other paradigm cells are language-specific. It is a property of the language, rather than a property of the formal architecture, at which level of the hierarchy (clause, phrase, word) complete morphosyntactic representations (henceforth, MSRs) are handed over to morphology. This modeling strategy maintains a unified view of morphology and syntax that cuts across the separation between form and function, and draws the distinction between them according to the nature of the realization rules that spell out the function: syntax involves *recursion* to other paradigms, morphology maps functions directly to surface forms in the lexicon.

2.3.3 Independence Assumptions

There are different ways to specify realization rules that relate content cells in paradigms to sequences of content cells of dominated forms. This work proposes an approach prominent in morphology – to identify abstract units that can be combined to form new hypotheses and retain generalizations about how surface forms come about. These units need not be minimal Saussarian signs, but rather, they are different aspects of the complete form-function association. In order to identify these different aspects we isolate different dimensions of description for each local constituent, and point to independence assumptions between these dimensions. Such independence assumptions will lead straight forwardly to a probabilistic interpretation of the generative model. We articulate two independence assumptions for each paradigm cell: (i) the independence of form and function, and (ii) independence of different dimensions of realization.

[1] For the time being, we are discarding other means of realization such as intonation, prosody, etc.

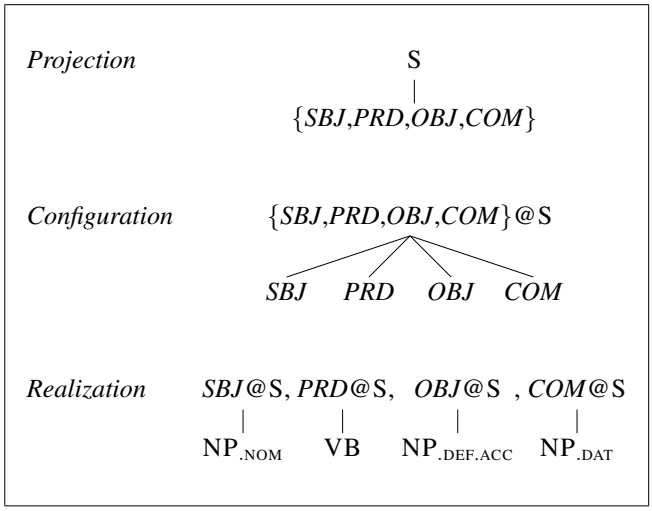


Figure 4: The Relational-Realizational (RR) backbone

These independence assumptions give rise to three generative phases for each rule, termed *projection, configuration* and *realization*, illustrated in figure 4.

The Projection Phase. The goal of the first phase in the realizational cycle is to pick out a content cell in the paradigm that specifies the function of a constituent. Let us assume a syntactic paradigm associated with a lexical predicate, a relational network, and a property-bundle. In the projection phase, a morphosyntactic representation of the syntactic category and the lexical predicate projects the set of grammatical relations that represents the set of arguments to be realized.

The Configuration Phase. Having picked out a cell in a syntactic paradigm, the remaining challenge is to spell out how it is realized. The configuration phase determines the ordering and juxtaposition of *relational slots*; these are slots in which different grammatical relations are realized. The configuration phase is at the same level of abstraction as the basic definition of word-order parameters in Greenberg (1963) – the order defines functions, not types of constituents. Furthermore, the realization of grammatical relations may be supplemented by additional elements such as auxiliaries or function words, co-heads, punctuation marks, and even slots reserved for modification or adjunction. So the configuration phase may reserve additional realizational slots spread out in between the relational labels.

The Realization Phase. The configuration phase allocated slots in which grammatical relations are to be realized as lower-level constituents. In order to realize these relations, we need to specify the syntactic category of the dominated constituents, and the features which are required to be marked within the scope of these constituents for realizing their function in the overall structure. Each of the relational and realizational slots is assigned a complete morphosyntactic representation that isolates a region in another syntactic paradigm.

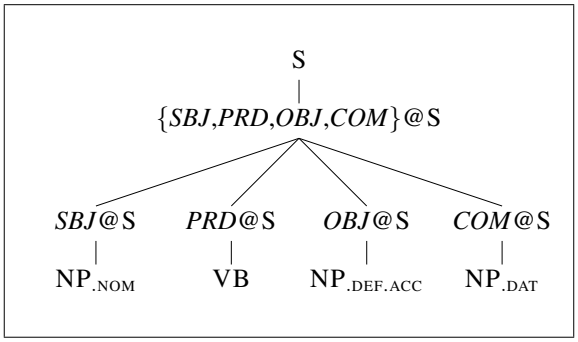


Figure 5: The Relational-Realizational (RR) representation

The properties delegated to the morphosyntactic representation may later be realized periphrastically (as a part of the configuration of a lower-level constituent) or morphologically (by delegating features to the morphosyntactic description of dominated constituents). As the structure unfolds, the morphosyntactic representation of constituents becomes increasingly specific until fully-specified descriptions may be handed over to the morphological component of the grammar for spell-out.

The Formal Model The resulting representational format of these three phases is a *linearly-ordered labeled tree* as depicted in figure 5. The complex labels of non-terminal nodes represent three distinct kinds of concepts: (i) grammatical relation labels (GRs) (ii) sets of grammatical relations, and (iii) sequences of morphosyntactic representations (MSRs) of constituents (marked here in indexed upper-case letters designating P(arent) and C(hildren)). We can identify in such trees context-free rules that correspond to the *projection*, *configuration* and *realization* phases which jointly spell-out syntactic constituents. The result of this process delivers MSRs to the next level of constituents. Morphological spell-out finally maps MSRs of pre-terminal constituents to surface forms.

(1)
- **Projection**
  $P \rightarrow \{GR_i\}_{i=1}^{n}$
- **Configuration**
  $\{GR_i\}_{i=1}^{n}@P \rightarrow \langle..[GR_{i-1} : GR_i], GR_i, [GR_i : GR_{i+1}]..\rangle$
- **Realization**
  - for Relational Slots
    $\{GR_i@P \rightarrow C_i\}_{i=1}^{n}$
  - for Realizational Slots
    $\{GR_{i-1} : GR_i@P \rightarrow C_{i_{1_i}} \ldots C_{i_{m_i}}\}_{i=1}^{n+1}$
- **Spell-Out**
  $C \rightarrow s$

The Probabilistic Model Assuming the formal generative device just described, we can define probability distributions by choosing dependencies between phases of generation. Here each phase is conditionally dependent on the previous one.

(2)
- The **Projection** Distribution
  $\mathbf{P}_{\text{projection}}(\{GR_i\}_{i=1}^{n} \mid P)$
- The **Configuration** Distribution
  $\mathbf{P}_{\text{configuration}}(\langle..[GR_{i-1}{:}GR_i], GR_i, [GR_i{:}GR_{i+1}]..\rangle \mid \{GR_i\}_{i=1}^{n}@P)$
- The **Realization** Distribution
  - for Relational Slots
    $\mathbf{P}_{\text{realization}}(C_i \mid GR_i@P)$
  - for Realizational Slots
    $\mathbf{P}_{\text{realization}}(C_{i_{1_i}} \ldots C_{i_{m_i}} \mid GR_{i-1}{:}GR_i@P)$
- The **Spell-out** Distribution
  $\mathbf{P}_{\text{spellout}}(s \mid C)$

Because of the local independence between these parameters, the probability distributions can be estimated using relative frequency estimates that maximize the likelihood of the data. The estimated RR-PCFG can be used for efficient exhaustive parsing. Because conditional dependencies allow us to specify systematic relations between form and function, the model can also be used for describing consistent and complete argument marking patterns, as we do next.

3 The Application: Modern Hebrew Morphosyntax

The Relational-Realizational (RR) architecture defined in section 2 can be straightforwardly applied to describing morphosyntactic phenomena in the Semitic language Modern Hebrew. We show that the RR representation can capture linguistic facts about word order and argument marking in Hebrew, and that individual parameters can be used to state linguistic generalizations in a probabilistic grammar.

Word-Order and Sentence Structure Modern Hebrew is an SVO language, like English and many other languages. Its unmarked, canonical word-order pattern is *subject, verb, object* as in example (3).

(3) a. דני נתן את המתנה לדינה.
dani natan et hamatana ledina.
Dani gave ACC the-present to-Dina.
"Dani gave the present to Dina."

On top of this basic word-order pattern, grammatical elements may be fronted, triggering an inversion of the unmarked Subject-Verb order (called *triggered inversion (TI)* in Shlonsky and Doron (1991)) as in (4a)-(4b). TI is similar to V2

constructions in Germanic languages. A *triggered inversion* stands in contrast with *free inversion*, in which subject-verb inversion may occur independently of such fronting (Shlonsky and Doron, 1991, footnote 2). Under certain information structuring conditions, *verb-initial* sentences are also allowed (VI in Melnik (2002)). A variation in the basic word order may also occur due to, e.g., *topicalization*, in which an element is fronted without triggering Subject-Verb inversion, as in (4c).

(4) a. את המתנה נתן דני לדינה.
et hamatana natan dani ledina.
ACC the-present gave Dani to-Dina.
"Dani gave the present to Dina."

b. לדינה נתן דני את המתנה.
ledina natan dani et hamatana.
to-dina gave Dani ACC the-present.
"Dani gave the present to Dina."

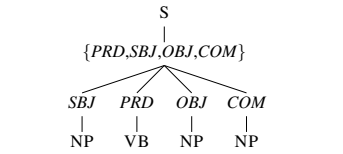
c. את המתנה, דני נתן לדינה.
et hamatana, dani natan ledina.
ACC the-present, Dani gave to-Dina.
"Dani gave the present to Dina."

The four alternative sentences in (3)–(4) only vary in their word-order pattern, due to triggering, inversion, and topicalization. The left hand side of figure 6 presents the RR representation of the structure of the different alternatives, and the right hand side shows the decomposition of the structure into generative phases, described as *parameters*. All sentences have type identical *projection* parameters and *realization* parameters, capturing the fact that their argument structure and argument marking patterns are exactly the same. The different sentences vary in the *configuration* parameters, reflecting the flexibility in the word-order pattern and additional realizational slots (e.g., punctuation). Learning these parameters from data can quantify exactly the production probabilities of the realization alternatives.

Differential Object Marking Core case marking in Hebrew displays sensitivity to the semantic properties of the phrase. This is a *differential* pattern of marking — objects in Hebrew are marked for accusativity if and only if they are also definite (Aissen, 2003). This pattern of marking is independent of the configurational positions of the different elements, as shown in (5).

(5) a. דני נתן את המתנה לדינה.
dani natan **et** **ha**matana ledina.
Dani gave **ACC** **DEF-**present DAT-Dina.
"Dani gave the present to Dina."

(3)



$\Rightarrow$

$\mathbf{P}_{\text{projection}}(\{PRD, SBJ, OBJ, COM\} \mid \text{S})$

$\boxed{(\mathbf{P}_{\text{configuration}}\langle SBJ, PRD, OBJ, COM\rangle \mid \{PRD, SBJ, OBJ, COM\}@\text{S})}$

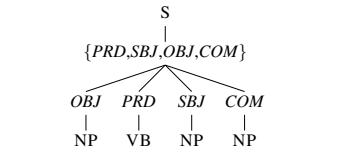
$\mathbf{P}_{\text{realization}}(\text{VB} \mid PRD@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid SBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid OBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid COM@\text{S})$

(4a)



$\Rightarrow$

$\mathbf{P}_{\text{projection}}(\{PRD, SBJ, OBJ, COM\} \mid \text{S})$

$\boxed{\mathbf{P}_{\text{configuration}}(\langle OBJ, PRD, SBJ, COM\rangle \mid \{PRD, SBJ, OBJ, COM\}@\text{S})}$

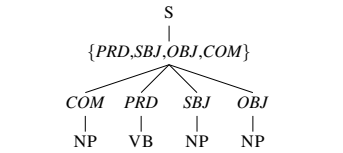
$\mathbf{P}_{\text{realization}}(\text{VB} \mid PRD@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid SBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid OBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid COM@\text{S})$

(4b)



$\Rightarrow$

$\mathbf{P}_{\text{projection}}(\{PRD, SBJ, OBJ, COM\} \mid \text{S})$

$\boxed{\mathbf{P}_{\text{configuration}}(\langle COM, PRD, SBJ, OBJ\rangle \mid \{PRD, SBJ, OBJ, COM\}@\text{S})}$

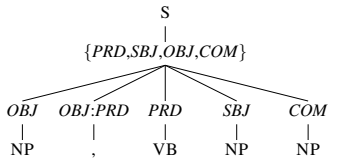
$\mathbf{P}_{\text{realization}}(\text{VB} \mid PRD@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid SBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid OBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid COM@\text{S})$

(4c)



$\Rightarrow$

$\mathbf{P}_{\text{projection}}(\{PRD, SBJ, OBJ, COM\} \mid \text{S})$

$\boxed{\mathbf{P}_{\text{configuration}}(\langle OBJ, OBJ : PRD, PRD, SBJ, COM\rangle \mid \{PRD, .., PRD\}@\text{S})}$

$\mathbf{P}_{\text{realization}}(\text{VB} \mid PRD@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid SBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid OBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid COM@\text{S})$

$\boxed{\mathbf{P}_{\text{realization}}(, \mid OBJ : PRD@\text{S})}$

Figure 6: Basic word order and sentence structure

b. את המתנה נתן דני לדינה.

et **ha**matana natan dani ledina.
ACC **DEF-present** gave Dani DAT-Dina.

"Dani gave the present to Dina."

Via a particular sort of Semitic construction, termed Construct State Noun (CSN), an object phrase may become arbitrarily long. Such CSNs are also subject to differential marking patterns, however accustivity is marked at the beginning of the CSN and definiteness is marked on the last form. This means that there can be an unbounded distance between these inter-dependent feature markers, which is again orthogonal to the configurational position of the object, as shown in (6).

(6) a. דני נתן את מתנת יום ההולדת לדינה.

dani natan [**et** matnat yom **ha**huledet] ledina.
Dani gave [ACC present-of day-of **DEF-**birth] to-Dina.

"Dani gave the birthday present to Dina."

b. את מתנת יום ההולדת נתן דני לדינה.

[**et** matnat yom **ha**huledet] natan dani ledina.
[ACC present-of day-of **DEF-**birth] gave Dani to-Dina.

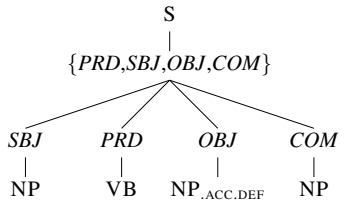
"Dani gave the birthday present to Dina."

The empirical facts are then as follows. Object marking in Hebrew requires reference to two overt markers, accusativity and definiteness. The contribution of the different markers is not independent, even though they appear on surface forms that are disjoint. This pattern of marking is orthogonal to the object position as well as to the way the object is spelled out (as a pronoun, noun, a CSN, etc.).

Let us consider sentences (5a)–(5b). The RR representation and parametrization of these constituents are presented in figure 7. Again, the difference between the parameter types lies at the configuration layer, but here we focus on the similarities. The two sentences share the *OBJ* relation parameter, $\mathbf{P}_{\text{realization}}(\text{NP} \mid \textit{OBJ}@\text{S})$. The label NP refers to an entire paradigm, but instead of NP we wish to indicate a morphosyntactic representation that isolates the functionally relevant region in the NP paradigm for an *OBJ* relation, so we specify $\mathbf{P}_{\text{realization}}(\text{NP}_{\text{DEF.ACC}} \mid \textit{OBJ}@\text{S})$.

The $\text{NP}_{.\text{DEF.ACC}}$ region then poses refined morphosyntactic requirements for this dominated constituent, regardless of its position. There are different ways in which these requirements can be filled. The $\text{NP}_{.\text{DEF.ACC}}$ MSR may be spelled out synthetically, for instance, using a pronoun marked for accusativity, gender, person, number and inherently definite; or it can be spelled out periphrastically, using the special accusative clitic את ('et') and a common noun marked for definiteness. It can also be spelled out syntactically, where the special clitic את attaches to an NP that has its own network of relations, e.g.,genitive constructions and CSNs, where the latter case is subject to a distinct feature-spreading pattern. In all cases, the isolated region in the NP paradigm makes sure that the realization is consistent and complete with respect to the delegated function.

(5a)



$\Rightarrow$ $\mathbf{P}_{\text{projection}}(\{PRD, SBJ, OBJ, COM\} \mid \text{S})$

$\boxed{\mathbf{P}_{\text{configuration}}(\langle SBJ, PRD, OBJ, COM\rangle \mid \{PRD, SBJ, OBJ, COM\}@\text{S})}$

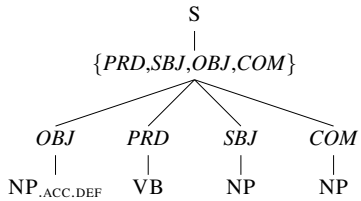
$\mathbf{P}_{\text{realization}}(\text{VB} \mid PRD@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid SBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP}_{.\text{ACC.DEF}} \mid OBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid COM@\text{S})$

(5b)



$\Rightarrow$ $\mathbf{P}_{\text{projection}}(\{PRD, SBJ, OBJ, COM\} \mid \text{S})$

$\boxed{\mathbf{P}_{\text{configuration}}(\langle OBJ, PRD, SBJ, COM\rangle \mid \{PRD, SBJ, OBJ, COM\}@\text{S})}$

$\mathbf{P}_{\text{realization}}(\text{VB} \mid PRD@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid SBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP}_{.\text{ACC.DEF}} \mid OBJ@\text{S})$

$\mathbf{P}_{\text{realization}}(\text{NP} \mid COM@\text{S})$

Figure 7: Differential Object-Marking (DOM)

	Model	*Plain*	*Head*	*Parent*	*ParentHead*
Base					
	SP-PCFG	67.61/68.77	71.01/72.48	73.56/73.79	73.44/73.61
	RR-PCFG	65.86/66.86	71.84/72.76	74.06/74.28	75.13/75.29
BaseDef					
	SP-PCFG	67.68/68.86	71.17/72.47	74.13/74.39	72.54/72.79
	RR-PCFG	66.65/67.86	73.09/74.13	74.59/74.59	76.05/76.34
BaseDefAcc					
	SP-PCFG	68.11/69.30	71.50/72.75	74.16/74.41	72.77/73.01
	RR-PCFG	67.13/68.01	73.63/74.69	74.65/74.79	**76.15/76.43**

Table 1: Differential Object Marking: F1 for sentences 1–500 in the Treebank. *Plain, Parent, Head* are syntactic splits. *Base, Def, Acc* are morphological splits.

4 Evaluation

Tsarfaty and Sima'an (2008), Tsarfaty et al. (2009) and Tsarfaty (2010) report a series of experiments that learn RR descriptions from treebank data and use them for wide coverage statistical parsing of Modern Hebrew. Here we limit the discussion to the methodological outline and to highlighting the main results. Interested readers are encouraged to follow up on the detailed analysis in the original articles.

All of the reported experiments use data from the Modern Hebrew treebank. The models are trained on sentences 500-5500 and tested on sentences 1-500. An automatic procedure is used to read off RR parameters from phrase-structure trees augmented with functional and morphological features. The paradigmatic representation of constituents uses the treebank labels' set, morphological information, and the relation labels' set {SBJ,PRD,OBJ, COM, I-COM,CNJ}. The lexical category of the predicate is percolated to each syntactic constituent in the representation. The training procedure uses simple relative frequency estimates and rare-words distribution for lexical smoothing. A general purpose CKY implementation is used for parsing, and all experiments are evaluated using Parseval on trees in canonical form (i.e., all RR-specific information is removed prior to evaluation).

Tsarfaty and Sima'an (2008) show that RR versions of the treebank grammar perform at the same level or significantly better than PCFGs that use history-based conditioning context. Moreover, morphological information is seen to contribute significant improvements for an RR treebank grammar, while leading to performance degradation with other PCFGs, as recapitulated in table 1. Tsarfaty et al. (2009) show that an RR grammar augmented with differential-object marking features significantly outperforms different versions of Head-Driven treebank grammars à la Collins (2003). The RR grammars in Tsarfaty et al. (2009) are more economic than Head-Driven ones learned for the same set of data. Both Tsarfaty and Sima'an (2008) and Tsarfaty et al. (2009) guess the PoS tags of words. Tsar-

Model		*Base*	*BaseGen*	*BaseDefAcc*	*BaseGenDefAcc*
SP-AGR	*Plain*	79.77	79.55	80.13	*80.26*
		(3942)	(7594)	(4980)	*(8933)*
RR-AGR	*Plain*	80.23	81.09	81.48	**82.64**
		(3292)	(5686)	(3772)	**(6516)**
SP-AGR	*Parent*	*83.06*	82.18	79.53	80.89
		(5914)	(10765)	(12700)	(11028)
RR-AGR	*Parent*	83.49	83.70	83.66	**84.13**
		(6688)	(10063)	(12383)	**(12497)**
SP-AGR	*Parent Head*	*76.61*	64.07	75.12	61.69
		(10081)	(16721)	(11681)	(18428)
RR-AGR	*Parent Head*	**83.40**	81.19	83.33	80.45
		(12497)	(22979)	(13828)	(24934)

Table 2: Differential Object Marking and Agreement for gold PoS tagged input. *Plain, Parent, Head* are syntactic splits. *Base, Def, Acc* are morphological splits.

faty (2010) reports parsing results for an extended set of features, including DOM features and gender agreement, when parsing gold PoS-tagged input. These results are summarized in table 2. The best result here (F_1 84.13) constitutes the best parsing result reported so far for Hebrew in the gold PoS-tags setting.

Tsarfaty (2010) finally shows that the parameter tables read off from the treebank can provide an immediate probabilistic interpretation for typological descriptions of the language. For instance, a probability distribution over production probabilities at the left of table 3 confirms the observation that Hebrew is primarily an SVO language, while allowing for word-order variation. The probability distribution over the realization of objects captures, for different types of lexical heads, the DOM pattern discussed in section 3, with a sharp distribution. Probability tables showing sharp distributions for morphological realization parameters and less sharp distributions for configuration parameters, reflect the *less-configurational* nature of Hebrew. If we are to estimate the probability distributions of RR parameters for different languages, comparing the empirical distributions we obtain may provide us with a precise way to quantify different levels of *nonconfigurationality*.

5 Conclusion

The idea presented here, viewing syntactic categories as designating paradigms and augmenting them with realization rules, provides for a powerful modeling strategy which can be developed into a complete architecture of specifying and statistically learning syntactic descriptions. The Relational-Realizational (RR) architecture developed herein is particularly adequate for languages that exhibit rich morphosyntactic interactions. The RR architecture is simple in the sense that it alternates three

Probability	*Configuration*	
1 %	□ *PRD* □ *SBJ OBJ* □	VSO
1.3%	*SBJ* □ *PRD OBJ* □	SVO
1.7%	□ *PRD OBJ SBJ* □	VOS
1.7%	□ *SBJ PRD* □ *OBJ* □	SVO
3%	*OBJ PRD SBJ* □	OVS
3.7%	□ *PRD SBJ* □ *OBJ* □	VSO
4.1%	*SBJ* □ *PRD* □ *OBJ* □	SVO
6.5%	□ *SBJ PRD OBJ* □	SVO
10.3%	*SBJ* □ *PRD OBJ* □	SVO
12.3%	□ *PRD SBJ OBJ* □	VSO
15.6%	*SBJ PRD* □ *OBJ* □	SVO
35.3%	*SBJ PRD OBJ* □	SVO

Probability	*Realization*
5.8%	$\mathrm{NP}_{\mathrm{DEF.ACC}}\langle\mathrm{PRP}\rangle$@S
6.5%	$\mathrm{NP}_{\mathrm{DEF.ACC}}\langle\mathrm{NNT}\rangle$@S
6.7%	$\mathrm{NP}_{\mathrm{DEF.ACC}}\langle\mathrm{NN}_{\mathrm{DEF}}\rangle$@S
7.4%	$\mathrm{NP}_{\mathrm{DEF.ACC}}\langle\mathrm{NNP}\rangle$@S
8.8%	$\mathrm{NP}\langle\mathrm{NNT}\rangle$@S
14.7%	$\mathrm{NP}_{\mathrm{DEF.ACC}}\langle\mathrm{NN}\rangle$@S
43.5%	$\mathrm{NP}\langle\mathrm{NN}\rangle$@S

Table 3: Word-Order and Object-Marking Parameter Tables ($\mathbf{P}(x) > 1\%$)

phases of generation for each constituent, it can be specified in a fully formal way, it is robust in the sense that it can be easily applied to treebank data, and it can be used to automatically learn treebank grammars for efficient and accurate parsing. The probabilistic parameters learned by the model are easily interpretable as indicating morphological and structural dimensions of typological variation, which can potentially be developed into empirical measures of the level of nonconfigurationality of different languages. At the same time, the models presented here use a particular set of independence assumptions which conceptualizes morphology as orthogonal to positions. Future versions will also explore different assumptions, for learning rules that spell-out the morphological and syntactic realization jointly.

References

Aissen, Judith. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* 49, 435–483.

Anderson, Stephen R. 1992. *A-Morphous Morphology*. Cambridge.

Aronoff, Mark. 1994. *Morphology By Itself*. Cambridge: The MIT Press.

Blevins, James P. 2010. *Word and Paradigm Morphology*. Oxford University Press.

Bloomfield, Leonard. 1933. *Language*. Holt, Rinehart and Winston Inc.

Bresnan, Joan. 2000. *Lexical-Functional Syntax*. Blackwell.

Cahill, Aoife, Burke, Michael, O'Donovan, Ruth, Riezler, Stefan, van Genabith, Josef and Way, Andy. 2008. Wide-Coverage Deep Statistical Parsing using Automatic Dependency Structure Annotation. *Computational Linguistics* 34(1).

Charniak, Eugene. 1996. Tree-Bank Grammars. In *AAAI/IAAI, Vol. 2*.

Collins, Michael. 2003. Head-Driven Statistical Models for Natural Language Parsing. *Computational Linguistics* 29(4).

Greenberg, Joseph H. 1963. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph H. Greenberg (ed.), *Universals of Language*, pages 73–113, MIT Press.

Hockenmaier, Julia and Steedman, Mark. 2002. Acquiring Compact Lexicalized Grammars from a Cleaner Treebank. In *Proceedings of LREC*.

Kübler, Sandra. 2008. The PaGe Shared Task on Parsing German. In *Proceedings of the ACL Workshop on Parsing German*.

Matthews, Peter H. 1974. *Morphology*. Cambridge University Press.

Matthews, Peter H. 1981. *Syntax*. Cambridge University Press.

Melnik, Nurit. 2002. Verb-Initial Constructions in Modern Hebrew. Ph.D. Thesis. University of Califiornia at Berkeley.

Miyao, Yusuke and Tsujii, Jun'ichi. 2008. Feature-Forest Models for Probabilistic HPSG Parsing. *Computational Linguistics* 34(1), 35–80.

Pike, Kenneth L. 1963. A Syntactic Paradigm. *Language* 39(2), 216–230.

Postal, Paul M. and Perlmutter, David M. 1977. Toward a Universal Characterization of Passivization. In *BLS 3*.

Sapir, Edward. 1921. *Language: An Introduction to the Study of Speech*. Brace and company.

Shlonsky, Ur and Doron, Edit. 1991. Verb Second in Hebrew. In Dawn Bates (ed.), *The Proceedings of the Tenth West Coast Conference on Formal Linguistics*.

Stump, Gregory T. 2001. *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge Studies in Linguistics, No. 93, Cambridge University Press.

Tsarfaty, Reut. 2010. Relational-Realizational Parsing. Ph.D. Thesis. University of Amsterdam.

Tsarfaty, Reut, Seddah, Djame, Goldberg, Yoav, Kübler, Sandra, Candito, Marie, Foster, Jenifer, Versley, Yannick, Rehbein, Ines and Tounsi, Lamia. 2010. Statistical Parsing of Morphologically Rich Languages: What, How and Whither. In *Proceedings of NA-ACL workshop on Parsing Morphologically Rich Languages*.

Tsarfaty, Reut and Sima'an, Khalil. 2008. Relational-Realizational Parsing. In *Proceedings of CoLing*.

Tsarfaty, Reut, Sima'an, Khalil and Scha, Remko. 2009. An Alternative to Head-Driven Approaches for Parsing a (Relatively) Free Word Order Language. In *Proceedings of EMNLP*.

REVERSING F-STRUCTURE REWRITING FOR GENERATION FROM MEANING REPRESENTATIONS

Sina Zarrieß and Jonas Kuhn
Universität Stuttgart

Proceedings of the LFG10 Conference

Miriam Butt and Tracy Holloway King (Editors)

2010

CSLI Publications

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

Abstract

We describe the design of an LFG-based generation system that provides a framework for empirical studies on the choice among grammatical paraphrases (i.e. syntactic alternations), as an effect of interacting soft constraints. To be able to study the relevant variation, we extend the XLE generation architecture so it no longer departs from standard f-structures, but from a more abstract level of (meaning) representation. This representation is constructed by means of XFR term-rewrite rules. We discuss the design of the meaning representation in light of the surface realisation task. In particular, we address the problem of obtaining a transfer grammar that reverses meaning construction, taking into account the generation performance.

1 Introduction

In this paper, we describe the design of an LFG-based generation system that provides a framework for studying soft constraints on grammatical paraphrases, i.e. syntactic alternations. These alternations have recently attracted interest in theoretical linguistic research, motivating models of grammar that assume statistical preferences to be guiding the use of certain linguistic constructions. By way of illustration, we cite an example from Bresnan and Ford (2010):

Given the following linguistic context in a dialogue:

(1) And I said, I want a backpack.
I told him, if you want to give me a present for Christmas ...

What is the most likely continuation of the sentence?

(2) a. ... give me a backpack.
b. ... give a backpack to me.

The alternatives in (2) illustrate the English dative alternation. Bresnan et al. (2007) show that speakers prefer one over the other construction depending on the discourse context and the discourse accessibility of the verb's arguments. For (2), the speaker in the dataset chose (2-a). This can be explained by the fact that the speakers statistically prefer first-person, pronominal, discourse-given recipients (*me*) to precede nominal, discourse-new themes (*backpack*).

Interestingly, the insight that discourse properties of referents are an informative factor in modelling linguistic preferences among grammatical variants is corroborated by computational research using generation with implemented broad-coverage grammars – where the relevant distinctions are subject to complex interactions of multiple factors and information sources. Cahill and Riester (2009) use the generator integrated in the XLE system to generate syntactic alternations (mainly word order variations) from given corpus sentences. They address the task of ranking these alternations, i.e. finding the appropriate realisation in context, by training a log-linear statistical model to replicate the actual realization choices for

corpus data from a treebank. Their experiments show that a model approximating discourse properties of the referents in a sentence improves the results of the realisation ranking model.

There are a number of lingustically interesting alternations that the work by Cahill and Riester (2009) could not study as participating in the ranking process, e.g. argument or voice alternations. F-structures are usually underspecified at the level of word order, but not at the more abstract semantic level encoding the realisation of predicate arguments. For instance, an LFG grammar would usually assign different f-structure representations to the active and passive realisation of a sentence. To be able to include these alternations, we need to extend the current XLE generation architecture so it departs from a level of representation abstracting away from syntactic alternations.

XLE supports generation from partially underspecified feature structure representations.[1] So, in principle, one could design a brand new feature representation for the intended level of abstraction. However, a level of representation normalising the relevant alternations has already been designed and related to f-structures from the ParGram LFG grammars, in the context of textual entailment and question answering tasks: Crouch and King (2006). Since our experiments are aimed at capturing interaction effects in real corpus data, it is important to achieve broad coverage of syntactic, morphological and lexical phenomena relatively fast. So, the most natural way to go is to adapt the existing representation and mapping mechanism for our purposes.

Crouch and King (2006) use the term-rewrite transfer system included in the XLE system (the "XFR system"), for mapping f-structures to flat semantic representations. Originally designed for machine translation, the system has proven highly useful from a practical point of view, since it supports rapid data-oriented engineering for various kinds of format conversion. The resulting transfer rule sets are generally very robust, since it is easy to include catch-all rules (and override them for specific data instances). It is also relatively straightforward to port an XFR transfer grammar from one ParGram grammar to another, taking advantage of the carefully controlled parallel f-structure geometry across languages.

The XFR system is unidirectional, so it cannot be reversed directly. This means that for our project of building semantics-based generation taking advantage of existing work on meaning construction, we have to address two questions: (1) what should be the design for our meaning representation (which parts of the entailment-oriented shallow semantics do we want to take over, etc.), and (2) how can the reverse mapping from the meaning representation to (a packed representation of all possible) f-structures be realized.

We introduce the task of surface realisation ranking in more detail and discuss the motivations of this work in Section 2. In Section 3, we provide a brief overview

[1]There are limits posed by theoretical results showing that the generation from underspecified features structures is undecidable in the general case (Wedekind, 1999) – but the XLE generator takes advantage of the constructive approach of Kaplan and Wedekind (2000).

of the extended generation architecture proposed in this paper. Section 4 describes the design of the meaning representation and discusses several adaptations for the surface realisation task. Finally, in Section 5, we treat the problem of obtaining a transfer grammar that reverses the meaning construction and point out its relation to generation performance.

2 Surface Realisation Ranking in the LFG Architecture

2.1 Ranking in the LFG Architecture

LFG grammars implemented in the XLE framework are generally reversible so that they can be used in parsing and generation. In both scenarios, one has to deal with disambiguation, i.e. ranking problems. Formally, the disambiguation problem amounts to the selection of the (or a) contextually appropriate analysis/realisation from a set of candidates that is characterised by underspecification in the shape of a "packed" LFG representation. In parsing, all candidate analyses share a common surface string; in generation, the candidate realisations share (a partial specification of) an underlying input representation, typically a partial f(unctional) structure. The two dual choice problems are illustrated on the left-hand side of Figure 1.

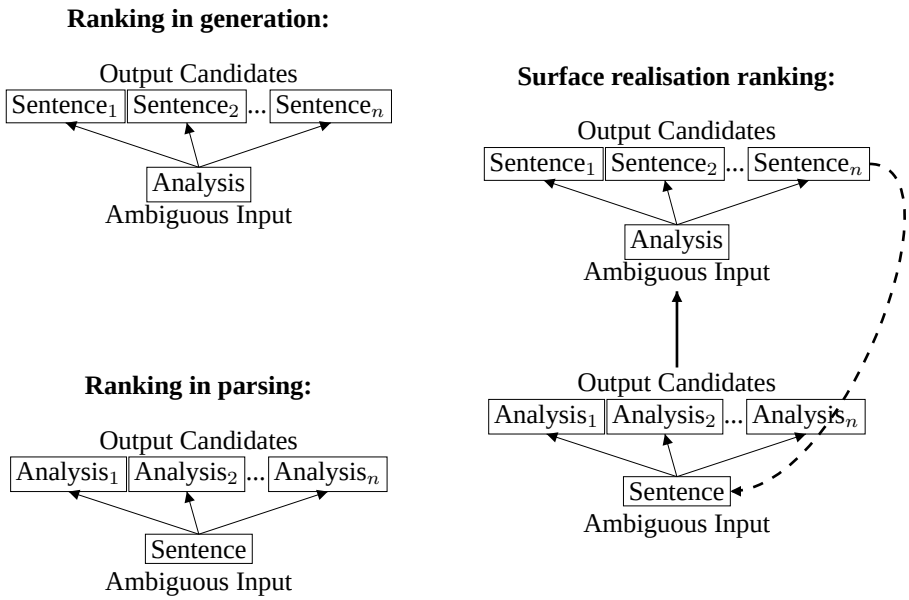


Figure 1: Ranking in a reversible grammar architecture

In both scenarios, log-linear statistical models for ranking the candidates have proven successful for modelling the preferred choice based on corpus data modelling the linguistic experience of a speaker.[2] Formally, the set-up is very similar

[2]In the log-linear ranking approach, each candidate structure is represented as a vector of its

to the Optimality Theoretic (OT) LFG architecture (Bresnan, 1996, 2000), which can be based on the same reversible framework of candidate generation (Kuhn, 2000, 2001, 2003).[3]

From a generation perspective, an LFG f-structure can be considered an abstract syntactic representation that is underspecified with respect to word order and certain aspects of morphological/lexical word choice (Cahill et al., 2007). Therefore, in mapping f-structures onto surface sentence strings, one usually obtains various possible, truth-conditionally equivalent realisations.

The quality of a mechanism choosing a particular surface realisation can be evaluated straightforwardly in a corpus-based setting. The typical design for testing a surface realisation component against realistic corpus data is displayed on the right-hand side of Figure 1 (going from the bottom to the top): First, a corpus sentence is parsed and mapped to a linguistic, underspecified analysis. Second, a generator maps this analysis to all possible surface realisations which have to be ranked by a realisation ranking model. Finally, the output of the ranking is compared against the original corpus sentence. There are multiple ways and measures to assess this comparison, such as automatic measures from Machine Translation evaluation or human judgements (Cahill, 2009).

2.2 Ranking for Free Word Order Languages

The ranking problems described in the previous section are especially challenging in languages with free word order. Consider the following example from German:

(3) Maria schenkt Thomas ein Buch.
Maria.NOM gives Thomas.DAT a book.ACC.

Sentence (3) illustrates a verb with three case-ambigous arguments, such that in parsing the sentence receives four possible analyses (*Mary* can be the subject, and the direct and indirect object, *Thomas* can also have all three functions, *book* can be the subject and the direct object). When we generate from an arbitrary f-structure for Sentence (3), we obtain the set of surface realisations in (4) which amounts to the set of all permutations of the three arguments.

(4) a. Maria schenkt ein Buch Thomas.
b. Maria schenkt Thomas ein Buch.
c. Ein Buch schenkt Thomas Maria.
d. Ein Buch schenkt Maria Thomas.

contextually relevant properties. The property weights (corresponding to the relative ranks of the constraints in an OT setting) can be discriminatively trained on corpus data using numerical optimization algorithms, which ensure that the weights for the various properties are set in such a way that the observed analyses/realisations are ranked the highest (Riezler et al., 2002; Cahill et al., 2007).

[3]The close relationship between an OT constraint ranking approach and log-linear models (which is just a different name for Maximum Entropy models) is discussed by Goldwater and Johnson (2003) and Jäger (2004).

e. Thomas schenkt ein Buch Maria.
f. Thomas schenkt Maria ein Buch.

If we were able to generate from an f-structure underspecified for voice, we would additionally obtain the surface realisations in (5) illustrating all possible permutations in passive voice (where in German only the theme argument can be turned into the passive subject).

(5) a. Maria wird von Thomas ein Buch geschenkt.
Maria.DAT is by Thomas a book.NOM given.
b. Maria wird ein Buch von Thomas geschenkt.
c. Ein Buch wird Maria von Thomas geschenkt.
d. Ein Buch wird von Thomas Maria geschenkt.
e. Von Thomas wird Maria ein Buch geschenkt.
f. Von Thomas wird ein Buch Maria geschenkt.

To our knowledge, the impact of syntactic alternations like voice on realisation ranking in free word order languages has so far not been investigated in computational frameworks working with reversible grammars. Velldal (2008) reports on HPSG-based generation experiments for English where he constrasts generation from meaning representations that are underspecified and specified for voice and topicalisation. As one would expect, the underspecified representations trigger much more (about twice as many) surface realisation candidates and the ranking task becomes much harder.

While it is difficult to compare surface realisation experiments based on different grammars and languages, one would, at least theoretically, expect that the status or function of syntactic alternations differs between languages like English and German, since German has more options available for achieving a particular ordering and hence, conveying subtle information structural differences. In English, the use of syntactic alternations (e.g. the dative alternation) is often attributed to statistical word order patterns. Bresnan et al. (2007) base their explanation of the dative alternation on the finding that "animate, pronominal, short, discourse-acessible arguments tend to precede inanimate, nonpronominal, long arguments." In German, the situation is less clear, since these precedence patterns are not constrained by the word order restrictions.

2.3 Surface Realisation and the Problem of Input Representation

Before moving on to the design of the extended generation architecture, we briefly point out an additional, independent advantage of using a more abstract shallow meaning representation instead of a standard LFG f-structure.

Grammar-based generators are a good basis for focussed studies on surface realisation (or "tactical" generation), since these systems (usually) produce grammatical output, and are actually able to produce all grammatical realisations of a given abstract input. However, an obvious limitation of grammar-based genera-

tors is that they require a very specific input representation which corresponds to the internal specification of the grammar. Depending on the system context of the surface realiser, this input representation is often hard (or almost impossible) to predict in external applications (see Section 4). As a consequence, grammar-based generators are rarely used in real-life generation applications. A further disadvantage of grammar-specific input for generation is the fact that the results obtained by different generators based on different grammars or input representations are difficult to compare (Belz et al., 2010).[4]

We extend the XLE generation set-up to take a more shallow representation as input, using an added conversion step at the beginning. This can be seen as a first step towards making the grammar-based XLE generator applicable in traditional NLG domains, like e.g. text summarisation, where the input representation can be expected to be more abstract or underspecified than fully-fledged LFG f-structures. The initial conversion step can be re-engineered fairly easily to adapt it to the relevant system context.

3 System Overview

The work presented in this paper investigates the feasibility of interfacing the XLE generator with a preprocessing step, which produces a packed underspecified f-structure representation of the f-structures compatible with a shallow meaning representation, abstracting away from morpho-syntactic alternations. As pointed out in Section 1, practical engineering considerations lead us to assume that this shallow input representation is most conveniently built by means of transfer rules, reusing a good deal of the work on meaning representations in Crouch and King (2006), a.o.

The generation architecture we propose is illustrated in Figure 2. First, an input corpus sentence is parsed and mapped to a flat semantic representation. Note that the subject of the passive f-structure is mapped to a "semantic object" in the meaning representation. In the reverse mapping from meaning representation to f-structures, the generator produces an f-structure chart that, besides the original f-structure, realises its meaning-equivalent syntactic paraphrases, e.g. voice alternations. This f-structure chart is then mapped to all its corresponding surface sentences by means of the standard XLE generator. Finally, a ranking model selects the most appropriate surface realisation.

Thus, our surface realisation testing architecture is very similar to Cahill et al. (2007). We just extend their generation pipeline by intermediate steps of further

[4]One reason for the lack of comparable tools for surface realisation is the lack of standardised resources annotated with semantic representations. Bohnet et al. (2010) present statistical generation experiments on the CoNLL'09 data which integrates semantic annotations from PropBank. However, they face the problem that this semantic annotation is far from complete, i.e. the relations between certain words are missing (e.g. adjectival modifiers). As a solution, Bohnet et al. (2010) add the missing semantic relations based on some handcrafted rules and the underlying dependency tree which results in semantic representations very similar to syntactic representations.

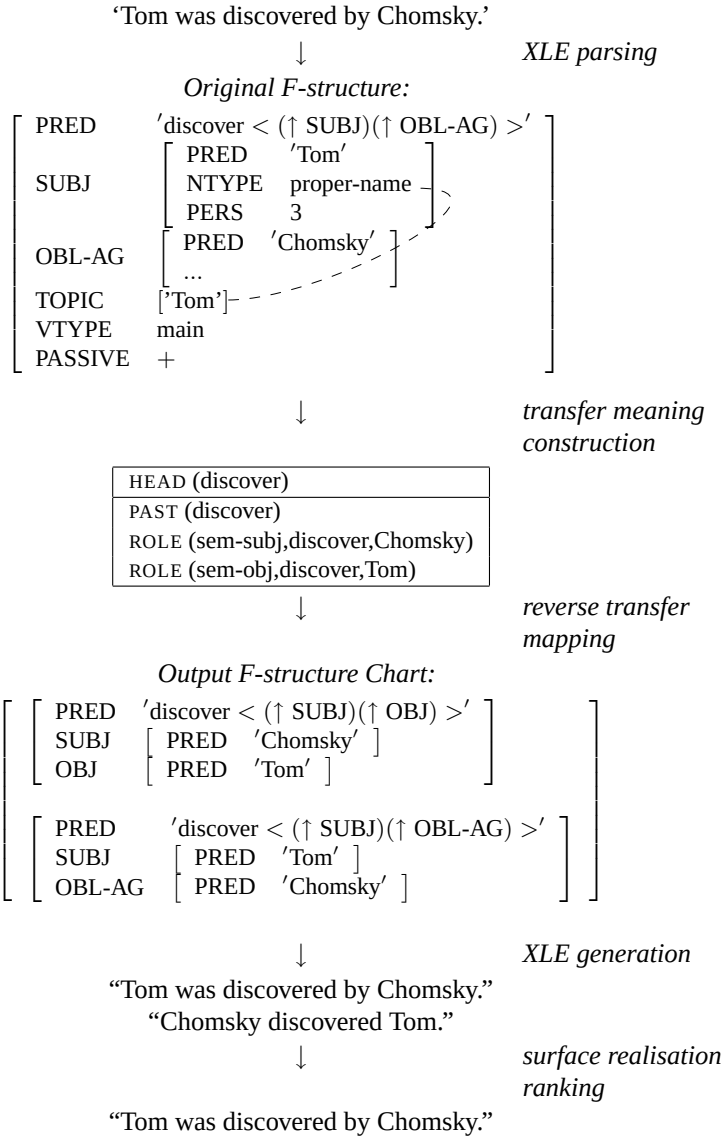


Figure 2: Generation via meaning representations

analysis, followed by generation of a broader f-structure (chart) representation, such that we do not directly regenerate from an f-structure obtained for a corpus sentence. This work focuses on the added intermediate steps in the testing pipeline, i.e. the bidirectional mapping between f-structures and a more abstract meaning representation. We leave examination of the final realisation ranking in the new setting for future work.

By using the XLE grammar-based generator in our architecture, we make sure that the final output of the generation system (if there is one) is a grammatical

sentence. However, it is important to note that in the additional generation step from semantics to f-structure (charts), the wellformedness of the transfer output is not guaranteed or checked since the transfer rules can produce arbitrary sets of f-structure terms as output. We use the grammar-based generator as a filter (similar to Crouch et al. (2004)) that only maps those f-structures to surface sentences that correspond to the definition of the underlying grammar.

4 A Meaning Representation for Surface Realisation

The standard meaning construction approach for the ParGram LFG grammars implemented in the XLE framework is the transfer semantics system developed by Crouch and King (2006). It has been ported to German by Zarrieß (2009). The system exploits the XLE transfer module to map LFG f-structures to shallow meaning representations on the basis of an ordered list of term-rewrite rules. In this section, we will discuss the design of the representation and its usefulness for generation. In the next section, we will discuss the technical aspects of reversing the rules for generation.

4.1 Normalising Paraphrases

The main purpose that a meaning representation for surface realisation serves is to normalise the analyses of truth-conditionally equivalent syntactic structures. As the result of this normalisation, syntactic alternations get assigned an identical meaning representation. In the generation step, the surface realiser will then map the meaning representation to all its possible syntactic alternations.

The semantic representation we want to generate from was originally designed for a textual entailment application (Crouch and King, 2006). To capture the entailment relation between, e.g., active and passive realisations of a verb, the representation assigns a uniform analysis to these alternations. As an example, the sentences in (6-a) and (6-b) would both be assigned the meaning representation in (6-c). The subject of the active and the oblique agent of the passive verb are both normalised to a "semantic subject". Such a normalised meaning representation is exactly what we need in generation.

(6) a. Peter saw Mary.
b. Mary was seen by Peter.
c.

HEAD `(see)`
PAST `(see)` ROLE `(sem-subj,see,Peter)` ROLE `(sem-obj,see,Mary)`

The meaning construction mechanism from Crouch and King (2006) implements a number of further normalisation operations for other types of paraphrases

or alternations that are both interesting for entailment and surface realisation. The implemented normalisations include the following:

(7) a. Attributive vs. predicative modifiers
   (i) Peter reads a good book.
   (ii) Peter reads a book that is good.
   b. Clefts
   (i) It is a book that Peter reads.
   (ii) Peter reads a book.
   c. Genitives
   (i) the building's shadow
   (ii) the shadow of the building
   d. Nominalisations vs. verbal realisations
   (i) Peter regrets the destruction of the city.
   (ii) Peter regrets that the city was destroyed.

4.2 Implicit Syntactic Information

An important aspect of the paraphrase normalisation is that many syntax-internal features are removed from the meaning representation. In practice, the f-structures that correspond to a certain pair of meaning-equivalent sentences, e.g. active and passive alternations, do not only differ in their argument frame and passive feature. The f-structures usually also specify a lot of other, e.g. morphological, features of the involved noun phrases and the verb that differ between the alternation constructions. An example alternation pair and its corresponding f-structure pair is given in Table 1.

Tom sieht Marie.
Tom sees Mary.

$$\begin{bmatrix} \text{PRED} & \text{'sehen} < (\uparrow ...)(\uparrow ...) > \text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Tom'} \\ \text{CASE} & \text{nom} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'Marie'} \\ \text{CASE} & \text{acc} \end{bmatrix} \\ \text{CHECK} & \begin{bmatrix} \text{AUX-SELECT} & \text{'haben'} \end{bmatrix} \\ \text{TOPIC} & \begin{bmatrix} \text{'Tom'} \end{bmatrix} \\ \text{PASS} & - \end{bmatrix}$$

Marie wird von Tom gesehen.
Mary is by Tom seen.

$$\begin{bmatrix} \text{PRED} & \text{'sehen} < (\uparrow ...)(\uparrow ...) > \text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Marie'} \\ \text{CASE} & \text{nom} \end{bmatrix} \\ \text{OBL-AG} & \begin{bmatrix} \text{PRED} & \text{'Tom'} \\ \text{CASE} & \text{dat} \end{bmatrix} \\ \text{CHECK} & \begin{bmatrix} \text{AUX-SELECT} & \text{'sein'} \\ \text{PARTICIPLE} & \text{'perfect'} \end{bmatrix} \\ \text{TOPIC} & \begin{bmatrix} \text{'Marie'} \end{bmatrix} \\ \text{PASS} & + \end{bmatrix}$$

Table 1: F-structure pair for passive-active alternation

For generation, it is important to remove these syntax-internal features because they would implicitly disambiguate an abstract semantic representation of an alternation. For instance, if the meaning representation would not underspecify the case of a noun phrase, the surface realiser would have implicit syntactic information about the original sentence realisation.

In the LFG ParGram grammars, many of these syntax-internal features are conventionally subsumed under the technically motivated CHECK-feature. So these can easily be detected and removed when constructing the semantic representation. In the mapping from semantics to f-structure, these features do not need to be reconstructed since the XLE generator can deal with underspecified f-structures (see Section 5.1).

However, in certain problematic cases, the f-structures for an alternation pair contain implicit features that are not syntax-internal. As an example, consider the sentence pair in Table 2. The analyses are produced by a German LFG grammar whose lexicon does not have an entry for the proper noun *Karthago*. XLE provides a "guessing" mechanism for unknown words. In this case, the German grammar has been set up to assume that unknown capitalized word forms are proper names, leaving the gender and number feature unspecified (since there are proper names for all genders and in singular and plural – like *Beatles*).

As a consequence, the f-structure for *Karthago* in the passive sentence does not have a NUM feature since the number of the noun cannot be inferred from the syntax. By contrast, the f-structure for *Karthago* in the active sentence does have a NUM feature which comes from the inflectional morphology of the verb. So the two sentences have different meaning representations (if the meaning construction takes number into account).

Such types of implicit information in the f-structure are not easy to deal with in generation. First, it is difficult in practice to foresee such problems and debug them when they occur. Second, the XLE generator is very sensitive to slight changes in the f-structure input. If the surface realiser were to add a NUM feature to the f-structure in the passive sentence in Table 2 (which may seem to be a reasonable move), the generator would fail (because the structure that the grammar assigns to the sentence is no longer subsumed by the input representation). On the other hand, one would drastically change the output of the surface realisation if the NUM feature was generally underspecified (in this case, the generator would produce the singular and plural realisation for each noun phrase in a given input f-structure).[5]

While the above type of grammar-internal, implicit information may suggest we are dealing with more of a technical than a principled problem, similar cases of indirect disambiguation of a meaning representation *do* occur in situations that are

[5]This problem with syntax-internal, atomic features has also been noted in other applications, e.g. Machine Translation. Graham (2010) reports drastically varying performance of their MT system depending on the quality of atomic feature translation. She also reports that grammar coverage of the generator varies between 12% and 41% depending only on the translation quality of the atomic features. This corroborates the aforementioned claim that grammar-based generators can be hard to use in external applications.

Rom wurde von Karthago erobert.
Rome was by Carthage conquered.

$$\begin{bmatrix} \text{PRED} & \text{'erobern} < (\uparrow \ldots)(\uparrow \ldots) > \text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Rom'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \end{bmatrix} \\ \text{OBL}_{AG} & \begin{bmatrix} \text{PRED} & \text{'Karthago'} \\ \text{PERS} & 3 \end{bmatrix} \\ \text{PASS} & + \end{bmatrix}$$

Karthago eroberte Rom.
Carthage conquered Rome.

$$\begin{bmatrix} \text{PRED} & \text{'erobern} < (\uparrow \ldots)(\uparrow \ldots) > \text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Karthago'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'Rom'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \end{bmatrix} \\ \text{PASS} & - \end{bmatrix}$$

Table 2: F-structure pair for passive-active alternation: the features for *Karthago* are asymmetric

fully motivated linguistically. These structures need to be addressed in the meaning construction. For (8-a), the normalised meaning representation (8-b) contains implicit information that its original sentence must have been realised in active voice. This is because the subject of the sentence is the generic pronoun *man* which cannot be used as an oblique agent in a prepositional phrase, i.e., (8-c) is ungrammatical. Thus, if the realiser derives an f-structure where the generic pronoun is realised as the oblique agent, the grammar-based generator rules will not produce a surface sentence for this input.

(8) a. Man hat Maria im Park gesehen.
One has Mary in the park seen.

b.

HEAD `(see)`
PAST `(see)` ROLE `(sem-subj,sehen,man)` ROLE `(sem-obj,sehen,Maria)`

c. *Maria wurde von man im Park gesehen.
Mary was by one in the park seen.

d. Maria wurde von jemandem im Park gesehen.
Mary was by somebody in the park seen.

In order to be able to generate a passive paraphrase from Sentence (8-a), the meaning representation would have to abstract away from the lexical realisation of the pronoun such that the generator could realise the subject as a different pronoun, e.g. *jemand* (*somebody*), as in (8-d). As a consequence, the surface realisation step would be extended from word order and structural choice to lexical choice, which is usually considered as a separate step of generation (Bateman and Zock, 2003).

A similar and very frequent type of implicit syntactic information occurs in coordinated sentences. For instance, in sentence (9), the noun phrase *Tom* is the

subject of two verb phrases. At the moment, the meaning representation keeps the information about the lexical identity of the two subjects in a lexical index (marked as integers in (9-b)). If the generator "knows" that the two subjects have to be realised by the same noun phrase, it cannot produce a passive paraphrase for one of the verb phrases due to syntactic constraints. However, if we interpret the representation as even more abstract and allow the realiser to generate a pronoun for *Tom* in one of the verb phrases (such as in (9-c)), we introduce a completely new type of generation problem (i.e. the generation of referring expressions) into our system.

(9) a. Tom sieht Marie und schenkt ihr einen Apfel.
Tom sees Marie and gives her an apple.

b.

HEAD (sehen)
ROLE (sem-subj,sehen,Tom:1)
ROLE (sem-obj,sehen,Marie:2)
ROLE (sem-subj,schenken,Tom:1)
ROLE (sem-obj,schenken,Apfel:3)
ROLE (recipient,schenken,sie:4)

c. Marie wurde von Tom gesehen und bekam von ihm einen Apfel geschenkt.
Marie was by Tom seen and got by him an apple given.

Finally, the type of implicit syntactic information that needs to be added or removed in paraphrase normalisation is also dependent on the complexity of the underlying alternation. For instance, the meaning representation normalises relative clauses and deverbal attributive adjuncts, such as (10-a-b). However, the non-finite verb in (10-a) does not carry any tense information whereas the finite verb in (10-b-c) does. Thus, in order to generate a relative clause paraphrase for (10-a), the meaning construction needs to include rules that infer the tense of *laughing*.

(10) a. Peter saw a laughing girl.
b. Peter saw a girl who was laughing.
c.

HEAD (see)
PAST (see)
PAST (laugh)
ROLE (sem-subj,see,Peter)
ROLE (sem-obj,see,girl)
ROLE (sem-subj,laugh,girl)

All these examples show that the boundaries between lexicalisation, grammaticalisation and surface realisation in generation get blurred rather quickly. Thus, the design decisions made at the level of meaning representation will greatly influence the difficulty and the outcome of the final surface realisation task. Moreover, we have seen that the meaning representations and the f-structures of a given alternation pair have to be carefully examined in a variety of syntactic contexts in order to produce well-formed input for the grammar-based generator.

5 Reversing Meaning Construction

This section addresses the issue of mapping meaning representations as discussed in Section 4 to f-structure chart representations from which the standard XLE generator is able to generate – given the fact that the XFR system is not directly reversible. We already mentioned the engineering advantage of re-using existing resources as much as possible – in particular in view of the multilingual setting of ParGram, which will make it relatively easy to port solutions to other languages. Hence, our approach is to develop XFR rules for the backward mapping from meaning representations to f-structures that draw upon the forward mapping rules as much as possible.[6] In Section 5.1, we show that if the meaning construction is restricted to a specific type of normalisation rules and if the generation of syntax-internal features is left to the grammar-based generator, the reverse transfer grammar can be easily derived.

A second important issue raised by our surface realisation architecture is the computational complexity and runtime performance of generation. The f-structure output produced by a reverse meaning construction is formally more complex than the f-structures that have been used in surface realisation experiments so far: whereas Cahill et al. (2007) generate from single f-structures that represent one possible analysis of a sentence, we will generate from f-structure charts which represent all the possible realisations of a syntactic alternation. Moreover, the f-structures used by Cahill et al. (2007) are almost completely specified, i.e., they contain all the syntax-internal features needed by the grammar. In our case, as already mentioned in Section 4, the f-structures will necessarily by underspecified to a certain degree since not all syntax-internal features can and should be reconstructed from the meaning representation. These properties of the f-structure input will have a noticeable effect on generation performance, which we will discuss in Section 5.2.

5.1 Transfer Rules and Bidirectionality

The XFR term rewrite system has been used in a variety of system contexts: f-structure based machine translation (Riezler and Maxwell, 2006), sentence condensation (Crouch et al., 2004), and textual entailment oriented shallow meaning construction (Crouch and King, 2006). See Crouch et al. (2004) for a detailed illustration of the XFR system.

According to Emele et al. (1996), term rewrite rules can be defined as follows:

(11) a. $\langle \textit{LHS Set} \rangle \# \langle \textit{LHS Conds} \rangle \leftrightarrow \langle \textit{RHS Set} \rangle \# \langle \textit{RHS Conds} \rangle$
b. $\langle LHS\ Set \rangle \# \langle \textit{LHS Conds} \rangle \rightarrow \langle \textit{RHS Set} \rangle$

[6] As an alternative option, one could consider a system that automatically learns the mapping between these structures, in the style of Bohnet et al. (2010). However, we feel that such a purely statistical approach ignores much of the implicit knowledge given in the forward meaning construction grammar and that it risks producing output incompatible with the XLE generator.

c. $\langle LHS\ Set \rangle \leftarrow \langle$*RHS Set*$\rangle \# \langle$*RHS Conds*$\rangle$

The most general definition in (11-a) specifies a transfer rule as a bidirectional rewrite relation between a set of left hand side terms and a set of right hand side terms. The rewrite can be conditioned on a set of terms on both sides of the rule. The corresponding unidirectional rule definitions are given in (11-b-c). In a unidirectional transfer rule, only one rule side can have rewrite conditions.

The XFR system represents an f-structure internally as a set of two-place terms.[7] By this means, one can formulate rewrite rules on f-structures that perform arbitrary lexical and structural transformations. An example rewrite rule is given in (12). The sample rule applies to f-structures that have a PASSIVE and VTYPE feature as well as an oblique agent, mapping the oblique agent to a "logical subject" (i.e., using the f-structure of active clauses as the prototypical representation).

(12) $+$*VTYPE*$(\%V, \%\%)$, $+$*PASSIVE*$(\%V, +)$, *OBL-AG*$(\%V, \%$*LogicalSUBJ*$)$
$==>$ *SUBJ*$(\%V, \%$*LogicalSUBJ*$)$.

As a unidirectional system, the XFR syntax allows conditions only on the left hand side of rules. Other transfer systems, such as Emele et al. (1996) from the Verbmobil project, implement a bidirectional syntax for rewrite rules. However, Emele et al. (1996) also mention that the implementation of a bidirectional transfer grammar is difficult in the case of large sets of rules. They report that unidirectional rules are more effective in practice since the grammar writer does not have to keep track of the bidirectional rule conditions.

In the case of meaning construction, it would presumably be even more difficult to specify bidirectional rewrite rules than for machine translation. One reason is that the meaning construction deletes a lot of syntax-internal features from the f-structure, e.g., CASE, PERS, or TOPIC (see the discussion on syntax-internal features in Section 4). An example for such a deletion rule is given in (13). The rule simply deletes every CASE feature from its input.

(13) *CASE*$(\%\%, \%\%) ==> 0$.

A bidirectional version of the deletion rule in (13) would have to be much more elaborate since it would need to specify exactly the contexts in which a CASE feature appears in an f-structure (essentially duplicating constraints from the grammar and the lexicon). Similarly, when we want to reverse unidirectional meaning construction rules at a fully general level, we cannot expect to find an automatic procedure that uses only the information in the forward rules.

[7] The term's name represents the f-structure attribute; the first argument is the f-structure under which the attribute is embedded (where f-structures are referenced by variables var(0), var(1), ..., which have a fixed reference for the full analysis); the second argument is the attribute value, either an atomic value (e.g., CASE`(var(1),acc))`, or an embedded f-structure node OBJ`(var(0),var(1))`. The rule syntax for terms to be rewritten vs. conditions is as follows: A prefixed $+$ on left hand rule side turns a term into a (positive) condition, which is not consumed during rule application. Identifiers starting with a % are variables.

Instead of deriving the formally exact reverse counterpart of the meaning construction transfer, we opt for an approximate transfer reversal. We do not need to generate full-fledged f-structures from the meaning representations because the XLE generator can handle underspecified input (Crouch et al., 2004) and will use the appropriate constraints from the grammar and lexicon to navigate the space of possibilities. By allowing the generator to add CASE features with arbitrary values, it can essentially follow the exact grammatical and lexical restrictions on this feature. We thus avoid a redundant (and presumably error-prone) duplication of this knowledge in the backward rewrite rules.

Leaving the generation of syntax-internal features to the generator, the general problem of reversing normalisation transfer rules is substantially simplified. As an example, consider the three rules (14). This is a typical rule set for normalisation: several sets of left hand terms, which correspond to meaning-equivalent syntactic structures, are mapped to an identical set of right hand terms. The normalisation rules in (14-a-b) are conditionned on the syntax-internal PASSIVE feature (in (14-a) it has to have the value $-$, in (14-b) the value $+$). After normalisation, the syntax-internal feature is deleted in (14-c).

(14) a. $+\mathit{PASSIVE}(\%V, -), \mathit{SUBJ}(\%V, \%\mathit{SUBJ})$
$==> \mathit{AGENT}(\%V, \%\mathit{SUBJ}).$
b. $+\mathit{PASSIVE}(\%V, +), \mathit{OBL\text{-}AG}(\%V, \%\mathit{SUBJ})$
$==> \mathit{AGENT}(\%V, \%\mathit{SUBJ}).$
c. $\mathit{PASSIVE}(\%\%, \%\%) ==> 0.$

Given that we do not need to reconstruct the syntax-internal features in the mapping from semantics to f-structure, one can straightforwardly derive a reverse version of the transfer rule sequence in (14), which is given in (15). The set of terms corresponding to the normalised partial meaning representation is optionally mapped to all its possible syntactic realisations (the $?=>$ operator stands for optional rewrite). The deletion rule in (14-c) and the rule conditions in (14-a-b) can be ignored.

(15) a. $\mathit{AGENT}(\%V, \%\mathit{SUBJ}) ?=> \mathit{SUBJ}(\%V, \%\mathit{SUBJ})$
b. $\mathit{AGENT}(\%V, \%\mathit{SUBJ}) ?=> \mathit{OBL\text{-}AG}(\%V, \%\mathit{SUBJ})$

Of course, in the general case, the transfer rules used for meaning construction from f-structures are not constrained to the format exemplified in (14). The grammar implemented by Crouch and King (2006) is actually far more complex and notably integrates recursive rules that rearrange the embeddings of the f-structure nodes. However, for our current work we can restrict attention to the type of simple normalisation rules, essentially a subset of the rules used by Crouch and King (2006).[8]

[8] We also implemented inspection tools for keeping track of the flow of information during term rewrite transfer, in order to isolate the relevant rules quickly.

5.2 F-Structure Charts in Transfer and Generation

Having discussed a way of constraining transfer rules for easy reversal, we show in this section that we need even stricter constraints on the transfer rules in order to keep the generation feasible with respect to performance.

In the reverse mapping from meaning construction to f-structure, nothing guarantees that we actually generate an f-structure that is within the coverage of a given LFG grammar. In our generation architecture (Figure 2), we rely on the fact that the XLE generator will select from the chart those f-structures that comply with the grammar specification. However, if the generator has to deal with f-structure charts that comprise a huge number of f-structures that cannot be generated from, it will often time out or fail.

By way of illustration, we contrast generation from an identical meaning representation based on two different reverse transfer grammars that generate active and passive alternations for transitive and ditransitive verbs.[9]

Our meaning representation here is simply an f-structure that abstracts from the voice of the verb, i.e. predicate arguments are mapped to semantic roles, and passive and verb morphology features are deleted from the f-structure. Depending on the formulation of the normalisation rules, the reverse generation rules may potentially look very different. In (16) and (17), we present excerpts from two transfer grammars that perform the same f-structure mappings in different ways. The transfer grammar in (16) incorporates a notion of argument frames: the semantic roles are not mapped to syntactic roles independent of each other. The naive reverse grammar in (17) on the other hand *does* employ an independent mapping rule for each semantic role.

(16) a. $\mathit{AGENT}(\%V, \%\mathit{Agent}), \mathit{THEME}(\%V, \%\mathit{Theme}),$
$\mathit{RECIPIENT}(\%V, \%\mathit{Recipient})$
$?{=}{>}\ \mathit{SUBJ}(\%V, \%\mathit{Agent}), \mathit{OBJ}(\%V, \%\mathit{Theme}),$
$\mathit{OBJ\text{-}TH}(\%V, \%\mathit{Recipient}).$

b. $\mathit{AGENT}(\%V, \%\mathit{Agent}), \mathit{THEME}(\%V, \%\mathit{Theme}),$
$\mathit{RECIPIENT}(\%V, \%\mathit{Recipient})$
$?{=}{>}\ \mathit{OBL\text{-}AG}(\%V, \%\mathit{Agent}), \mathit{SUBJ}(\%V, \%\mathit{Theme}),$
$\mathit{OBJ\text{-}TH}(\%V, \%\mathit{Recipient}).$

[9] Note that in German, there are two types of passive that a ditransitive verb can undergo: (1) regular passive, turning the direct object into the passive subject, and (2) *bekommen* passive, turning the indirect object into the passive subject. In the latter case, the passive is constructed with the special auxiliary *bekommen* (lit. "get"); see Example (i).

(i) a. Die Frau schenkt Maria ein Buch.
The woman.NOM gives Maria.DAT a book.ACC.

b. Ein Buch wird Maria von der Frau geschenkt.
A book.NOM is Maria.DAT by the woman given.

c. Maria bekommt ein Buch von der Frau geschenkt.
Maria.NOM gets a book.NOM by the woman given.

(17) a. *AGENT*$(\%V, \%$*Agent*$)$?=> *SUBJ*$(\%V, \%$*Agent*$)$.
b. *AGENT*$(\%V, \%Agent)$?=> *OBL-AG*$(\%V, \%AG)$
c. *THEME*$(\%V, \%$*Theme*$)$?=> *OBJ*$(\%V, \%$*Theme*$)$.
d. *THEME*$(\%V, \%$*Theme*$)$?=> *SUBJ*$(\%V, \%$*Theme*$)$.

Grammar (16) will mostly produce f-structures that are well-formed and that can be generated from, whereas grammar (17) will produce a lot of f-structures that are not compatible with LFG assumptions or specific grammatical/lexical constraints, e.g., f-structures with two subjects or without a subject. In the final surface realisation, these f-structures will not produce any surface sentence; however they substantially slow down the generation process.

For our generation experiment, we considered a set of 156 German sentences extracted from the HGC, a huge German corpus of newspaper text.[10] In Table 3, we report generation performance based on two different inputs for the surface realiser, one that was produced by means of the naive transfer rules in (17), and one that was produced by means of the lingustically informed rules in (16). The timeout parameter was set to 500 seconds. As can be seen, the generator cannot easily deal with the f-structure chart input that contains a lot of illformed structures. It times out in 30% of the cases and the average generation time is dramatically increased compared to generation from mostly well-formed input.

	# f-structures	avg. generation time (excl. timeouts)	# timeouts
Naive Rules	156	246.14 (110.68)	53
Informed Rules	156	36.20 (27.04)	3

Table 3: Generation performance depending on the transfer rules that produced the f-structure input

These results add an important aspect to the discussion about transfer grammar reversibility in Section 5.1. Even if we had a method that could automatically reverse any given transfer grammar, the f-structure charts produced by that reverse grammar would not necessarily be usable in generation experiments on actual corpus sentences.

Moreover, in Table 4, we compare the number of surface realisations that are produced in generation from meaning representations and generation from usual f-structures. In both cases, the total average of surface realisations is very high due to some very long sentences in our test set. If we compare the number of realisations sentence-wise, the picture is more realistic: In generation from meaning representations that abstract from the voice of a verb, the number of realisations increases by a factor of 2.8 on average. However, in 40% of the sentences, the number of surface realisations did not increase at all, i.e. no alternations could be

[10] All contain a ditransitive verb that instantiates its three arguments, such that it should generally be possible to generate several voice alternations. We did not include special rules for specific constructions like coordinations or generic pronouns (see Section 4), such that, in these contexts, the grammar will rule out the automatically generated alternations.

generated. This suggests that a more abstract meaning representation (as discussed in Section 4) would have a huge impact on the surface realisation output.

Avg. number of realisations for semantic input	25092.16
Avg. number of realisations for syntactic input	14168.57
Avg. increase of realisations per sentence	284%
Sentences with no increase in realisations	64
Total number of sentences	156

Table 4: Number of surface realisations produced in generation from meaning representations

6 Conclusion

In Sections 1 and 2, we outlined the two main motivations for implementing an LFG-based surface realisation system that generates syntactic alternations from meaning representations. First, this generation architecture provides a framework for studying the interplay of multiple soft constraints on the basis of complex corpus data, taking advantage of high-quality linguistic grammars that have broad coverage at the same time. Hence, a topic of great theoretical linguistic interest can be addressed from a computational perspective. Second, this work has demonstrated the usability of the grammar-based XLE generator in a setting where the (underspecified) input representation is not directly produced by the grammar, thus taking a first step towards making the generator applicable in a wider range of natural language generation domains.

In light of the discussions and experiments presented in this paper, we can conclude that our architecture is definitely suited for carrying out targeted linguistic studies of a well-delimited set of syntactic alternations. For instance, with the help of our system, it is possible to do large-scale surface realisation experiments focussing on specific phenomena, comparing them to the smaller-scale and more controlled experiments in theoretical linguistic research, e.g. by Bresnan et al. (2007). It is also possible to empirically study the complex interaction of two or three factors known to play a role in surface realisation, e.g., word order, voice and discourse status of argument phrases.

In addition, Section 4 on the design of the meaning representation showed that by doing actual surface realisation studies, it is more likely that residual issues with a particular level of abstraction chosen as the input representation will be brought to our attention. An example is the implicit exclusion of a passive realization due to a particular lexical choice for the agent argument, or the question whether or not a tense feature is included in the abstract input representation.

Concerning the second motivation, our conclusion is more cautious. In Section 4 and 5, we have seen several difficulties with the mapping between a (more or less) grammar-external meaning representation and an f-structure input that can be dealt with by the XLE generator. The main problems are that (a) the generator is

very sensitive to slight changes in the f-structure input and the underspecification mechanism does not always remedy this problem, and (b) the generator can be used to filter illformed f-structures. However, if the input contains a massive number of illformed structures, the performance decreases dramatically. In the case of well-delimited linguistic studies, both of these rather technical problems can be addressed through careful manual design of the transfer rules that map between semantics and f-structure. However, interfacing the grammar-based generator with an arbitrary semantic representation seems to require a more elaborate generation architecture.

References

Bateman, John, and Michael Zock. 2003. Natural Language Generation. In R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Belz, Anja, Mike White, Josef van Genabith, Deirdre Hogan, and Amanda Stent. 2010. Finding common ground: Towards a surface realisation shared task. In *Proceedings of the 6th International Natural Language Generation Conference (INLG'10)*.

Bohnet, Bernd, Leo Wanner, Simon Mill, and Alicia Burga. 2010. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *To appear in Proceedings of COLING-2010*, Beijing, China.

Bresnan, Joan. 1996. LFG in an OT setting: Modelling competition and economy. In M. Butt and T. H. King (eds.), *Proceedings of the First LFG Conference*, CSLI Proceedings Online.

Bresnan, Joan. 2000. Optimal syntax. In J. Dekkers, F. van der Leeuw, and J. van de Weijer (eds.), *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford University Press.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the Dative Alternation. In G. Boume, I. Kraemer, and J. Zwarts (eds.), *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan, and Marilyn Ford. 2010. Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. *Language* 86(1):186–213.

Cahill, Aoife. 2009. Correlating Human and Automatic Evaluation of a German Surface Realiser. In *Proceedings of the 47th Annual Meeting of the ACL*, pp. 97–100, Suntec, Singapore. Association for Computational Linguistics.

Cahill, Aoife, Martin Forst, and Christian Rohrer. 2007. Stochastic realisation ranking for a free word order language. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pp. 17–24, Saarbrücken, Germany. DFKI GmbH. Document D-07-01.

Cahill, Aoife, and Arndt Riester. 2009. Incorporating Information Status into Generation Ranking. In *Proceedings of the 47th Annual Meeting of the ACL*, pp. 817–825, Suntec, Singapore. Association for Computational Linguistics.

Crouch, Dick, and Tracy Holloway King. 2006. Semantics via F-Structure Rewriting. In M. Butt and T. H. King (eds.), *Proceedings of the LFG06 Conference*.

Crouch, Richard, Tracy Holloway King, John T. Maxwell III, Stefan Riezler, and Annie Zaenen. 2004. Exploiting F-structure Input for Sentence Condensation. In M. Butt and T. H. King (eds.), *Proceedings of the LFG04 Conference*, University of Canterbury.

Emele, Martin C., Michael Dorna, Anke Lüdeling, Heike Zinsmeister, and Christian Rohrer. 1996. Semantic-based transfer. In *Verbmobil: Foundations of Speech-to-Speech Translation*.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, and Ö. Dahl (eds.), *Proceedings of the Stockholm Workshop on 'Variation within Optimality Theory. April 26-27, 2003 at Stockholm Univ. Sweden*, pp. 111–120.

Graham, Yvette. 2010. *Deep Syntax in Statistical Machine Translation*. PhD thesis, Dublin City University, School of Computing.

Jäger, Gerhard. 2004. Maximum entropy models and stochastic optimality theory. Manuscript, University of Potsdam.

Kaplan, Ronald M., and Jürgen Wedekind. 2000. LFG generation produces context-free languages. In *Proceedings of COLING-2000*, pp. 297–302, Saarbrücken.

Kuhn, Jonas. 2000. Faithfulness violations and bidirectional optimization. In M. Butt and T. H. King (eds.), *Proceedings of the LFG 2000 Conference, Berkeley, CA*, CSLI Proceedings Online, pp. 161–181.

Kuhn, Jonas. 2001. Generation and parsing in Optimality Theoretic syntax – issues in the formalization of OT-LFG. In P. Sells (ed.), *Formal and Empirical Issues in Optimality-theoretic Syntax*, pp. 313–366. Stanford: CSLI Publications.

Kuhn, Jonas. 2003. *Optimality-Theoretic Syntax—A Declarative Approach*. Stanford, CA: CSLI Publications.

Riezler, Stefan, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, III, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the ACL*, pp. 271–278, Morristown, NJ, USA. Association for Computational Linguistics.

Riezler, Stefan, and John Maxwell. 2006. Grammatical Machine Translation . In *Proceedings of HLT-NAACL'06*, New York.

Velldal, Erik. 2008. *Empirical Realization Ranking*. PhD thesis, University of Oslo, Department of Informatics.

Wedekind, Jürgen. 1999. Semantic-driven generation with LFG- and PATR-style grammars. *Computational Linguistics* 25:277–281.

Zarrieß, Sina. 2009. Developing German semantics on the basis of parallel LFG grammars. In *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks (GEAF 2009)*, pp. 10–18, Suntec, Singapore. Association for Computational Linguistics.