



Proceedings of the LFG 05 Conference

University of Bergen

Editors: Miriam Butt and Tracy Holloway King

2005

CSLI Publications

ISSN 1098-6782

The Proceedings of the LFG'05 Conference

University of Bergen

Editors: Miriam Butt and Tracy Holloway King

2005 CSLI Publications

ISSN 1098-6782

Editors' Note

The program committee for LFG'05 were Aoife Cahill and Tara Mohanan. 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. We particularly thank the local organizing committee, Helge Dyvik, Victoria Rosén, Koenraad de Smedt and Helge Lødrup, who put on a conference that was splendid both in terms what was offered and how it was organized. In particular, Victoria Rosén (and then, later, Koenraad de Smedt) guided the pre-conference hike through a truly magically dense fog past lakes (only to be heard) and steep cliffs!

The table of contents lists all the papers presented at the conference and some that were accepted but could not be presented. Some papers were not submitted to the proceedings. For these papers, we suggest contacting the authors directly.

Hard Copy: All of the papers submitted to the LFG05 proceedings are available in one large pdf file, to be viewed and printed with Adobe Acrobat. Use the Show Bookmarks option to jump between papers in the table of contents.

Table of Contents

| | |
|--|---------|
| Figueiredo de Alencar, Leonel and Carmen Kelling Are Reflexive Constructions Transitive or Intransitive? Evidence from German and Romance | 1-20 |
| Alsina, Alex, Tara Mohanan and KP Mohanan How to get rid of the COMP | 21-41 |
| Beermann, Dorothee, Jonathan Brindle, Lars Hellan, Solomon Tedla, Florence Bayiga, Janicke Furberg, Yvonne Otoo and Mary Esther Kropp Dakubu A Comparison of Comparisons | 42-53 |
| Börjars, Kersti and Nigel Vincent Position vs. Function in Scandinavian Presentational Constructions | 54-72 |
| Broadwell, George Aaron It ain't necessarily S(V)O: Two kinds of VSO Languages | 73-83 |
| Burke, Michael, Aoife Cahill, Josef van Genabith and Andy Way Evaluating Automatically Acquired F-structures against PropBank | 84-99 |
| Dipper, Stefanie German Quantifiers: Determiners or Adjectives? | 100-115 |
| Estigarribia, Bruno Direct Object Clitic Doubling in OT-LFG: A New Look at Rioplatense Spanish | 116-135 |
| Falk, Yehuda Open Argument Functions | 136-153 |
| Forst, Martin, Jonas Kuhn and Christian Rohrer Corpus-Based Learning of OT Constraint Rankings for Large-Scale LFG Grammars | 154-165 |
| Fortmann, Christian On Parentheticals (in German) | 166-185 |
| Judge, John, Michael Burke, Aoife Cahill, Ruth O'Donovan, Josef van Genabith and Andy Way Strong Domain Variation and Treebank-Induced LFG Resources | 186-204 |
| Kibort, Anna The INs and OUTs of the Participle-Adjective Conversion Rule | 205-225 |
| King, Tracy Holloway Cliticizing LFG | 226-237 |

| | |
|---|---------|
| Kokkonidis, Miltiadis | 238-252 |
| Why Glue a Donkey to an F-structure When You Can Constrain and Bind it Instead? | |
| Luís, Ana and Ryo Otoguro | 253-270 |
| Morphological and Syntactic Well-formedness: The Case European Portuguese Clitics | |
| Mchombo, Sam, Yukiko Morimoto and Caroline Féry | 271-293 |
| Partitioning Discourse Information: A Case of Chichewa Split Constituents | |
| Mittendorf, Ingo and Louisa Sadler | 294-312 |
| Numerals, Nouns and Number in Welsh NPs | |
| Mycock, Louise | 313-333 |
| 'Wh'-in-situ in Constituent Questions | |
| O'Donovan, Ruth, Aoife Cahill, Josef van Genabith and Andy Way | 334-352 |
| Automatic Acquisition of Spanish LFG Resources from the CAST3LB Treebank | |
| Rákos, György and Tibor Laczkó | 353-370 |
| Verbal Category and Nominal Function: Evidence from Hungarian Subject Clauses | |
| Rosén, Victoria, Paul Meurer and Koenraad de Smedt | 371-387 |
| Constructing a Parsed Corpus with a Large LFG Grammar | |
| Schneider, Gerold | 388-407 |
| A Broad-Coverage, Representationally Minimalist LFG-like Parser: Chunks and F-structures are Sufficient | |
| Sells, Peter | 408-428 |
| The Peripherality of the Icelandic Expletive | |
| Spencer, Andrew | 429-446 |
| Case in Hindi | |
| Strunk, Jan | 447-467 |
| Pro-Drop in Nominal Possessive Constructions | |
| Wescoat, Michael | 468-486 |
| English Nonsyllabic Auxiliary Contractions: An Analysis in LFG with Lexical Sharing | |
| Wier, Thomas | 487-497 |
| On Pivots and Subjects in Georgian | |

ARE REFLEXIVE CONSTRUCTIONS TRANSITIVE OR INTRANSITIVE? EVIDENCE
FROM GERMAN AND ROMANCE

Leonel F. de Alencar
Universidade Federal do Ceará

Carmen Kelling
Universität Konstanz

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

Despite the substantial literature dedicated to it, the status of the reflexive is still controversial. Among others, the question whether reflexive constructions are transitive or intransitive, and if intransitive, whether they are unaccusative or unergative, has been intensively investigated. In this paper, we discuss this issue with data from German and Romance. We argue against the intransitivity hypothesis, showing that the reflexive behaves like a direct object. We implement the transitivity hypothesis in LMT, explaining why some reflexivized verbs behave like unaccusatives, while others show unergative-like behavior.

1 Introduction¹

The last decades have seen much work on reflexives in a wide variety of languages under different theoretical perspectives. In this paper, we reconsider certain properties of reflexives in German and Romance within the framework of LFG, especially from the perspective of Lexical Mapping Theory (LMT), its argument linking module (e.g., Bresnan 2001, Falk 2001). The primary focus lies on the reflexives examined also by Grimshaw (1982), i.e., reflexive verbs (1), reflexives in decausative (i.e. anticausative) constructions (2), and intrinsic reflexives (3).

- | | | | |
|-----|----|---|----------|
| (1) | a. | Max rasiert sich. Max shaves REFL | (German) |
| | b. | Max se rase. Max REFL shaves 'Max shaves.' | (French) |
| (2) | a. | Die Tür öffnet sich. the door opens REFL | (German) |
| | b. | La porte s'ouvre. the door REFL opens 'The door opens.' | (French) |
| (3) | a. | Max schämt sich. Max ashamed REFL 'Max is ashamed.' | (German) |
| | b. | Max s'évanouit. Max REFL faints 'Max faints.' | (French) |

The literature about reflexives is abundant. We distinguish three different positions concerning the argument status of the reflexive pronoun: the Strong Uniform Approach, the Weak Uniform Approach, and the Forked Approach (Alencar 2005). The first position maintains that the reflexive pronoun is a syntactic argument of the verb in all constructions, e.g. Selig (1998), Turley (1999), Steinbach (2002), and Kaufmann (2003 a, 2003 b). As far as the verbs in (1) to (3) are concerned, this view amounts to the *transitivity hypothesis*. The opposite view is taken by the second approach, the Weak Uniform Approach, according to which the reflexive is not a syntactic argument of the verb in any construction, e.g. Grimshaw (1982), Wehrli (1986), Alsina (1996), and Reinhart and Siloni (2004, 2005).² In the case of the

¹ We are indebted to Anette Frank, Ingrid Kaufmann, Françoise Kerleroux, Judith Meinschaefer, and Christoph Schwarze for comments and suggestions. Special thanks are due to Martine Lorenz-Bourjot and Nicole Nicaise for helping us with the French data and to Bruce Mayo for proofreading. Needless to say, all remaining errors are our responsibility.

² Sells, Zaenen and Zec (1987) espouse the Weak Uniform Approach for German reflexive *sich*, but the Strong Uniform Approach for Dutch *zich*.

reflexivization of monotransitive verbs, this analysis equals what we will refer to in this paper as the *intransitivity hypothesis*. Finally, defenders of the Forked Approach assume that the reflexive is a syntactic argument of the verb only in some constructions. For these scholars, the verb in (1) instantiates a transitive entry, while (2) and (3) instantiate intransitive entries, e.g. Helbig and Buscha (1991), Oesterreicher (1992), Butt et al. (1999), and Waltereit (2000).

Within derivational generative grammar, the advocates of the intransitivity hypothesis are divided into two opposing camps, as to whether reflexive constructions represent unaccusative or unergative entries. This, in turn, depends on whether the intransitive predicate's argument is internal or external. In LFG, as in Lexical Decompositional Grammar (LDG) (Kaufmann 1995), this syntactic configurational-based distinction is reinterpreted semantically: unaccusatives have a theme/patient argument, while unergatives have an agent argument. Recast in terms of LMT, the intransitivity approach has to decide whether the sole element in the argument structure of reflexivized verbs is an agent or a theme/patient argument, marked, respectively, as [-o] and [-r].

In this paper, we address questions (i) and (ii) concerning the status of the *se/sich* element and (iii) concerning the argument structure of the reflexive constructions from (1) to (3):

- i. Is it a syntactic argument of the verb or just a grammatical marker of valency reduction without argument status?
- ii. Is it a decausativity marker, or is it a detransitivity marker?
- iii. Does a reflexivized verb instantiate an unaccusative or an unergative argument structure?

We will argue for a uniform treatment of reflexives as syntactic arguments in all constructions from (1) to (3) and, consequently, against analyses that deny their argument status. In the following, we deconstruct, first, the main arguments in favor of the intransitivity hypothesis, including the asymmetry between reflexive verbs and transitive verbs in French causative constructions. The value of this asymmetry as evidence for intransitivity has, so far as we know, never been challenged. Second, we add evidence in favor of the transitivity hypothesis to that which has been proposed in the literature so far. Finally, we implement the transitivity analysis within the framework of LFG/LMT. The argument structures assigned to the various types of reflexivized verbs explain naturally why some of them show unaccusative-like behavior, while others show unergative-like behavior. This analysis reveals that the reflexive is a thematic direct object in (1) and an expletive and hence non-thematic direct object in the constructions (2) and (3). In the case of (2), this expletive can also be seen as a marker of decausativization, since it results from the application of a lexical rule which suppresses the agent role from the verb's LCS.

This paper is structured as follows. First, we show in section 2 that the arguments in favor of the intransitivity hypothesis are flawed. In section 3 positive evidence is presented supporting the transitivity hypothesis. Section 4 implements this analysis in terms of the LMT formalism. Finally, the main conclusions from the paper are drawn in section 5.

2 Against the intransitivity hypothesis

Those scholars who have supported the intransitivity hypothesis in Romance, among them Grimshaw (1982), cite the parallel behavior of French reflexive verbs and intransitive verbs in causative constructions and in NP extraposition as evidence for their analysis. As far as German is concerned, the apparently analogous behavior of German reflexive verbs and intransitive verbs, for instance in impersonal passive constructions, has long been considered

evidence of the non-argument status of the reflexive (e.g. Sells, Zaenen and Zec 1987, Butt, King et al. 1999). However, all these behaviors must be weighed against other important evidence.

2.1 French causative constructions

The asymmetry between (5) and (7) and the parallel between (4) and (6) have long been taken as evidence of the intransitivity of the reflexive. Examples (4) to (6) were taken from Reinhart and Siloni (2004:162, 2005:393), but similar data have been cited by proponents of the intransitivity hypothesis since Kayne (1975). The grammaticality of (6), though, is controversial, as we will see later.

| | | | | | |
|-----|----------------|---------|--------------------------------|--|--------------|
| (4) | Je ferai | | courir Paul. | | intransitive |
| | I make.1PS.FUT | | run Paul | | |
| | | | ‘I’ll make Paul run.’ | | |
| (5) | Je ferai | | laver Max à Paul. | | transitive |
| | I make.1PS.FUT | | wash Max to Paul | | |
| | | | ‘I’ll make Paul wash Max.’ | | |
| (6) | Je ferai | se | laver Paul. | | reflexive |
| | I make.1PS.FUT | himself | wash Paul | | |
| | | | ‘I’ll make Paul wash himself.’ | | |
| (7) | *Je ferai | se | laver à Paul. | | reflexive |
| | I make.1PS.FUT | himself | wash to Paul | | |

From (4) to (7) we draw a completely different conclusion, namely that the asymmetrical behavior of the reflexive verb in these examples is not due to its alleged intransitive status. Instead, we show below that it follows either from binding constraints on the reflexive anaphor or from linking constraints in biclausal causative constructions. In order to demonstrate this, we must first take a brief look at the grammatical regularities of complex predicate formation in the case of causative constructions such as (4) and (5).

2.1.1 Complex predicate formation: the monoclausal construction

With Alsina (1996), Abeillé, Godard and Miller (1997), and Butt (1997), among others, we assume that complex predicate formation in examples like (4) and (5) as well as (8) involves the creation of a new argument structure out of the argument structures of the causative and the embedded predicate. The resulting f-structure is monoclausal, instantiating a single domain of predication.

| | | | | | | |
|-----|-------------|--------|---|-----------------|-----------|-------------------|
| (8) | J’ai fait | écrire | une lettre | au directeur | par Paul. | (Comrie 1981:172) |
| | I have made | write | a letter | to the director | by Paul. | |
| | | | ‘I made Paul write a letter to the director.’ | | | |

The assignment of grammatical relations to this complex argument structure follows a cross-linguistically well observed pattern, which we call, for convenience only, the Default Causativization Paradigm (DCP), schematized in (9) to (11). Grey shading indicates the argument structure contributed by the embedded predicate. Bold type and dark grey shading highlight grammatical function change in the embedded verb’s valency frame.

- (9) Causativization of an intransitive verb (cf. (4))

| | | |
|-------------|-------|---------------|
| LCS | Agent | Causee |
| a-structure | x | y |
| f-structure | SUBJ | OBJ |

- (10) Causativization of a transitive verb (cf. (5))

| | | | |
|-------------|-------|------------------------|-------|
| LCS | Agent | Causee | Theme |
| a-structure | x | y | z |
| f-structure | SUBJ | OBJ_θ | OBJ |

- (11) Causativization of a ditransitive verb (cf. (8))

| | | | | |
|-------------|-------|------------------------|-------|------------------|
| LCS | Agent | Causee | Theme | Beneficiary |
| a-structure | x | y | z | w |
| f-structure | SUBJ | OBL_θ | OBJ | OBJ _θ |

In the DCP, the causer (the main predicate's agent) is always assigned the SUBJ, while the causee (the participant caused to do something) maps either onto the OBJ or a lower grammatical function on the Relational Hierarchy in (12), depending on the base verb's valency. The DCP is driven by the Uniqueness Condition (Falk 2001:115), which rules out multiple instantiations of a single grammatical function in the same domain of predication.³

- (12) Relational Hierarchy (Bresnan 2001:96)
SUBJ > OBJ > OBJ_θ > OBL_θ

According to the DCP, if the embedded verb is intransitive, the causee is mapped onto OBJ (cf. (9)). In the case of a transitive verb, the assignment of the OBJ relation to the causee would violate Uniqueness. To avoid this, the causee is mapped instead onto OBJ_θ (cf. (10)), the next available lower function in (12). Through an analogous strategy, the causee maps onto OBL_θ in the causativization of a ditransitive verb (cf. (11)).

The DCP is the most widespread pattern cross-linguistically, but not a grammatical universal. For example, the agent of a transitive verb may sometimes be realized as an OBL_θ (introduced by *par* 'by') instead of OBJ_θ (marked by *à* 'to') (cf. (13)).⁴

- (13) Jean a fait manger les pommes à/par Paul.
John has made eat the.PL apples to/by Paul
'John has made Paul eat the apples.'

In the following section, we consider a second type of causative construction.

2.1.2 Another causative construction: the biclausal construction

Analogously to Urdu (Butt 1997), Romance languages do not have only the monoclausal causative construction of the type *faire laver* 'make wash' (which is structurally similar to the Urdu Permissive). In our view, (6) exemplifies the so-called biclausal Equi construction, prototypically represented by the verb *laisser* 'let', exemplified in (14) (Kroeger

³ Multiple instantiations of OBJ_θ and OBL_θ are licensed if θ is differently instantiated (Falk 2001:106, FN 11).

⁴ Some speakers prefer *par* 'by' to express the causee in this case. It is not clear whether this alternation is syntactically or semantically conditioned, as Abeillé and Godard (2003:134, FN 17) point out.

2004:223). As Abeillé, Godard and Miller (1997) propose, French causative *faire* ‘make’ has both the monoclausal and the biclausal construction.

- (14) Marie a laissé Paul lire ces romans.
 Marie has let Paul read these novels
 ‘Marie let Paul read these novels.’

In the biclausal causative construction, the main verb does not form a complex predicate with the embedded verb instantiating a single domain of predication. Instead of a flat f-structure, we have a complex f-structure with two domains of predication in (14), as in the Urdu Instructive (cf. Butt 1997). In French as well as in Portuguese, German, etc. the causee is realized in this case as the main clause OBJ, independently of the valency of the embedded verb. This type of causative is an object control verb with the valency frame <SUBJ OBJ XCOMP>, as schematized in (15).

- (15) The biclausal causative construction

| LCS | Agent | Causee | Caused action |
|-------------|-------|--------|---------------|
| a-structure | x | y | p |
| f-structure | SUBJ | OBJ | XCOMP |

The distinction between the biclausal and the monoclausal construction in French correlates with clitic positioning, as evidenced in (16) and (17). As in (6), in (16) the clitic is adjacent to the verb of the embedded clause, separating it from the causative verb. In contrast, in the monoclausal construction, the causative and the embedded verb must be adjacent. A clitic may not separate them, as shown in (17).

- (16) a. Marie a laissé Paul les lire.
 Marie has let Paul them.ACC read
 ‘Marie let Paul read them.’
 b. *Marie les a laissé Paul lire.
 Marie them.ACC has let Paul read (Kroeger 2004:223)
- (17) a. Je le ferai laver à Paul.
 I him.ACC make.1PS.FUT wash to Paul
 ‘I’ll make Paul wash him.’
 b. *Je ferai le laver à Paul.
 I make.1PS.FUT him.ACC wash to Paul

In the next section, we reconsider the examples (6) and (7) in order to analyze their (un)grammaticality.

2.1.3 Explaining the behavior of reflexive verbs in causative constructions

With the distinction of the two different causative constructions, we can now explain the data in (6) and (7). Contrary to advocates of the intransitivity hypothesis, who consider the reflexive in (6) to be a mere intransitivization marker of a verb in a monoclausal causative construction of type (9), we analyze (6) as a biclausal causative construction, as suggested in the previous section. The reflexive instantiates, then, the OBJ of the embedded clause, which in turn functions as the main clause XCOMP. The NP *Paul* (i.e. the causee) instantiates the main clause OBJ, according to (15).

Cross-linguistically, reflexive anaphors are subject to item-specific binding constraints (Dalrymple 1993). For Romance *se*, its antecedent is constrained to be a SUBJ in its *Minimal*

Complete Nucleus, that is, the nucleus that contains the reflexive and a SUBJ that outranks it (Bresnan 2001:218-219). From the perspective of biclausal causativization and binding theory, we can see why (6) is possible. Here, the binding constraint is obeyed: the reflexive that is realized as the XCOMP OBJ function is bound by the implicit XCOMP SUBJ (realized as the main clause control verb's OBJ, i.e. *Paul*).

Turning to (7), there are two possibilities of explaining its non-grammaticality by means of independent principles without resorting to the intransitivity hypothesis. First, (7) can be seen as instantiating the same complex predicate *faire laver* 'make wash' of (5). In this case, this example is unacceptable due to a violation of the binding constraint on the reflexive. In fact, the Romance anaphor *se*, realizing an OBJ, may not be bound by a non-SUBJ co-argument (cf. (18)), let alone by a more oblique one, according to the Relational Hierarchy in (12) (cf. Dalrymple 1993). The PP *à Paul* 'to Paul' in (7) cannot bind the anaphor, since this PP is an OBJ_θ.

- (18) *Jean _{s_i'} est montré l'enfant_i.
 Jean REFL_i is shown the child_i

Secondly, one could analyze (7) as biclausal, since the reflexive separates the causative and the embedded verb. In this case, there is no violation of binding principles, because the reflexive is not bound by the OBJ_θ, but by the XCOMP SUBJ, which is functionally controlled by the main clause OBJ_θ. Under this analysis, the ungrammaticality of the construction is due to the lack of justification for realizing the causee as OBJ_θ. Recall that in biclausal causative constructions, the causee maps by default onto OBJ (cf. (15)).

From the analysis of (6) and (7) as biclausal, however, it seems to follow that (19) should be grammatical and (20) ungrammatical. This prediction, however, is not borne out. While (20) is not fully acceptable by all speakers, as Abeillé and Godard (2003:174) observe for parallel examples, (19) is invariably rejected.

- (19) *Je ferai se laver les mains Paul.
 I make-1PS.FUT REFL wash the hands Paul
 (20) ?Je ferai se laver les mains à Paul.⁵
 I make-1PS.FUT REFL wash the hands to Paul
 'I will make Paul wash his hands.'

The asymmetry between (19) and (20) seems, then, to support the monoclausal analysis of (6) as proposed within the intransitivity hypothesis. On the one hand, the prohibited realization of the causee in (19) as an OBJ would follow from the Uniqueness Condition. On the other hand, its realization as an OBJ_θ, paralleling (5), would follow from the Relational Hierarchy constraint.

Speakers of French, though, do accept (21 a), where the causee is realized as a clitic OBJ, along with (21 b), where it surfaces as a clitic OBJ_θ. The former construction is even preferred over the latter by some speakers.

⁵ We are grateful to Alex Alsina (p. c.) for suggesting these two examples.

- (21) a. Je le ferai se laver les mains.
 I him.ACC make.1PS.FUT REFL wash the hands
 b. Je lui ferai se laver les mains.
 I him.DAT make.1PS.FUT REFL wash the hands
 ‘I will make him wash his hands.’

Therefore, it seems that the unacceptability of (19) is not due to the causee being realized as an OBJ in itself, something that would cast serious doubt on our claim that this element realizes an OBJ. Instead, the problem with (19) lies in the unusual placement of the OBJ of the main clause. In biclausal constructions in French, a main clause OBJ canonically comes just after the main verb (cf. (22) and (23)). By contrast, the main OBJ surfaces only after the embedded OBJ in (19).

- (22) Je laisserai Paul se laver les mains.
 I let.1PS.FUT Paul REFL wash the hands
 (23) *Je laisserai se laver les mains Paul.
 I let.1PS.FUT REFL wash the hands Paul
 ‘I will let Paul wash his hands.’

As a matter of fact, speakers who accept both (21 a) and (21 b) reject (6), where the main clause OBJ, according to our analysis, is also displaced. They find (24), though, fully grammatical.

- (24) Je le ferai se laver.
 I him make.1PS.FUT himself wash
 ‘I’ll make him wash himself.’

These judgements about (6) and (24) mirror the opinion of Abeillé, Godard and Miller (1997), who reject (25), while accepting (26). This is structurally parallel to (24), with the clitic *le* instead of the reflexive realizing the embedded clause OBJ. This construction, though, is rejected as non-standard by some speakers, who prefer the monoclausal version in (27).

- (25) *Le professeur fera le lire les secondes.
 The professor make.FUT.3PS it read the.PL assistants
 ‘The professor will make the assistants read it.’
 (26) Le professeur nous a fait le lire. biclausal
 The professor us has made it read.
 ‘The professor made us read it.’
 (27) Le professeur nous l’a fait lire. monoclausal
 The professor us it has made read.
 ‘The professor made us read it.’

Note that proponents of the intransitivity approach, from Kayne (1975) to Reinhart and Siloni (2004, 2005), treat (6) as a structurally exact parallel to (4). Under this analysis, both sentences instantiate the ordering SUBJ <_f PRED <_f OBJ (where <_f symbolizes f-precedence, see section 3.1). However, there is a strong acceptability contrast between these two structures among native speakers. This is completely unexpected if we consider (6) a monoclausal construction and the reflexive just a valency reduction marker without argument status. In contrast, if we view the reflexive as the OBJ of the embedded clause of a biclausal construction, we can predict that speakers who reject (25), as reported by Abeillé, Godard and

Miller (1997), will also consider (6) to be ungrammatical, and this has in fact been corroborated by our own informants.

2.1.4 Summary

To sum up, we assume two different constructions for the French verb *faire* ‘make’. The first construction is the more common monoclausal construction (cf. (5) and (4)), the second is the biclausal construction (cf. (6)). The existence of these two different constructions and their interactions with binding and linking constraints explain the observed asymmetries between non-reflexivized and reflexivized verbs and intransitive verbs. We do not have to postulate reflexives to be intransitive in order to explain these data. Besides, the intransitive hypothesis leads to wrong predictions about the behavior of reflexivized transitives and intransitives in causativization. By contrast, the transitivity hypothesis accounts for the observed asymmetry, showing that reflexivized transitives behave like transitives.

2.2 NP extraposition

As we will show in this section, the argument in favor of the intransitivity hypothesis based on asymmetrical behavior of reflexives and transitives in French NP extraposition does not stand up to close scrutiny. As one can see from the examples (28) to (31) from Grimshaw (1982:112-116), NP extraposition is licensed for unaccusatives and decausative reflexive verbs (type of (2)), as shown in (28) and (29), respectively, but completely impossible with transitive verbs, as evidenced in (30) and (31).

- | | | | | | | | |
|------|-------------------------------|---------|--------------|--------------|------------|-----------------------|--------------|
| (28) | Il | passé | un | train | toutes les | heures. | unaccusative |
| | it | passes | a | train | all.PL the | hours | |
| | ‘A train goes by every hour.’ | | | | | | |
| (29) | Il | se | brisera | beaucoup de | verres. | decausative reflexive | |
| | it | REFL | break.FUT | many of | glasses | | |
| | ‘Many glasses will break.’ | | | | | | |
| (30) | *Il | mangera | cette tarte | trois | filles. | transitive | |
| | it | eat.FUT | this pie | three | girls | | |
| (31) | *Il | mangera | trois filles | cette tarte. | | | |
| | it | eat.FUT | three girls | this pie | | | |

However, native speakers of French do not accept NP extraposition for example (32) (from Martin 1970:380) with a reflexivized transitive verb (type of (1)) and for (33) with an unergative verb:

- | | | | | | | |
|------|-----|------|--------------|-------------------|---------------|-------------------------|
| (32) | *Il | se | fardait | un acteur dans sa | loge. | reflexivized transitive |
| | it | REFL | make-up.PAST | an actor in his | dressing-room | |
| (33) | *Il | a | dormi | trois filles | dans ce lit. | unergative |
| | it | has | slept | three girls | in this bed | |

It follows that the asymmetrical behavior of the sentences in (28) to (31) cannot be attributed to the contrast between transitivity and intransitivity, but to semantic factors.

2.3 Auxiliary selection in Romance

In Romance, auxiliary selection was long taken as evidence in favor of the intransitivity hypothesis, because reflexive verbs in French and Italian always select ‘be’ instead of ‘have’. Notwithstanding this fact, one must consider that Portuguese, Spanish and Catalan select

‘have’ for all verbs, including reflexivized verbs. Besides, even proponents of the intransitivity hypothesis currently reject auxiliary selection as a criterion for pairing reflexivized and intransitive verbs. As Reinhart and Siloni (2004:168), who espouse the intransitivity hypothesis, put it, auxiliary selection “is an intricate matter, which is not yet well understood”. Selection of ‘be’ by French and Italian reflexive verbs is not tied to intransitivity, as Schwarze (1998:103-104) shows, but to reflexivity (i.e. the mere syntactic presence of the reflexive pronoun), since there are intransitives in both languages which select ‘have’. Non-reflexive intransitive verbs in French, such as *réussir* ‘succeed’ and *rougir* ‘blush’, whose translation equivalents in Italian take ‘be’, most commonly select ‘have’ (Schwarze 1998:103).

As far as auxiliary selection of reflexive verbs is concerned, we have a clear contrast between German, on the one hand, and French and Italian, on the other. In German, as we will see in section 3, selection of ‘have’ by all reflexivized verbs implies that they are transitives, since all transitives in German select this auxiliary. In French and Italian, however, the fact that reflexive verbs always select ‘be’ does not mean necessarily that they are intransitives, since one cannot predict the auxiliary of a particular verb from its transitivity status. We can conclude that selection of ‘be’ by French and Italian reflexive verbs cannot be taken as evidence that these verbs are intransitive.

2.4 Impersonal passive constructions

Sells, Zaenen and Zec (1987) conclude from the asymmetry in (34) to (36) that the reflexive does not have OBJ status. As with other arguments in favor of the intransitivity hypothesis, this one also does not stand up (Alencar 2005).

- (34) Gestern wurde getanzt.
 yesterday was danced
 ‘Yesterday dancing took place.’
- (35) Jetzt wird sich gewaschen!
 now is oneself washed
 ‘Now one must wash oneself!’
- (36) a. Jetzt wird der Brief geschrieben.
 now is the letter.NOM written
 b. *Jetzt wird den Brief geschrieben.
 now is the letter.ACC written

We assume the passive operation to be twofold (Kroeger 2004:54, Eisenberg 1999:126): first, the active verb’s SUBJ is demoted, cf. (34) and (36 a); then the active verb’s OBJ is promoted to SUBJ of the passive construction, cf. (36 a). From a typological perspective, while the first operation is obligatory, the second is optional (that is, subject to parametric variation), as Kroeger (2004:54) observes. In the Finnish passive construction (37 b) (cf. active version in (37 a)), which parallels the German example (34), only SUBJ demotion applies.

- (37) a. Äiti jätti hänet kotiin. (Finnish)
 mother left him.ACC to.home
 b. Hänet jätettiin kotiin. (Finnish)
 him.ACC was.left to.home
 ‘He was left at home.’ (Kroeger 2004:54)

But if the reflexive is an OBJ, why is (35) licensed in German, while (36 b) is impossible? The answer is simple: there is no nominative reflexive in German (cf. Eisenberg

1999:129). Besides, as Bresnan (2001:7) suggests, reflexive subjects are ruled out by Universal Grammar (UG) principles. In fact, a reflexive subject would violate the Relational Hierarchy constraint (cf. (12)) on reflexive anaphors (cf. Berman and Pittner 2004:138). Assuming an optimality-theoretic account of constraints (for the combination of LFG and Optimality Theory, cf. Falk 2001:195), we claim that OBJ promotion is obligatory in the passivization of German (and of Romance, for that matter) transitive verbs. This explains the asymmetry between (36 a) and (36 b). OBJ promotion, though, is prevented from being applied in (35), because this would violate a higher order constraint, namely the invariant UG prohibition of reflexive subjects.

3 In favor of the transitivity analysis

3.1 Distribution of the reflexives

Evidence for an analysis of the reflexives in (1) to (3) as direct objects comes from distributional facts in German (Steinbach 2002) and in Romance. Reflexives are subject to exactly the same c-structure ordering constraints as object pronouns in German, as the following examples in (38) to (47) show:

- (38) weil er sie plötzlich geöffnet hat
 because he.SUBJ it.OBJ suddenly opened has
 ‘because he opened it suddenly’
- (39) weil sie sich plötzlich geöffnet hat
 because it.SUBJ REFL.OBJ suddenly opened has
 ‘because it opened suddenly’

In LFG, generalizations about precedence relations between constituents are captured by means of rules which resort to the notion of *f-precedence* (Falk 2001:67), i.e. precedence on the f-structure level. In German, we have a rule like (40) that constrains the relative ordering of subject and object pronouns in the so called *Mittelfeld*, which comprises the positions between the auxiliary and the main verb:

- (40) SUBJ $<_f$ OBJ

This rule accounts for the non-grammaticality of (41), where the OBJ *f*-precedes the SUBJ:

- (41) *weil sie er plötzlich geöffnet hat
 because it.OBJ he.SUBJ suddenly opened has
 ‘because he opened it suddenly’

Since the reflexive *sich* in (39) is an OBJ, inversion of the subject and reflexive in this construction is not possible:

- (42) *weil sich sie plötzlich geöffnet hat
 because REFL.OBJ it.SUBJ suddenly opened has
 ‘because it opened suddenly’

In European Portuguese, postverbal placement of pronominal clitics constitutes the default case (Luís and Otoguro 2004). In this respect, non-reflexive and reflexive objects behave alike (cf. (43) and (44)).

- (43) O João tinha- a chateado.
 João had her.OBJ annoyed
 ‘João had annoyed her.’
- (44) A Maria tinha- se chateado.
 Maria had REFL.OBJ been annoyed
 ‘Maria had been annoyed.’

We assume with Luís and Otaguro (2004) that preverbal placement of clitics is triggered by f-precedence rules which we generalize as (45), where X stands for an arbitrary element of the class of proclitic triggers, including quantified subjects and adverbs such as *também* ‘also’.

- (45) (\uparrow X) $<_f$ (\uparrow OBJ(θ))

The following examples show that both non-reflexive and reflexive clitics are subject to rule (45):

- (46) Alguma coisa a tinha chateado. (Mateus et al. 1989:332)
 something her.OBJ had annoyed
 ‘Something had annoyed her.’
- (47) Também se tinha chateado.
 also REFL.OBJ had been annoyed
 ‘She also was annoyed.’

We can conclude that data from German and Portuguese concerning the linearization order of grammatical functions constitute strong evidence that reflexives are treated by the syntax as objects, since they are subject to the same rules as their non-reflexive counterparts. If reflexives were treated as non-arguments, one would have to formulate separate rules to account for their placement. The intransitivity hypothesis leads, then, to a less parsimonious account of the grammar than the transitivity hypothesis.

3.2 Case and auxiliary selection in German

As Bierwisch (1996) and Kaufmann (2003 b) have suggested, the correlation between case and auxiliary selection in German also shows that the reflexive is an object.

- (48) Er hat sich vor dem Hund erschrocken.
 he has REFL of the dog frightened.
 ‘He became frightened of the dog.’
- (49) Er ist vor dem Hund erschrocken.
 he is of the dog frightened.
 ‘He became frightened of the dog.’

German transitive constructions force selection of ‘have’. Hence, the alternation from (48) with ‘have’ to (49) with ‘be’, where the meanings and semantic argument structures are virtually identical, can only be attributed to the fact that the reflexive *sich* is an object.

3.3 Past participle agreement in French

As stated in section 2.3, complex tenses in French are built with either the auxiliary *avoir* ‘has’ or *être* ‘be’ and the past participle of the main verb, which agrees in gender and number with

the subject or direct object, or remains uninflected (i.e. in the masculine singular form), depending on a complex of factors. Since direct objects can trigger past participle agreement in some contexts, this constitutes an important criterion for analyzing an element as a direct object. We will show in this section that past participle agreement in French is elegantly accounted for if, according to the transitivity hypothesis, all reflexive pronouns are analyzed as objects.

Examples (50) and (51) show that past participle agreement, in the case of intransitive verbs, is sensitive to the unergative and unaccusative distinction. With unergatives, there is no agreement with the subject (cf. (50)). With unaccusatives, though, the past participle must agree with the subject (cf. Berman and Frank 1996:130).

- (50) Elle a dancé.
 She.FEM.SUBJ has danced.MASC
 ‘She has danced.’
- (51) Elle est arrivée.
 She.FEM.SUBJ is arrived.FEM
 ‘She has arrived.’

As can be seen in (52), there is no agreement with the subject if the direct object (*sa fille* ‘her daughter’ in this case) does not precede the verb:

- (52) La mère a lavé sa fille.
 the mother has washed.MASC her daughter
 ‘The mother has washed her daughter.’

However, agreement with direct objects that precede the verb is obligatory, no matter whether it is a clitic pronoun (53), a relative pronoun (54), or the inverted direct object of a question (55).

- (53) La mère l’ a lavée.
 the mother her.FEM.OBJ has washed.FEM
 ‘The mother has washed her.’
- (54) la fille que la mère a lavée
 the daughter_i.FEM that_i.OBJ the mother has washed.FEM
 ‘the daughter that the mother has washed’
- (55) Combien de bouteilles ton frère a-t-il achetées?
 how many bottles.FEM.OBJ your brother has he bought.FEM.PL
 ‘How many bottles did your brother buy?’

We assume that agreement has the same source in the case of reflexives, i.e. the preceding reflexive clitic, which we analyze as a direct object:

- (56) La mère s’ est lavée.
 the mother herself.OBJ is washed.FEM
 ‘The mother has washed herself.’

That participle agreement is indeed triggered by the preceding direct object reflexive, and not by the subject as with unaccusatives in (51), is shown by the agreement behavior of indirect reflexive constructions in (57).

- (57) La mère s' est lavé les mains.
 the mother herself.OBJ_θ is washed.MASC the hands.OBJ.
 'The mother has washed her hands.'

As we can see in (57), there is neither agreement with the OBJ_θ reflexive clitic nor agreement with the subject, which would give *lavée*.FEM. Hence, with reflexive constructions, there is only agreement with a preceding direct object.

One could be tempted to explain this complex agreement pattern by postulating two rules: (i) the past participle agrees with the subject of an intransitive unaccusative verb or (ii) with a preceding direct object. If one classifies, as Grimshaw (1990) does, reflexive constructions as unaccusatives, then the agreement facts exemplified above are accounted for. However, the unaccusative analysis of reflexives has been much discredited in the last few years by proponents of the intransitivity hypothesis themselves (cf. e.g. Alsina 1996, Reinhart and Siloni 2004).

Under the transitivity hypothesis, though, a much more straightforward unified account of the parallel behavior of unaccusative verbs, reflexive verbs and transitive verbs with extracted objects is possible. We have to posit just one rule, namely that the past participle agrees with a preceding argument marked as [-r]. The past participle does not agree with unergatives, because the preceding argument (i.e. the subject) is a [-o] argument. As we will see in section 4, reflexive direct objects, just as non-reflexive direct objects and unaccusative subjects, are marked as [-r].

To sum up, concerning past participle agreement in complex tenses, reflexive constructions show the same behavior as transitive verbs. We therefore conclude that reflexive constructions are transitive and that the reflexive in (56) is an OBJ.

3.4 Behavior in participial constructions

The behavior of reflexives in participial constructions matches that of transitives, not that of intransitives (cf. (58)).

- (58) a. die sich öffnende Tür
 the REFL opening door
 'the opening door'
 b. die geöffnete Tür
 the opened door
 c. *die sich geöffnete Tür
 the REFL opened door

In (58 a), the reflexive *sich* is preserved in the present participle construction, just like any OBJ, cf. (59):

- (59) ein die Tür öffnender Mann
 a the door opening man
 'a man who is opening the door'

In the past participle construction in (58 c), though, the reflexive cannot surface. This asymmetry does not find an explanation under the intransitivity hypothesis. However, the transitivity hypothesis can explain why the reflexive is prohibited from being realized in (58 c): past participles, just like passive verbs, do not subcategorize for an OBJ (Berman and Frank 1996:183). As the expletive reflexive is an OBJ, it is incompatible with this construction.

3.5 Argument linking regularities

Finally, argument linking regularities, like (60) from Portuguese, show the reflexives to be direct objects: the reflexive blocks the possibility of realizing the theme argument *a viagem* ‘the trip’ as direct object. This alternation pattern also holds in other languages such as German (e.g. *fürchten* ‘to fear’).

- (60)
- | | | |
|----|----------------------------|--------------|
| a. | Lembrei-me | da viagem. |
| | remember.1SG.PAST-REFL.OBJ | of the trip |
| | ‘I remembered the trip.’ | |
| b. | Lembrei | a viagem. |
| | remember.1SG.PAST | the trip.OBJ |
| | ‘I remembered the trip.’ | |
| c. | *Lembrei-me | a viagem. |
| | remember.1SG.PAST-REFL.OBJ | the trip.OBJ |

This shows that the reflexive should be analyzed as an object.

4 LMT Analysis

In the following we sketch the mapping of the constructions under consideration. (61) shows the mapping pattern from LCS onto a-structure for typical transitive/causative verbs like *raser* ‘shave’ and *ouvrir* ‘open’. According to LMT principles, the a-structure configuration <[-o] [-r]> maps onto <SUBJ OBJ> in f-structure.

| | | | | |
|------|-------------|---------------------|-------|-------|
| (61) | LCS | | agent | theme |
| | a-structure | <i>raser/ouvrir</i> | < x | y > |
| | | ‘shave/open’ | [-o] | [-r] |
| | f-structure | | SUBJ | OBJ |

4.1 Reflexive verbs

Taking into account LFG’s binding principles, we assume the structure in (62) for reflexive verbs as *se raser* ‘shave oneself’ (cf. Kelling 2005). Binding principles predict the mapping of the theme argument onto a reflexive OBJ. Note that binding occurs on every level from LCS to f-structure.

| | | | | |
|------|-------------|-----------------|--------------------|-------------------------|
| (62) | LCS | | agent _i | theme _i |
| | a-structure | <i>se raser</i> | < x _i | y _i > |
| | | ‘shave oneself’ | [-o] | [-r] |
| | f-structure | | SUBJ _i | OBJ _i = REFL |

4.2 Decausative verbs

For decausative constructions, we assume the decausative operation in (63), which could also be argued to underlie the intransitive inchoative variant of a transitive causative verb like *to open*.

- (63) DECAUSATIVE OPERATION: agent → ∅

By means of (63) the agent argument is suppressed, and mapping principles ensure the mapping of the theme argument onto the SUBJ function. (64) shows the result of the application of this rule on transitive *open*. While in English the same verb stem is used in both variants, the decausative operation may be associated with a special morphological marking in languages like Modern Greek and Hebrew (Haspelmath 1993).

| | | | |
|------|-------------|---------------------|--------------|
| (64) | LCS | | theme |
| | a-structure | <i>open</i> (intr.) | < y > |
| | f-structure | | [-r] SUBJ |

However, unlike English and languages such as Modern Greek and Hebrew that code the operation morphologically, in German (and also Norwegian, cf. Dalrymple 1993:30-31) as well as in Romance, the [-r] feature associated with the least prominent argument of the causative verb variant remains at a-structure in form of a non-thematic (expletive) argument bound to the most prominent argument. As a consequence, this non-thematic argument must be realized as a reflexive OBJ, as is shown in (65).⁶

| | | | | |
|------|-------------|-----------------|--------------------|---|
| (65) | LCS | | theme _i | |
| | a-structure | <i>s'ouvrir</i> | < y _i > | - _i |
| | | 'open' | [-r] | [-r] |
| | f-structure | | SUBJ _i | OBJ _i = REFL _{EXPL} |

The [-r] feature refers to an unrestricted syntactic function, the kind of function which is not restricted as to its semantic role in the sense that it need not have any semantic role at all, that is, it can be an expletive (non-thematic argument), as for the italicized subject and object of (66) and (67), respectively (Bresnan 2001:308).

- (66) *It* is obvious that the world is flat.
 (67) I take *it* that the world is flat. (Falk 2001:107)

In the LFG literature, expletive subjects and expletive objects are considered to exist also in German (e.g. Butt et al. 1999:77, Berman 2003:64-67), cf. (68) and (69).

- (68) ... weil es keine Hoffnung gibt.
 because it no hope gives
 '... because there is no hope.' (Berman 2003:67)
- (69) Sie hat es eilig.
 she has it quickly
 'She's in a hurry.' (Berman 2003:64, FN 18)

Expletives are represented outside the angled brackets in which the verb's thematic arguments are listed, cf. the valency frame '<OBJ> SUBJ' for *geben* 'there to be' in (68), and '<SUBJ> OBJ' for *eilig haben* 'be in a hurry' in (69) (Berman 2003:67).

⁶ Two anonymous reviewers have remarked that (65) represents a violation of the so called Asymmetrical Object Parameter – AOP (Bresnan 2001:310, Falk 2001:114), which prohibits the configuration <... [-r] ... [-r] ...>. In fact, German and Romance languages are not symmetrical object languages, where the configuration <... [-r] ... [-r] ...> is licensed. With Bresnan (2001:310), we interpret the AOP, though, as a constraint on the mapping from LCS onto a-structure, not as a constraint on a-structures per se. Hence (65) does not fall under the AOP constraint.

Note that in (65) there is no semantic correlate to the OBJ in LCS. As observed by Steinbach (2002), this accounts for certain syntactic effects in German such as the asymmetry between (70) and (71), which was for a long time mistakenly regarded as evidence in favor of the intransitivity hypothesis (e.g. Helbig and Buscha 1991):

- (70) *Sich öffnet die Tür.
REFL opens the door
(71) Sich rasiert er.
REFL shaves he

From (65), one can predict that the non-thematic syntactic argument may be, at least in principle, suppressed, since it is negatively specified (Bresnan 2001:310). This prediction is borne out by data from Brazilian Portuguese (BP), where the expletive reflexive OBJ may be deleted in some dialects (cf. Monteiro 1994, Camacho 2003), as shown in (72).

- (72) a. A porta abriu-se. (European Portuguese/Standard BP)
the door opened REFL
b. A porta abriu. (Nonstandard BP)
the door opened

The suppression of the non-thematic [-r] argument produces the argument structure (73), which is identical to that of English decausative verbs (cf. (64)).

- (73) LCS theme
a-structure *abrir* (intr.) <y>
'open' [-r]
f-structure SUBJ

4.3 Intrinsic reflexive verbs

Intrinsic reflexive verbs are assigned either the a-structure <[-o]>[-r] or <[-r]>[-r], where the [-r] argument outside the angled brackets maps onto the expletive OBJ.

The a-structure <[-o]>[-r] is exemplified by verbs like French *se désister* 'to desist', German *sich beschweren* 'to complain', or Portuguese *queixar-se* 'to complain', cf. (74).

- (74) LCS agent;
a-structure *queixar-se* <x_i> _i
'to complain' [-o] [-r]
f-structure SUBJ_i OBJ_i=REFL_{EXPL}

As expected from this feature configuration, these verbs behave like unergatives (cf. Zaenen 1993):

- (75) a. *a mulher dançada (Portuguese)
the woman danced.PART.FEM
b. *a mulher queixada (Portuguese)
the woman complained.PART.FEM

The a-structure <[-r]>[-r] is exemplified by verbs like French *s'évanouir* 'to faint', German *sich verlieben* 'to fall in love', or Portuguese *arrepender-se* 'to repent', cf. (76).

| | | | | |
|------|-------------|----------------------|--------------------|--|
| (76) | LCS | | theme _i | |
| | a-structure | <i>arrepender-se</i> | <y _i > | - _i |
| | | 'to repent' | [-r] | [-r] |
| | f-structure | | SUBJ _i | OBJ _i =REFL _{EXPL} |

These verbs are thus correctly predicted to behave like unaccusatives:

| | | | | |
|------|----|-----------------------|----------------------|--------------|
| (77) | a. | a mulher | desaparecida | (Portuguese) |
| | | the woman | disappeared.PART.FEM | |
| | | 'the missing woman' | | |
| | b. | a mulher | arrependida | (Portuguese) |
| | | the woman | repented.PART.FEM | |
| | | 'the repentant woman' | | |

In this way, intrinsic reflexive verbs, even though they constitute semantically one-place predicates, are syntactically transitive verbs similar to either reflexive verbs (cf. (62)), or decausative verbs (cf. (65)). Parallel to the decausative construction, co-indexing appears only on the a-structure and f-structure levels. However, in contrast to decausatives, there is no derivational relationship to a causative transitive verb.

4.4 Summary

We have analyzed all reflexive constructions as syntactically transitive. Consequently, the reflexive never functions as a detransitivity marker. While reflexive verbs are semantically transitive, decausative reflexive verbs and intrinsic reflexive verbs are semantically intransitive. In this case, the reflexive is an expletive OBJ. In decausatives, this non-thematic syntactic argument can be seen as a marker of decausativization. German and Romance contrast, in this way, with languages like Modern Greek and Hebrew, on the one hand, and English, on the other, where decausativization is not syntactically marked by an expletive reflexive, but it is either morphologically marked or not marked at all on the verb.

5 Conclusion

In this paper, we have discussed the argument status of the reflexive in reflexive constructions, focusing on German and Romance reflexive verbs, decausative reflexives, and intrinsic reflexives. We have argued that in these constructions, the reflexive is best analyzed as a direct object. We have shown that the main arguments in favor of the intransitivity hypothesis do not hold. In particular, we have shown that the asymmetry between reflexive verbs and non-reflexive transitive verbs in French causative constructions, which for at least three decades was taken as evidence for the intransitive status of reflexive constructions, can be explained consistently as a result of binding and linking constraints without invoking intransitivity. We have shown that the arguments in the literature in favor of the transitivity hypothesis in German can be extended to the Romance languages, and we have presented additional evidence favoring the transitivity hypothesis. In the last part of the paper, we analyzed the reflexive constructions within the LMT formalism, explaining not only why reflexives behave like direct objects, but also why some reflexive verbs pattern with unergatives, while others are similar in behavior to unaccusatives. The syntactic and interpretational properties of reflexive verbs can only be accounted for if a distinction is drawn between syntactic valency and syntactic/semantic argument structure, as is the case in LFG/LMT.

References

- Abeillé, Anne, Danièle Godard, and Philip Miller. 1997. Les causatives en français: un cas de compétition syntaxique. *Langue Française* 115. 62-74.
- Abeillé, Anne and Danièle Godard. 2003. Les prédicats complexes dans les langues romanes. In *Les Langues Romanes: Problèmes de la Phrase Simple*, ed. Danièle Godard. Paris: Editions du CNRS. 125-184.
- Alencar, Leonel Figueiredo de. 2005. A discrepância entre valência sintática e semântica nas construções anticausativas alemãs. In *Dar a Palavra à Língua em Homenagem a Mário Vilela*, ed. Graça-Rio Torto, Olívia Figueiredo, and Fátima Silva. Porto: Editorial da Faculdade de Letras da Universidade do Porto. [to appear]
- Alsina, Alex. 1996. *The Role of Argument Structure in Grammar: Evidence from Romance*. Stanford, California: CSLI Publications.
- Berman Judith und Anette Frank. 1996. *Deutsche und französische Syntax im Formalismus der LFG*. Tübingen: Niemeyer.
- Berman, Judith. 2003. *Clausal Syntax of German*. Stanford, California: CSLI Publications.
- Bierwisch, Manfred. 1996. Fragen zum Beispiel. In *Wenn die Semantik arbeitet: Klaus Baumgärtner zum 65. Geburtstag*, ed. Gisela Harras and Manfred Bierwisch. Tübingen: Niemeyer. 361-378.
- Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.
- Butt, Miriam. 1997. Complex predicates in Urdu. In *Complex Predicates*, ed. Alex Alsina, Joan Bresnan, and Peter Sells. Stanford, California: CSLI Publications. 107-149.
- Butt, Miriam, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook*. Stanford, California: CSLI Publications.
- Camacho, Roberto Gomes. 2003. Em defesa da categoria de voz média no português. *D.E.L.T.A.* 19:91-122.
- Comrie, Bernard. 1981. *Language Universals and Linguistic Typology*. Oxford: Blackwell.
- Dalrymple, Mary. 1993. *The Syntax of Anaphoric Binding*. Stanford, California: CSLI Publications.
- Eisenberg, Peter. 1999. *Grundriss der Deutschen Grammatik*. Stuttgart: J. B. Metzler.
- Falk, Yehuda N. 2001. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Stanford, California: CSLI Publications.
- Grimshaw, Jane. 1982. On the lexical representation of romance reflexive clitics. In *The Mental Representation of Grammatical Relations*, ed. Joan Bresnan. Cambridge, Massachusetts: The MIT Press. 87-148.
- Grimshaw, Jane. 1990. *Argument Structure*. Cambridge, Massachusetts: The MIT Press.
- Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In *Causatives and Transitivity*, ed. Bernard Comrie and Maria Polinsky. Amsterdam: Benjamins. 87-120.
- Helbig, Gerhard and Joachim Buscha. 1991. *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht*. Leipzig: Verlag Enzyklopädie.
- Kaufmann, Ingrid. 1995. *Konzeptuelle Grundlagen semantischer Dekompositionsstrukturen: Die Kombinatorik lokaler Verben und prädikativer Komplemente*. Tübingen: Niemeyer.
- Kaufmann, Ingrid. 2003 a. Reflexive Verben im Deutschen. In *Arbeiten zur Reflexivierung*, ed. Lutz Gunkel, Gereon Müller, and Gisela Zifonun. Tübingen: Niemeyer. 135-155.
- Kaufmann, Ingrid. 2003 b. Infinitivnominalisierungen von reflexiven Verben: Evidenz gegen Argumentstrukturvererbung? In *(A)Symmetrien – (A)Symmetries: Beiträge zu Ehren von Ewald Lang – Papers in Honor of Ewald Lang*, ed. Claudia Maienborn. Tübingen: Stauffenburg. 203-232.
- Kayne, Richard S. 1975. *French Syntax: The Transformational Cycle*. Cambridge, Massachusetts: The MIT Press.

- Kelling, Carmen. 2005. Diathèses et structure argumentale. In *Diathesen im Französischen / Les Diathèses en Français*, ed. Carsten Sinner and Georgia Veldre. Frankfurt: Peter Lang. 99-113.
- Kroeger, Paul R. 2004. *Analyzing Syntax: A Lexical-Functional Approach*. Cambridge: Cambridge University Press.
- Luís, Ana and Ryo Otaguro. 2004. Proclitic contexts in European Portuguese and their effect on clitic placement. In *Proceedings of the LFG '04 Conference*, University of Canterbury, Christchurch, New Zealand, ed. Miriam Butt and Tracy Holloway King. Stanford, California: CSLI Publications. 334-352.
- Martin, Robert. 1970. La transformation impersonnelle. *Revue de Linguistique Romane* 34:377-394.
- Mateus, Maria Helena Mira, Ana Maria Brito, Inês Duarte, and Isabel Hub Faria. 1989. *Gramática da Língua Portuguesa*. 2. ed. Lisboa: Caminho.
- Monteiro, José Lemos. 1994. *Pronomes Pessoais*. Fortaleza: EUFC.
- Oesterreicher, Wulf. 1992. SE im Spanischen: Pseudoreflexivität, Diathese und Prototypikalität von semantischen Rollen. *Romanistisches Jahrbuch* 43:237-260.
- Pittner, Karin and Judith Berman. 2004. *Deutsche Syntax: Ein Arbeitsbuch*. Tübingen: Narr.
- Reinhart, Tanya and Tal Siloni. 2004. Against the unaccusative analysis of reflexives. In *The Unaccusativity Puzzle: Explorations of the Syntax-Lexicon Interface*, ed. Artemis Alexiadou, Elena Anagnostopoulou, and Martin Everaert. Oxford: Oxford University Press. 159-180.
- Reinhart, Tanya and Tal Siloni. 2005. The Lexicon-Syntax Parameter: Reflexivization and other arity operations. *Linguistic Inquiry* 36:389-436.
- Schwarze, Christoph. 1998. A lexical-functional analysis of Romance auxiliaries. *Theoretical Linguistics* 24:83-105.
- Selig, Maria. 1998. Pseudoreflexivität im Altitalienischen: Voraussetzungen und Richtungen eines Grammatikalisierungsprozesses. In *Transitivität und Diathese in romanischen Sprachen*, ed. Hans Geisler and Daniel Jacob. Tübingen: Niemeyer. 21-42.
- Sells, Peter, Annie Zaenen, and Draga Zec. 1987. Reflexivization variation: relations between syntax, semantics, and lexical structure. In *Working Papers in Grammatical Theory and Discourse Structure: Interactions of Morphology, Syntax, and Discourse*. CSLI Lecture Notes 11, ed. Masayo Lida, Stephen Wechsler, and Draga Zec. Stanford, California: CSLI Publications. 169-238.
- Steinbach, Markus. 2002. *Middle Voice: A Comparative Study in the Syntax-Semantics Interface of German*. Amsterdam: Benjamins.
- Turley, Jeffrey S. 1999. The creation of a grammaticalization chain: The story of Spanish decausative, passive, and indeterminate reflexive constructions. *Southwest Journal of Linguistics* 18:101-138.
- Waltereit, Richard. 2000. What it means to deceive yourself: The semantic relation of French reflexive verbs and their corresponding transitive verbs. In *Reflexives: Forms and Functions*, ed. Zygmunt Frajzyngier and Traci S. Curl. Amsterdam: Benjamins. 257-278.
- Wehrli, Eric. 1986. On some properties of French clitic *se*. In *The Syntax of Pronominal Clitics. (Syntax and Semantics 19)*, ed. Hagit Borer. Orlando: Academic Press. 263-283.
- Zaenen, Annie. 1993. Unaccusativity in Dutch: Integrating syntax and lexical semantics. In *Semantics and the Lexicon*, ed. James Pustejovsky. Dordrecht: Kluwer Academic Publishers. 129-161.

HOW TO GET RID OF THE COMP

Alex Alsina
Universitat Pompeu Fabra

KP Mohanan
National University of Singapore

Tara Mohanan
National University of Singapore

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

1. Introduction

This paper argues for reducing the inventory of grammatical functions (GFs) by eliminating the GF COMP, standardly assumed in LFG to be assigned exclusively to clausal categories. We show that this move is desirable not only because it results in a simpler framework (a framework with fewer constructs), but also because it yields simpler and more perspicuous analyses.

Let us assume that phrasal categories can be classified as nominal or clausal (among other possibilities) depending on whether the lexical head of which they are a projection is a noun or a verb. Adopting Grimshaw's (1997:376) notion of *extended projection* ("a unit consisting of a lexical head and its projection plus all the functional projections erected over the lexical projection"), clausal categories would include VP, as the smallest verbal projection, and also IP and CP as extended projections of V, and, likewise, nominal categories would include NP, and also DP and PP. In this paper we focus on CP, when we refer to clausal categories (but see section 5). We note that, with respect to non-subject arguments in early LFG, the grammatical functions of nominal and clausal categories are in complementary distribution:

| | | | | |
|-------|-----|-------------------|-------------------|------|
| (1) | OBJ | OBJ _{th} | OBL _{th} | COMP |
| NP/PP | √ | √ | √ | * |
| CP | * | * | * | √ |

Motivated by this redundancy in the framework, Alsina, Mohanan and Mohanan (1996) proposed that COMP be dropped from the inventory of GFs in LFG, since all references to COMP can be replaced by direct reference to CP complements of a predicate (taking "complement" as a non-subject argument).

Dalrymple and Lødrup 2000 (henceforth D&L) argue, contra Alsina et al., that an empirically adequate account of certain languages including English ("mixed language") requires a distinction between two kinds of clausal complements, one exhibiting and the other lacking object properties. In our earlier proposal (Alsina et al. 1996), we eliminated the redundancy by eliminating COMP and assuming that CPs were always OBJ/SUBJ. Based on the empirical inadequacy of this proposal, D&L make an alternative proposal: to retain COMP and allow CPs to be either COMP or OBJ. Under this proposal, we cannot predict the GF of an argument from the category CP. However, this does not entirely eliminate the redundancy: we can still predict the category of an argument from the GF COMP. D&L do not show that the patterns they account for in terms of OBJ vs. COMP are not attributable to other distinctions already available in standard LFG, namely, OBJ vs. OBL_θ, or OBJ vs. OBJ_θ. Their arguments for the retention of COMP in the inventory are therefore incomplete.

The proposal in this paper eliminates the redundancy completely by abandoning COMP and assuming that clausal categories have the same range of complement GFs as nominal categories, as shown in (2).

| | | | |
|-------|-----|------------------|------------------|
| (2) | OBJ | OBJ _θ | OBL _θ |
| NP/PP | √ | √ | √ |
| CP | √ | √ | √ |

Our proposal also reduces the redundancy implicit in the other GFs. In standard LFG, an OBJ and an OBJ_θ are always NP; in our proposal, they can be NP or CP. This move results not only in a smaller inventory of GFs, but also in a simpler and empirically more adequate theory for predicting the relevant facts involving clausal complements in Catalan, Spanish, Malayalam, and English.

Section 2 looks at different clausal complements in Catalan and shows that, if we were to assume that clausal complements that do not behave like objects are COMPS, the description of the facts would be considerably complicated. The facts can be explained in a simple way, without the GF COMP, if we assume that obliques (OBL_θ) can alternatively be realized as PPs or CPs. The difference between Catalan and Spanish regarding the possibility of clausal complements being introduced by a preposition can be explained through the interaction of competing constraints. Section 3 investigates different types of clausal complements in Malayalam and argues that all the relevant facts can be explained by appealing to independently required semantic distinctions, making the OBJ vs. COMP

distinction unnecessary. Section 4 shows that English clausal complements are best analyzed without assuming the GF COMP. Section 5 presents the main conclusions of the paper.

2. Catalan

2.1 The Catalan facts: Two types of clausal complements

In Catalan, verbs that select a particular preposition on their complement when the complement is nominal do not allow any preposition on the complement when it is a clause. This contrast is illustrated in (3) and (4). (3a) and (4a) contain two different predicates that select a different preposition for their NP complement; in (3b) and (4b), the same predicates take a clausal complement without a preposition.

- (3) a. *M' heu de convèncer de les seves possibilitats.*
 me have-2ndPL to convince of the 3POSS possibilities
 'You have to convince me of his possibilities.'
- b. *M' heu de convèncer (*de) que torni a casa.*
 me have-2ndPL to convince of that return-1stSG to home
 'You have to convince me to return home.'
- (4) a. *Estàvem d' acord en alguns punts.*
 were-1stPL of agreement on some points
 'We agreed on certain points.'
- b. *Estàvem d' acord (*en) que ens apugessin el sou.*
 were-1stPL of agreement on that us raised-SUBJ-3rdPL the salary
 'We agreed that they should raise our salary.'

Under standard assumptions in LFG, the PP/CP alternation in Catalan would have to be explained by assuming an alternative subcategorization frame. While the PP in examples like (3a) and (4a) would clearly be an oblique, the GF of clausal complements like those in (3b) and (4b) is not so clear. There is a generally held belief that an oblique has to be overtly marked by a preposition or a semantic case marker; the clausal complement in (3b) and (4b) has neither a preposition nor a case marker. Consequently, the standard position in LFG would be that this clausal complement is not an oblique.

D&L reject the possibility that the CP complement in such examples could be an OBL, like the corresponding PP. Since OBLs are normally realized as PPs, D&L claim, we would need to posit a principle of preposition deletion for an OBL to be realized as a CP, without a P. This suggests to D&L that we would have a PP with an unexpressed head: the clausal complement in such cases would be an OBL and a PP and we would thus expect it to behave just like a PP with an overt preposition. They observe that this expectation is contradicted by certain phenomena in German that show asymmetries between the PP and the CP realization of the same argument. In addition, D&L find that positing deletion operations or unpronounced elements does not fit well with a declarative theory of grammar such as LFG. (As we shall see, it is possible to assume that a clausal complement is an OBL, without assuming that it is a PP, that there is a process or P-deletion resulting in an unexpressed P.)

Following D&L, it seems we have a theoretical choice regarding the grammatical function of the CP complement in examples like (3b) and (4b): depending on its behavior, it can either be an OBJ or a COMP. For Catalan, we can take cliticization (pronominalization by means of the appropriate verbal clitic) and passivization as objecthood diagnostics.

Cliticization: Direct objects, or expressions bearing the OBJ function and having accusative or non-dative case, are pronominalized by means of a series of clitics including the third person singular *el* and *la* (and morpho-phonologically conditioned alternants) coreferential with masculine and feminine NPs respectively, and the so-called neuter *ho*, coreferential with a proposition, such as a clause:

- (5) a. *(La teva explicació) no l' he entesa.*
 the 2ndSG-POSS explanation not pron-3rdSG-FEM have-1stSG understood-FEM
 '(Your explanation_i) I didn't understand it_i.'
- b. *(Que hakis arribat tan tard) no ho he entès.*
 that have-2ndSG arrived so late not pron-3rdSG-PRO have-1stSG understood
 '(That you should have arrived so late_i) I didn't understand it_i.'

If we try to pronominalize the clausal complements of (3b) and (4b) by means of *ho*, we get ungrammatical results:

- (6) a. * *(Que torni a casa) ho heu de convèncer en Martí.*
 that return-1stSG to home pron-3rdSG-PROP have-2nd-PL of convince the Martí
 '(To return home) you have to convince Martin.'
- b. * *(Que ens apugessin el sou) ho estàvem d'acord.*
 that us raised-SUBJ-3rdPL the salary pron-3rdSG-PROP were-1stPL of agreement
 '(That they should raise our salary) we agreed on.'

The contrast between (5b) and (6) can be explained by assuming that the clitic *ho* functions as an OBJ coreferential with a proposition. Since the verb in (5) takes an OBJ, this OBJ can be encoded as the clitic *ho* when it is coreferential with a proposition. If we assume, on the other hand, that the verbs or predicates in (6) do not take an OBJ, but can instead take a COMP, there is no OBJ in these sentences that can be encoded as the clitic *ho*.

Passivization The verbs that can passivize are a subset of those that take OBJ in the active form. Not all verbs that take an OBJ can passivize: for example, possessive *tenir* 'have' or stative *pesar* 'weigh' take an OBJ in the active form, but cannot passivize. But the observation that only verbs that take an OBJ in the active form can passivize seems to be correct. Thus, *entendre* 'understand' (see (5)) can passivize with its clausal complement as the subject, as shown in (7a), whereas the clausal complement of *convèncer* and *estar d'acord* cannot be the passive subject of these verbs, as in (7b-c):

- (7) a. *Que votessin a favor de la proposta no va ser entès per una part del públic.*
 that vote-SUBJ-2ndSG in favor of the proposal not PAST-3rdSG be understood
 by a part of-the audience
 'That you should have voted in favor of the proposal was not understood by part of the audience.'
- b. * *Que tornés a casa va ser convençut en Martí.*
 that return-SUBJ-3rdSG to home PAST-3rdSG be convinced the Martí
 'That he return home was convinced Martín.'
- c. * *Que ens apugessin el sou va ser estat d'acord per tothom.*
 that us raised-SUBJ-3rdPL the salary PAST-3rdSG be been of agreement by everyone
 'That they should raise our salary was agreed on by everybody.'

The cliticization facts and the passivization facts shown above would follow from the assumptions below, using the OBJ vs. COMP distinction:

- clausal complements can be an OBJ or COMP.
- the clitic *ho* satisfies the OBJ function, but not the COMP function, and
- an argument can alternatively take the OBJ and SUBJ functions, by being specified as [-r].
- but an argument assigned the COMP function cannot alternately take the SUBJ function.

If we assume that the clausal complement of *entendre* is OBJ while that of *convèncer* and *estar d'acord* is COMP, it follows from the above that (i) the former, not the latter, permits the clitic *ho*, and (ii) the argument assigned the OBJ function in the active form of *entendre* can be assigned the SUBJ function in its passive form. Since the clausal complement of *convèncer* and *estar d'acord* is not an OBJ function, it cannot be assigned the SUBJ function in the passive form.

The proposal that a clausal complement can be either an OBJ or a COMP depending on the governing predicate seems to account for the facts observed so far. However, it is, in fact, an obstacle for an adequate description when a fuller range of facts is taken into consideration.

2.2 Four problems for the analysis using OBJ vs. COMP

Problem 1: The most important problem that this analysis encounters is the fact that the clausal complements we have just designated as COMPs alternate with either *en* or *hi*, both pronominal clitics, depending on the governing predicate. Thus, the clausal complement of *convèncer* can be expressed as *en*, whereas the clausal complement of *estar d'acord* can be expressed as *hi*:

- (8) a. *Me n' heu de convèncer.* b. *Hi estàvem d' acord.*
 me EN have-2nd-PL of convince HI were-1stPL of agreement
 ‘You have to convince me of that.’ ‘We agreed on that.’

The two pronominal clitics are not interchangeable and, so, replacing the one by the other in (8) creates ungrammatical structures. Thus, it seems we need to assume some abstract feature of the theory to distinguish those clausal complements expressible by *en* and those expressible by *hi*. Following the argument in the preceding paragraphs according to which we posit a GF distinction (OBJ vs. COMP) to predict which clausal complements are expressible by the clitic *ho* and which are not, we might want to posit two different GFs — say, COMP1 vs. COMP2 — to predict which clausal complements are expressible by the clitic *en* and which by the clitic *hi*.

If we took this approach, we would have to say that, for example, COMP1 can be encoded as the clitic *en* and COMP2 can be encoded by the clitic *hi*. Thus, a predicate like *convèncer* would have two alternative subcategorization frames: (a) <SUBJ OBJ OBL_{de}> and (b) <SUBJ OBJ COMP1>. And a predicate like *estar d'acord* would have the following two: (a) <SUBJ OBL_{en}> and (b) <SUBJ COMP2>. In both cases, the arguments involved are the same, so that the argument that can be realized as an OBL of *convèncer* can also be realized as a COMP1. We would also need to say that the clitic *en* can satisfy the OBL_{de} or COMP1 functions and that the clitic *hi* can satisfy the OBL_{en} or COMP2 functions.

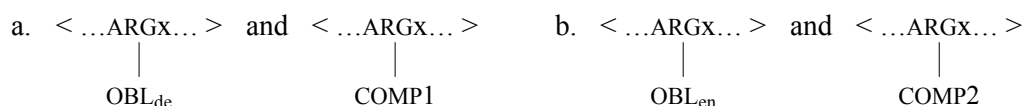
Problem 2: The COMP function (or, given the facts presented here, the COMP1 and COMP2 functions) is never the only possible GF that a given argument can bear, at least, in Catalan. All arguments that can be assigned a COMP1 or COMP2 function also have an alternative assignment of either OBL_{de} or OBL_{en}. In the approach that assumes the COMP function, this obligatory alternation is hard to explain and has to be stipulated for every predicate that subcategorizes for this function.

Furthermore, unlike what happens with other grammatical function alternations, such as the SUBJ-OBJ alternation in active-passive pairs, no verbal morphology is involved in the COMP1-OBL_{de} or COMP2-OBL_{en} alternation. Also, unlike what happens with the causative alternation in English, which also involves a SUBJ-OBJ alternation, this alternation does not involve any semantic difference. It is just a free alternation that depends on no feature or property of the governing predicate.

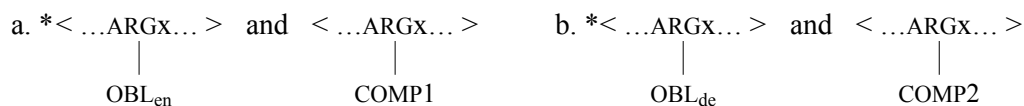
Problem 3: Whether the clausal complement of a particular verb is replaceable by *en* or by *hi* is not independent of the alternative forms of expression that the argument in question may have. Specifically, if an argument can be realized as a PP introduced by the preposition *de* or as a clausal complement, it can also be realized as the clitic *en*. And if an argument can be realized as a PP introduced by the prepositions *a*, *en* or *amb* or as a clausal complement, it can also be realized as the clitic *hi*. The PP/CP alternation is shown in (3) and (4) for the two classes of verbs and the possibility of the argument being expressed by means of the appropriate clitic is illustrated in (8).

If we tried to express this correlation in terms of the grammatical functions posited in the preceding paragraphs, we would somehow have to say that a predicate can have the alternative subcategorization frames in (9) but not those in (10):

(9) Possible alternative assignments of GF to a given argument:



(10) Impossible alternative assignments of GF to a given argument:



If verbs exemplifying (10a) existed in Catalan, they would require the preposition *en* on NP complements and the clitic *en* for clausal complements. (10b) corresponds to verbs that would take *de* on NP complements but the clitic *hi* for clausal complements. The non-existence of such verbs suggests that multiplying the number of GFs is not perhaps the right way to go.

The distributional patterns illustrated in (9) and (10) do not follow from anything in the theory and it is hard to imagine what kind of principle would explain the idea that a given argument can alternatively be assigned the GFs OBL_{de} and COMP1, but not the GFs OBL_{de} and COMP2, or that it can alternatively be assigned the GFs OBL_{en} and COMP2, but not the GFs OBL_{en} and COMP1. Thus, the facts—or, rather, the artifacts—shown in (9) and (10) are a problem for a theory that attempts to explain the properties of clausal complements that are not OBJ by assuming that they bear a special grammatical function such as COMP (or, possibly, COMP1 and COMP2).

Problem 4: Positing the COMP function (or COMP1 and COMP2) does not by itself explain why, in Catalan, the oblique functions (OBL_{de} or OBL_{en}) cannot be realized as a PP in which the head preposition is followed by an S (or CP). In order to account for the ungrammaticality of (11), we need to posit a constraint or principle excluding such structures.

- (11) a. * *M' heu de convèncer de que torni a casa.*
 me have-2ndPL to convince of that return-1stSG to home
 'You have to convince me to return home.'
- b. * *Estàvem d'acord en que ens apugessin el sou.*
 were-1stPL of agreement on that us raised-SUBJ-3rdPL the salary
 'We agreed that they should raise our salary.'

This could be achieved either with c-structure rules that do not generate structures such as [_{PP} P CP], but only generate PPs in which the P precedes an NP, or by allowing c-structure rules to generate these structures but then having a constraint ruling them out. Either of these options has a language-particular component, since there are languages that do allow [_{PP} P CP]. Spanish is an example of such a language, as we shall see shortly.¹

Summary: Positing COMP as part of the inventory of GFs and assuming it to be the GF assigned to clausal complements that cannot be analyzed as OBJ leads to a very complex description of the facts in Catalan. The complications are the following: we need to (a) make a distinction between COMPs expressed by means of the clitic *en* and those expressed by means of the clitic *hi* (possibly as COMP1

¹ Interestingly, Catalan also allows this structure when the preposition is not a governed preposition, but is, instead, the head of an adjunct phrase. *Sense* 'without' is an example of a preposition that can be used to introduce a CP, as in *sense que ho sabés* 'without his knowing it.' The idea is that in Catalan, a P is allowed to precede a CP when it heads an adjunct, but not when it heads a complement. If the c-structure rules did not generate the P-CP structure, we would incorrectly rule out this kind of adjuncts. In the OT approach to be presented later, an explanation is possible for the observation that the P-CP structure is possible, although only for adjuncts. When the "offending" preposition is governed, it can be left out because its features are recoverable from the governing predicate; when it is not governed, its features are not recoverable from the governing predicate and, so, it cannot be left out without losing semantically relevant information.

and COMP2 or in some other way); (b) stipulate that COMP1 alternates with OBL_{de} but not with other OBLs and that COMP2 alternates with certain OBLs including OBL_{en} but not with OBL_{de}; (c) stipulate that the two COMP functions (COMP1 and COMP2) always alternate with an OBL functions; and (d), in addition to positing the two COMP functions, assume a principle ruling out the structure [_{pp} P CP].

2.3 A solution without COMP

An adequate explanation of the Catalan facts involving clausal complements that avoids the problems just noted requires assuming the following:

- A predicate like *convèncer* in Catalan consistently maps its third argument (call it the theme) onto an OBL function, whether it is expressed as a PP, a CP or a pronominal clitic such as *en*.
- Whereas some languages allow a governed preposition to immediately precede a CP, as in Spanish, some languages do not, as in Catalan.
- If a language rejects the [_{pp} P CP] structure, an OBL function may correspond to a CP, without null or empty prepositions or headless PPs.
- Predicates like *convèncer* and *estar d'acord* require a specific case feature on their oblique argument. This case feature is carried by (or realized by) the appropriate preposition or the appropriate pronominal clitic.

The facts that need explaining can be given a simple explanation consistent with these assumptions by adopting an Optimality-Theoretic view of constraint interaction. Let us assume the existence of two universal constraints: a markedness constraint rejecting a PP consisting of a P and a CP, and a faithfulness constraint requiring the features in the f-structure (the output) to correspond to lexically specified features (the input) in the corresponding c-structure.

(12) **No P+CP:** star a structure containing the c-structure tree [_{pp} P CP].

(13) **C-to-F Faithfulness:** features in the f-structure must be lexically specified by the elements in the c-structure.

C-to-F Faithfulness is violated whenever a feature in a given f-structure is not part of the information carried by the set of lexical items in the c-structure that map onto that f-structure. As we shall see, different rankings of the two constraints (12) and (13) yield languages that have CPs introduced by prepositions, complying with (13) but not (12), such as Spanish, and languages that do not have CPs introduced by prepositions, complying with (12), but not (13), such as Catalan and English.

Another important constraint that we need to take into account is what we might call *Completeness*, formulated as follows:²

(14) **Completeness:** the requirements of argument structure must be satisfied in the f-structure.

This constraint requires that any feature that argument structure specifies must be found in the f-structure. We will assume that Completeness is a high-ranking constraint, and for present purposes, that there are no well-formed structures that violate it. The relevance of this constraint is apparent when we consider predicates like *convèncer* or *estar d'acord*, which require that one of their arguments bear a particular case feature.

Let us assume that the argument structure of *convèncer* is as follows:

(15) 'convince < [Ext] [Int] [CASE gen] >'

Implicit in this representation is the idea that arguments are ordered by prominence in the argument structure and are classified according to certain features such as "Ext" (designating the external

² This formulation of Completeness diverges formally from formulations of Completeness available in the literature (such as Bresnan 2001: 72: "every GF designated by a PRED must be present in the f-structure of that PRED."), although it plays a similar role. Here Completeness is taken to constrain the mapping between argument structure, represented as the value of the feature PRED, and grammatical functions.

argument) or “Int” (designating an internal argument). While this is not particularly relevant to the analysis being developed here, what is important for this analysis is the idea that certain arguments are required to have a particular case feature in their corresponding f-structure. In the example in (15), the third argument is required to have the feature [CASE gen] (genitive case). This means that the f-structure corresponding to this argument must include the feature, failing which Completeness would be violated. Such a case specification on an argument can be taken to be like a constraining equation: the corresponding f-structure must include the specified feature to satisfy Completeness.

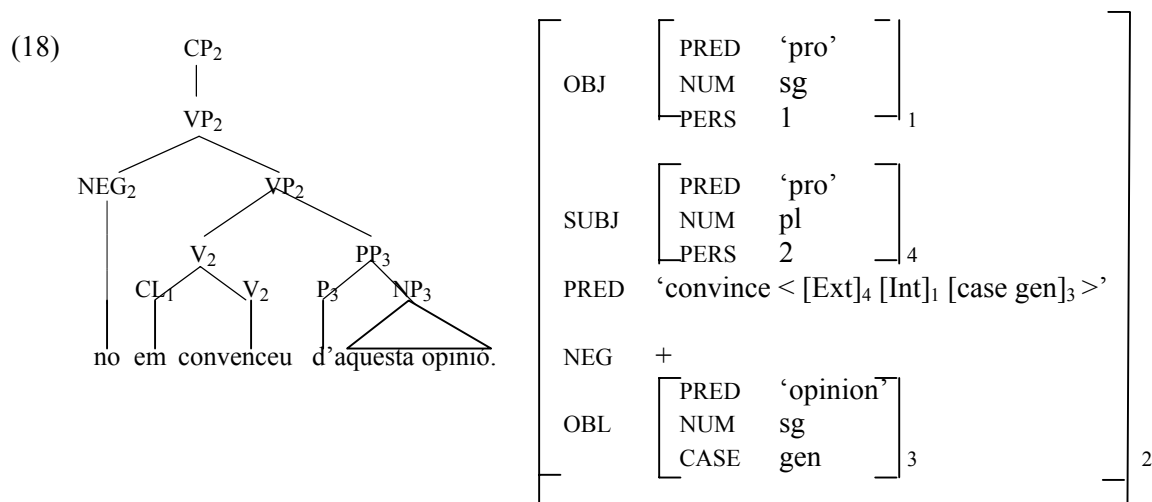
The case feature specified in (15) in Catalan may be provided by two different vocabulary items: the preposition *de* and the pronominal clitic *en*. The relevant lexical entries are given in (16):

- (16) a. *de*: P [CASE gen]
 b. *en*: cl $\begin{bmatrix} \text{PRED} & \text{pro} \\ \text{CASE} & \text{gen} \end{bmatrix}$

Given Completeness, a predicate with the argument structure in (15) requires the f-structure corresponding to its third argument to have the feature [CASE gen]. In order to respect C-to-F Faithfulness, the presence of this feature in an f-structure requires its c-structure correspondent to include a constituent that has this information in its lexical entry. This means that one of the two lexical entries in (16) has to be used in order to provide the required case feature for the third argument of (15). If the preposition in (16a) is used, it projects a PP structure and the PP occupies the expected positions for a PP. If the clitic is used, its position is that of a verbal affix: it attaches to the appropriate verb along with other clitics, if there are any.

The pair of c- and f-structures corresponding to an example like (17), in which the PP structure is used, are as in (18):

- (17) *No em convenceu d'aquesta opinió.*
 not me convince-2nd PL of this opinion
 ‘You are not convincing me of this opinion.’



The correspondence between the c-structure and the f-structure of the same expression is notated by subscripting each c-structure node and the corresponding f-structure with the same index; likewise regarding the correspondence between argument structure and grammatical functions. The correspondence between the c-structure and the f-structure is not regulated by annotations on the c-structure or the c-structure rules, but by general correspondence principles between the two structures (see Alsina 1996:21-34 for a proposal along these lines).

Both c-structure and f-structure are subject to well-formedness conditions applying internally to each structure; and the pairing of c-structure and f-structure is subject to the appropriate mapping constraints. For example, a well-formedness condition involved in the f-structure in (18) states that the GF attribute whose value is an f-structure containing the CASE feature ‘gen’ (and other “semantic”

case values such as ‘en’ or ‘a’) is OBL. By this condition, the structure [OBL [CASE gen]] is well-formed, whereas other possible structures, such as [OBJ [CASE gen]] or [SUBJ [CASE gen]], are not.

Some of the relevant constraints and conditions are those given in (12)-(14). If we evaluate the paired structures in (18) in relation to these three principles, we see that those structures satisfy the three principles. Completeness (14) is satisfied, because all of the requirements of the argument structure in (15) are met in its f-structure, particularly the requirement that its third argument have the case feature ‘gen’. C-to-F Faithfulness (13) is also satisfied, at least with respect to this argument, since this required case feature is contributed by a lexical item in the c-structure, which has this information as part of its lexical entry. And the c-structure constraint (12), No P+CP, is also satisfied because the c-structure does not contain the offending structure.

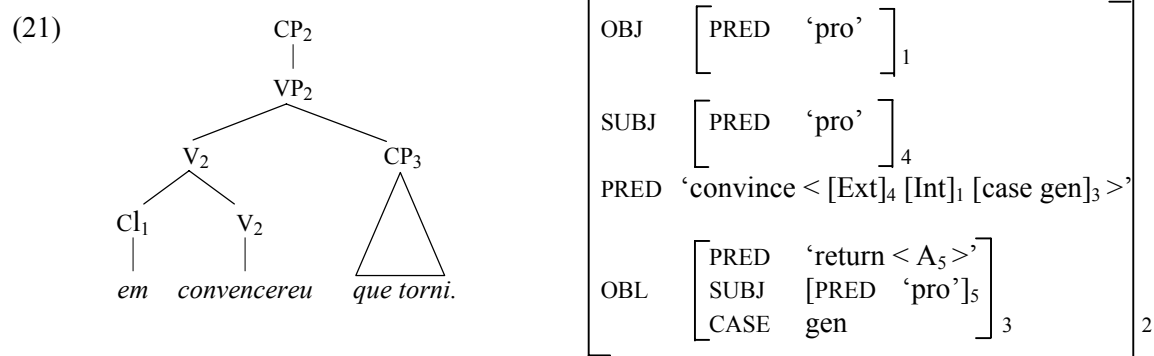
The relevance of the Optimality-Theoretic conception of constraint interaction becomes apparent when one of the constraints proposed cannot be satisfied. This situation arises when an argument, such as the third argument of *convèncer* in Catalan, is realized in the c-structure by a CP. In order to satisfy the markedness constraint (12), the c-structure cannot include the offending P+CP structure. Yet, in order to satisfy the faithfulness constraint (13), the c-structure must include the preposition *de* introducing the CP complement, since this preposition is the only lexical item that has this feature and can be part of a phrasal constituent mapping onto the required OBL function. On the assumption that Completeness must be satisfied and therefore the GF mapping onto the third argument of *convèncer* must include the required case feature, either the markedness constraint (12) or the faithfulness constraint (13) must be violated. Let us assume that the relative ranking of these two constraints in languages that do not allow a CP to be introduced by a governed preposition, such as Catalan and English, is as shown in (19):

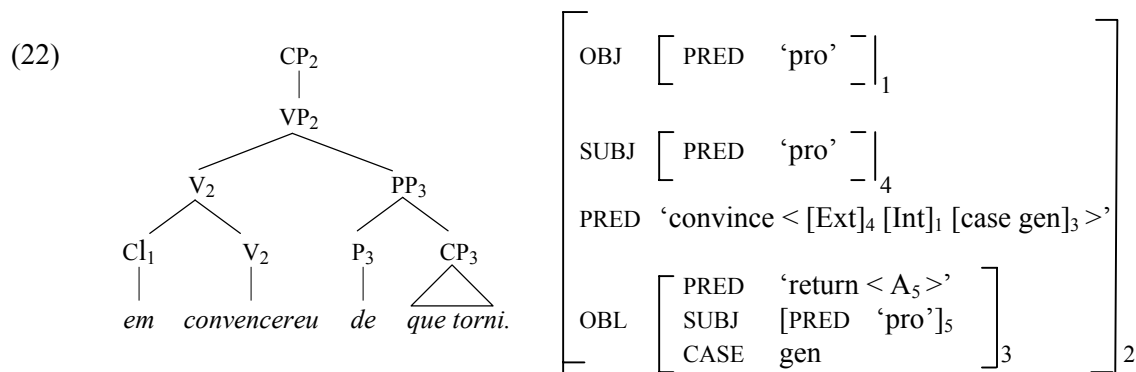
(19) Constraint ranking in Catalan and English: (12) » (13)

The consequence of this ranking is that the faithfulness constraint (13) can be violated in order to avoid the [_{PP}P CP] structure rejected by (12). Therefore, the OBL complement of a verb like *convèncer*, when it is clausal, cannot be realized by a PP, but by a CP (without a preposition). Let us consider the contrast in the Catalan examples in (20):

(20) a. *Em convencereu que torni.* b. * *Em convencereu de que torni.*
 me convince-FUT-2ndPL that return-1stSG
 ‘You will convince me to come back.’

The corresponding structures (ignoring some amount of irrelevant information in the f-structures) are (21) for (20a) and (22) for (20b):





The two f-structures are identical, in spite of the difference in c-structure, resulting from the presence of the preposition *de* in (22) and its absence in (21). Although there is no preposition *de* in (21) to license the case feature of the oblique (clausal complement), a competitor of (21) that lacked the feature [CASE gen] would violate Completeness. There is no reason to posit an empty P or a PP node in (21), as it would not add anything to the structure: such additional structure would be ruled out by a principle such as Economy of Expression (Bresnan 2001). Let us see how the two paired structures (21) and (22) are evaluated taking into account the two relevant constraints (12) and (13):

(23) Competition between (21) and (22), in Catalan:

| | Completeness | (12) No P+CP | (13) C-to-F Faith |
|---------|--------------|--------------|-------------------|
| a. (21) | | | * |
| b. (22) | | *! | |

Thus, given the ranking of constraints (12) and (13) proposed for Catalan, the structure in which the oblique clausal complement is expressed as a CP without a preposition, (21), emerges as the optimal structure. Consequently, an analysis that does not assume the GF COMP and does not exclude the possibility that a CP may map onto the GF OBL is possible for the facts under consideration, and is much simpler than an analysis that assumes that non-object clausal complements are COMPs.

2.4 The Spanish facts

Unlike Catalan, Spanish allows clausal complements to be introduced by a preposition. In fact, it requires the governed preposition in these contexts, as the following Spanish examples illustrate, where (24b) and (25b) contain a preposition followed by a CP:

(24) a. *Lo tenéis que convencer de sus posibilidades.*
 him have-2ndPL to convince of 3POSS possibilities
 ‘You have to convince him of his possibilities.’

b. *Lo tenéis que convencer *(de) que vuelva a casa.*
 him have-2ndPL to convince of that return-3rdSG to home
 ‘You have to convince him to return home.’

(25) a. *Estábamos de acuerdo en algunos puntos.*
 were-1stPL of agreement on some points
 ‘We agreed on certain points.’

b. *Estábamos de acuerdo *(en) que nos subiesen el sueldo.*
 were-1stPL of agreement on that us raised-SUBJ-3rdPL the salary
 ‘We agreed that they should raise our salary.’

Following the ideas in D&L, one would have to say that Spanish differs from Catalan not only in lacking the constraint prohibiting the [_{PP}P CP] structure, but also in not having verbs that subcategorize for COMP. Thus, there would be two parameters of variation distinguishing Spanish from Catalan: whether or not the constraint No P+CP is active, and whether or not the language has COMP (or COMP1

and COMP2). This second difference is not just stated once in the grammar (for instance, as the presence or absence of a constraint or as a difference in constraint ranking), but in fact is distributed throughout the lexicon. We need to stipulate for every lexical entry of a predicate that takes an OBL that can be realized by a CP (whether preceded by a preposition, as in Spanish, or not, as in Catalan) that the predicate in Spanish takes an argument that bears an OBL function not alternating with any other function and that the corresponding predicate in Catalan takes an argument that alternatively bears an OBL function and a COMP function.

Given the Optimality Theory approach taken above, all we need to assume is that Spanish differs minimally from Catalan (and English) regarding clausal complements in having a different ranking of the relevant constraints ((12) and (13)). Instead of (19), Spanish has the following constraint ranking:

(26) Constraint ranking in Spanish: (13) » (12)

As a consequence of this ranking, clausal complements bearing the OBL function in Spanish must be PPs: the preposition allows the structure to satisfy the faithfulness constraint (13), even though it violates the markedness constraint (12).

Let us compare the Catalan examples (20) with the equivalent examples in Spanish, which would have the same f-structures as the corresponding Catalan examples and whose c-structure is partially indicated in (27). The tableau showing the competition between these two structures is given in (28).

(27) a. * Me convenceréis [CP que vuelva]. b. Me convenceréis [PP [P de] [CP que vuelva]].
 me convince-FUT-2ndPL that return-1stSG
 ‘You will convince me to come back.’

(28) Competition between (27a) and (27b), in Spanish:

| | Completeness | (13) C-to-F Faith | (12) No P+S |
|----------|--------------|-------------------|-------------|
| a. (24a) | | *! | |
| b. (24b) | | | * |

Thus, the simple re-ranking of two highly motivated constraints allows us to explain the fact that oblique functions alternate as PPs and CPs in Catalan and are consistently expressed as PPs in Spanish. This allows us to dispense with the grammatical function COMP, which, as we saw earlier, creates massive complications for the description of the facts.

3. Malayalam

3.1 Three declarative finite clause constructions

Our second source of evidence for abandoning the GF COMP comes from Malayalam, which has three declarative finite subordinate clause constructions schematically given in (29a-c):

(29) a. S-enn̄ḍ b. S-ennat̄ḍ c. S-at̄ḍ
 ... that ... that it ... it

The three types of clauses in (29) are illustrated in (30a-c) respectively:

(30) a. [iwiTe weLLam unD̄ḍ enn̄ḍ] enik'k'ḍ ariyunnuNDaayirunnilla.
 here water is that I-DAT know-PAST-NEG
 I didn't know that there is water here.

b. [iwiTe weLLam unD̄ḍ enn̄at̄ḍ] enik'k'ḍ ariyunnuNDaayirunnilla.
 here water is that-it I-DAT know-PAST-NEG
 I didn't know (the fact) that there is water here.

c. [iwiTe weLLam uLLat̄ḍ] enik'k'ḍ ariyunnuNDaayirunnilla.
 here water is-it I-DAT know-PAST-NEG
 I didn't know about there being water here.

Complement selection involves these constructions, as illustrated in (31a-c). The examples provide a glimpse into some of the distributional restrictions that the constructions exhibit:

- (31) a. [kuTTi ciriccuwenn̄] amma wicaariccu / *kaNTu / *niSeedhiccu.
 child smiled-that mother thought saw denied
 Lit: Mother thought/ *saw/ *denied that the child smiled. (Mother thought that the child smiled.)
- b. [kuTTi ciriccuwennat̄] amma *wicaariccu / *kaNTu / niSeedhiccu.
 child smiled-that-it mother thought saw denied
 Lit: Mother *thought/ *saw/ denied it that the child smiled. (Mother denied that the child smiled.)
- c. [kuTTi ciriccat̄] amma *wicaariccu / kaNTu / *niSeedhiccu.
 child smiled-it mother thought saw denied
 Lit: Mother thought/ *saw/ *denied it the child smiled. (Mother saw the child smiling.)

Of the three types of clausal complements, the verb *wicaarikk* ‘think’ takes only S-*that* ((31a)), *niSeedhikk* ‘deny’ takes only S-*that-it* ((31b)), and *kaaN* ‘see’ takes only S-*it* ((31c)).

3.2. An OBJ-vs-COMP analysis

One way to account for these asymmetries would be to assume, à la D&L, that the verbs in (31) select different grammatical functions as their complements. For instance, one may stipulate that the verbs in (31b) and (31c) take OBJ, while that in (31a) selects COMP. As the S-*that* clause in (31a) is a CP, while both the S-*that-it* clause ((31b)) and S-*it* clause ((31c)) are NPs (or DPs), this suggestion appears quite reasonable. And its plausibility increases when we see that S-*that* clauses cannot be the object of a postposition ((32a), unlike S-*that-it* and S-*it* (32b) and (32c):

- (32) a. * [kuTTi ciriccuwenn] - ineppatti * [kuTTi ciriccuwenn̄] koNT̄
 child smiled-that about child smiled-that because of
- b. [kuTTi ciriccuwennat] - ineppatti [kuTTi ciriccuwennat̄] koNT̄
 child smiled-that-it about child smiled-that-it because of
 about (the statement) that the child smiled because of (the statement) that the child smiled
- c. [kuTTi ciriccat] - ineppatti [kuTTi ciriccat̄] koNT̄
 child smiled-it about child smiled-it because of
 about the child’s smiling because of the child’s smiling

An analysis based on the distinction between OBJ and COMP appears further confirmed by examples of passives ((33)), and of subject clauses (((34))):

- (33) a. * [kuTTi ciriccuwenn̄] ammayaal wicaarikkappeTTu.
 child smiled-that by the mother think-PASS-PAST
- b. [kuTTi ciriccuwennat̄] ammayaal niSeedhikkappeTTu
 child smiled-that-it by the mother deny-PASS-PAST
 That the child smiled was denied by the mother.
- c. ? [kuTTi ciriccat̄] ammayaal kaaNappeTTU
 child smiled-it by the mother see-PASS-PAST
 That the child smiled was seen by the mother.
- (34) a. * [kuTTi ciriccuwenn̄] ammaye santooSippiccu
 child smiled-that mother-ACC happy-CAUSE-PAST
- b. [kuTTi ciriccuwennat̄] ammaye santooSippiccu
 child smiled-that-t mother-ACC happy-CAUSE-PAST
 That the child smiled pleased the mother.
- c. [kuTTi ciriccat̄] ammaye santooSippiccu
 child smiled-it mother-ACC happy-CAUSE-PAST
 That the child smiled pleased the mother.

The asymmetry in (32) would follow from the assumption that in Malayalam, a P cannot take a COMP, and the facts of passives in (33) from the assumption that a COMP cannot alternate with a SUBJ. The asymmetry in (34) can be explained by assuming that Malayalam disallows CP subjects. This can be generalized as a constraint that CPs cannot be associated with SUBJ or OBJ (unrestricted [-r] functions).

3.3 Problems for the OBJ vs. COMP analysis

The contrast in Malayalam between the *S-that* clauses on the one hand, and the *S-that-it* and *S-it* clauses on the other, are eminently amenable to an analysis in terms of the GF distinction between OBJ and COMP. However, the analysis breaks down when we explore a bit further. We first note that it is indeed possible for some instances of *S-that* clauses to be subjects, as in (35):

- (35) a. [kuTTi ciriccuwennō] naataake parannu.
 child smiled-that throughout the land spread
 That the child smiled spread throughout the land.
- b. [kuTTi ciriccuwennō] urappaNō.
 child smiled-that certain is
 That the child smiled is certain.

Likewise, the idea that the non-passivizability of *S-that* and the passivizability of *S-that-it* and *S-it* stem from their being instances of COMP and OBJ respectively breaks down in (36)-(38):

- (36) a. [kuTTi ciriccuwennō] amma sthaapiccu
 child smiled-that mother establish-PAST
 Mother established that the child smiled.
- b. [kuTTi ciriccuwennō] ammayaal sthaapiikkappeTTu.
 child smiled-that by the mother establish-PASS-PAST
 That the child smiled was established by the mother.
- (37) a. [kuTTi ciriccuwennatō] amma nyuuspepparil kaNTu
 child smiled-that-it mother newspaper-in see-PAST
 Mother saw in the newspaper that the child (had) smiled.
- b. * [kuTTi ciriccuwennatō] ammayaal nyuuspepparil kaaNappeTTu
 child smiled-that-it by the mother newspaper-in see-PASS-PAST
- (38) a. [kuTTi ciriccatō] amma kaNNaaTiyil kaNTu
 child smiled-it mother mirror-in see-PAST
 Mother saw the child having smiled in the mirror.
- b. ? [kuTTi ciriccatō] ammayaal kaNNaaTiyil kaaNappeTTu
 child smiled-it by the mother mirror-in see-PASS-PAST

The application of D&L's other diagnostics also reveal clusterings that fail to converge on the OBJ-COMP distinction. The pronoun *atō* 'it' can occur as the complement of any of the verbs in (31), as in (39), taking as its antecedent any of the three clause types:

- (39) amma atō wicaariccu / kaNTu / niSeedhiccu
 mother it thought/ saw/ denied
 Mother thought/saw/denied **it**.

For each clause type, it can only conjoin with a clause of the same type. In (40d-f), the coordinate structures are unacceptable when the conjuncts differ in clause type, regardless of their order.

- (40) a. [kuTTi ciriccuwennum] [amma karaññuwennum]
 child smiled-that-and mother cried-that-and
 That the child smiled and the mother cried
- b. [kuTTi ciriccuwennatum] [amma karaññuwennatum]
 child smiled-that-it-and mother cried-that-it-and
 (The facts) that the child smiled and the mother cried

| | | | | |
|------|---|--------------------------|---------------|--------------------------|
| c. | [<i>kuTTi</i> | <i>ciriccatum</i>] | [<i>amma</i> | <i>karaññatum</i>] |
| | child | smiled-it-and | mother | cried-it-that |
| | The child's smiling and the mother's crying | | | |
| d. * | [<i>kuTTi</i> | <i>ciriccuwennum</i>] | [<i>amma</i> | <i>karaññuwennatum</i>] |
| | child | smiled-that-and | mother | cried-that-it-and |
| e. * | [<i>kuTTi</i> | <i>ciriccuwennum</i>] | [<i>amma</i> | <i>karaññatum</i>] |
| | child | smiled-that-and | mother | cried-it-and |
| f. * | [<i>kuTTi</i> | <i>ciriccuwennatum</i>] | [<i>amma</i> | <i>karaññatum</i>] |
| | child | smiled-that-it-and | mother | cried-it-and |

To summarize:

- Distribution within a PP suggests an OBJ-COMP distinction between *S-that-it* and *S-it* on the one hand, and *S-that* on the other.

However, passivization, replaceability with a pronoun, and conjoinability show otherwise:

- some instances of all three clause types can occur as the SUBJ of a passive, others cannot;
- all three clause types can be replaced with a pronoun; and
- no clause type can be conjoined with another type.

Clearly, it is hard to tell a coherent OBJ-COMP story for the three clausal complement constructions.

3.4 The semantics of *S-that*, *S-that-it* and *S-it*

We would like to suggest a different story that assumes all three constructions to be instances of OBJ. To make this move, we draw on independently required semantic distinctions, which make the distinction between OBJ and COMP redundant.

Our appeal to semantics is in the spirit of Kiparsky and Kiparsky's (1971) proposal for [fact] as a semantic construct that interacts with syntactic patterns. Central to their proposal are two points:

- Verbs like *know* and *regret*, unlike verbs like *believe* and *consider*, carry the presupposition on the part of the speaker that the proposition expressed by the clausal complement is true; and
- Verbs that carry this presupposition disallow the subject-to-object raising construction.

Thus, rather than making a lexical stipulation directly in terms of syntax ((41a)), Kiparsky and Kiparsky state a general constraint on the syntax-semantics pairing ((41bii)), alongside the independently required lexical specification of the semantics of the verb ((41bi)):

- (41) a. *regret*: does not permit S-to-O raising.
- b. (i) *regret*: presupposition that the PROP of the clausal complement is true.
(ii) Constraint: Factive verbs do not allow S-to-O raising.

As we will see below, an investigation of the semantics of the three clausal complement constructions in Malayalam, though unrelated to the issue of raising, reveals the following generalizations:

- (42) a. *S-that* expresses a propositional function (i.e., a proposition, question, request, or wish);
b. *S-that-it* expresses (i) a [+def] proposition, along with
(ii) the presupposition that the proposition is true (=factive); and
c. *S-it* expresses an event/situation.

The *S-it* clause: Consider the following examples:

- (43) a. [*kuTTi* *aanaye* *nuLLi* *ennat̪*] *s'ariyalla*
child elephant-ACC pinched that-it right-is-NEG
It is not true that the child pinched the elephant.
(Not: It is wrong of the child to have pinched the elephant.)

- b. [kuTTi aanaye nuLLiyat̪] s'ariyalla
 child elephant-ACC pinched-it right-is-NEG
 It is wrong of the child to have pinched the elephant.
 (Not: It is not true that the child pinched the elephant.)

The word *s'ari* 'right' is ambiguous between an epistemic interpretation (true) and a pragmatic/moral interpretation (appropriate). Only the epistemic interpretation is available for (43a) as *S-that-it* expresses a proposition. In contrast, only the pragmatic/moral interpretation is available for (43b) as *S-it* expresses an event/situation. The examples given below highlight this contrast further:

- (44) a. [iwiTe weLLam unD̪ ennat̪] satyam / nuNa aaN̪
 here water is that-it truth / falsehood is
 (The statement) that there is water here is true/false.
- b. * [iwiTe weLLam uLLat̪] satyam / nuNa aaN̪
 here water is-it truth / falsehood is
 There being water here is true/false.

Truth and falsity apply to propositions, not events/situations. The unacceptability of (44b) follows from its embedded clause being an event/situation rather than a proposition.

The S-that and S-that-it clauses: The following examples show that an *S-that* clause allows the embedded clause to express not only assertions, but also questions, wishes, and requests:

- (45) a. [kuTTi rooD̪ kroos ceytuwoo ennat̪] enikk̪ ariyilla.
 child road cross do-PAST-QUES that to me know-NEG
 I don't know if the child crossed the road.
- b. [aar̪ rooD̪ kroos ceytu ennat̪] enikk̪ ariyilla.
 who road cross do-PAST that to me know-NEG
 I don't know who crossed the road.
- c. [kuTTi onn̪ rooD̪ kroos ceytenkil ennat̪] amma moohiccu.
 child one road cross do-past-IF that mother wished
 Lit: The mother wished that 'if only the child would cross the road!'
- d. [kuTTi onn̪ rooD̪ kroos ceyyu ennat̪] amma paraññu.
 child one road cross do-IMP that mother said
 Lit: The mother said that 'Child, please cross the road!'

This latitude is not available to *S-that-it* clauses, as shown in (46):

- (46) a. * [kuTTi rooD̪ kroos ceytuwoo ennat̪] enikk̪ ariyilla.
 child road cross do-PAST-QUES that-it to me know-NEG
 I don't know if the child crossed the road.
- b. * [aar̪ rooD̪ kroos ceytu ennat̪] enikk̪ ariyilla.
 who road cross do-PAST that-it to me know-NEG
 I don't know who crossed the road.
- c. * [kuTTi onn̪ rooD̪ kroos ceytenkil ennat̪] amma moohiccu.
 child one road cross do-past-if that-it mother wished
 Lit: The mother wished that 'if only the child would cross the road!'
- d. * [kuTTi onn̪ rooD̪ kroos ceyyu ennat̪] amma paraññu.
 child one road cross do-imp that-it mother said
 Lit: The mother said that 'Child, please cross the road!'

If we define 'propositional function' (PROP-F) as including propositions, questions, requests, and wishes, we may say that *S-that* expresses a PROP-F, while *S-that-it* expresses only a proposition.

There is a further property associated with *S-that-it*, that the proposition it expresses is [+def], i.e., an already existing assertion, shared by the speaker and listener; (47 shows this aspect of the clause:

- (47) a. [kuTTi rooD̂ kroos ceytu enn̂] amma prastaawiccu
 child road cross do-PAST that mother declared.
 Mother declared the child crossed the road.
- b. * [kuTTi rooD̂ kroos ceytu ennat̂] amma prastaawiccu
 child road cross do-PAST that-it mother declared.
- c. [kuTTi rooD̂ kroos ceytu ennat̂] amma niSeediccu
 child road cross do-PAST that-it mother denied.
 Mother denied (the statement) that the child crossed the road.
- d. * [kuTTi rooD̂ kroos ceytu ennat̂] amma prastaawiccu
 child road cross do-PAST that-it mother declared.

The asymmetry in the acceptability of (48c) below in the context of (48a) vs. (48b) illustrates the contrast between S-*that* and S-*that-it* in terms of the factive presupposition of the latter:

- (48) a. [patt̂ kuTTikaL warum enn̂] ñaan pratiikSiccirunnu.
 ten children will come that I was expecting
 I was expecting that ten children would come.
- b. [patt̂ kuTTikaL warum ennat̂] ñaan pratiikSiccirunnu.
 ten children will come that-it I was expecting
 I was expecting it that ten children would come.
- c. pakSe naal̂ kuTTikaLee wannuLLu
 but four children-only came-MOD
 But only four children came.

The sequence of (48a) and (48c) forms a coherent discourse. After (48b), however, (48c) is unacceptable. This unacceptability can be explained as resulting from the logical contradiction that combining (48b) and (48c) yields: (48b) carries the presupposition that the proposition expressed by the embedded clause is true, while (48c) asserts that this proposition is false. (48a) carries no such presupposition, and hence the combination is unproblematic. Interestingly, there are speakers for whom the English glosses for the Malayalam sentences exhibit the same contrast, depending on whether or not the pronoun *it* is present (Menzel (1973)).

The following examples illustrate the same contrast when these constructions appear as SUBJ:

- (49) a. [mukhyamantri warunnuND̂ enn̂] naaTaake paranniTTunT̂.
 chief minister is coming that land-all has spread
 That the Chief Minister is coming has spread all over the land.
- b. [mukhyamantri warunnuND̂ ennat̂] naaTaake paranniTTunT̂.
 chief minister is coming that-it land-all has spread
 It (= the news) that the Chief Minister is coming has spread all over the land.
- c. pakSe [warilla enn̂] enikk̂ urappaaN̂.
 but will come-NEG that to me certainty is
 But I am certain that (he) will not come.

(49c) can follow (49a) as a piece of continuous text, but it cannot follow (49b), as it asserts the opposite of what the embedded clause in (49b) presupposes. This presupposition is absent in (49a).

A caveat is in order at this point. Even though the S-*that-it* construction as a SUBJ carries the factive presupposition in examples like (48b) and (49b), it does not do so in the examples in (50):

- (50) a. [kuTTi rooD̂ kroos ceytu ennat̂] satymalla / nuNayaaN̂.
 child road cross do-PAST that-it truth-is-NEG falsehood-is
 (The statement) that the child crossed the road is not true/is false.
- b. [patt̂ kuTTikaL warum ennat̂] satymalla / nuNayaaN̂.
 ten children will come that-it truth-is-NEG falsehood-is
 (The statement) that ten children will come is not true/is false.

Likewise, the *S-that-it* clausal complement in (51b) does not carry the factive presupposition:

- (51) a. [kuTTi rooDð kroos ceytu ennð] ñaan wis 'wasikkunnilla.
 child road cross do-PAST that I believe-PRES-NEG
 I do not believe that the child crossed the road.
- b. [kuTTi rooDð kroos ceytu ennatð] ñaan wis 'wasikkunnilla.
 child road cross do-PAST that-it I believe-PRES-NEG
 I do not believe (the statement) that the child crossed the road.

The only semantic difference between (51a) and (51b) is that (51b) carries the presupposition of the existence of the proposition (as a claim that someone has made, for instance) in the relevant discourse context. We are faced here with what looks like an inconsistency in the data.

However, a comparison of the examples in (51) with those in (52) offers a clue to what is happening:

- (52) a. [kuTTi rooDð kroos ceytu ennð] awan wis 'wasikkunnilla.
 child road cross do-PAST that he believe-PRES-NEG
 He does not believe that the child crossed the road.
- b. [kuTTi rooDð kroos ceytu ennatð] awan wis 'wasikkunnilla.
 child road cross do-PAST that-it he believe-PRES-NEG
 He does not believe (the statement) that the child crossed the road.
- c. ñaanum wis 'wasikkunnilla.
 I-also believe-PRES-NEG
 I don't believe (it) either.

Parallel to what we saw in (48) and (49), (52c) can only follow (52a) to yield a coherent piece of text. (52b) carries the presupposition that the child did cross the street, which (52c) contradicts.

The difference between (51b) and (52b) is that in (51b), the SUBJ of the matrix clause, explicitly negating the truth of the embedded clause, is also the speaker, unlike in (52b). To explain the absence of factivity in (51b), we will assume a special exemption when the matrix SUBJ explicitly denying the presupposition is the speaker.-

3.5 An explanation for the facts

Contrary to the hypothesis suggested in section 3.2, suppose we assume that *S-that*, *S-that-it* and *S-it* clauses can all be SUBJ or OBJ. If so, it should not be surprising that all three clause types allow being replaced by a pronoun ((39)). If they all have the same range of GFs, their differences in behavior must come from their categorial or semantic properties.

Taking categories first, as mentioned earlier, the *S-that* clause is a CP while the *S-that-it* and *S-it* clauses are NPs: it is only to be expected that a clause headed by *it* is an NP. Support for this position comes from the fact that *S-that-it* and *S-it* can be hosts of case inflections, but not *S-that*. This categorial distinction is sufficient to account for their distribution in a PP ((32a-c)): the sister of a P in a PP must be an NP, i.e., the constraint “no P+CP” ((12)) is ranked higher than the constraint of C-to-F faithfulness ((13)):

- (53) In Malayalam:
 a. *S-that* is a CP; *S-that-it* and *S-it* are NPs.
 b. (12) >> (13)

Let us review the remaining facts that we saw in sections in 3.1-3.3 in the light of the semantic properties of these clauses that we saw in (42).

The non-conjoinability of different clause types (40d-f) can be explained by assuming that the constituents of a conjoined expression must have the same propositional features, i.e., EVENT vs. PROP-F, PROPOSITION vs. REQUEST/QUESTION/WISH, and FACTIVE vs. NON-FACTIVE.

Turning to the non-passivizability of examples like (33a), it is worth noting that explanations of passivizability in terms of GFs are illegitimate if we accept LMT. Granted that an argument specified as [+r] cannot be the SUBJ or OBJ, and hence is disallowed from being the subject of a passive, we still need to explain why some clauses are assigned [-r] and others [+r]. Saying that the former are OBJ and the latter COMP is simply a restatement of what needs to be explained. Hence, regardless of the solutions we come up with, non-passivizability does not constitute an argument in support of COMP.

Furthermore, passivizability does not distinguish any clause type from the others. While the S-*that* clause is not passivizable in ((33a)), it is indeed passivizable in (36b). Likewise, the S-*that-it* clause is passivizable in ((33b)), but not in (37b). Precisely how these asymmetries are to be explained, we will leave to a more fine-grained account of semantics that tells us when an argument that is eligible for objecthood (i.e., a [-r] argument) is also eligible to be a passive subject.

For an explanation of the facts of complement selection illustrated in (30) and (31), we need to go no further than events and factivity. For instance, consider the following specifications in the lexical semantics of these verbs:

- (54) a. *ariy-* ‘know’: OBJ clause: PROP-FUNCTION/FACTIVE PROP/EVENT
 b. *wicaarikk-* ‘think’: OBJ clause: PROPOSITION
 c. *niSeedhikk-* ‘deny’: OBJ clause: [+def] PROPOSITION
 d. *kaaN-* ‘see’: OBJ clause: EVENT
kaaN- ‘infer through seeing’: OBJ clause: PROPOSITION

Given these specifications, it would follow that *ariy-* ‘know’ can take all the three clause types ((30)); *wicaarikk-* ‘think’ takes only S-*that* ((31a)); *niSeedhikk-* ‘deny’ takes only S-*that-it* ((31b)); *kaaN-* ‘see’ takes only S-*it* ((31c)); and *kaaN-* ‘infer through seeing’ takes only S-*that-it* ((37a)). Nothing further needs to be said about the facts of complement selection associated with these clause types.

What remains to be explained is the asymmetry between (34a) on the one hand, and (35a,b), (36b) and other similar examples on the other. A possible clue to a solution for the distribution of S-*that* as SUBJ in these sentences may be found in the following examples in English:

- (55) a. *That John flunked the test has spread far and wide. I happen to know that he didn't flunk, though.*
 b. *That John flunked the test upset him. *I happen to know that he didn't flunk, though.*

The above contrast indicates that the SUBJ clause of *upset*, like the OBJ clause of *regret*, carries the presupposition that the proposition it expresses is true. Taking the contrast in (55) as a clue, we propose that:

- (56) If: the grammatical system of a language marks the semantic type of embedded clauses in structural terms,
 then: in that language,
 a verb that carries the presupposition that its SUBJ/OBJ is factive will only allow a clause marked for factivity in that position, and
 a verb that takes a [+def] proposition will allow only a clause marked for [+def] proposition in that position.

The verb in (34a) is the causative *santooSippikk* ‘make x happy’. If a causer must be a person, thing, event or a definite proposition, it follows that the subject of *santooSippikk*, if clausal, must be an EVENT or [+def] PROP. By (56), then, it cannot be S-*that*. Hence the ungrammaticality of (34a).

With direct access to the semantics of arguments in complement selection and other grammatical phenomena in Malayalam, COMP once again becomes redundant.

4. English

Once we accept the idea that a clause may bear the same range of GFs that nominal structures are assumed to bear, most of the arguments presented in D&L for distinguishing COMP from OBJ in

languages like English can be straightforwardly reinterpreted as arguments for distinguishing OBJ from OBL. The fact that certain clausal complements alternate with nominal complements, whereas others don't, is taken in D&L as evidence for the claim that the former are OBJ and the latter COMP. This contrast is illustrated in (57) (from D&L: 107):

- (57) a. *I believe [that the earth is round] / it.*
 b. *I hope [that it will rain] / *it.*

If we assume that *believe* takes an OBJ, it is to be expected that, semantics permitting, the OBJ should be alternatively a CP or an NP. If, on the other hand, we assume that the complement of *hope* is a COMP, as in D&L, then it follows that this complement should not be expressed as an NP, given the claim that COMPs are always CPs. However, if we assume that *hope* takes an OBL (only) as its complement, it also follows that this complement should not be expressed as an NP, given the claim that OBL in English cannot be a prepositionless NP.

The contrast in passivizability in examples like (58) (from D&L: 108–109) has been argued to be evidence for the distinction between OBJ and COMP.

- (58) a. *That the earth is round was not believed.*
 b. * *That it would rain was hoped.*

D&L assume that OBJ alternates with SUBJ because both are [–r] arguments and that COMP does not alternate with SUBJ because it is a [+r] argument. Thus, the grammaticality of (58a) follows from assuming that the internal argument of *believe* is [–r] and can, therefore, be either OBJ, or SUBJ (in a passive), and the ungrammaticality of (58b) follows from assuming that the internal argument of *hope* is [+r] and, therefore, not compatible with either OBJ or SUBJ, but is compatible with COMP. However, it is clear that a [+r] argument is also compatible with OBL. Consequently, the contrast in (58) can just as well be taken as evidence for the distinction between OBJ and OBL. Furthermore, in D&L there is an indeterminacy as to whether a [+r] argument should be assigned the OBL or the COMP function, since both are compatible with [+r]. How does the mapping theory discriminate between [+r] arguments to be assigned an OBL function and those to be assigned a COMP function? In the present proposal, this indeterminacy disappears because, without COMP, all [+r] arguments would be OBLs.

The observation that nouns and adjectives cannot take NP complements, but do take CP complements is interpreted by D&L as evidence for the existence of the GF COMP: if we assume a restriction against OBJ appearing in f-structures headed by N or A, the fact that CPs can be complements of these categories can be explained by assuming that these CPs are COMPs. The contrast between NPs and CPs as complements of Ns and As is illustrated in (59): the verb *fear* takes an OBJ, which can either be an NP or a CP, as in (59a), whereas the noun *fear*, in (59b), and the adjective *scared*, in (59c), allow only a CP complement, not an NP complement:

- (59) a. *Tim fears {that he may be found out / thunderstorms}.*
 b. *Tim's fear {that he may be found out / *thunderstorms}.*
 c. *Tim is scared {that he may be found out / *thunderstorms}.*

The prohibition against having OBJ assigned to an argument of an N or A does not imply that these categories should not take clausal complements, according to D&L, since clausal complements can bear the GF COMP. However, positing the GF COMP is not the only way to explain the fact that Ns and As can take clausal complements. Since it is clear that these categories can take OBLs as complements, typically expressed as PPs, as shown in (60), the assumption that clausal complements may also bear the GF OBL readily explains the possibility of a clausal complement of these categories.

- (60) a. *Tim's fear of thunderstorms.*
 b. *Tim is scared of thunderstorms.*

All we need to assume is that nouns and adjectives only take OBLs as complements, by default introduced by the preposition *of*. The OBL is either a PP or a CP, and it is predictable that, when the complement is clausal, there is no preposition introducing the oblique phrase because of the constraint excluding a PP in which the preposition precedes a CP ((12)).

Finally, the contrast between CP complements that can be topicalized, or, more generally, enter into an unbounded dependency, and those that cannot has also been taken by D&L as evidence for the distinction between OBJ and COMP. The examples in (61) (from D&L: 109 and Dalrymple 2001: 81) illustrate this contrast:

- (61) a. *That it would rain, everybody believed.*
b. * *That Chris yawned we weren't aware.*

According to D&L, all that needs to be assumed to explain this contrast is that the clausal complement of *believe* is an OBJ, whereas that of *aware* is a COMP, and that there is a stipulation that a COMP cannot enter into an unbounded dependency. It is not immediately obvious how this contrast is to be accounted for in the proposal that dispenses with COMP. Given the fact that *aware* can take either a CP or PP complement, as shown in (62) (from Dalrymple 2001: 81), we would have to assume that *aware* takes an OBL, which, as explained, is realized alternately as a CP or as a PP (in contrast to Dalrymple's (2001) assumption that it has an alternative subcategorization with a COMP and with an OBL):

- (62) a. *We weren't aware that Chris yawned.*
b. *We weren't aware of the problem.*

It is not possible to stipulate that an OBL cannot enter into an unbounded dependency, since this is clearly incorrect, as shown by an example like (63):

- (63) *Of the problem we weren't aware.*

Furthermore, topicalization of the clausal complement of *aware* is possible provided the preposition that introduces the oblique is retained, stranded in post-verbal position (from Dalrymple 2001: 81):

- (64) *That Chris yawned we weren't aware of.*

The contrast between (61b) and (64) seems to show that preposition stranding, which is normally an option alongside preposing of the entire PP, is obligatory when the phrase that enters into an unbounded dependency is a CP. Optimality Theory provides a simple way to account for that contrast without introducing additional constraints. Let us assume that the options of preposition stranding and of preposition pied-piping are in competition and are equally optimal (in English and other languages with these options), unless one of the options violates a constraint that the other does not. Normally, then, both options are possible, e.g., the pied-piping option in (63), and the P-stranding option in (65).

- (65) *The problem we weren't aware of.*

So, how do we explain that only the P-stranding option is allowed when an oblique CP is involved in the unbounded dependency? The answer is that the structure without the preposition, corresponding to example (61b), violates C-to-F Faith, whereas the competing structure *with* the preposition, corresponding to example (64), does not violate this constraint. Let us assume that an adjective like *aware* subcategorizes for an OBL with the case feature 'of'. This case feature is only provided by the lexical item *of*; therefore, we expect this preposition to appear, introducing the oblique complement of *aware*. Otherwise, C-to-F Faith is violated. However, in English, as in Catalan, the constraint No P+CP ranks higher than C-to-F Faith, accounting for the failure of the preposition to appear just in case the preposition should precede a CP.

In a topicalization structure (or, more generally, in an unbounded dependency construction), the topicalized phrase bears the discourse function TOPIC (in other cases, FOCUS), which is functionally identified with an in-clause, or non-discourse, function in its f-command domain. When an oblique is involved in an unbounded dependency, either the entire oblique or just the object of the oblique preposition is functionally identified with the phrase bearing the discourse function (and is, therefore, missing from its expected position). The first case requires preposition pied-piping and the second case requires P-stranding, as otherwise the structure would violate C-to-F Faith. However, when the oblique is a CP, preposition pied-piping is not possible, as the structure would violate No P+CP, which is worse than a violation of C-to-F Faith. That leaves P-stranding as the only option: the CP bears the discourse function TOP and is functionally identified with OBJ in the f-structure of the oblique.

In conclusion, all of the facts of English concerning clausal complements can be explained without appealing to the GF COMP, and in a simpler way than if COMP is assumed. In essence, these facts are explained without positing any constraint that is not independently required. The most important constraint involved is ‘No P+CP’, which is assumed even in theories that posit COMP, such as D&L.

5. Concluding remarks

To summarise, we have argued that, given the restricted assignment of GFs to nominal vs. clausal categories shown in (1), the GF COMP is redundant in LFG. In a framework without COMP, the regularities attributed to COMP can be restated in terms of CP OBJ or CP OBL_θ. Our reanalysis of the English facts cited as evidence in support of COMP serves as an illustration of how this can be done. The facts of Catalan and Malayalam show that the strategy of increasing the number of grammatical functions beyond what is necessary and appealing to these distinctions to express all aspects of complement selection fails to provide satisfactory grammars.

Clausal complements in languages like Catalan exhibit a contrast between OBJ and OBL_θ, showing that many of the phenomena that are appealed to in motivating the alleged distinction between OBJ and COMP cannot be expressed solely in terms of grammatical functions, but need reference to f-structure features that constrain the realization of the arguments. Clausal complements in languages like Malayalam show that many of the phenomena appealed to in motivating the distinction between OBJ and COMP must be expressed in terms of the semantics of the arguments, the relevant distinctions in this case being EVENT vs. PROP-F, PROP vs. REQUEST/QUESTION/WISH, [+/-def], and FACTIVE.

Given such independently required distinctions, the attribute COMP becomes redundant in the inventory of GFs, particularly within the multi-dimensional co-present architecture of LFG. If we allow CP complements to be associated with the same range of GFs as NP complements, the patterns that allegedly motivate the postulation of COMP, but are not attributable to independent categorial and semantic distinctions, can be explained in terms of functional contrasts already available within LFG.

If we abandon the function COMP in LFG, the obvious question is, what about the function XCOMP? Given that they are both clausal complements, and that XCOMP may be considered a special case of COMP, XCOMP should probably go the same way as COMP. The label signals that the unit in question is a clause whose subject is obligatorily controlled. Instead of such a diacritic, what we need is a theory of control that tells us under what conditions the SUBJ of a clause is obligatorily controlled. The elimination of COMP and XCOMP from the inventory of GFs has a desirable offshoot. Current mapping theory, with two features, [+/-r] and [+/-o], can express only four distinctions among subcategorizable GFs, namely, [-r,-o] (SUBJ), [-r,+o] (OBJ), [+r,+o] (OBJ_θ) and [+r,-o] (OBL_θ). This feature system does not provide for COMP and XCOMP. And rightly so.

References

- Alsina, A. 1996. *The role of argument structure in grammar*. Stanford: CSLI Publications.
- Alsina, A., T. Mohanan, and K. P. Mohanan. 1996. Untitled submission to the LFG List. 3 Sept 1996.
- Bresnan, J. 2001. *Lexical-functional syntax*. Oxford: Blackwell.
- Dalrymple, M. and H. Lødrup. 2000. The grammatical functions of complement clauses. In M. Butt and T. H. King (eds.), *Proceedings of the LFG00 Conference*. CSLI Publications.
- Dalrymple, M. 2001. *Lexical Functional Grammar*. San Diego: Academic Press.
- Grimshaw, J. 1997. Projections, heads and optimality. *Linguistic Inquiry* 28: 373–422.
- Kiparsky, C. and P. Kiparsky. 1971. Fact. In D. Steinberg and L. Jakobovitz (eds.), *Semantics*. Cambridge: Cambridge University Press.
- Menzel, P. 1973. A constraint on the deletion of embedded subjects. *Papers in Linguistics* 6:141–179.

A COMPARISON OF COMPARATIVES

Dorothee Beermann, Jonathan Brindle, Lars Hellan,
Solomon Tedla, Janicke Furberg, Florence Bayiga, Yvonne Otoo (all: NTNU),
Mary Esther Kropp Dakubu (University of Ghana)

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

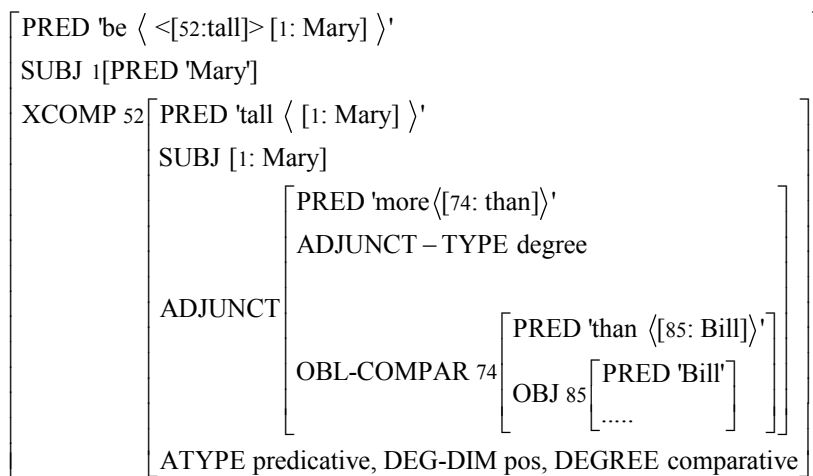
1 Introduction*

In the Pargram grammar of English (cf. Butt et al. 1999), comparatives of both of the types in (1)

- (1) a. Sarah is more intelligent than John
 b. Mary is taller than Bill

are analyzed in f-structure by adjoining a phrase headed by the PRED 'more' to the element expressing the dimension of the comparison (above, the adjectives 'intelligent' and 'tall'); the adjunct in turn takes an OBL-COMPAR object, introducing the 'than'-clause and its object. The f-structure for (1b) is shown, simplified and in XLE format, in (2):

(2)



A principle holding of f-structures is that of *cross-linguistic validity*: e.g., for transitive constructions, no matter whether subject- and objecthood are marked by linear order, case, or other means, the same attributes are used in the f-structures across languages. For comparatives, this is to say that if (2) is an analysis of (1b) as a construction in English, it should be valid also of constructions equivalent to (1b) in other languages, as far as the factors represented in (2) are concerned. These are the factors of *comparison*, that the comparison reflects a 'positive' dimension, and that there is an *oblique* constituent also somehow involved in comparison - the representation itself doesn't say how, but in the case in question, it is understood that it reflects a term of the comparison. Also said in (2) is that the expression of comparison enters the construction as an *adjunct*.

We may note that (2) plus the invariance principle do *not* claim that all comparison take the form of adjunction: a paraphrase of (1b) could be *Mary's height exceeds Bill's height*, where the key expression of comparison is the main verb, and not an adjunct; we still don't want to see this as a counterexample to the invariance principle. The point is made in Stassen 1985 that all languages, alongside their typical way of expressing comparison, may use a strategy like the one just exemplified; and one may agree that as far as the notion 'comparative construction' is

* This note grew out of a course on LFG at NTNU during Fall 2003 and Spring 2004, where most of the authors took part. We thank the editors of this volume, Miriam Butt and Tracy King, for helpful comments.

concerned, the task of grammar is to address the 'typical' patterns. The interest of (2), then, is as a template of 'typical' comparative constructions,¹ not just in English, but across languages.

An obvious proviso to this point is the possible restrictedness of (2) to simple adjectival comparison: more complex types of content - as may be expressed in comparison of quantities, of manners, and more - may require different constructional patterns, with different schemata. However, again, one will hold that each such type of f-structure schema will be cross-linguistically invariant for the type of content in question. In this note, we will focus on simple adjectival comparison, like what is expressed in (1).

With these preliminaries, let us say where we want to go with the present note. On the one hand, we want to investigate to what extent the schema instantiated in (2), and under the invariance principle, is actually true: when we go to typologically different languages, will simple adjectival comparison still take forms representable with f-structures as in (2)? Here we will look at one language from this point of view, the West-African language Ga, which uses serial verb constructions for the expression of comparison; this will be a topic in section 3.

On the other hand, as noted, to delineate a class of constructions as being 'the same' relative to comparison, it may be relevant to take the semantics of the construction into account. Moreover, to properly construe the sense of the attribute OBL-COMP in (2), we may want to be more precise about exactly what are the 'terms' of a comparison. Such points may be most perspicuously achieved if we can supply an explicit semantic representation going along with each comparative construction, i.e., co-define a semantic structure together with a c- and an f-structure. Adopting the formal construct 's-structure' as defined in Halvorsen 1995, Halvorsen and Kaplan 1995, Fenstad et al. 1985, we will propose a format of semantic representation applied to the analysis of comparatives, and also a way of co-defining this format of representation with other structures. This will be the topic of section 2.

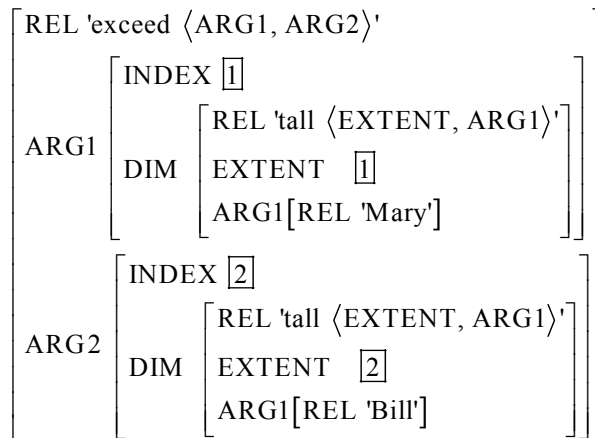
The constructs defined in section 2 will be carried along to the analysis of constructions in Ga, so that a counterpart of (1b) in Ga will indeed be provided with the same s-structure as (1b) has. We assume that this constitutes a formal marking of the necessary equivalence between these constructions, to warrant the question whether they share f-structure properties, and we provide a tentative assessment of this in section 3. We here, in turn, point to a counterpart in still another language, where the answer may be different. Our purpose in this note being only to open for the type of investigation here sketched, we leave this case for further investigation.

2 The semantics of comparatives

In proposing a format of semantic representation applied to the analysis of comparatives, we want to accommodate standardly recognized features of the semantics of these constructions, as reflected, e.g., in Klein 1980, Seuren 1973, Hellan 1981, Heim 2000. Using the format of s-structure, a semantic representation of (1b) can be given as in (3):

¹ 'Template' in the following way: in the format of a declarative grammar formalism, one does not state "if a construction expresses simple adjectival comparison, then it takes a form involving...". One rather supplies just the template instantiated by (2) as a format for the encoding of such comparison, and no other schema, thereby enforcing this as the only 'channel' of expression as far as factors reflected in f-structure are concerned.

(3) S-structure of *Mary is taller than Bill*:



This representation assumes an explicit 'exceed' relation interconnecting the degree to which Mary is tall and the degree to which Bill is tall. Each of these degrees is associated with a predication involving 'tall', introduced under the attribute DIM (for 'dimension of comparison') together with the relevant participant and with the attribute EXTENT indicating the *extent* to which the relation in question obtains. This extent may be thought of as standing in a one-to-one relation with whatever 'degree' unit may be invoked in substantiating the comparison. This degree unit is here introduced by the attribute INDEX, for simplicity exposed as identical to the extent, although in principle, it is only functionally related to it

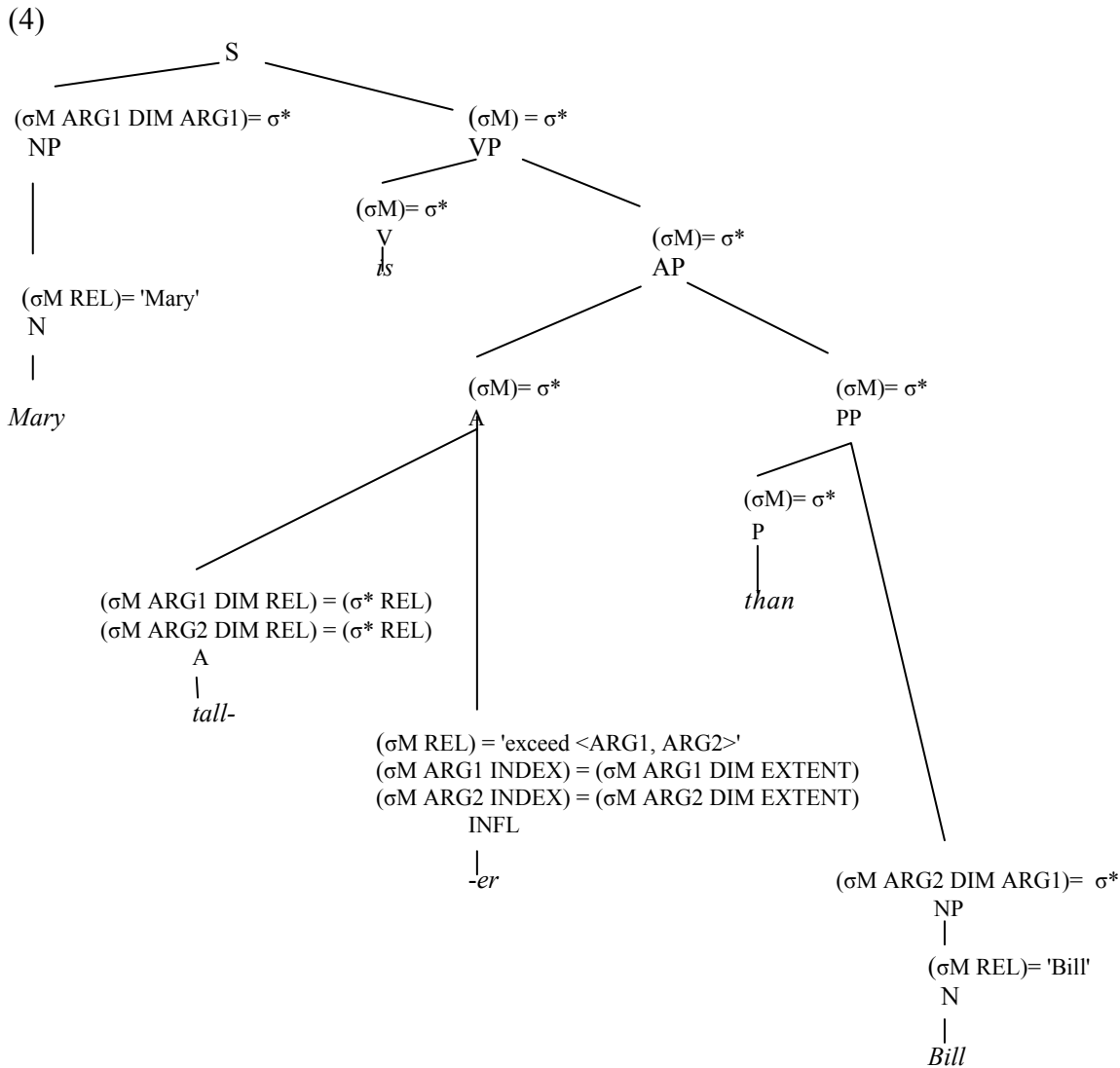
'REL' has two possible values as far as comparatives are concerned, 'exceed' and 'equate' (for *as tall as*), both taking ARG1 and ARG2. 'Exceed' refers to whichever directed dimension of comparison is expressed by the DIM predicate: if the adjective is *small*, then the direction is one of increasingly smaller amounts of height, and if *tall*, increasingly higher amounts.²

The values of the paths ARG1 | DIM | REL and ARG2 | DIM | REL need not be the same: in an example like *the lamp is taller than the window is wide*, degrees of height and width are compared. Conversely, the values of the paths ARG1 | DIM | ARG1 and ARG2 | DIM | ARG1 need not be distinct: in *the door is taller than wide*, the ARG1s are the same. All of this variation is allowed by the formalism.

(4) is a display of annotations on the c-structure of (1b) whereby (3) can be obtained. 'σ' is the function from c-structure nodes to s-structure specifications. 'σM' stands for 's-structure of the mother of the current node', and 'σ*' stands for 's-structure of the current node itself'. Shortcutting considerations of adequacy of morphological representation, we here assign a semantic contribution to the affix *-er* directly.³

² We thereby avoid the situation induced by the use of the attribute DEG-DIM in (2), which seems to presuppose that every adjective comes as a member of a pair constructible along a 'positive-negative' dimension.

³ This stays close to early works like Bresnan (1973) and Davis and Hellan (1975). However, a more correct way, both for the capturing of morphological generalization (given the highly different ways in which comparative morphology can be realized - as *-er*, as *more*, or through suppletion) and for adherence to lexical integrity, would be to base this part of the semantics on morphological features already accommodating the morphological variation.



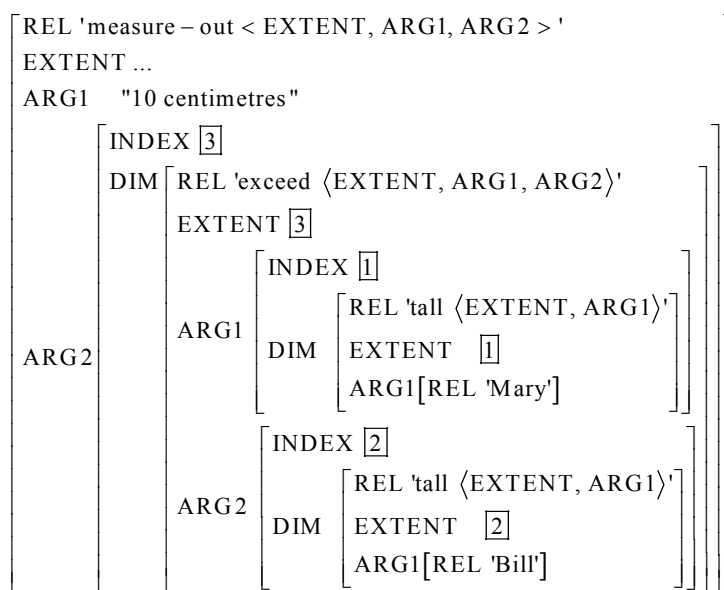
As will be noted, the inflected adjective *taller*, through the impact of the comparative morphology, acts as the semantic head of the construction, and thereby defines the main frame of the AVM in (3), leaving for *tall*, *Mary* and *Bill* to contribute their parts in a compositional fashion. *Than* is treated as semantically empty; any contributions that one might want to associate with it are here carried by *-er*, although a plausible alternative could be to co-allocate some of the specifications on *taller* to *than*, subject to unification in the compositional assembly.

Note that in order for such an annotated tree to extend beyond the particular structure in (1b), some further annotation is needed. The case where *tall* can unproblematically be taken to serve as DIM|REL of both of the degree arguments - as induced by the specification $(\sigma M \text{ ARG1 DIM REL}) = \sigma^* \text{ REL}$ and $(\sigma M \text{ ARG2 DIM REL}) = \sigma^* \text{ REL}$ in (4) - is restricted to those occurrences of *than* where it is followed only by an NP. As soon as something else follows, as a more or less truncated clause, allowance must be made for this part to include an adjective, as in *the door is taller than it/the window is wide*. Following Hankamer 1973, one can distinguish two variants of *than* in accordance with this, one being a preposition and the other a complementizer, and in the tree annotation, make the specifications $(\sigma M \text{ ARG1 DIM REL}) = \sigma^* \text{ REL}$ and $(\sigma M \text{ ARG2 DIM REL}) = \sigma^* \text{ REL}$ for the adjective non-conditional only in the case where *than* is a P. Analogously, for cases like *the door is taller than wide*, one must have the option of specifying the subject NP as providing the REL also for the second degree's DIM|ARG1: this is *allowed* when *than* is a complementizer, but not required. We will not try to spell out here the exact disjunctive and

conditional specifications needed to cover this array of possibilities, or, alternatively, explore the possibility of capturing these cooccurrence patterns through differential meaning assignments to *than*.

The role of EXTENT in (3) may be seen as that of 'measuring out' - it measures out the height of Mary, and the height of Bill. In cases like *Mary is 10 centimetres taller than Bill*, one may say that the role of *10 centimetres* is in turn to measure out the *exceed* relation holding between Mary's height and Bill's height. Accordingly, we will regard the presence of an EXTENT as generally associated with any relation, and represent the sentence *Mary is 10 centimetres taller than Bill* in the way of (5):

(5) S-structure of *Mary is 10 centimetres taller than Bill*:



Here the value of ARG2 | DIM | EXTENT records the extent to which Mary's height exceeds Bill's height, and is reentered as INDEX of the ARG2 of *measure-out*. The more precise details of the representation of *10 centimetres* we leave open for now, the point presently made being only that a general treatment of 'degree recursion', as further exemplified in *Mary is almost 10 centimetres taller than Bill*, *Mary is more than 10 centimetres taller than Bill*, *Mary is 10 centimetres less than half a meter taller than Bill*, etc., can be obtained through exploiting an EXTENT attribute along similar lines as illustrated in (5).⁴

We have now indicated how a semantics of comparatives can be given expression using the formalism of s-structure,⁵ and indicated a mapping algorithm between c-structure and s-

⁴ This applies to composition of s-structure specifications; for f-structure specification, these types of construction are fully accounted for in the English PARGRAM grammar.

⁵ We make no proposal here concerning superlatives. In an s-structure representation, they may conceivably be quite like comparatives, since they too express an exceed relation. The second term of this type of comparison is typically expressed through a partitive-like PP (as in *tallest of the boys*), which supplies a set of which the first term in the comparison is a member (or subset). This is a relation between the values of the paths ARG1 | DIM | ARG1 and ARG2 | DIM | ARG1, and once formalized, e.g., through a relation *instantiate*, little more need be said in s-structure distinct from what is said for comparatives. It will then seem reasonable to have a marking of 'superlative' be part of the f-structure representation, corresponding to the sub-specification 'DEGREE comparative' in (2).

As REL-values in the s-structure representations of comparatives we have so far proposed using *exceed*, *equate*, *measure-out* and now *instantiate* in the case of superlatives (and presumably partitives); although this is little more than a preliminary sketch of an analysis, are we in a position to say whether this brings us close to a complete list of notions involved in this area of analysis? Presumably, 'yes', twice - as for what to expect.

structure adequate for simple adjective comparison. Once representations like (3) and (5) are available at the level s-structure, a question may be whether some of the f-structure attributes in (2), such as DEGREE and DEG-DIM, might be redundant; this is a possibility, but not one that we will explore here.⁶ Another issue is how mappings to f-structure and s-structure may interact: in this example, they may seem fairly independent, but in, e.g., transitive structures, the assignment of status as ARG1 and ARG2 relative to the verb in s-structure will clearly be dependent on f-structure information about grammatical functions and diathesis; thus, in a more concise outline of the design envisaged, this aspect of interaction between mappings clearly will need to be stated.

An aspect of the f-structure (2) which is definitely not shared with the s-structure (5) is the specification of the comparative as an *adjunct*; and this is a point we address in the next section. We here turn to a pattern of comparison instantiated in the West-African language Ga. This pattern deviates from the comparative construction in English in two respects: it is a multiverb construction, and the comparative meaning is expressed through a verb with a meaning 'exceed'. The latter, as we will note, is a widely used pattern, also for the 'typical' construction of comparison in a language.

3 'Exceed'-comparative languages

Stassen (1985) observes that cross-linguistically, one of the major strategies for expressing comparison is using a free-standing lexical item with a meaning like 'exceed'; the strategy used in most Indo-European languages, such as the one in English, is in the larger perspective less prevalent. For example, the 'exceed' type comparative construction is the typical pattern in the West-African languages, and for illustration, we will look at comparatives from the Kwa language Ga, spoken in the Accra area of Ghana

3.1 Comparatives in Ga

Examples of comparatives in Ga are given in (6). These examples employ the verbs *fe* "surpass, exceed, be more than" and *tamɔ* "resemble, be like". Both belong to a limited class of verbs which have been called *verbids* (cf. Dakubu 2004), whose distinguishing feature is the ability to occur as the second part of a multiverb construction while *not* being subject to a requirement of argument sharing and tense/aspect agreement with the preceding verb. In effect, verbids, which are morphologically fullfledged verbs, have a function very much like that of event-modifying prepositions in a language like English:

(6)

Verbid *fe* "surpass, exceed, be more than"

- a. Ado ke fe Kofi
Ado be.tall exceed Kofi
'Ado is taller than Kofi'
- b. Ado é-!kéèè fè Kofi
Ado NEG-be.tall.IMPERF exceed Kofi
'Ado is not taller than Kofi'
- c. Ado ye-ɔ yeɛ pii fe Kofi
Ado eat-HAB yam much exceed Kofi
'Ado eats more yam than Kofi'

⁶ Some opinions have already been stated in footnotes 2 and 5, however.

- d. Nyé !é Àdo ye yeɛ pii fe Kofi
 Yesterday TOP Ado ate yam much exceed Kofi
 'Yesterday Ado ate more yam than Kofi did'
- e. Ado ye-ɔ yeɛ pii fe bɔ-ní Kofi yè-ɔ amádaá
 Ado eat-HAB yam much exceed manner-REL Kofi eat-HAB plantain
 'Ado eats more yam than Kofi eats plantain'
- f. Ado ye-ɔ fufuí òyáòyáì fe Kofi
 Ado eat-HAB fufu fast exceed Kofi
 'Ado eats fufu faster than Kofi does'

Verbid *tamɔ* “resemble, be like”

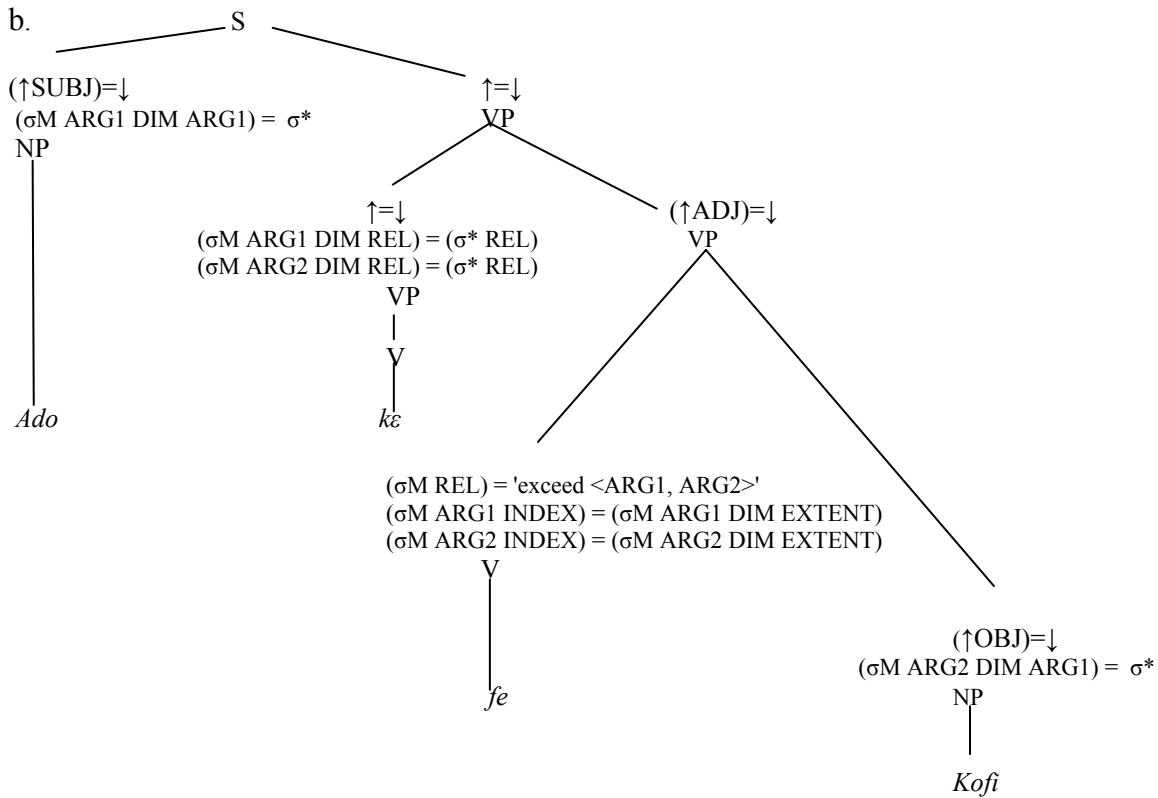
- g. Ado ke tamɔ Kofi
 Ado be.tall resemble Kofi
 'Ado is as tall as Kofi'
- h. Ado é-!kéɛ tàmɔ Kofi
 Ado NEG-be.tall.IMPERF resemble Kofi
 'Ado is not as tall as Kofi'
- i. Ado ye-ɔ yeɛ tamɔ Kofi
 Ado eat-HAB yam resemble Kofi
 'Ado eats yam as much as/ the way that Kofi does'
- j. Ado ye-ɔ yeɛ pii tamɔ Kofi
 Ado eat-HAB yam much resemble Kofi
 'Ado eats as much yam as Kofi does'
- k. Ado ye-ɔ fufuí hwàḡhwaḡ tamɔ Kofi
 Ado eat-HAB fufu greedily resemble Kofi
 'Ado eats fufu as greedily as Kofi does'

3.2 Assignment of *s*-structure to Ga comparatives

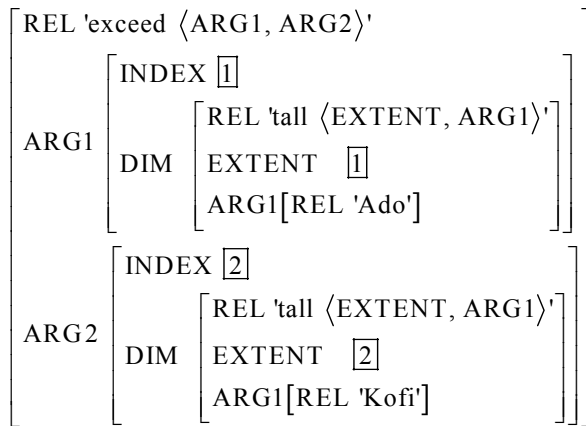
A semantics of comparatives centered around a relation 'exceed' will appear quite natural for a language of this type. The way in which this 'naturalness' can be formally spelled out is through an annotation like the one associated with *-er* above, but with the verbid head of the second verb phrase as the carrier of the equations. The following thus illustrates the Ga analogue of the annotated *c*-structure for English given in (4), for the sentence (6a) repeated as (7a); (8) below is in turn the associated *s*-structure. As argued in Dakubu (op.cit.) and Dakubu and Hellan (2003), it is reasonable to treat the verbid VP as an adjunct relative to the preceding VP. We reflect this assumption as well in the annotation in (7b), which combines the sigma- and the phi-functions, the latter written with the standard up- and down-arrows.

(7)

- a. (= (6a)) Ado ke fe Kofi
 Ado be.tall exceed Kofi
 Ado is taller than Kofi



(8) S-structure of (7a):

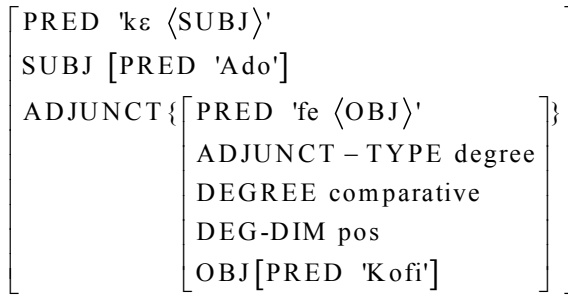


3.3 Invariance of *f*-structures

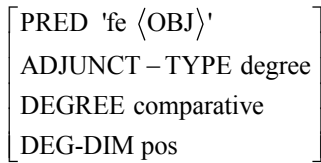
By the annotations in (7b), the *f*-structure for (7a) will be as in (9a), if we assume that in addition to the lexical sigma-specifications given in (7c), we have a phi-specification for *fe* as in (9b) (now using a more informal notation than in (2); we leave open here to what extent it will be motivated to classify a verbid with exactly the same attributes as have been used in the Pargram grammar for the English comparative morphology - for the purpose of comparison, we minimize these differences):

(9)

a. possible f-structure for (7a):



b.



The main difference between (2) and (9a) is that in (2), the predication of tallness is exposed as embedded in a 'raising'-like structure, whereas in (9a), it sits at the outermost layer. Aside from this, however, they both expose the comparative as an adjunct. In this critical respect, the f-structure representations of simple adjectival⁷ comparison in Ga and English thus have the same structure, in conformity with the invariance principle of f-structures.

3.4 A potentially different case

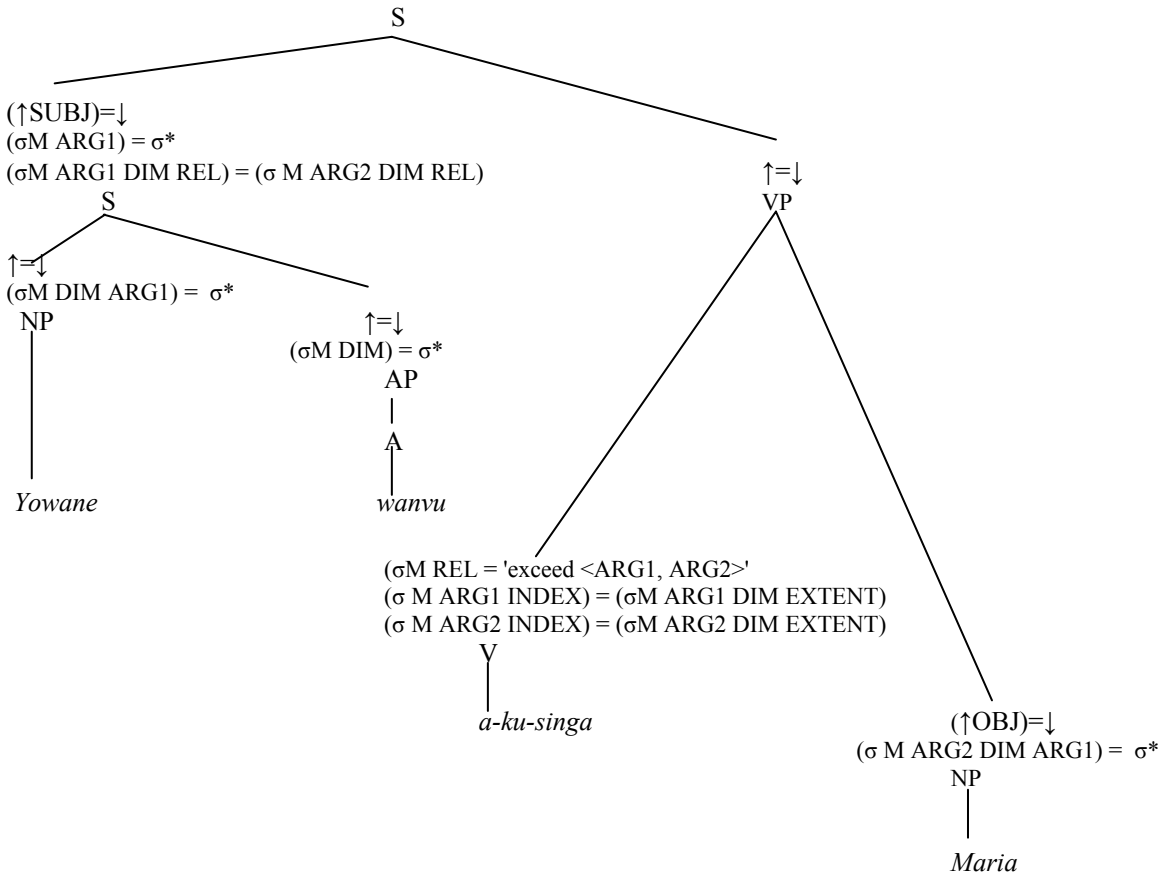
Comparative constructions in Luganda, a representative of the Bantu family, employs the same 'exceed'-verb strategy as Ga, with one exception: the verbal constituent preceding the 'exceed'-verb acts like a subordinate clause. An example is given in (10a), and an approximate annotated c-structure in (10b):

(10)

a. Yowane mu-wanvu a-ku-singa Maria
 John(NL.1) NL.1-tall IV-INF-exceed Maria
 John is taller than Mary

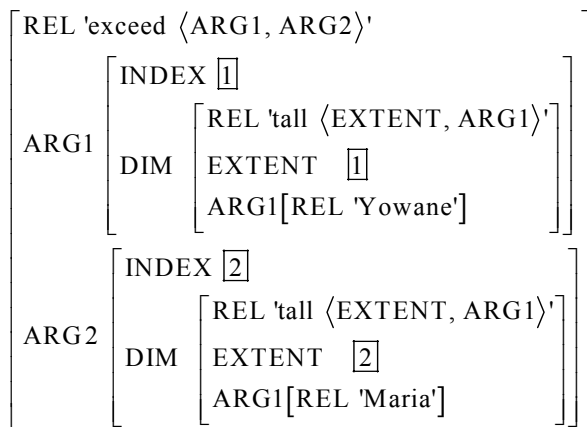
⁷ In Ga, the relevant item is actually a verb, but this is not a matter of consequence here.

b.



In this structure, as indicated by the functional annotation, the ‘exceed’ verb is the head of the whole construction, whereas the predicate ‘tall’ is part of the SUBJ argument relative to this head verb. The s-structure induced by the sigma-annotation will be exactly the same as we have seen earlier, i.e., in this case (11) below, while in this case, it is doubtful whether the comparative expression can be counted as an adjunct.

(11) S-structure for (10a):



Pursuing a more careful analysis of the relevant patterns in Luganda exceeds the bounds of this note, so we leave this case as a potential challenge to the invariance principle of f-structures.

4 Summary

We have outlined a format for the specification of the semantics of comparatives, using s-structure assigned on the basis of c-structure annotation. The s-structure representations are intended to be uniform across languages, highlighting the universal notion of comparison as interrelating extents. In the current setting, for simple adjectival comparatives of three languages with highly diverse c-structures, we have induced such structures through c-structure annotation.

More essential still to the LFG design is the invariance of f-structures relative to construction types across languages. The parallel analyses of English and Ga comparatives suggest that at least for this pair of languages, whose strategies for expressing comparison on the surface (and in c-structure) appear very different, an interesting degree of invariance can be argued to hold. Further research will show whether this will prevail throughout the 'exceed' type languages, and, of course, throughout further typological diversity.

References

- Bresnan, Joan 1973. Syntax of the Comparative Clause Construction in English. *Linguistic Inquiry*, 275-344.
- Butt, Miriam, Tracy Holloway King, Maria-Eugenia Nini and Frederique Segond. 1999. *A Grammar-writer's Cookbook*. Stanford: CSLI Publications.
- Dakubu, Mary Esther Kropp. 2004. Clauses without Syntactic subject. *Journal of African Languages and Linguistics*.
- Dakubu, Mary Esther Kropp, and Lars Hellan. 2003. The "verbid" construction in Ga: a VP with adjunct function. Beermann, D. and L. Hellan (eds) On-line Proceeding TROSS 2003.
- Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III, Annie Zaenen (eds), 1995. *Formal Issues in Lexical-Functional Grammar*. Stanford: CSLI Publications.
- Davis, Charles and Lars Hellan 1975. The Syntax and Semantics of Comparatives. Unpubl. ms., Univ. of Trondheim/ Univ. of Notre Dame.
- Fenstad, Jens E., Per-Kristian Halvorsen, Tore Langholm, and Johan van Benthem. 1985. Equations, schemata and situations: A framework for linguistic semantics. Technical Report 29. Stanford University: CSLI.
- Halvorsen, Per-Kristian and Ronald M. Kaplan. 1995. Projections and Semantic Descriptions in Lexical-Functional Grammar. In Dalrymple, M. et al. (eds) 1995.
- Halvorsen, Per-Kristian 1995. Situation Semantics and Semantic Interpretation in Constraint-Based Grammars. In Dalrymple, M. et al. (eds) 1995.
- Hankamer, Jorge. 1973. Why there are two *than*'s in English. CLS 9.179-91.
- Heim, Irene. 2000. Degree operators and scope. In Jackson, B and Matthews T. (eds) *Proceedings of Semantics and Linguistic Theory 10*. Ithaca, NY: CLC Publications.
- Hellan, Lars. 1981. *Towards an Integrated Analysis of Comparatives*. Tuebingen: Guenter Narr Verlag.
- Klein, Ewan. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy* 4, 1-45.
- Thomason (ed) *Formal Philosophy*. New Haven: Yale University Press.
- Seuren, P.A. (1973) The comparative. In Kiefer, F. and Ruwet, N. (eds). *Generative Grammar in Europe*. Dordrecht: Reidel.
- Stassen, Leon. 1985. *Comparison and Universal Grammar*. Oxford: Basil Blackwell.

POSITION VS FUNCTION IN SCANDINAVIAN PRESENTATIONAL CONSTRUCTIONS

Kersti Börjars and Nigel Vincent

The University of Manchester

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

In some theoretical approaches, grammatical relations are assumed to be defined structurally, so that the crucial clue to the grammatical relation of an element is its position in the tree. Lexical Functional Grammar, in contrast, does not assume a universal one-to-one mapping between structural position and grammatical relation — though grammatical relations may well be defined structurally in some languages. This means that in languages which do not rely solely on structure, the grammatical relation of a particular element has to be established on grounds other than structure. In this paper, we look in particular at the association between postverbal position and objects. We consider postverbal noun phrases in an information-structurally marked construction in the Scandinavian languages, often referred to as the presentational construction. These postverbal noun phrases have been analysed as objects — largely on positional grounds — in transformational theories and also within LFG analyses. Analysing them as objects does, however, raise a number of problems, in particular in that they lack some crucial object properties and have some properties typical of subjects. In this paper, we provide evidence against an object analysis and formulate an analysis within which the postverbal noun phrase is a subject.

1. Introduction*

In this paper, we are interested in the general issue of the relation between functions and positions. In particular, we are concerned with constructions in which some element is found in a position which is non-canonical given its grammatical function and where this is motivated by the special information-structural conditions which hold for that element. We will look at the specific example of the so-called presentational construction in Swedish and Mainland Scandinavian more generally.

2. The issue: grammatical relations in presentational constructions

Presentational constructions in Swedish can occur with monotransitive verbs, both unaccusative (1a) and unergative (1b), with passive transitive verbs, as in (2), and certain ditransitive verbs, as in (3) (cf. Platzack (1983), Falk (1989, 1993) and Vikner (1995) for comparative data across the Scandinavian languages).¹

- (1) a. **Det** sitter **en kackerlacka** på locket. PAR
EXPL sit.PRES a cockroach on lid.DEF
'There is a cockroach on the lid.'
- b. **Det** hade bråkat **folk** på hennes buss. MA
EXPL have.PST causing.trouble.PRES people on her bus
'People had been causing trouble on her bus.'

* We are grateful to the participants of LFG05 for their comments, in particular Joan Bresnan, Ron Kaplan, Helge Lødrup, Irina Nikolaeva, Bjarne Ørsnes and Stephen Wechsler. We would also like to thank David Andréasson, Neil Ferguson and Helge Hoel, who helped with data, each in their own way.

¹ When an example has been taken from a corpus, a text available on the web or an article, this is indicated by the example. If there is no indication, the example is constructed. Similarly, unless there is an indication to the contrary, examples are from Swedish. A list of web sources is provided after the bibliography.

- (2) a. **Det** sätts **ett tak** på den norska laxproduktionen. PAR
 EXPL put.PRES.PASS a ceiling on the Norwegian salmon production
 ‘A ceiling is put on the Norwegian salmon production.’
- b. **Det** äts **många och långa middagar** här. AOS
 EXPL eat.PRES.PASS many and long dinner.PL here
 ‘Many a long dinner is eaten here.’
- (3) **Det** väntade mig **någon pinsam episod.** HIST
 EXPL wait.PST I.ACC some embarrassing episode
 ‘Some embarrassing episode was lying in wait for me.’

A major concern for any syntactic analysis of such constructions is what grammatical relations to assign to the two noun phrases in bold in each example. This question will be the focus of our paper and in order not to prejudice the issue, we will use terms neutral with respect to grammatical relations to distinguish the two: EXPL(itive)NP and P(ost)V(erb)NP, respectively.

In traditional approaches to sentences such as those in (1) to (3), subject properties are assumed to be distributed over both EXPLNP and PVNP and terms like FORMAL SUBJECT vs ACTUAL SUBJECT are used to contrast the two (e.g. SAG (4:53) and Faarlund, Lie & Vannebo (1997:827–8)).² In generative approaches, on the other hand, PVNP is often assumed to be the object, for instance by Platzack (1983), Askedal (1986) and Vikner (1995) for Scandinavian. Similar assumptions have been made by Lødrup (1999), who provides an analysis within OT-LFG (see also Bresnan & Kanerva 1989, Bresnan 1994 and van der Beek 2003, for arguments in a similar vein relating to other languages). In these approaches, it is assumed that only one of the two phrases can be the subject and hence two lines of arguments can be used to support the idea that the PVNP is the object. Firstly, by arguing that EXPLNP is the subject, it is implicitly argued that PVNP is an object and secondly, these authors claim that PVNP displays object properties.

Both approaches give rise to problems for descriptive as well as theoretical accounts of the data. In one instance, there appear to be two noun phrases filling the subject function. In the other, the noun phrase which fills the object function lacks some typical object properties and displays some other properties which are highly untypical of objects. The problems illustrated here by the presentational focus construction in the Scandinavian languages are by no means unique to these languages. Lambrecht (2000) provides a cross-linguistic study of similar constructions and concludes that the PVNP is, in fact, a subject which has absorbed object properties, hence capturing the conflict we have just described in a different way. In this paper, we will examine the status of EXPLNP and PVNP with respect to Swedish data, with a view to establishing how grammatical relations are distributed. We will also discuss the possibility that some properties which are assumed to be associated with certain grammatical relations are actually in a sense “meta-properties”, in that they are associated with a particular information-structural role which is frequently filled by the grammatical relation in question.

3. The status of the expletive

One of the subject properties of EXPLNP is that it appears to show agreement in the same way that a subject would. Finite verbs in Swedish do not show agreement at all, but participles do and the data in (4) suggests that the participle in a presentational focus construction agrees with the subject.

² Throughout this paper, we will use SAG to refer to Teleman, Hellberg & Andersson (1999).

- (4) a. **Det** blev **inlagt** fyra trafikoffer i går. SAG 4:385
 EXPL(NT.SG) become.PST admitPRT.SG.NT four traffic casualty.PL yesterday
 ‘Four traffic casualties were admitted yesterday.’
- b. **Det** blev **skjutet** en älg. Lødrup 1999:206
 expl(NT.SG) become.pst shoot.prt.SG.NT a.SG.COM moose(COM)
 ‘A moose was shot.’

However, as we shall see in §4, Lødrup’s statement that ‘The facts are especially clear in Swedish [...]’ (1999:206) is not entirely accurate; even for Swedish, the agreement data is more complex than this.³

It is worth pointing out here also that the neuter singular can be considered the default number–gender combination in Swedish (cf. Vincent & Börjars, To appear), so that the behaviour in (4) can be described as a failure to agree, rather than as agreement with the EXPLNP, which can also be described as neuter by default. This is indeed the way in which it is described in SAG (4:385). We will have reason to return to this in §4.

A further argument for the subject status of EXPLNP relates to question formation. In Swedish, as in other verb second languages, phrases fulfilling most functions within the clause can occur clause initially. This is illustrated for subjects in (5a) and for a direct object in (5b). As the parallel examples in (6) show, even though the subject and the object have the same position in the declarative, only the subject can be involved in question formation by inverting with the finite verb.

- (5) a. Oscar gillar honom.
 Oscar like.FIN he.ACC
 ‘Oscar likes him.’
- b. Honom gillar Oscar.
 he.ACC like.FIN Oscar
 ‘Oscar likes him [TOP/FOC].’
- (6) a. Gillar Oscar honom?
 like.FIN Oscar he.ACC
- b. *Gillar honom Oscar?
 like.FIN he.ACC Oscar
 ‘Does Oscar like him.’

As the examples in (7) show, the EXPLNP in presentational constructions behaves like a subject in this respect (compare with (1a) and (2a)).

- (7) a. Sitter det en kackerlacka på locket?
 sit.PRES EXPL a cockroach on lid.DEF
 ‘Is there a cockroach on the lid?’
- b. Sätts det ett tak på den norska laxproduktionen?
 put.PRES.PASS EXPL a ceiling on the Norwegian salmon production
 ‘Is a ceiling put on the Norwegian salmon production?’

³ Christensen & Taraldsen (1989:58–59) claim that the data is clearcut also in Norwegian, though dialects divide into two types, with some dialects having *det* as the expletive and a non-agreeing participle and other dialects having the expletive *der* and an agreeing participle.

As exemplified in (8), the expletive in these constructions can also function as the subject of a raising verb such as *verka* ‘seem’. It is generally assumed that only the subject of the lower clause can do this (cf. Askedal 1986:27).⁴

- (8) a. Det verkar sitta en kackerlacka på locket.
 EXPL seem.PRES sit.INF a cockroach on lid.DEF
 ‘There seems to be a cockroach sitting on the lid.’
- b. Det verkar sättas ett tak på den norska laxproduktionen?
 EXPL seem.PRES put.INF.PASS a ceiling on the Norwegian salmon production
 ‘Is a ceiling put on the Norwegian salmon production?’

Further subject properties associated with EXPLNP have been discussed in the literature, but given that the status of the EXPLNP is less disputed than that of the PVNP, we will concentrate the discussion on the latter here.⁵

3. The status of the postverbal NP

3.1 Object properties of PVNP suggested in the literature

The argument for object status of the PVNP which is used most commonly in the literature is its position. Askedal (1986:31), for instance, states that their position shows PVNPs ‘quite unequivocally to be direct objects’. In transformational approaches, grammatical relations are derived directly from structural positions and arguments based on position are then to be expected (cf. for instance Platzack (1983) and Åfarli (1992)). In an approach like LFG, on the other hand, where grammatical relations need not be defined configurationally, the object status of the PVNP need not follow from its position. There are, however, languages where functions are defined positionally, English is one example. Lødrup (1999), in his LFG-OT analysis of presentational constructions, assumes this holds also for Scandinavian languages and hence he uses the PVNP’s position to argue for its object status. His argument is not based solely on the PVNP’s position with respect to the verb, but also to the indirect object: ‘A direct object in a presentational focus sentence has the same position as in a non-presentational sentence, it is the sister of V-position after the indirect object if there is one.’ (Lødrup 1999:206) The Norwegian examples he adduces can be found in (9).

- (9) a. **Det** ventet meg **problemer** Lødrup 1999:205
 EXPL await.PST me problem.PL
 ‘Problems were awaiting me.’

⁴ The facts are slightly complicated here by the fact that objects and other functions can appear in initial position, preceding a raising verb. However, given that these languages are verb second languages, occurring in initial position is not sufficient for subject status. The crucial property of the initial element in (8) is that it functions as a subject for instance with respect to question formation.

⁵ Faarlund, Lie & Vannebo (1997:833) mention in passing some criteria which have not occurred in the theoretical literature as far as we are aware, for instance the fact that whereas EXPLNP can occur in a finite subordinate clauses, it shares with most other subjects the property that it does not occur in non-finite clauses. Also, there is a type of tag question in both Norwegian and Swedish which treats the EXPLNP as the subject.

- b. **Det** ble overrakt barna **en liten gevinst** Lødrup 1999:206
 EXPL become.PST award.PRT child.PL.DEF a small prize
 ‘A small prize was given to the children.’

However, it has been argued by Sells (2001) and more strongly by Börjars, Engdahl & Andréasson (2003) that word order in the so-called midfield is not driven purely by syntactic constraints, but that it is heavily influenced by information-structural constraints. In particular, Börjars, Engdahl & Andréasson show that known information tends to precede new information and that this is a more important consideration than the desire for the subject to occur immediately after the verb when something else occupies the pre-verbal slot (2003:54–55). This would then open up an alternative interpretation of the data in (9), namely that the bold PVNP is the subject, but that it follows the indirect object for more widely applicable information-structural reasons, in particular because it is a weak pronoun. This would indeed make examples such as (9) parallel to the example in (10), for which one would not want argue that *Tutankhamun* functions as anything but subject.

- (10) **Därför** gav dem Tutanchamons förbannelse ingen ro. MA
 therefore give.PST they.ACC Tutanchamon.POSS curse no peace
 ‘For this reason, Tutankhamun’s curse did not give them any peace.’

If the bold noun phrases in (9) are regular direct objects which follow indirect objects, in line with general word order constraints in the language, then we should expect to find parallel examples where the indirect object is a full noun phrase. However, a search of corpora and the web shows first that constructions parallel to the Norwegian examples in (9) are quite rare in Swedish and second that the indirect objects which occur between the verb and the PVNP in presentational constructions are weak pronouns. Examples are provided in (11).

- (11) a. **Det** gavs dem **ingen tid** till gottgörelse. Portugalien
 EXPL give.PST.PASS them no time for recompense
 ‘They were given no time to make amends.’
- b. **Det** gavs honom **goda möjligheter att utarbeta sitt system.** Tektid
 EXPL give.PST.PASS him good opportunities INF draw up his system
 ‘He was given good opportunities to work out his system.’

In the majority of presentational constructions which involve a recipient/benefactor, this is expressed through a PP, which follows the PVNP, as in (12) (cf. also examples in SAG 4:385 fn1). If the order in (10) and (11) is simply the result of a general tendency for indirect objects to precede direct objects, then this is surprising. If, on the other hand, the PVNP is a subject, then the fact that the indirect object can only precede it when it is a weak pronoun follows straightforwardly. Example (12b) is especially interesting, since the PVNP is “heavy” and might therefore be expected to occur as far to the right as possible (cf. ‘heavy NP shift’), still the alternative in which the recipient occurs as a pronoun NP immediately following the verb is dispreferred in this case.

- (12) a. **Det** gavs **ingen hjälp** för honom. PAR
 EXPL give.PASS no help for him
 ‘No help was given to him.’

- b. **Det** överräcktes **färdryck av bubbelkaraktär samt** TRK
 EXPL hand over.PST.PASS travel drink of bubble character as well as
älgar av den mjukare sorten till oss.
 elk of the softer kind to us
 ‘We were presented with a glass of sparkling wine for the road and some cuddly
 toy elks.’

We take this to be evidence that the pronouns (and occasionally other non-rhematic NPs) which occur immediately after the verb and precede the PVNP do so for information-structural reasons and no conclusions can be drawn on the basis of this about the grammatical relation associated with PVNP.⁶ We conclude then that positional evidence for object status is not reliable for Swedish and we suspect this extends to V2 languages more generally (*pace* van der Beek 2003).⁷

A further argument for the object status of PVNP relates to the non-existence of presentational constructions containing active transitive verbs (cf. Askedal 1986, Lødrup 1999, Mikkelsen 2002). This is illustrated here from Norwegian in (13), but parallel examples in Swedish are also ungrammatical.

- (13) *Det spiste en mann en kake Lødrup 1999:207
 EXPL eat.PST a man a cake
 ‘There was a man eating a cake.’

If PVNPs are themselves direct objects, it is argued, then the requirement for there to be one unique filler of that function would rule out active transitive presentational constructions.⁸

There are, however, factors which weaken this argument. First, for verbs which are optionally transitive, the presentational focus construction is ruled out even when there is no object. It would be inappropriate to analyse a sentence such as (14a) as having the object function filled. This would amount to a kind of object pro-drop analysis which is otherwise unmotivated for Swedish. The ungrammaticality of (14b) can then not be accounted for by reference to a syntactic constraint relating to the co-occurrence of two direct objects. Given that it would be desirable to explain the ungrammaticality of (13) and (14b) in the same terms, relying for an explanation on the assumption that *en mann* is an object in (13) is not satisfactory.⁹

⁶ Vikner (1995) and Mikkelsen (2002) both give examples of constructions of the type [EXPLNP — V — non-pronominal IO — PVNP] also in Danish with verbs which have cognates in Swedish; *tillföll* ‘accrue to.PST’ and *skänktes* ‘donate.PST.PASS’. Searches of available corpora and the web have not thrown up any examples of the relevant construction with these verbs. This could be taken as evidence that such constructions do not occur naturally because the information-structural constraints which allow the indirect object to precede the PVNP do not apply to a full noun phrase in the sense that non-rhematic noun phrases would tend to be expressed by a pronoun.

⁷ Van der Beek uses the notion of ‘object position’ in her argumentation, but states ‘I will call the complement OBJ, even though I realize that it is not a regular object’ (2003:25).

⁸ In some accounts, the same generalisation is stated in terms of Case; only one noun phrase can be assigned Accusative Case (e.g. Mikkelsen 2002:15).

⁹ The issue is further complicated by the fact that certain non-object complements behave in peculiar ways with respect to presentational constructions too. Even though *vara* ‘be’ is permitted in presentational constructions — indeed even in languages which are restrictive with respect to presentational constructions, BE usually permits it — there are some odd restrictions, compare (i) and (ii):

- (i) a. En intervju med Hans Blix var på TV igår.
 a interview with Hans Blix be.PST on TV yesterday
 b. Det var en intervju med Hans Blix på TV igår
 EXPL be.PST a interview with Hans Blix on TV yesterday

- (14) a. Ett litet barn åt i köket.
a.NT small.NT.SG child(NT) eat.PST in kitchen.DEF
- b. *Det åt ett litet barn i köket.
EXPL eat.PST a.NT small.NT.SG child(NT) in kitchen.DEF
‘A small child was eating in the kitchen.’

It is not clear to us what the restriction is in these cases, but a prohibition against two direct objects does not capture the generalisation correctly and hence this is not an argument for the direct object status of PVNP.

Consider in this context also English locative inversion sentences, which are similar to the presentational focus sentences in the sense that a subject occurs in a non-canonical postverbal position because of marked information-structural properties. Bresnan (1994) argues that the PVNP is the object in these construction, where peculiar effects on argument structure can also be observed. Only intransitive or passivised transitive verbs can occur in locative inversion in English, but it is interesting to see that in the latter case, a *by*-phrase expressing the agent cannot be present as illustrated in (15), from Bresnan (1994:78). This could obviously not be a restriction on the occurrence of two objects.

- (15) ??Among the guests of honour was seated my mother by my friend Rosie.

Askedal argues that VP pronominalisation also provides evidence of the object status of the PVNP: ‘In cases of VP Pronominalization, the indefinite NP is suppressed in exactly the same way as objects of transitive verbs (whereas *det*, being a syntactic subject, is of course retained)’ (Askedal 1986:29). The Norwegian examples to support the argument are found in (16).¹⁰

- (16) a. (Spiser han kaker?) Ja, han gjør det. Lødrup 1999:207
eat.PRES he biscuit.PL yes he do.PRES it
‘(Is he eating biscuits?) Yes he is.’
- b. (Arbeider det en manni skogen?) Ja, det gjør det.
work.PRES EXPL a man in forest yes EXPL does it
‘(Is there a man working in the forest?) Yes there is.’

However, whereas we consider this evidence that *det* in (16b) is a subject, just like *han* is in (16a), we would argue that examples such as these do not reveal anything about what type of arguments are contained within the VP. Indeed, as the Norwegian example in (17) shows, the construction is possible also when there is no argument within the VP.

- (17) (Regner det?) Ja, det gjør det.
rain.PRES EXPL yes EXPL do.PRES it
‘(Is it raining?) Yes it is.’

-
- ‘There was an interview with Hans Blix on TV yesterday.’
- (ii) a. En intervju på TV igår var så tråkig att somnade.
an interview on TV yesterday be.PST so boring that I fell asleep
- b. *Det var en intervju på TV igår så tråkig att jag somnade.
EXPL be.PST an interview on TV yesterday so boring that I fell asleep
‘There was an interview on TV yesterday which was so boring that I fell asleep.’

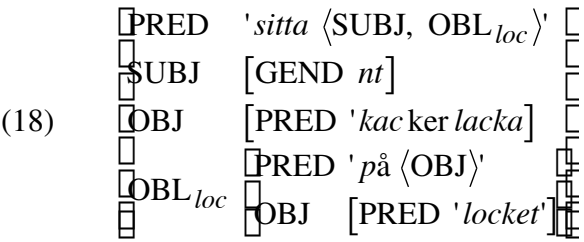
¹⁰ We use the examples from Lødrup (1999:207) rather than Askedal (1986), since there is a clearer parallelism in Lødrup’s examples.

In conclusion, having surveyed some of the arguments used in the literature to support the assumption that the PVNP is the object, we find none of them convincing.

3.2 Problems with assigning PVNP object status

The analysis of the PVNP as an object also raises a number of problems for any analysis. First, as has been pointed out in the literature, if the PVNP is the object in sentences such as (1b), the agentive role associated with the verb is mapped to an object, something which is typologically extremely rare.¹¹ This issue and how to resolve it is at the core of Lødrup (1999). These objects would also fail to show a number of properties normally associated with objects; for instance in that they resist passivisation. Furthermore, as we shall see in §3.3, the PVNP may agree with the verb in Swedish, and this in a language which does not otherwise show object agreement.

Serious issues also arise with the match between selectional properties of the verb and the way in which arguments are realised. We discuss this issue here in terms of LFG, but it will arise within any theory albeit expressed in different terms. The f-structure associated with a sentence such as (1a) would have to be something akin to the simplified representation of given in (18).



In general terms, the problem here is that an argument function required by the verb, namely the subject, does not have a filler with a semantic value and there is an object which is not selected by the verb. In LFG terms, the f-structure in (18) violates Completeness in that the SUBJ function which expresses an argument of the verb does not contain a PRED feature. There is also a violation of Coherence in that the OBJ function found in the f-structure is not selected by the PRED of the verb.

3.3 Subject properties of PVNP

A further problem with analysing the PVNP as an object is the fact that the noun phrase actually shows a number of subject properties. For instance, the PVNP can show agreement with the predicate. Although finite verbs in Swedish do not show agreement, in certain constructions a participle may agree with the subject.

The tendency is that if the participle follows the PVNP, then it usually agrees with the PVNP, as in (19a), whereas if the participle precedes the PVNP, as in (4), then it usually occurs in

¹¹ Kroeger (1993) argues that Tagalog has agentive objects, but it should be pointed out that grammatical relations have a very different status in Tagalog.

neuter singular.¹² In a constituent question in which the PVNP occurs in clause initial position, either form of the participle is frequent, as (19b) indicates.¹³

- (19) a. Det blev fyra trafikoffer inlagda igår. SAG 4:385
 EXPL become.PST four casualty.PL admit.PRT.PL yesterday
 ‘Four traffic casualties were admitted yesterday.’
- b. Hur många brev blev det SAG 4:385
 how many letter.PL become.PST EXPL
 skrivet / skrivna igår?
 write.PRT.NT.SG write.PRT.PL yesterday
 ‘How many letters were written yesterday?’

This split agreement behaviour is then similar to that found in English with *There is/are three blokes in the living room*.¹⁴

Case marking has been used in the literature to argue for the object status of the PVNP. In particular, analyses of Danish and Norwegian within which the PVNP is considered the object appeal to the fact that the PVNP occurs in the object case (cf. Mikkelsen (2002:10) for Danish and Askedal (1986:31) for Norwegian). However, this should not necessarily be taken as evidence that the PVNP is the object in Danish and Norwegian; languages vary as to what case they assign to grammatical relations in non-canonical positions. Furthermore, in Swedish, the PVNP does in fact have to occur in its subject form, as in (20).

- (20) Det var bara hon / *henne hemma.
 EXPL be.PST only she.NOM she.ACC at home
 ‘Only she was at home.’

Given the close similarities in the properties of the PVNPs in these languages, it would seem inappropriate to draw two different conclusions based on this limited data. It is more likely to be attributable to the languages dealing with the notion of default case in different ways.¹⁵

The PVNP may also show subject-like behaviour in that it can bind reflexives. Swedish has a reflexive determiner *sin*, which can only be bound by a subject and not by a direct object as in (21a). As (22b) shows, it can be bound by the PVNP.¹⁶

¹² Note that the tendency for the participle to be more likely to agree with the PVNP when it follows it may lead one to consider this a separate type of construction in which the PVNP and the participle form a “small clause” (see also discussion in Mikkelsen 2002:16).

¹³ Christensen & Taraldsen (1989:70–72) and Lødrup (1999:206) do discuss Swedish non-agreement, but only compare the agreeing past participle (used with passive *bli*) with the non-agreeing supine form (used with the perfective *have*). They do not discuss the neuter versus non-neuter agreement.

¹⁴ The agreement with *there*, or maybe better the lack of agreement with PVNP in English is often assumed to be a feature of modern informal English, but both types of agreement are mentioned for instance in Kruisinga and Erades (1911) and Jespersen (1924:155).

¹⁵ Note for instance the difference in responses to sentences such as *Who wants a beer?*, where Danish would use the object form of the pronoun (like English), whereas Swedish would use the subject form.

¹⁶ Askedal (1986:29) gives examples similar to (18b), marks them with a question mark and comments in a footnote that ‘possibly sentences like (14) [parallel to (18b), KEB/NBV] are so to speak “acceptable by default”’. His argument rests on the fact that the non-reflexive possessive would be totally unacceptable in the same position. This does not, however, strike us as a strong line of argument given that it is a fundamental property of *sin* that it is in complementary distribution with the parallel non-reflexive pronoun and hence always occurs where *hans* is unacceptable.

- (21) a. Peter_i kittlade Oscar_j med sin_{i/*j} / hans_{*ilj} fjäder.
 Peter tickle.PST Oscar with his.REFL.COM his feather
 ‘Peter tickled Oscar with his feather.’
- b. Det kom en man_i med sin_i / *hans_i fru.
 EXPL come.PST a man with his.REFL his wife
 ‘There came a man with his (own) wife.’

The ability to control subjects of non-finite clauses is commonly assumed to be a property unique to subjects. This is borne out for Swedish by examples like (22a), where the subject of the non-finite verb can only be interpreted as co-referential with the subject of the main clause. As the indices in (22b) shows, the PVNP can function as the controller in such constructions.

- (22) a. Oscar_i såg en man_j och pro_{i/*j} vinkade.
 Oscar see.PST a man and wave.PST
 ‘Oscar saw a man and waved.’
- b. Det kom en man_i och pro_i pratade med mig.
 EXPL come.PST a man and talk.PST with me
 ‘There came a man and talked to me.’

With reference to Platzack (1983), Askedal (1986) and Áfarli (1992), Lødrup (1999:207) states that ‘It should be mentioned that this analysis [of the PVNP as a direct object KEB/NBV] is not controversial in Scandinavian generative grammar.’ Our conclusion on the basis of the data we have surveyed is quite the contrary, the status of the PVNP as the object gives rise to a large number of problems. Instead if the PVNP is analysed as a subject, then its unusual properties can be accounted for in terms of its unusual information-structural properties.

4. Some potential LFG analyses

4.1 The generalisation

In the foregoing we have argued that both the EXPLNP and the PVNP share properties which are usually associated with subjects, and that the PVNP is only object-like in its apparent occupancy of the post-verbal position while in most other respects it does not display prototypical object properties. Indeed, we would suggest that the respects in which the PVNP displays object-like properties, these are not properties of the object per se, but rather properties associated with non-topics and here serve to mark the subject as a non-topic (Lambrecht 2000). Hence we argue that in a sentence such as (1) — repeated here for convenience — the expected subject does not occur in its canonical position because it has marked information-structural properties, and the EXPLNP *det* inherits many but not all of its subject properties.

- (1) a. **Det** sitter **en kackerlacka** på locket. PAR
 EXPL sit.PRES a cockroach on lid.DEF
 ‘There is a cockroach on the lid.’

b. **Det** hade bråkat **folk** på hennes buss. MA
 EXPL have.PST causing.trouble.PRES people on her bus
 ‘People had been causing trouble on her bus.’

This claim is in accord with what traditional grammarians have claimed. Thus SAG says: ‘The expletive pronoun *det* functions as the subject a number of clause types. In most cases, the construction appears to be a way of allowing a phrase which would otherwise have functioned as the subject associated with the predicate a place in the rhematic part of the clause.’ (SAG 4:53)¹⁷ However, it is also in line with the idea behind some modern theoretical analyses as expressed by Falk (1993: 250): ‘I will argue that topic *det* has a pragmatic function of signalling a non-topic subject...’.¹⁸

The consequence of these arguments is that we reject an analysis in which EXPLNP is assigned the SUBJ function and the PVNP the OBJ function, at least for Swedish and in all likelihood by parity of argument for the other Mainland Scandinavian languages as well.¹⁹ In this section we review some alternative analyses, and consider their implications for the architecture of LFG.

4.2 PRESFOC and anticipatory pronoun: a syntactic account

On what we will call the syntactic account, the EXPLNP *det/der* is assigned the SUBJ function. The PVNP is also assigned the function SUBJ. The latter does not occur in its canonical position because it has information-structural properties not compatible with being on the left edge of the clause. To capture this additional property, we assign a discourse function PRESFOC to PVNP.²⁰ The f-structure link between SUBJ and PRESFOC is established by the following rules, which in turn assume the general clause structure argued for in Börjars, Engdahl & Andréasson (2003):

¹⁷ Translated from Swedish: ‘Det expletiva pronomenet *det* fungerar som subjekt i olika satstyper. I de flesta fallen verkar konstruktionen vara ett sätt att ge det led som annars skulle ha fungerat som subject vid satsens predikat en plats i satsens rematiska del.’

¹⁸ The historical development is as follows. In OSw, *det* occurs frequently as an introduction to a new story; ‘*Det* clearly has the function of an introductory marker, not only of a clause with a non-topical subject, but of a whole story.’ (Falk 1993:256) At this stage, as Falk shows, *det* occurs only with the verb ‘be’ and the NP is often followed by a relative clause. By the early Modern Swedish stage, *det* has developed into a marker of non-topical subjects and the construction is no longer restricted to one particular verb. At this stage, Falk (1993:260–263) argues that *det* does not yet function as a subject, for instance in that it is not obligatory. However, when Swedish develops into a non-pro-drop languages, *det* does develop into an expletive subject.

¹⁹ See Bresnan (1982: 72-80) for an early LFG account along these lines of the English *there* construction. In conference discussion Joan Bresnan reiterated her preference for this analysis citing in support the fact that in English the NP in the *there* construction cannot trigger deletion of a co-referential co-ordinate subject, hence the ungrammaticality of **there were two children in the room and played quietly together*. How, she asks, is this to be explained if *the children* is treated as SUBJ? The solution to this apparent paradox, we suggest, is once again to be found in information structure. Although *two children* here is SUBJ, informationally it has a different status from the subject of the following clause. It is plausible to suppose that it is this lack of informational parallelism which blocks deletion in these circumstances.

²⁰ The term PRESFOC is a convenient but in some respects arbitrary label, and another term may be more appropriate if this analysis were to be developed within a more fully worked out approach to information structure. For the sake of simplicity, we will treat PRESFOC as an attribute within the f-structure. However, it may well be more appropriately located in a separate and more elaborate i(nformation)-structure or d(iscourse)-structure (compare for instance King 1997; Choi 1999; Dahlstrom 2003; O’Connor 2004). This is an issue which we will not pursue here.

- (23) a. $F'' \sqsupset F', NP$
 $\uparrow = \square$ ($\uparrow GDF$) = \square
- b. $F' \sqsupset F, \dots, NP$
 $\uparrow = \square$, \dots , ($\uparrow PRESFOC$) = \square \dots
 $(\uparrow SUBJ) = \square$

This account is actually a schematic way of representing our approach. Assuming the general principles argued for in Börjars, Engdahl & Andréasson (2003) with respect to the interaction between c-structure rules and alignment constraints, in a more articulated version of this analysis, the most appropriate way of accounting for the positioning of the PVNP, may not be to explicitly introduce its status as PRESFOC through the c-structure rule in (21). Instead, the particular information-structural properties of this subject would be accounted for in terms of alignment constraints referring both to grammatical relations and discourse functions.

We are proposing then that in presentational focus sentences like the ones we deal with here, both the EXPLNP and the PVNP contribute to the subject function; *det* is introduced by (23a) and the GDF is instantiated as SUBJ and the PVNP is introduced by (23b). This may appear to lead to problems in that a feature clash would be expected as two elements in the c-structure are associated with one f-structure function. The PVNP, being a full noun phrase, will have a semantic value (expressed in LFG as a value for the PRED feature) as well as agreement features such as number and gender. If the EXPLNP has conflicting features, a clash would clearly arise. However, in section 2, we pointed out that the form which has been described in the literature (e.g. Lødrup 1999) as agreeing with the expletive, is in fact the default number and gender form, so that it could also be described as a non-agreeing form. A default form is the form in which a category, in this case an adjective or a participle, occurs when the filler of the function with which it would normally agree does not carry the relevant feature. For instance, a clause would not normally be assumed to have gender and number features and an adjective agreeing with a clause would then go into the default form, in Swedish singular neuter.²¹ Now, if we claim that the participles in (3) do in fact not agree, but occur in a default, then this amounts to claiming that the element it might be expected to agree with does not have the relevant features.²² Of course, since the expletive *det* is homophonous to the third person singular neuter pronoun in Swedish, it is tempting to say that there is agreement. However, just like adjectives which would in principle agree with a clause occurs in the default gender and number form, so does a pronoun which refers to a clause. Furthermore, in some varieties of Swedish and Norwegian and in Danish, the expletive used is *där* (*der*), homophonous with a distal adverb (cf. SAG 4:54). In these cases, one would not want to ascribe the features singular and neuter to the expletive. In fact, since this expletive can be combined with a proximal adverb in such dialects, as in (24), it is clear that it does not have a proximal/distal feature either.

- (24) a. Där finns gott om plats här.
EXPL exist.PRES good of place here
‘There is plenty of space here.’

²¹ For further discussion of adjectives and default forms in Swedish, see Vincent & Börjars (To appear).

²² For a very different analysis of Norwegian which also relies on the assumption that the neuter singular forms are there as default forms, see Christensen & Taraldsen (1989:61–63). Note, however, that they assume that *det* in Norwegian has inherent gender and number features, whereas we assume it lacks them. Their analysis of Swedish also differs from ours with respect to default and agreeing forms (1989:70–72).

There are then a number of reasons to assume that the expletive *det* is not the third person singular neuter pronoun, but instead it is a pronoun with no person, number or gender features. This then means that it would not contribute any agreement features to the SUBJ function and hence it would not cause a clash. If we assume also that expletives do not have any semantic content, *det* would not have a PRED feature value, again avoiding a clash with the features of the PVNP. The expletive is still appropriately described as the subject, since it has the positional features associated with a subject, for instance in that it is the element which can invert with the finite verb to form a question. Preverbal elements which are not subjects cannot do this.²³

The analysis we have proposed here is akin to the focus analysis which could be proposed for English sentences with a fronted focus and a resumptive pronoun; *Beans, I like them*. A big difference with the constructions we have analysed here is that in the presentational focus, the relation between the two elements which contribute to the same function is always local and limited to subjects. Hence an analysis in terms of functional uncertainty is not required.

4.3 OT accounts

Two other OT accounts of this construction exist in the literature. The first is proposed by Lødrup (1999) and depends on marking the presentational focus in the input and then postulating a constraint, REALIZE FOCUS, which ‘says that an argument that is marked as presentational focus in the input should be realized in the presentational focus position, which is object position in Norwegian.’ (Lødrup 1999:213). Independently of the empirical arguments we have adduced against this analysis in §3 above, there is clearly some circularity here since the behaviour we are seeking to explain is written in both to the input and to the definition of the constraint. The fact that this analysis leads to agents being realised as objects is dealt with by assuming for the grammar of Norwegian a ranking in which REALIZE FOCUS takes precedence over *AGENTIVE OBJ, which universally disprefers the association of Agent with Object.

The second OT-based analysis, by Mikkelsen (2002), is not conceived within LFG but the leading ideas translate straightforwardly enough. Her account differs from Lødrup’s in that what she proposes to mark in the input is not presentational focus but definiteness. This feeds into the following constraint ranking (somewhat simplified here for expository purposes):

(25) SUBJ » *EXPLETIVE » *SUBJ/DEF

In other words, it is more important that there should be a subject than that the language should avoid using expletives, but when there is a definite subject available it is better to avoid expletives. From this it follows that: ‘Where previous analyses posit a restriction on what can appear in the pivot position, the present analysis treats the definiteness effect as an epiphenomenon arising from the interaction of constraints governing the subject position’ (Mikkelsen 2002: 29).

From our point of view, Mikkelsen’s account is an improvement on Lødrup’s since she does not fall into the trap of circularity, since for her presentational focus now follows interpretively and is not stated as a separate DF.²⁴ Instead what she appears to have done is to replace the definiteness effect in postverbal position with indefiniteness effect in subject position, which is

²³ We will also assume that there should be no need to stipulate that *det* cannot bear a DF (contrast Sells (this volume) on Icelandic) since this should follow from the properties we assume are associated with the expletive.

²⁴ Since she is working within a Chomskyan framework the problem of agentive objects does not directly arise as such. Instead she has to ensure that accusative Case is checked on the DP in postverbal position.

equally stipulative. The epiphenomenal effect which she rightly seeks is, we would argue, better handled in the account we offer in §4.2 where the linear ordering of constituents follows from the fact that there is a contradiction between being definite and expressing rhematic information.

4.4 Discontinuous subjects: a morphological account?

A near relation of the analysis we have proposed in §4.2 treats *det* and its ‘associate’ as a discontinuous realization of the single GF SUBJ, in some respects similar to clitic doubling. In effect such an analysis moves the resolution of the problem into the domain of morphology and sees the initial expletive as a kind of anticipatory agreement. If both *det* and its associate are SUBJ, then we once again get the problem of feature clashing in that *det* is neuter but the subject may well not be. In Danish and Norwegian but not in Swedish there is also a case clash since *det* is presumably nominative but the postverbal argument is accusative (*der var kun hende/*hun tilbage* ‘there was only her/*she left’; Mikkelsen 2002: 10). This clashing feature problem can however be resolved in the same way as we propose under 4.2.

4.5 Resumption/anticipation and resource management: a semantic account?

So far we have discussed alternative analyses which work at different levels, involving one or more of syntax, morphology and discourse. If only for completeness’ sake, it is natural to ask in addition whether a semantic account might not also be possible. Consider as a point of departure the treatment of resumptive pronouns in Asudeh (2004). Cross-linguistically these occur in long distance dependency constructions such as relative clauses and constituent questions, as in the Swedish example (26) analysed in detail by Asudeh (2004: chap 7.1)

- (26) Vilken elev_i trodde Maria att han_i skulle fuska?
‘Which student did Maria think that (he) would cheat?’

The problem here is to avoid both the questioned constituent *vilken elev* and the resumptive pronoun *han* making independent contributions to the overall interpretation of the sentence since this would lead to a violation of uniqueness. Instead of handling this problem, as is traditionally done, at the level of syntax, Asudeh proposes an account of resumptive pronouns which treats them at f-structure simply as pronouns with their own PRED feature but with the resource logic removing them from the computation of the semantic value of the clause.

The construction we have been discussing shows some clear analogies with this situation since the pronominal *det* has what might be called as a ‘presumptive’ or cataphoric function but does not make an independent contribution to the argument structure of the clause. Although we have neither the space nor the time to work out the details, it is possible that the presentational construction would yield to a similar, resource logic based account. The crucial thing from the present perspective is that once again *det* and the PVNP would both be assigned the f-structure role SUBJ and so the apparent paradox of agentive objects would disappear.

5. Consequences of alternative analyses

In this section we highlight the consequences of our arguments for our understanding of the presentational constructions both in the modern Mainland Scandinavian languages and more widely, while in the next section we will briefly review the implications for LFG as a model of natural language structure.

A first issue concerns the status of the arguments within presentational clauses, and in particular whether it is legitimate in this instance to talk of agentive objects. As we have already noted, there is a strong cross-linguistic dispreference for realizing agents as objects. In the literature the Scandinavian presentational construction is one of those most frequently cited in support of the idea that nonetheless there are special circumstances in which an agent can surface as an object. While we cannot claim to have refuted potential instances in other languages, we believe that we have provided sufficient evidence that the items in question are not objects.

It is also possible to question whether they are even agents, that is to say whether movement to the postverbal position does not in fact trigger a process of deagentivization. The fact that agent oriented adverbs are odd in *there* constructions have been used as evidence for this claim. However, as Engdahl (To appear, 37–38) points out, this cannot be a true semantic constraint since agentive modification can be added in a second clause, as in (27b).

- (27) a. *Det arbetar motvilligt 5000 lärare på universitetet. Engdahl To app:37
 EXPL work.PRES reluctantly 5000 teacher.PL at university.DEF
 ‘5000 teachers work reluctantly at the University.’
- b. Det arbetar 5000 lärare på universitetet, Engdahl To app:38
 EXPL workPRES 5000 teachers at the.university
 flera av dem ganska motvilligt.
 several of them rather reluctantly
 ‘5000 teachers work at the University, several of them reluctantly.’

See also Lødrup (1999:211–2) for references and critical evaluation of deagentivisation.

Once it is accepted that the PVNP is not an object but that both it and the EXPLNP are in some partial and complementary respects subjects, then the question that naturally arises is: what are the principles that trigger the construction? For Lødrup it is a matter of a special type of so-called presentational focus, whereas for Mikkelsen the question relates to definiteness. There are genuine insights in both these approaches. We would argue that our approach captures the best of both in that it allows the information structure, and the alignment conditions relating it, to derive the linear positioning. Properties such as definiteness can then be related to information-structural notions relating to newness as appropriate and hence have an indirect relation with the conditions on the construction.

6. Implications for LFG

Perhaps the most obvious implication for LFG lies in the treatment of expletives, which on our analysis are allowed to share their grammatical function with a substantive and semantically complete item (i.e one which has its own PRED feature) elsewhere in the clause. This is not a property which is limited to subjects; consider for example the sentences in (28):

- (28) a. I hate you(r) being rude to your uncle.
 b. I hate *(it) when you are rude to your uncle
 c. Konstbevattningen gjorde det möjligt PAR
 irrigation.DEF make.PST EXPL possible
- att i stället få ut tre risskördar varje år
 INF instead get out three rice harvests every year
 ‘The irrigation made it possible instead to get three harvests of rice per year.’

On the assumption that the string *you(r) being rude to your uncle* is to be assigned the function COMP it would be natural to treat *it* as a COMP-expletive just like *det* is a SUBJ-expletive. A similar argument holds for (28c) and parallel constructions in a number of languages.²⁵ This in turn would require a generalized mechanism for connecting expletives and their ‘associates’.

Casting the net wider, there is scope for further research into what exactly constitutes an expletive and how it relates to other types of pronoun. We have assumed here that an expletive lacks features altogether. In particular, we have treated an expletive as having no PRED feature but Asudeh’s (2004) account of resumptive pronouns, which in some respects are similar to expletives, assigns all pronouns the PRED feature ‘pro’ and deals with the consequences of this move at the level of the resource logic. We have also suggested that even the apparent positive specifications of such pronouns for number and gender are in fact the consequences of featural absence. However, as Louise Mycock (p.c.) reminds us, some expletives seem to need to have other morphosyntactic features such as the [+ wh] property in so-called *wh*-expletive constructions (see Mycock 2004 for references and discussion).

Finally, our arguments add further weight to the case for exploring and refining the ways grammatical and discourse functions interact within the architecture of LFG, and how in turn these functions map onto linearly defined positions within the clause.²⁶

References

- Åfarli, Tor 1992. *The syntax of Norwegian passive constructions*. Amsterdam: John Benjamins.
- Alsina, Alex 1996. *The role of argument structure in grammar. Evidence from Romance*. Stanford: CSLI.
- Alsina, Alex, Tara Mohanan & K.P. Mohanan This volume. How to get rid of the COMP.
- Askedal, John Ole 1986. On ergativity in modern Norwegian. *Nordic Journal of Linguistics* 9:25–45.
- Asudeh, Ash. 2004. Resumption as resource management. PhD Dissertation. Stanford University.
- Börjars, Kersti, Engdahl, Elisabet, & Andréasson, Maia. 2003. Subject and object positions in Swedish. In *The Proceedings of the LFG’03 Conference*, eds. Miriam Butt and Tracy Holloway King. Stanford, Ca: CSLI Publications. 43–58
- Bresnan, Joan 1994. Locative inversion and the architecture of Universal Grammar. *Language* 70:72–131.

²⁵ The alternative analysis would treat clausal complements as bearing functions like SUBJ and OBJ, as proposed in Alsina, Mohanan & Mohanan (this volume).

²⁶ Compare in this connection the work of Dalrymple & Nikolaeva (2005).

- Bresnan, Joan & Kanerva 1989. Locative inversion in Chichewa: a case study of factorization in grammar. *Linguistic Inquiry* 20: 1-50.
- Choi, Hye-Won 1999. Optimizing structure in context: scrambling and information structure. Stanford, Ca: CSLI Publications.
- Christensen, Kirsti Koch & Taraldsen, Tarald 1989. Expletive chain formation and past participle agreement in Scandinavian dialects. In *Dialect variation and the theory of grammar*, ed. Paola Benincà. Dordrecht: Foris. 58–83.
- Dahlstrom, Amy 2003. Focus constructions in Meskwaki (Fox). In *The Proceedings of the LFG'03 Conference*, eds. Miriam Butt and Tracy Holloway King. Stanford, Ca: CSLI Publications. 144–163.
- Dalrymple, Mary 2001. *Lexical Functional Grammar*. San Diego & London: Academic Press.
- Dalrymple, Mary & Irina Nikolaeva 2005. Agreement and discourse function. Paper presented as part of the workshop on *Agreement and its role in grammar at Lexical-Functional Grammar 05*, Bergen, July 2005.
- Engdahl, Elisabet To appear. Semantic and syntactic patterns in Swedish passives. In *Demoting the agent*, eds Torgrim Solstad & Benjamin Lyngfeldt. Amsterdam: Benjamins.
- Falk, Cecilia 1989. On the existential construction in the Germanic languages. *Working Papers in Scandinavian Syntax* 44: 45–59.
- Falk, Cecilia 1993. *Non-referential subjects in the history of Swedish*. PhD thesis, Lund University.
- Keenan, Edward 1976. Towards a universal definition of ‘subject’. In *Subject and topic*, Ed Charles Li. New York: Academic Press. 303–333.
- King, Tracy Holloway 1997. Focus domains and information structure. In *The Proceedings of the LFG'97 Conference*, eds. Miriam Butt and Tracy Holloway King. Stanford, Ca: CSLI Publications.
- Kroeger, Paul 1993. *Phrase structure and grammatical relations in Tagalog*. Stanford, Ca: CSLI Publications.
- Lambrecht, Knud 2000. When subjects behave like objects: an analysis of the merging of S and O in sentence-focus constructions across languages. *Studies in Language* 24: 611–682.
- Lødrup, Helge. 1999. Linking and optimality in the Norwegian presentational focus construction. *Nordic Journal of Linguistics* 22:205–230.
- Mikkelsen, Line 2002. Reanalyzing the Definiteness Effect: evidence from Danish. *Working Papers in Scandinavian Syntax* 65:1–75.
- Mycock, Louise 2004. The *wh*-expletive construction. Proceedings of the LFG04 Conference University of Canterbury, Miriam Butt and Tracy Holloway King (Editors) Stanford, CA: CSLI Publications, pp. 370-390. <http://csli-publications.stanford.edu/>
- O'Connor, Robert 2004. Information structure in Lexical-Functional Grammar: the discourse-prosody correspondence in English and Serbo-Croatian. PhD thesis, The University of Manchester. [Under revision.]
- Platzack, Christer 1983. Existential sentences in English, Swedish, German and Icelandic. In *Papers from the Seventh Scandinavian Conference of Linguistics*, ed. Fred Karlsson. Helsinki: Department of Linguistics, Helsinki University. 80–100.
- Safir, Kenneth J. 1987. What explains the definiteness effect? In *The representation of (in)definiteness*, ed Alice G.B. ter Meulen. Cambridge, Ma: MIT Press. 72–97.

- Sells, Peter 2001. *Structure, alignment and optimality in Swedish*. Stanford, Ca: CSLI Publications.
- Sells, Peter This volume. The peripherality of the Icelandic expletive.
- Teleman, Ulf, Staffan Hellberg & Erik Andersson 1999. *Svenska Akademiens grammatik*. Stockholm: Nordstedts. (SAG)
- van der Beek, Leonoor. 2003. The Dutch it-cleft construction. In *The Proceedings of the LFG'03 Conference*, eds. Miriam Butt and Tracy Holloway King. Stanford, Ca: CSLI Publications. 22–42
- Vikner, Sten 1995. *Verb movement and expletive subjects in the Germanic subjects*. Oxford: Oxford University Press.
- Vincent, Nigel & Kersti Börjars To appear. Feature resolution and the content of features. In *Festschrift for Joan Bresnan*, eds Jane Simpson & Annie Zaenen. Stanford, Ca: CSLI Publications.

Sources

- AOS: *Allt om Stockholm – Stora Bostadsguiden*, at <http://info.aos.se/E/F/STOSE/0000/01/78/5.2.html>, 8 July 2005.
- HIST: *Historietter*, available electronically as part of Projekt Runeberg <http://runeberg.org>.
- MA: Maia Andréasson (pc), observed examples
- PAR: Parole corpus of Swedish, available at <http://spraakbanken.gu.se>.
- Portugalien: Selma Lagerlöf, 1914, *Kejsaren av Portugalien*. available electronically as part of Projekt Runeberg <http://runeberg.org>.
- Tektid: *Teknisk Tidskrift*, 1932, *Väg- och vattebyggnadskonst*. available electronically as part of Projekt Runeberg <http://runeberg.org>.
- TRK: Report on trip by the drum corps of the Swedish army at <http://www.fomusc.mil.se/trk/article.php?id=9092>, 8 July 2005.

IT AIN'T NECESSARILY S(V)O:
TWO KINDS OF VSO LANGUAGES

George Aaron Broadwell

University at Albany, State University of New York

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract: Some VSO languages, such as Welsh, show evidence for a VP constituent, with VSO order obtained by positioning the verb in a higher functional projection outside S. However, in other VSO languages, such as Zapotec, constituency tests show no evidence for a VP, and indeed seem to provide evidence against such a constituent. Lexical-Functional Grammar allows us to give different syntactic analyses of these two types of VSO languages that capture their fundamental diversity.

1. Do VSO languages have a VP node?

The issue of whether VSO languages have a VP node is an important one for grammatical theory. If some languages lack VP nodes, then grammatical relations such as SUBJECT and OBJECT cannot be defined in terms of phrase-structure configuration, as has been a frequent assumption of syntactic theory since at least Chomsky (1965). Other syntactic phenomena as well – such as anaphora and incorporation – in many theories also depend on a structural asymmetry between the subject and object.

As a result, many syntacticians have sought evidence for an underlying VP node in VSO languages, with the surface VSO order derived by movement of the verb (Anderson and Chung 1977, McCloskey 1983). For some VSO languages – particularly the Celtic languages – such analyses seem to be essentially correct. These languages show various constituency tests that point to the existence of a VP node. For example, Welsh shows sentences of the following sort, in which a VP constituent is fronted:

- 1) [Adeiladu tai ym Mangor] a wnaeth o.
build houses in Bangor PART do:PST:3SG he

‘He built houses in Bangor.’ (focus on VP)

There is no theoretical obstacle to positing analyses of this sort in Lexical-Functional Grammar. Bresnan (2001:126-131) adopts an extended head analysis of Welsh along the following lines:

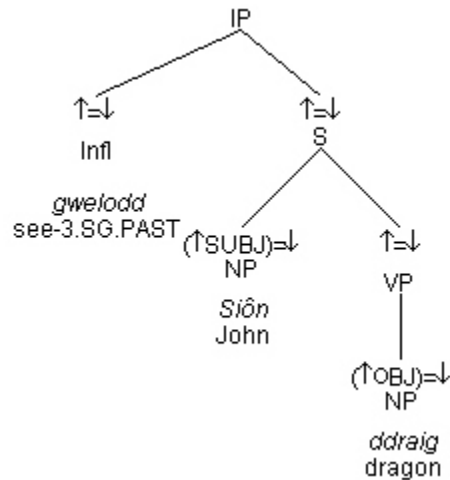


Figure 1 Welsh structure proposed by Bresnan (2001)

This analysis recognizes that Welsh has a VP, and that V (appearing in the Infl position) is the extended head of this VP. This analysis also has a ‘vacated S’ constituent which consists of the SUBJ and OBJ.

However from the evidence that *some* VSO languages have an underlying VP, it clearly does not follow that *all* VSO languages are best analyzed in this way. In this paper, I discuss one VSO language, San Dionisio Ocotepc Zapotec (SDZ), which fails all constituency tests for VP.¹ I argue that it is best treated with a ‘flat’ VSO structure like the following:²

¹ SDZ is an Otomanguan language spoken in San Dionicio Ocotepc, Oaxaca, Mexico by 2,000 - 3,000 people. I thank Cheryl Black, Pamela Munro, and Yuching Tseng for useful discussion of this material. Special thanks to Luisa Martínez, who provided all the SDZ data.

The orthography for SDZ is adapted from the practical orthographies for other Zapotec languages spoken in the Valley of Oaxaca. In the SDZ orthography symbols have their usual phonetic values, with the following exceptions. <x> = /ʃ/ before a vowel and /ʒ/ before a consonant, <xh> = /ʃ/, <dx> = /d ʒ/, <ch> = /tʃ/, <c> = /k/ before back vowels, <qu> = /k/ before front vowels, <rr> = trilled /r/, and <eh> = /ɛ/. Doubled vowels are long. SDZ is a language with four contrastive phonation types: breathy <Vj>, creaky <V’V>, checked <V’>, and plain <V>.

Glosses use the following abbreviations: a=animal, aff = affirmative, cer = certainty, com = completive aspect, con = continuative aspect, cs = causative, def = definite future aspect, dem = demonstrative, foc = focus, hab = habitual aspect, neg = negative, p = possessed, plur = plural, pot = potential aspect, q = question, r=respect, ref=reflexive, rel = relative, stat= stative aspect, top=topic.

² There is also a higher CP projection which contains complementizers and interrogative foci. I have omitted it from this discussion for reasons of space.

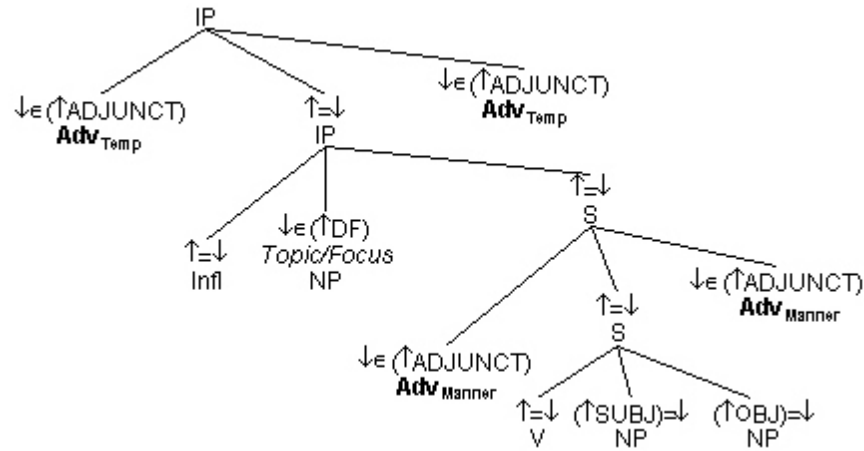


Figure 2 Proposed structure for San Dionicio Ocotepc Zapotec

2. Coordination and flat structure

One extremely useful constituency test in SDZ is coordination. Every phrase shown in the figure above is available for coordination in Zapotec. Consider the following example, which shows coordination of S:

2) [Ù-zíí' Juàány gèhèht]_S chì'í
com-buy Juan tortilla and

[ù-dàw Màríí lèh'èhn]_S
com-eat Mary them

'Juan bought tortillas and Mary ate them.'

However, there is no coordination of any smaller constituent headed by a verb. In particular there is no coordination of the VP or of the 'vacated S' constituents which a verb-movement or extended head analysis posits. Consider the following ungrammatical attempts at coordination of VP and 'vacated S' constituents:

3) *Juàány [ù-zíí' gèhèht]_{VP} chì'í [ù-dàw lèh'èhn]_{VP}.
Juan com-buy tortilla and com-eat them

(Juan bought tortillas and ate them.)

- 4) *Ù-dàw [Juààny bèjl] chì'í [Màríí bè'l].
com-eat Juan fish and Maria meat

(Juan ate fish and Maria meat.³)

The only grammatical coordination pattern for sentences repeats the entire S with a clitic pronoun in the second conjunct:

- 5) [_S Û-zíí' Juààny gèhèht] chì'í [_S ù-dàw=bì lèh'èhn]
com-buy Juan tortilla and com-eat=3 them

'Juan bought tortillas and he ate them.'

The extended head analysis posits two constituents (VP and 'vacated S') which are unavailable for coordination in Zapotec. The flat S hypothesis correctly predicts that coordination is only available for the entire clause.

3. Adverb position and flat structure

The proposed tree for Zapotec in figure (2) shows the positions for manner and temporal adverbs. Ernst (2002) shows that manner adverbs usually adjoin to VP and temporal adverbs to IP. In Zapotec, there are no adverbs which may adjoin to the VP or 'vacated S' constituents posited by the extended head analysis:

- 6) *Ù-dàw bèh'cw ngàngá' [_{VP},bèh'l].
com-eat dog slowly meat
(The dog ate the meat slowly)

- 7) *Ù-dàw ngàngá' [_S,bèh'cw bèh'l].
com-eat slowly dog meat
(The dog ate the meat slowly.)

Instead manner adverbs adjoin to S and temporal adverbs adjoin to IP. We can distinguish the adjunction sites because S follows the focus/topic position and IP precedes the same position. Consider the following examples of adjunction of manner adverbs to the S:

³ This example would involve across-the-board movement of the verb from both conjuncts.

- 8) [IP Bèh'cw ngàngá' [S ù-dàw bèh'l]].
 dog slowly com-eat meat
 'The dog (*topic/focus*) ate the meat slowly.'
- 9) [IP Bèh'cw [S ù-dàw bèh'l] ngàngá'].
 dog com-eat meat slowly
 'The dog (*topic/focus*) ate the meat slowly.'
- 10) *Ngàngá' [IP bèh'cw [S ù-dàw bèh'l]].
 slowly dog com-eat meat
 ('The dog (*topic/focus*) ate the meat slowly.)

The following examples show that temporal adverbs are adjoined to IP:

- 11) *[IP Bèh'cw ná'í [S ù-dàw bèh'l]].
 dog yesterday com-eat meat
 (The dog (*topic/focus*) ate the meat yesterday.)
- 12) Ná'í [IP bèh'cw [S ù-dàw bèh'l]].
 yesterday dog com-eat meat
 'The dog (*topic/focus*) ate the meat yesterday.'

The extended projection analysis posits phrase boundaries that ought to be adjunction sites for adverbs. The flat S analysis predicts adjunction to S and IP only, and thus makes the correct predictions about adverb positions in Zapotec.

4. Auxiliaries and VSO

Another important difference between Welsh and Zapotec is found in the order of the auxiliary and main verb. In Welsh, the order is **Aux S V O**, while in Zapotec it is **Aux V S O** (Broadwell 2003). Consider the following Zapotec examples, which show the position of auxiliaries:

- 13) B-yàlòò ù-dòàb Juáany gèhjs.
com-stop com-smoke Juan cigarette

‘Juan stopped smoking.’

- 14) *B-yàlòò Juáany ù-dòàb gèhjs.
com-stop Juan com-smoke cigarette

(Juan stopped smoking.)

Welsh auxiliary order frequently has V and OBJ adjacent to each other, and so a Welsh language learner is exposed to constructions with overt surface VPs. In contrast, Zapotec never shows an order where V and OBJ are adjacent – with the exception of sentences with topicalized or focussed SUBJ constituents. Thus a Zapotec language learner has little evidence to favor a VP constituent.

5. The function of IP in Zapotec

The extended head analysis is not correct for most Zapotec clauses. However in the *definite future* aspect, the verb **is** positioned in Infl⁴. SDZ has two aspects -- the potential and the definite future -- which are both translated into English with the future:

- 15) S-àw báád bèhld yù'ù.
def-eat duck snake earth

‘The duck is going to eat a worm.’

- 16) G-âw báád bèhld yù'ù.
pot-eat duck snake earth

‘The duck is going to eat a worm.’

It is difficult for speakers to explain the difference in meaning between these two sentences, but (as the label implies) the definite future seems to entail a stronger speaker commitment to the truth of the proposition.

Despite the similar translations, however, there are striking syntactic differences. In the

⁴ My analysis here is influenced by the movement-based account given by Lee (1999) for the related language San Lucas Quiavini Zapotec. Lee shows for that language that topicalization is unavailable for sentences with verbs in the definite future aspect.

definite future, the preverbal topic/focus position becomes unavailable. Yet there is no difficulty in using the preverbal topic/focus position with the potential aspect:

- 17) *Báád s-âw bèhld yù'ù.
duck def-eat snake earth

(‘The duck (*topic/focus*) will eat the worm.’)

- 18) Báád g-âw bèhld yù'ù.
duck pot-eat snake earth

‘The duck (*topic/focus*) will eat the worm.’

The definite future also differs from the potential in the behavior of adverbs. Manner adverbs may normally adjoin to either the left or right of S.

- 19) Diáp g-ú'ld Màrì.
strongly pot-sing Maria

‘Maria will sing strongly/loudly.’

- 20) G-ú'ld Màrì diáp.
def-sing Maria strongly

‘Maria will sing strongly/loudly.’

In the definite future aspect, only right adjunction of adverbs is possible.

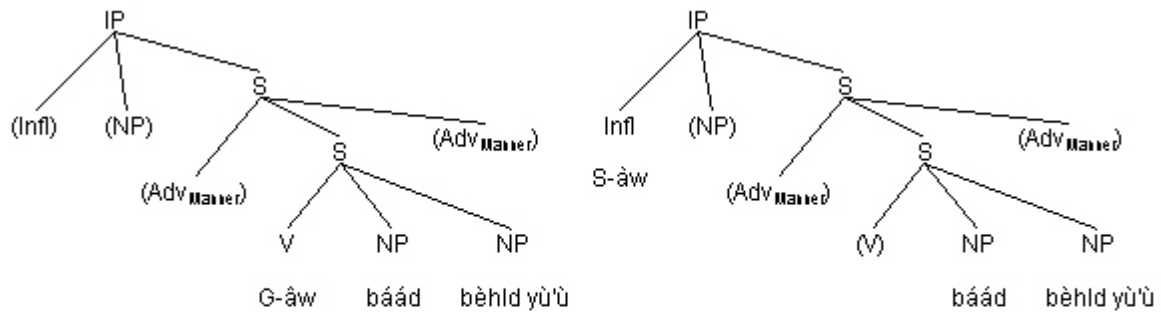
- 21) S-ù'ld Màrì diáp.
def-sing Maria strongly

‘Maria will sing strongly/loudly.’

- 22) *Diáp s-ù'ld Màrìi.
strongly def-sing Maria

(Maria will sing strongly/loudly.)

Both the topicalization and adverb placement facts follow if the potential and definite future aspects have the structures below.



'The duck will eat worms.'
(potential aspect)

'The duck will eat worms.'
(definite future aspect)

Figure 3 Potential and definite future aspect in Zapotec

These trees show that when the verb is in the definite future aspect, it is in Infl, and precedes both the manner adverb position and the [Spec, IP] (topic/focus) position. Thus the initial manner adverb position and the [Spec, IP] (topic/focus) now both follow the verb. This accounts for the ungrammaticality of the following two examples (repeated from above):

- 23) *Diáp s-ù'ld Màrìi.
strongly def-sing Maria

(Maria will sing strongly/loudly.)
- 24) *Báád s-àw bèhld yù'ù.
duck def-eat snake earth

(‘The duck (*topic/focus*) will eat the worm.’)

One might ask whether it is possible in such a case for the adverb or the topic/focus to follow a verb in the definite future aspect. In fact such cases are also ungrammatical:

- 25) *S-ù'ld diáp Màrii.
def-sing strongly Maria

(Maria will sing strongly/loudly.)

- 26) *S-àw báád bèhld yù'ù.
def-eat duck snake earth

(‘The duck (*topic/focus*) will eat the worm.’)

The ungrammaticality of these cases seems to follow from an independent requirement for adjacency between the verb and its subject. Such a restriction on possible orders has been noted for other VSO languages by researchers such as McCloskey (1996) and Black (2000).

6. The diversity of VSO structures

The fact that Zapotec does show some extended head structures supports the idea that syntactic theory must allow such a mechanism. However, the contrast between definite future aspect and other aspects also argues that not all instances of VSO are due to extended head structures.

In most cases, Zapotec VSO is due to a flat S structure. Only in the definite future is it plausible to suggest that the V is in a higher functional position (such as Infl.) And even in cases where V is analysed as occurring in a higher position, there is still no evidence that the underlying structure contains a VP.

Lexical-Functional Grammar allows both flat and extended head analyses of VSO – even for different structures in the same language. This theoretical flexibility accords well with the Zapotec facts.

In contrast, current Principles and Parameters/Minimalist analyses of VSO explicitly reject the possibility of flat structure, and force raising of the V or some phrase containing V, such as VP (Carnie and Guilfoyle 2000). All VSO languages in these analyses derive from underlying SVO. The Zapotec evidence for flat VSO structures presents a problem for this approach, and favors a theory like LFG which allows for the possibility of two kinds of VSO languages.

7. References

Anderson, Stephen R. and Sandra Chung. 1977. On grammatical relations and clause structure in verb-initial languages, in P. Cole and J. Saddock, eds. *Grammatical relations: Syntax and semantics* 8:1-25. New York: Academic Press.

Black, Cheryl. 2000. *Quiégolani Zapotec syntax: A principles and parameters approach*. Dallas TX: Summer Institute of Linguistics

Bresnan, Joan. 2000. *Lexical-Functional syntax*. Oxford: Blackwell.

Broadwell, George Aaron. 2003. Optimality, Complex Predication, and Parallel Structures in Zapotec. *Proceedings of LFG 2003*.

Carnie, Andrew and Eithne Guilfoyle. 2000. *The syntax of verb initial languages*. Oxford University Press.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Ernest, Thomas. 2002. *The syntax of adjuncts*. Cambridge University Press.

Lee, Felicia Ann. 1999. *Antisymmetry and the syntax of San Lucas Quiaviní Zapotec*. Ph.D. thesis. UCLA

McCloskey, James. 1983. A VP in a VSO language? In Gerald Gazdar, Ewan Klein & Geoffrey K. Pullum (eds.) *Order, concord and constituency*, 9-55. Dordrecht: Foris.

McCloskey, James. 1996. On the Scope of Verb Movement in Irish. *Natural Language and Linguistic Theory* 14: 47–104.

EVALUATING AUTOMATICALLY ACQUIRED F-STRUCTURES AGAINST PROPBANK

Michael Burke, Aoife Cahill, Josef van Genabith and Andy Way

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

An automatic method for annotating the Penn-II Treebank (Marcus et al., 1994) with high-level Lexical Functional Grammar (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001) f-structure representations is presented by Burke et al. (2004b). The annotation algorithm is the basis for the automatic acquisition of wide-coverage and robust probabilistic approximations of LFG grammars (Cahill et al., 2004) and for the induction of subcategorisation frames (O’Donovan et al., 2004; O’Donovan et al., 2005). Annotation quality is, therefore, extremely important and to date has been measured against the DCU 105 and the PARC 700 Dependency Bank (King et al., 2003). The annotation algorithm achieves f-scores of 96.73% for complete f-structures and 94.28% for preds-only f-structures against the DCU 105 and 87.07% against the PARC 700 using the feature set of Kaplan et al. (2004). Burke et al. (2004a) provides detailed analysis of these results. This paper presents an evaluation of the annotation algorithm against PropBank (Kingsbury and Palmer, 2002). PropBank identifies the semantic arguments of each predicate in the Penn-II treebank and annotates their semantic roles. As PropBank was developed independently of any grammar formalism it provides a platform for making more meaningful comparisons between parsing technologies than was previously possible. PropBank also allows a much larger scale evaluation than the smaller DCU 105 and PARC 700 gold standards. In order to perform the evaluation, first, we automatically converted the PropBank annotations into a dependency format. Second, we developed conversion software to produce PropBank-style semantic annotations in dependency format from the f-structures automatically acquired by the annotation algorithm from Penn-II. The evaluation was performed using the evaluation software of Crouch et al. (2002) and Riezler et al. (2002). Using the Penn-II Wall Street Journal Section 24 as the development set, currently we achieve an f-score of 76.58% against PropBank for the Section 23 test set.

1 Introduction

Recent research (Burke et al., 2004b) has presented a method for automatically annotating the Penn-II treebank (Marcus et al., 1994) with Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001) f-structure representations. The automatic f-structure annotation algorithm is a central component in a larger project which automatically acquires wide-coverage and robust probabilistic approximations of LFG grammars (Cahill et al., 2004) and induces LFG lexical resources (O’Donovan et al., 2004; O’Donovan et al., 2005). Annotation quality is, therefore, extremely important and to date has been evaluated, using the methodology and software presented in (Crouch et al., 2002) and (Riezler et al., 2002), against the DCU 105¹ and the PARC 700 Dependency Bank (King et al., 2003). The annotation algorithm achieves f-scores of 96.73% for complete f-structures and 94.28% for preds-only² f-structures against the DCU 105 and 87.07% against the PARC 700 using the feature set of Kaplan et al. (2004). Burke et al. (2004a) provides further analysis of these results and describes the conversion software used in the PARC 700 evaluation process.

In this paper we present an evaluation of the f-structures produced by the annotation algorithm for Penn-II treebank trees against PropBank (Kingsbury and Palmer, 2002). PropBank was developed independently of any grammar formalism and provides a platform for making more meaningful comparisons between parsing technologies than was previously possible. PropBank has been used for the evaluation of CCG

¹Available from <http://www.computing.dcu.ie/research/nclt/gold105.txt>.

²Preds-only f-structures consider only paths in f-structures ending in a PRED feature-value pair.

(Gildea and Hockenmaier, 2003) and HPSG (Miyao and Tsujii, 2004) parsers. The methodology presented in this paper will allow the parsing technology of Cahill et al. (2004) to eventually be evaluated against PropBank and for direct comparisons with CCG, HPSG and other parsers to be made. Whereas previous evaluations of the annotation algorithm have been against syntax-based gold standards (DCU 105 and PARC 700), evaluating against PropBank provides a semantic evaluation of the automatically acquired f-structures. Using PropBank also allows a much larger scale evaluation than was previously possible. The quality of the f-structure annotation algorithm can eventually be evaluated against PropBank data for the *entire* Penn-II treebank.

PropBank adds semantic information to the syntax trees of Penn-II, identifying predicates and their semantic arguments. To give a simple example, for the sentence *Both companies rejected the offers*, PropBank identifies *rejected* as the predicate with *both companies* as ARG0 and *the offers* as ARG1. Figure 1 provides the f-structure produced by the annotation algorithm for the example sentence, (a subset of the) triples extracted from that f-structure and the corresponding PropBank triples. A simple mapping of SUBJ to ARG0 and OBJ to ARG1 is sufficient to obtain the semantic annotations provided by PropBank in this example, but clearly a more elaborate mapping is required to extract PropBank-style semantic annotations from more complex automatically f-structure-annotated Penn-II trees.

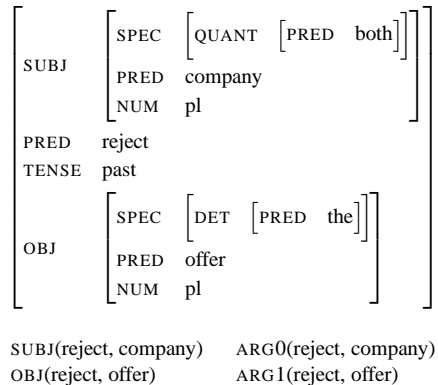


Figure 1: F-structure produced by annotation algorithm for *Both companies rejected the offers* with some extracted LFG triples and the expected PropBank triples

Section 2 introduces the automatic f-structure annotation algorithm. Section 3 provides an overview of PropBank and the process of converting the PropBank semantic annotations into a dependency format. Section 4 describes the conversion software required to systematically convert the triples extracted from the automatically generated f-structures for evaluation against PropBank. Section 5 presents and analyses the results of the evaluation process. Using the Penn-II Wall Street Journal Section 24 as the development set, currently we achieve an f-score of 76.58% against PropBank for the Section 23 test set. Section 6 summarises and provides possibilities for future work.

2 Automatic F-Structure Annotation Algorithm

The automatic f-structure annotation algorithm (Burke et al., 2004b; Cahill et al., 2004; O’Donovan et al., 2004; O’Donovan et al., 2005) is modular (Figure 2). The first module, *Left-Right Context Rules*, head-lexicalises the treebank using a modified version of Magerman’s (1994) scheme. This process creates a bi-

partition of each local subtree, with nodes lying in either the left or right context of the head. An annotation matrix is manually constructed for each parent category in the treebank by analysing the most frequent CFG rule types that together give at least 85% coverage of rule tokens for that parent category in the treebank. For example, only the most frequent 102 NP rule types were analysed to produce the NP annotation matrix which generalises to provide default annotations for the complete set of 6,595 NP rule types in the treebank. Default annotations are read from these matrices by the annotation algorithm to annotate nodes in the left and right context of each subtree.

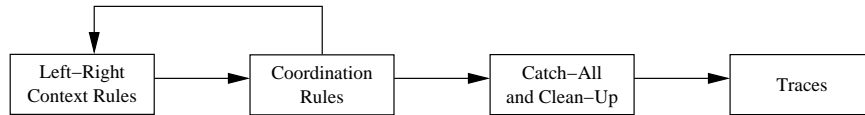


Figure 2: Annotation Algorithm modules

The annotation of co-ordinate structures is handled by a separate module in the annotation algorithm, because the relatively flat analysis of co-ordination in Penn-II would significantly complicate the *Left-Right Context Rules*, making them harder to maintain and extend. Once the elements of a co-ordination set have been identified, the *Left-Right Context Rules* module may be re-used to provide default annotations for any remaining unannotated nodes in the local subtree.

The *Catch-All and Clean-Up* module provides default annotations for remaining unannotated nodes that are labelled with Penn functional tags, e.g. -SBJ. A small amount of over-generalisation is accepted within the first two annotation algorithm modules to allow a concise statement of linguistic generalisations. Some annotations are overwritten to counter this problem and to systematically correct other potential feature clashes.

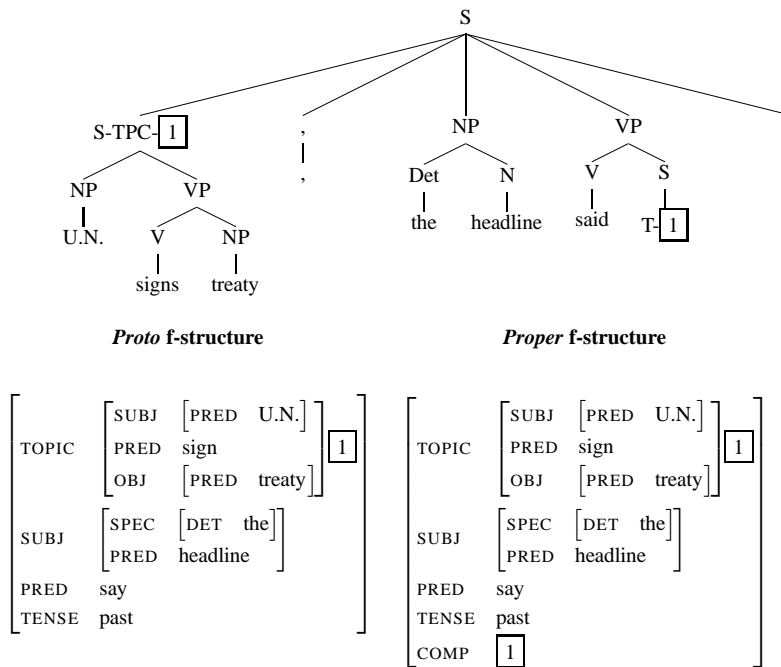


Figure 3: Penn-II style tree with LDD trace and corresponding re-entrancy in proper f-structure

The first three modules of the annotation algorithm produce *proto* f-structures which do not yet resolve non-local dependencies. To create *proper* f-structures, the *Traces* module uses the wide range of trace information encoded in Penn-II to capture dependencies introduced by topicalisation, passivisation, relative clauses and questions. Figure 3 illustrates a Penn-II style tree and corresponding *proto* and *proper* f-structures for the sentence *U.N. signs treaty, the headline said*. The *Traces* module translates the Penn-II trace and co-indexation information to capture the long-distance dependency (LDD) in terms of a re-entrancy in the proper f-structure which is absent from the corresponding proto f-structure.

The annotation algorithm achieves near complete coverage for the WSJ section of Penn-II with 99.82% of the 48K sentences receiving a single connected and covering f-structure. Table 1 provides a quantitative evaluation of the f-structures produced by the annotation algorithm. Feature clashes in the annotation of 85 trees result in no f-structure being produced for those sentences. Nodes left unannotated by the annotation algorithm in two trees caused two separate f-structure fragments for both sentences.

| # f-structures | # sentences | Treebank Percentage |
|----------------|-------------|---------------------|
| 0 | 85 | 0.176 |
| 1 | 48337 | 99.820 |
| 2 | 2 | 0.004 |

Table 1: Quantitative Evaluation

While achieving such wide coverage is important, the annotation quality must be of a high standard, particularly as the annotation algorithm plays a vital role in the generation of wide-coverage, probabilistic LFG parsing technology (Cahill et al., 2004) and lexical resources (O’Donovan et al., 2004; O’Donovan et al., 2005). Annotation quality has been measured in terms of precision, recall and f-score³ against the DCU 105 and PARC 700 Dependency Bank using the evaluation methodology and software presented in (Crouch et al., 2002) and (Riezler et al., 2002). The DCU 105 is a set of gold standard f-structures for 105 randomly selected sentences from Section 23 of the WSJ section of Penn-II. To create the gold standard f-structures the Penn-II trees were first automatically annotated and the annotations were then manually corrected and extended. The PARC 700 consists of dependency structures for 700 randomly selected sentences from Section 23 of the WSJ section of Penn-II. These sentences were automatically parsed by a hand-coded, deep LFG grammar of English using the XLE system (Maxwell and Kaplan, 1993). In cases where multiple parses were generated, the best parse was manually chosen. The f-structures of the best parses were then automatically converted to dependency format (triples) and manually extended and corrected by two independent reviewers.

The f-structure annotation algorithm currently achieves an f-score of 96.73% for complete f-structures and 94.28% for preds-only f-structures against the DCU 105 (Table 2). Burke et al. (2004a) presents conversion software developed to overcome some of the systematic differences in linguistic analysis, feature geometry and nomenclature between the automatically acquired f-structures and the PARC 700 dependency structures. The f-structures automatically acquired by the annotation algorithm and mapped by conversion software achieves an f-score of 87.07% for the feature set of Kaplan et al. (2004) against the PARC 700. Burke et al. (2004a) provides a detailed analysis of the evaluation process and the results.

³Precision, recall and f-score were calculated according to the following equations:

$$precision = \frac{\# \text{ of correct feature-value pairs in the automatically generated f-structure}}{\# \text{ of feature-value pairs in the automatically generated f-structure}}$$

$$recall = \frac{\# \text{ of correct feature-value pairs in the automatically generated f-structure}}{\# \text{ of feature-value pairs in the gold standard f-structure}}$$

$$f - score = \frac{2 \times precision \times recall}{precision + recall}$$

| | DCU 105 | | PARC 700 |
|-----------|----------------------------------|-------------------|--|
| | <i>All grammatical functions</i> | <i>Preds only</i> | <i>Feature set of Kaplan et al. (2004)</i> |
| Precision | 96.77 | 94.32 | 87.95 |
| Recall | 96.69 | 94.24 | 86.21 |
| F-score | 96.73 | 94.28 | 87.07 |

Table 2: Annotation quality evaluated against DCU 105 and PARC 700

3 PropBank

3.1 Overview

PropBank (Kingsbury and Palmer, 2002) adds a layer of semantic annotation to the syntactic structures of Penn-II. The process of semantic role annotation was semi-automatic. The output of a rule-based automatic argument tagger which encodes class-based mappings between grammatical and semantic roles was manually corrected and extended. The tagger achieved 83% accuracy. PropBank contains a set of semantic frames for each Penn-II verb. The semantic frames define particular meanings for each verb and the roles played by their semantic arguments in each case. PropBank annotates Penn-II by identifying token verb occurrences, assigning a semantic frame to that verb and marking the semantic arguments of the verb. PropBank does not annotate or provide semantic frames for *be*.

3.2 Semantic Frames

PropBank assigns a set of semantic frames for every verb in Penn-II. Each semantic frame provides a definition for the semantic role labels relevant to that particular instance of the verb. Table 3 provides the three semantic frames for the predicate *yield*. The first semantic frame for *yield* defines the semantic role labels for the meaning *to result in*: ARG0 is the “thing yielding” and ARG1 is the “thing yielded”.

| | | | |
|------|-------------------------|------------------------|-------------------------------|
| | (yield.01) To result in | (yield.02) To give way | (yield.03) To give a dividend |
| ARG0 | thing yielding | thing giving way | thing providing a dividend |
| ARG1 | thing yielded | what’s lost | dividend, earnings |
| ARG2 | n/a | what’s preferred | recipient |

Table 3: PropBank semantic frame set for the predicate *yield*

An example sentence, annotated with semantic role labels, for this semantic frame is: $[_{ARG0} \text{A single acre of grapes}] \text{yielded } [_{ARG1} \text{a mere 75 cases}] [_{ARGM-TMP} \text{in 1987}]$. The semantic role label annotations indicate that in this example sentence *a single acre of grapes* is the “thing yielding” while *a mere 75 cases* is the “thing yielded”. The phrase *in 1987* is annotated as an optional modifier ARGM-TMP. Annotated example sentences for the three semantic frames for *yield* are:

- (1) Frame 1: “To result in”
 $[_{ARG0} \text{A single acre of grapes}] \text{yielded } [_{ARG1} \text{a mere 75 cases}] [_{ARGM-TMP} \text{in 1987}]$.
- Frame 2: “To give way”
 $[_{ARG0} \text{John}] \text{yielded } [_{ARG1} \text{the right-of-way}] \text{to } [_{ARG2} \text{the Mack truck}]$.
- Frame 3: “To give a dividend”

The Canadian government announced [_{ARG0} *a new, 12-year Canada Savings Bond issue*] *that will yield* [_{ARG2} *investors*] [_{ARG1} *10.5%*] [_{ARGM-TMP} *in the first year*].

3.3 Semantic Argument Annotation

PropBank provides a file of semantic annotations for Penn-II in the following format. The annotations first identify the relevant Penn-II tree by providing the Penn-II file name and line number, e.g. line 12 in `wsj/00/wsj_0004.mrg` identifies the tree shown in Figure 4 for the sentence *The top money funds are currently yielding well over 9%*. The annotation then identifies the verb being annotated and the relevant semantic frame for this occurrence of the verb, which in this case is “yield.01”, the frame “to result in” as outlined in Table 3. The semantic arguments are then listed in the form `terminal:node height-semantic role`. Terminals are numbered from left to right starting with zero.

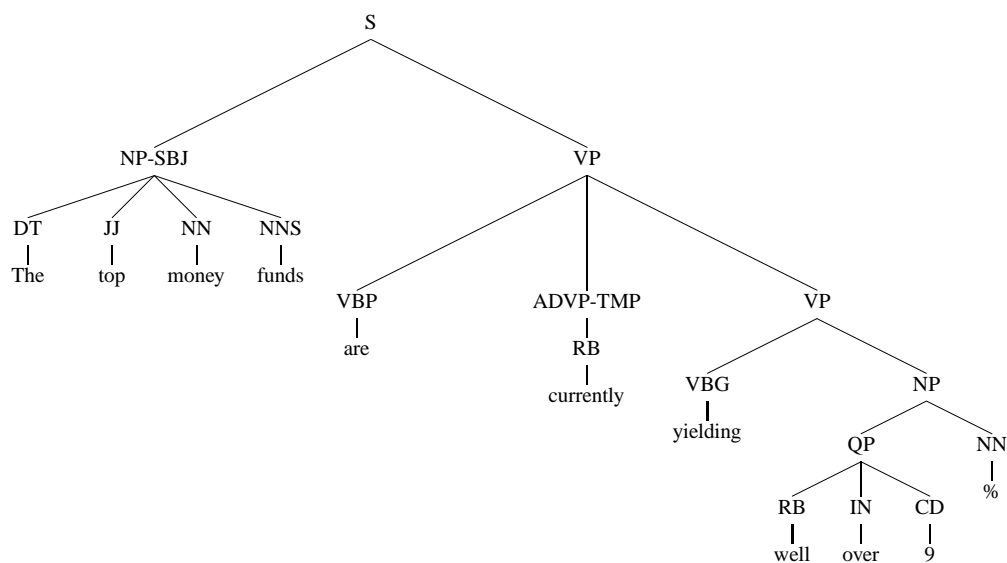


Figure 4: Penn-II tree for the sentence *The top money funds are currently yielding well over 9%*

PropBank annotates the semantic arguments of the verb *yield* in the example sentence as: `0:1-ARG0`, `5:1-ARGM-TMP` and `7:2-ARG1`. The annotation `0:1-ARG0` indicates that the node NP-SBJ which governs the noun phrase *The top money funds* is a semantic argument of the verb *yield* with the semantic role ARG0. This node is found in the tree of Figure 4 by starting with the POS tag of terminal 0 in the tree, i.e. DT, and traversing one node, i.e. `0:1`, upwards from that node. Similarly, the argument paths `5:1-ARGM-TMP` and `7:2-ARG1` indicate that the semantic roles ARGM-TMP and ARG1 are played by the nodes ADVP-TMP and NP governing *currently* and *well over 9%*, respectively.

3.4 Creating Gold Standard PropBank Dependencies

In order to evaluate the automatic f-structure annotation algorithm the PropBank semantic annotations were converted into a dependency format (triples). By also mapping the automatically generated f-structures into a set of semantic role triples, the methodology and software of Crouch et al. (2002) and Riezler et al. (2002) could then be used to evaluate the annotation algorithm in terms of precision, recall and f-score.

The PropBank semantic annotations were automatically converted into triples of the form: SEMANTIC ROLE(verb, argument). The Penn-II nodes representing the semantic roles were identified by automatically traversing the argument paths as outlined in Section 3.3. For each node, the head word of the subtree governed by that node was identified using the head-lexicalisation rules of the annotation algorithm, which are a modified version of the rule set of Magerman (1994). The verbs and head words were lemmatised with the XLE lemmatiser also used by the annotation algorithm. The PropBank semantic roles were conflated, removing the different subtypes of ARGM modifiers (Table 4), to the subset: ARG0, ARG1, ARG2, ARG3, ARG4, ARG5 and ARGM.

| | | | |
|-----|-----------------------|-----|-------------------|
| ADV | adverbial | MOD | modal verb |
| CAU | cause | NEG | negation |
| DIR | direction | PNC | purpose not cause |
| DIS | discourse connectives | PRD | predication |
| EXT | extent | REC | reciprocal |
| LOC | location | TMP | temporal |
| MNR | manner | | |

Table 4: PropBank ARGM subtypes

To create PropBank triples for the sentence *The top money funds are currently yielding well over 9%*, the head words of the nodes NP-SBJ, ADVP-TMP and NP (Figure 4) were automatically identified as *funds*, *currently* and *%*, respectively. After lemmatising all words and conflating the semantic roles, the triples ARG0(yield, fund), ARG1(yield, percent) and ARGM(yield, currently) were created. This process was applied to all trees in the treebank.

4 Converting F-Structures into Semantic Roles

We developed conversion software to produce PropBank-style semantic role annotations in the dependency format introduced in Section 3.4 from the f-structures automatically acquired by the annotation algorithm from Penn-II trees. Triples are extracted from the f-structures generated by the annotation algorithm and then post-processed by the conversion software to produce semantic role annotations. The conversion procedure employs default mappings from LFG feature names to PropBank semantic roles before handling the following phenomena which require more complex mappings:

- Particles
- Modal verbs
- Mapping to ARG3, ARG4 and ARG5
- Verbs deviating from default mapping patterns
- Filtering remaining unwanted triples

4.1 Default mappings

Default mappings are used to map LFG feature names to PropBank semantic role annotations. Table 5 lists these mappings for active verbs. Passive voice is identified by the annotation algorithm which results in

PASSIVE triples being extracted from the automatically generated f-structures. These triples are used by the conversion software to map the SUBJ triple of passive verbs to ARG1 (subjects of active verbs are mapped by default to ARG0), while oblique agents are mapped to ARG0.

| LFG feature name | PropBank semantic role |
|------------------|------------------------|
| SUBJ | ARG0 |
| OBJ | ARG1 |
| COMP | ARG1 |
| XCOMP | ARG1 |
| OBJ_THETA | ARG2 |
| OBL | ARG2 |
| OBL2 | ARG2 |
| ADJUNCT | ARGM |

Table 5: Default mappings from LFG feature names to PropBank semantic roles for active voice

The default mappings of Table 5 were applied to the automatically generated LFG triples for the active verb *yield* in the sentence *The top money funds are currently yielding well over 9%*. The resulting mapped PropBank-style triples and the original LFG triples are provided in Table 6. The default mappings are successful for this sentence, producing the required PropBank triples.

| Automatically generated LFG triples | Mapped PropBank-style triples |
|-------------------------------------|-------------------------------|
| SUBJ(yield, fund) | ARG0(yield, fund) |
| OBJ(yield, percent) | ARG1(yield, percent) |
| ADJUNCT(yield, currently) | ARGM(yield, currently) |

Table 6: Default mappings applied to automatically generated triples for *The top money funds are currently yielding well over 9%*

4.2 Particles

PropBank annotates phrasal verbs by grouping all nodes representing the phrasal verb and providing their semantic arguments as normal. When creating the gold standard PropBank triples we combined the grouped nodes to form a complex predicate for the phrasal verb. Phrasal verbs have a single triple for each semantic argument as with all other verbs. The third column of Table 7 provides the gold standard triples we extracted from PropBank for the phrasal verb *snap up* in the sentence *Earlier this year, Japanese investors snapped up a similar fund*. The first column provides the triples produced by the f-structure annotation algorithm for the same sentence, while the second column shows the PropBank-style triples produced by the application of the default mappings to the triples of column one.

An f-score of zero will be achieved for this sentence unless the complex predicate analysis is adopted for the phrasal verb. The Penn-II PRT (particle) tag is automatically annotated $\uparrow\text{PART}=\downarrow$, which results in the triple $\text{PART}(\text{snap}, \text{up})$ in this example. The conversion software uses the PART triple to create the complex predicate which replaces all occurrences of the bare verb in the mapped triples. This allows the desired gold standard triples to be produced by the mapping module.

4.3 Modal verbs

Modal verbs are represented in PropBank as optional arguments of the main verb. This treatment differs markedly from the cascading XCOMP analysis of the automatically generated f-structures and triples. Table

| Automatically generated LFG triples | Triples created by default mappings | Gold standard PropBank triples |
|-------------------------------------|-------------------------------------|--------------------------------|
| SUBJ(snap, investor) | ARG0(snap, investor) | ARG0(snap_up, investor) |
| OBJ(snap, fund) | ARG1(snap, fund) | ARG1(snap_up, fund) |
| ADJUNCT(snap, year) | ARGM(snap, year) | ARGM(snap_up, year) |
| PART(snap, up) | | |

Table 7: Triples for *Earlier this year, Japanese investors snapped up a similar fund*

8 provides the automatically generated LFG triples and gold standard PropBank triples for the sentence *France can boast the lion’s share of high-priced bottles*.

| Automatically generated LFG triples | Gold standard PropBank triples |
|-------------------------------------|--------------------------------|
| SUBJ(can, france) | |
| MODAL(can, +) | ARGM(boast, can) |
| XCOMP(can, boast) | |
| SUBJ(boast, france) | ARG0(boast, france) |
| OBJ(boast, share) | ARG1(boast, share) |

Table 8: Automatically generated LFG triples and gold standard PropBank triples for the sentence *France can boast the lion’s share of high-priced bottles*.

The annotation algorithm uses the Penn-II MD tag to annotate modal verbs. The MODAL triple triggers the creation of an ARGM triple in the mapping module. The cascading XCOMP triples are traversed from the modal verb to identify the main verb which is then modified by the new ARGM triple. Having created this new triple, all other triples associated with the modal verb are removed. This procedure, coupled with the default mappings, allows the gold standard PropBank analysis to be achieved.

4.4 Relative clauses

The gold standard triples extracted from PropBank do not contain relative pronouns. Instead, the head noun being modified by the relative clause takes the place of relative pronouns in the gold standard triples. As the default mappings are not sufficient to compute the desired PropBank-style triples from the automatically generated LFG triples for verbs embedded within relative clauses, a further mapping step handles relative pronouns.

The automatically generated LFG triples indicate the presence of a relative clause through RELMOD and TOPICREL triples. The first column of Table 9 provides the automatically generated LFG triples for the fragment *The rights, which expire Nov. 21*. The RELMOD triple indicates that the noun (lemmatised as *right*) is modified by a relative clause which has *expire* as its main verb. The value *pro* represents the relative pronoun, whose surface form *which* is provided by the PRON_FORM triple. The TOPICREL triple links the *pro* value to the verb, indicating which pronoun is the fronted element of the relative clause. The SUBJ triple indicates that the relative pronoun is the subject of the relative clause.

Applying the default mappings to the triple SUBJ(expire, pro) would produce the incorrect PropBank triple ARG0(expire, pro). To overcome this problem, the conversion software first locates RELMOD triples. A RELMOD triple indicates that a noun is modified by a relative clause and provides the main verb of that clause. The TOPICREL triple associated with that main verb is then found. This triple provides the relative pronoun. Every occurrence of that relative pronoun, in all triples for that sentence, is replaced with the noun from the RELMOD triple (Table 9, second column). With this step in place, the default mappings (in this case from SUBJ to ARG0) are used to achieve the correct analysis.

| Automatically generated LFG triples | LFG triples without relative pronouns | Gold standard PropBank triples |
|-------------------------------------|---------------------------------------|--------------------------------|
| RELMOD(right, expire) | RELMOD(right, expire) | |
| PRON_FORM(pro, which) | PRON_FORM(right, which) | |
| TOPICREL(expire, pro) | TOPICREL(expire, right) | |
| SUBJ(expire, pro) | SUBJ(expire, right) | ARG0(expire, right) |
| ADJUNCT(expire, november) | ADJUNCT(expire, november) | ARGM(expire, november) |

Table 9: Automatically generated LFG triples and mapped PropBank triples for the fragment *The rights, which expire Nov. 21*

4.5 Mapping to ARG3, ARG4 and ARG5

The mappings outlined so far will not generate any triples for the semantic roles ARG3, ARG4 and ARG5. While using the WSJ section 24 of Penn-II as a development set, it became clear that a significant number of ARG3 and ARG4 annotations occur in pairs with verbs taking two oblique prepositional phrases, headed by *from* and *to*. The PP headed by *from* was usually annotated ARG3, while the PP headed by *to* was annotated ARG4. This information was encoded in the conversion software to produce the desired ARG3 and ARG4 triples instead of mapping by default to ARG2. ARG5 occurs very infrequently (only 5 times in WSJ section 23). No mapping was developed for this semantic role.

4.6 Mappings for specific verbs

In many cases, even when the annotation algorithm generates a correct f-structure, there are no syntactic cues which can be used to produce the expected PropBank triples. The syntactic information available through the automatically generated f-structures and triples is insufficient for mapping the semantic roles of, for example, *climb*. Table 10 provides three sets of triples for the sentence *Net profit climbed to 30%*; (i) the triples produced by the f-structure annotation algorithm, (ii) the mapped triples produced using the conversion software described so far and (iii) the expected PropBank triples.

| Automatically generated LFG triples | Mapped triples | Gold standard PropBank triples |
|-------------------------------------|---------------------|--------------------------------|
| SUBJ(climb, profit) | ARG0(climb, profit) | ARG1(climb, profit) |
| ADJUNCT(profit, net) | | |
| OBL(climb, to) | ARG2(climb, to) | ARG4(climb, to) |
| OBJ(to, percent) | | |
| QUANT(percent, 30) | | |

Table 10: Automatically generated LFG triples, mapped triples and PropBank triples for *Net profit climbed to 30%*

Applying the default mappings to the automatically generated triples produces ARG0 and ARG2 triples which should actually be ARG1 and ARG4, respectively. Having reviewed the development set, this is the normal expected behaviour for the verb *climb*. There is no further syntactic information available which could be used in a general mapping rule to produce the correct triples in this case, without degrading the overall performance of the conversion software for most verbs. Instead of introducing a general rule to deal with this case, a specific rule was introduced for the verb *climb* mapping SUBJ to ARG1, OBJ to ARG2, OBL to ARG3 for prepositional phrases headed by *from* and to ARG4 for PPs headed by *to*.

Other verbs in the development set displayed the same behaviour as *climb*. On examination of the VerbNet classes containing *climb*, class 45.6 provided many verbs which required the mapping outlined above:

- (2) *appreciate, balloon, climb, decline, decrease, depreciate, differ, diminish, drop, fall, fluctuate, gain, grow, increase, jump, lessen, mushroom, plummet, plunge, rise, rocket, skyrocket, soar, surge, tumble, vary*

This list was amended on further analysis of the development set, with *lessen* removed and *return* added to the list of verbs mapped in the same manner as *climb*.

A number of other specific mappings were created for groups of verbs, e.g. VerbNet class 48.1.1:

- (3) *appear, arise, awake, awaken, break, burst, come, dawn, derive, develop, emanate, emerge, erupt, evolve, exude, flow, form, grow, gush, issue, materialize, open, plop, result, rise, spill, spread, steal, stem, stream, supervene, surge, wax*

For active occurrences of a subset of these verbs SUBJ is mapped to ARG1. The defaults and other general mappings are used for all other triples with these verbs.

4.7 Filtering

Penn-II verbal POS tags and phrasal bracketing cannot always be used to accurately predict which words are annotated by PropBank. Errors in Penn-II POS tagging would result in the annotation algorithm producing PropBank triples for words which are not annotated by PropBank. In some cases, words which are correctly tagged in Penn-II as verbs and bracketed as the head of a VP are not annotated by PropBank. The annotation algorithm would be punished in these cases for correctly producing PropBank-style triples.

The original version of the conversion software used the PropBank gold standard triples to overcome this problem. The gold standard triples were consulted to indicate which words were annotated as verbs in PropBank. The conversion software only produced PropBank-style triples for those lemmas. This procedure has since been removed and the conversion software no longer refers to the gold-standard triples, relying instead on Penn-II POS tagging and bracketing only.

For the purpose of evaluation, a CAT(egory) feature with the value is v is added to the f-structures produced by the annotation algorithm for all words POS-tagged in Penn-II as verbs and bracketed as the head of a VP, ADJP, PP or any category annotated with the Penn-II -PRD (predicative) functional tag. CAT triples are extracted from the automatically-generated f-structures and are used to filter the PropBank-style triples produced by the conversion software. PropBank-style triples are only produced for lemmas occurring with a CAT triple.

The new procedure is preferred to the original consultation of the gold-standard PropBank triples to identify the annotated verbs as it is more methodologically sound and the results presented in Table 11 are derived with the new procedure. The new procedure achieves an f-score which is only 0.32% lower than the original procedure.

5 Evaluation

5.1 Results

The 2,416 trees in the Wall Street Journal Section 23 of Penn-II were annotated by the automatic f-structure annotation algorithm. Triples were extracted from the resulting f-structures and passed through the conversion software outlined in this paper. These triples were evaluated against the gold standard triples extracted

from the PropBank annotations for the same sentences using the methodology and software presented in (Crouch et al., 2002) and (Riezler et al., 2002). Without specific verb mappings an f-score of 73.42% is achieved, with precision and recall at 75.14% and 71.77%, respectively. Including specific verb mappings sees the overall f-score increase to 76.58% as a result of improved precision and recall scores of 78.44% and 74.81%. Table 11 provides the results in terms of precision, recall and f-score for each semantic role both without and with specific verb mappings.

| Without Specific Verb Mappings | | | | With Specific Verb Mappings | | |
|--------------------------------|--------------|--------------|---------|-----------------------------|----------------|---------|
| | Precision | Recall | F-score | Precision | Recall | F-score |
| ARG0 | 3176/4289=74 | 3176/3708=86 | 79 | 3127/3887=80 | 3127/3708=84 | 82 |
| ARG1 | 3408/4297=79 | 3408/5009=68 | 73 | 3685/4506=82 | 3685/5009=74 | 77 |
| ARG2 | 349/775=45 | 349/1115=31 | 37 | 460/863=53 | 460/1115=41 | 47 |
| ARG3 | 25/28=89 | 25/173=14 | 25 | 54/60=90 | 54/173=31 | 46 |
| ARG4 | 24/28=86 | 24/102=24 | 37 | 50/54=93 | 50/102=49 | 64 |
| ARG5 | 0/0=0 | 0/5=0 | 0 | 0/0=0 | 0/5=0 | 0 |
| ARGM | 2978/3837=78 | 2978/3765=79 | 78 | 3006/3865 = 78 | 3006/3765 = 80 | 79 |
| Overall | 75.14 | 71.77 | 73.42 | 78.44 | 74.81 | 76.58 |

Table 11: Annotation quality measured against PropBank for WSJ Section 23 of Penn-II, with and without mappings for specific verbs

5.2 Analysis

The overall f-score of 76.58% achieved by the annotation algorithm against PropBank for WSJ section 23 of Penn-II is lower than the results in previous evaluation experiments. Against the DCU 105 an f-score of 96.73% was achieved for complete f-structures and 94.28% for preds-only f-structures, while against the PARC 700 Dependency Bank using the feature set of Kaplan et al. (2004) the f-score was 87.07%. When evaluating the automatically generated f-structures — a syntax-based resource — against a gold standard of semantic relations, lower results should be expected than in experiments evaluating the f-structures against syntax-based gold standards, such as the DCU 105 and PARC 700.

Overall, precision is higher than recall, indicating that our algorithm is more likely to produce a partial analysis than an incorrect one. The only semantic role with precision lower than recall is ARG0. The conversion software attempts to map the semantic arguments of specific verbs which deviate from the behaviour captured in the default mappings. Most mappings for specific verbs map the SUBJ triple to ARG1 instead of the default mapping to ARG0. These mappings result in an improvement in f-scores for ARG0 and ARG1 of 3% and 4%, respectively. However, the conversion software does not provide specific mappings for enough verbs which results in too many SUBJ triples still being incorrectly mapped to ARG0.

A further, albeit less significant, explanation for the lower precision score for ARG0 is the failure of the annotation algorithm in some cases to identify a verb occurrence as having passive voice. In a syntax-based evaluation, this results in a missing PASSIVE triple which lowers recall slightly and leaves precision unchanged. The impact is not so significant as there are a far more triples per sentence than in the semantic evaluation. A missing passive marker in this semantic evaluation means that the SUBJ triple is mapped by default to ARG0 instead of ARG1. This results in lower precision for ARG0 and lower recall for ARG1, which is reflected in the scores for ARG1; precision 82%, recall 74%.

The best results are achieved for the semantic roles ARG0, ARG1 and ARGM with f-scores of 82%, 77% and 79%, respectively. As these semantic roles are the most frequently occurring, accounting for 90% of all gold standard triples, the development of mappings for these triples was the main focus of this research.

However, when the conversion software does produce the less frequently occurring ARG3 and ARG4 triples they are usually correct, as shown by the high precision scores of 90% and 93%, respectively. The low recall scores of 31% and 49% indicate that far too few ARG3 and ARG4 triples are being mapped.

These infrequently occurring semantic roles do not have obvious default equivalent LFG feature names which makes them particularly difficult to map. The specific verb mappings allow significant improvements to be made: f-scores increase for ARG3 and ARG4 by 21% and 27%, respectively. A relatively conservative approach was taken when mapping these semantic roles which accounts for some of the shortfall. Another reason for the scarcity of these triples is that they are only produced through the mapping of OBL triples produced by the annotation algorithm. Distinguishing between obliques and adjuncts is an area fraught with difficulty for the annotation algorithm, which relies on the Penn-II -CLR and -DTV functional tags for the annotation of obliques. In the original Penn-II annotation, these functional tags were employed relatively inconsistently and infrequently which may contribute to the shortage of ARG3 and ARG4 triples. This fact also partially explains the poor results for ARG2, which has higher precision than recall, caused by ARG2 triples not being produced in sufficient volume. Obliques are one source of ARG2 triples.

No mappings have been developed to produce ARG5 triples as they occur too infrequently for any general pattern to be established.

6 Summary and Future Work

This paper has presented an evaluation of the automatic f-structure annotation algorithm (Burke et al., 2004b) against PropBank for the test set, WSJ section 23 of Penn-II. A dependency-format gold standard was extracted from PropBank to facilitate the evaluation process. The Penn-II trees were automatically annotated to produce LFG f-structures, from which triples were extracted. Conversion software was developed to map these triples to produce PropBank-style semantic annotations in dependency format. Section 24 of the WSJ section of Penn-II and PropBank was used as the development set for the mapping software. An f-score of 76.58% was achieved against PropBank for the test set. These results are lower than those achieved in previous syntax-based qualitative evaluation experiments. A detailed analysis of the results was provided.

As PropBank was developed independently of any grammar formalism, it provides a platform for making more meaningful comparisons between parsing technologies than was previously possible. However, given the format of the PropBank annotations and the need to convert these annotations to allow evaluation to take place, currently it is not straightforward to draw clear conclusions from such comparisons. There is a need for greater transparency in the evaluation process which could be achieved through collaboration on the development of a universal set of gold standard PropBank triples.

Evaluating the parsing technology of Cahill et al. (2004) against PropBank is one obvious area for the development of this research. However, the mapping software will have to be improved significantly in order to provide a fair evaluation of this technology. An alternative approach to the mapping process may be required, as there are clear limitations to the improvements which can be made to the current mapping software.

The evaluation process provides useful feedback on the quality of the automatic f-structure annotations. Greater focus needs to be placed on the analysis of the evaluation results for the purpose of improving the annotation algorithm itself and not just the mapping software. The analysis of the results to date has shown that the identification of passive voice is one area which needs to be improved. Further research into this area will allow improvements to be made to the annotation algorithm and parsing technology.

References

- Bresnan, J. 2001. *Lexical-Functional Syntax*. Blackwell, Oxford.
- Burke, M., A. Cahill, R. O'Donovan, J. van Genabith, and A. Way. 2004a. The Evaluation of an Automatic Annotation Algorithm against the PARC 700 Dependency Bank. In *Proceedings of the Ninth International Conference on LFG*, pages 101–121, Christchurch, New Zealand.
- Burke, M., A. Cahill, R. O'Donovan, J. van Genabith, and A. Way. 2004b. Treebank-Based Acquisition of Wide-Coverage, Probabilistic LFG Resources: Project Overview, Results and Evaluation. In *The First International Joint Conference on Natural Language Processing (IJCNLP-04), Workshop "Beyond Shallow Analyses - Formalisms and Statistical Modeling for Deep Analyses"*, Hainan Island, China [no page numbers].
- Cahill, A., M. Burke, R. O'Donovan, J. van Genabith, and A. Way. 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*, pages 320–327, Barcelona, Spain.
- Crouch, R., R. Kaplan, T. Holloway King, and S. Riezler. 2002. A Comparison of Evaluation Metrics for a Broad Coverage Parser. In *Proceedings of the LREC Workshop: Beyond PARSEVAL – Towards Improved Evaluation Measures for Parsing Systems*, pages 67–74, Las Palmas, Canary Islands, Spain.
- Dalrymple, M. 2001. *Lexical-Functional Grammar*. San Diego, CA and Academic Press, London.
- Gildea, D. and J. Hockenmaier. 2003. Identifying Semantic Roles Using Combinatory Categorical Grammar. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 57–64, Sapporo, Japan.
- Kaplan, R. and J. Bresnan. 1982. Lexical Functional Grammar, a Formal System for Grammatical Representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA, pages 173–281.
- Kaplan, R., S. Riezler, T. Holloway King, J. T. Maxwell, A. Vasserman, and R. Crouch. 2004. Speed and Accuracy in Shallow and Deep Stochastic Parsing. In *Proceedings of the Human Language Technology Conference and the Fourth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04)*, pages 97–104, Boston, MA.
- King, T. H., R. Crouch, S. Riezler, M. Dalrymple, and R. Kaplan. 2003. The PARC700 Dependency Bank. In *Proceedings of the EAACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, pages 1–8, Budapest, Hungary.
- Kingsbury, P. and M. Palmer. 2002. From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993, Las Palmas, Canary Islands, Spain.
- Magerman, D. 1994. *Natural Language Parsing as Statistical Pattern Recognition*. Ph.D. Thesis, Department of Computer Science, Stanford University, CA.

- Marcus, M., G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 110–115, Princeton, NJ.
- Maxwell, J. and R. Kaplan. 1993. The Interface between Phrasal and Structural Constraints. *Computational Linguistics*, **19**(4):571–589.
- Miyao, Y. and J. Tsujii. 2004. Deep Linguistic Analysis for the Accurate Identification of Predicate-Argument Relations. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1392–1397, Geneva, Switzerland.
- O’Donovan, R., M. Burke, A. Cahill, J. van Genabith, and A. Way. 2004. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II Treebank. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 368–375, Barcelona, Spain.
- O’Donovan, R., M. Burke, A. Cahill, J. van Genabith, and A. Way. 2005. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. *Computational Linguistics*, **31**(3):329–365.
- Riezler, S., T. King, R. Kaplan, R. Crouch, J. T. Maxwell, and M. Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02)*, pages 271–278, Philadelphia, PA.

GERMAN QUANTIFIERS:
DETERMINERS OR ADJECTIVES?

Stefanie Dipper
Institute of Linguistics
University of Potsdam

Proceedings of the LFG05 Conference
University of Bergen
Miriam Butt and Tracy Holloway King (Editors)
2005
CSLI Publications
<http://csli-publications.stanford.edu/>

Abstract

In this paper, I address the categorial status of quantifiers and similar expressions in German. Traditionally, they are assigned either of two classes: determiners and adjectives. I argue that German quantifiers in principle are ambiguous and can be realized alternatively as determiners or adjectives. The categorial status is mirrored by the declension of attributive adjectives following these quantifiers. I present an LFG analysis that accounts for the categorial ambiguity. The analysis also covers multiple quantifiers.

1 Introduction

This paper discusses the categorial status of quantifiers and similar expressions in German, as exemplified in (1).

- (1) a. Canonical quantifiers
manche / viele / alle / zwei Frauen
some many all two women
- b. Definite and indefinite articles
die / eine Frau
the a woman
- c. Demonstrative, interrogative, and possessive determiners
jene / welche / meine Frau
that which my woman
- d. Other quantifiers
allerlei / solcherlei Leute
various such people

In the remainder of this paper, I somewhat loosely use the term “quantifiers” to refer to the different kinds of expressions in (1).

The analysis presented here has been developed in the context of the Pargram Project (Butt et al., 2002) at the IMS Stuttgart. This project focuses on the c- and f-structural implementation of a German LFG grammar. Hence, what we are heading for is a *c-structural and f-structural analysis* of the quantifiers in the above examples that can serve as the base of a *robust and efficient implementation*.¹

We will see below that German grammarians often assume that there are “determiner-like” and “adjective-like” quantifiers in German. In my analysis, I come to a similar conclusion in that I classify quantifiers as expressions of category D or A. The criteria that I apply in the classification, however, are different from the grammarians’ criteria and, hence, quantifiers are grouped differently in my analysis.

The paper is organized as follows: In sec. 2, I survey the literature, focusing on categorial analyses of German quantifiers. I then introduce the notion of *declension* (sec. 3) and investigate this property with regard to our quantifiers (sec. 4). In sec. 5, I propose an analysis of German quantifiers, applying declension as the defining criterion for the categorial status D vs. A. Finally, I show how ambiguous and multiple quantifiers are integrated in my analysis (sec. 6).

¹Further details on the implementation can be found in Dipper (2003), which includes all DP-relevant rules and lexicon entries of the implementation.

2 Previous Analyses of the German DP

In the literature outside of LFG, quite a lot of work can be found on DP analyses in general and the DP in German in particular (for the German DP, cf. Bhatt (1990); Netter (1994); Olsen (1991); Pafel (1994)). An issue that is often discussed in the literature is whether there is a full DP projection even if no specifier or determiner is overtly expressed, as in the case, e.g., of mass nouns or predicatives (cf. the discussion in Bhatt (1990, ch. 9)).

The question as to the categorial status of quantifiers in German is addressed rather rarely in formal analyses. In descriptive work, three types of quantifiers are usually distinguished: “Artikelwörter” (article words), e.g. *alle* ‘all’, “Zahladjektive” (numerals), and “indefinite Zahladjektive” (indefinite numeral adjectives), e.g. *viele* ‘many’ (e.g., Helbig and Buscha (1993, ch. 5)). If one wants to interpret this distinction in terms of categorial status, article words seem to correspond to expressions of category D, and numerals and indefinite numeral adjectives to expressions of category A. Then the question arises how to formally define the classes of article words vs. indefinite numeral adjectives.²

Often it is assumed (sometimes implicitly) that the following criteria indicate adjectival status: (i) modification by adverbs such as *sehr* ‘very’, which is typical of adjectives (assumed, e.g., by Bhatt (1990, p. 213ff)); (ii) co-occurrence with the definite article.³

Testing a first candidate, e.g., *mehrere* ‘some’, for the criteria above, the data show that *mehrere* neither cooccurs with the definite article, cf. (2), nor does it allow for modification by an adverb (3).

- (2) a. *mehrere Menschen*
 some people
- b. **die mehreren Menschen*
 the some people
- (3) **sehr mehrere Menschen*
 very some people

In contrast to *mehrere*, *viele* ‘many’ is compatible with the definite article (4) and can be modified by an adverb (5).

- (4) a. *viele Menschen*
 many people

²The classification of Eisenberg (1999) is somewhat different. Besides a highly restricted class of article words, he distinguishes between numerals, pronouns, and quantifying adjectives. In his approach, the question is how to tell pronouns from quantifying adjectives.

³For instance, Bhatt classifies quantifiers like *beide* ‘beide’ as an adjective in (ia) and as a determiner in (ib)—apparently based on the presence/absence of the definite article.

- (i) a. *die beiden jungen Frauen*
 the both young women
 ‘both of the young women’
- b. *beide genannten Verfahren*
 both mentioned methods
 ‘both methods mentioned’

- b. *die vielen Menschen*
 the many people
 ‘the numerous people’
- (5) *sehr viele Menschen*
 very many people

Most of the quantifiers in German behave like *mehrere*, i.e. at first sight, the data seem to suggest that most of the quantifiers in German (including *mehrere*) are determiners, and that quantifying adjectives such as *viele* constitute an exceptional case.

However, the fact that a quantifier is incompatible with the definite article or with modifying adverbs may well be due to the semantics of the quantifier in question and need not be connected to its (syntactic) status as a determiner or adjective at all. Hence, the above criteria ought not to be applied to determine the categorial (adjectival) status of the quantifiers.⁴

In the context of LFG, details of the internal structure of nominal phrases are often left open. There is some literature about the analysis of the DP in Northern Germanic languages, cf. Börjars (1998); Börjars et al. (1999). They focus on the feature DEF, which in these languages can or must be expressed via a noun suffix.

⁴Other properties that are attributed to quantifiers are:

1. They occur at the left periphery of the DP, cf. (i).

(i) *die / alle / viele jungen Frauen* vs. **jungen die / alle / viele Frauen*
 the all many young woman young the all many woman

2. They “close” a DP, i.e., nouns that cannot represent a DP on their own form DPs when they are preceded by a quantifier, cf. (ii).

(ii) **Frau lachte* vs. *die / welche / manche Frau lachte*
 woman laughed the which some woman laughed

3. Semantically, they differ from (ordinary) adjectives in that they are not intersective, cf. (iii). Instead, they typically have little descriptive content and contribute information about the quantity or definiteness of the entities that are referred to by the head noun.

(iii) a. *junge Frauen* = $\lambda x [\text{woman}(x) \ \& \ \text{young}(x)]$
 young women
 b. *alle / viele Frauen* $\neq \lambda x [\text{woman}(x) \ \& \ \text{all/many}(x)]$
 all many women

However, properties 1 and 3 do not help us in telling article words from indefinite numeral adjectives since all quantifiers behave uniformly in these respects. Property 2 clearly involves semantic properties of the DP’s head noun and, moreover, does not hold for all quantifiers: due to their meaning, certain quantifiers cannot close arbitrary DPs but combine with mass or plural nouns only, compare (iva) and (ivb/c). Mass and plural nouns, however, can represent a DP on their own, in contrast to singular count nouns like *Frau* ‘woman’.

- (iv) a. **einige Frau*
 some woman[SG]
 b. *einiges Geld*
 some money[SG]
 c. *einige Frauen*
 many women[PL]

Among other things, this feature determines the declension of attributive adjectives within the DP: [DEF +] triggers so-called weak adjective agreement, [DEF –] triggers strong adjective agreement, cf. the Swedish example in (6).

- (6) a. *en röd bil*
 a red[ST] car
- b. *den röda bilen*
 this red[WK] car[DEF]

German, however, does not have a noun suffix that indicates definiteness. Furthermore, although German also has weak and strong adjective agreement, as we will see below, most (non-LFG) analyses of the German DP assume that definiteness plays no role in adjectival declension (for a different view, see Pafel (1994)). This is easily seen by the indefinite article *ein* ‘a’, which combines with strong or weak adjectival declension, depending on case, cf. (7).⁵ So, clearly, the German DP differs from DPs in Northern Germanic languages in important aspects.

- (7) a. *ein süßer/*süße Wein*
 a sweet wine
 [NOM] [ST/*WK]
- b. *einem *süßem/süßen Wein*
 a sweet wine
 [DAT] [*ST/WK]

While in German, strong/weak adjective declension does not correlate with definiteness, I argue in the next sections that it mirrors the structure of a DP and, hence, can be used to determine the c-structural status of quantifiers.

3 Agreement Patterns in German

In a German DP, determiners, adjectives, and nouns show agreement relations with respect to different features. I distinguish two types of agreement: (i) adjective–noun and determiner–noun agreement, concerning the features gender, number, case; (ii) determiner–adjective agreement, concerning the strong-weak feature declension.

3.1 Adj–N and D–N agreement (gender, number, case)

In attributive position, a German adjective agrees in gender, number, and case with its head noun, cf. (8).⁶

- (8) a. *süßer Wein*
 sweet wine
 [MASC,SG,NOM] [SG,NOM]

⁵Below I classify the form *ein*, as in (7a), as uninflected rather than marked for case.

⁶German nouns are inherently/lexically marked for gender. Note that due to massive case syncretism, many of the nouns in the examples could be dative or accusative. I only mark the reading(s) that are valid in the given context.

b. *süßes* *Bier*
 sweet beer
 [NEUT,SG,NOM/ACC] [SG,NOM/ACC]

c. *süße* *Weine*
 sweet wines
 [PL,NOM/ACC] [PL,NOM/ACC]

Likewise, a determiner agrees with its head noun (and with attributive adjectives, if present), cf. (9).

(9) a. *der* *Wein*
 the wine
 [MASC,SG,NOM] [SG,NOM]

b. *das* *Bier*
 the beer
 [NEUT,SG,NOM/ACC] [SG,NOM/ACC]

3.2 D-Adj agreement (declension)

Besides gender, number, and case, a fourth parameter is involved, “declension”. Both determiners and adjectives show declension, but in different ways.

Determiners Determiners come in two declension types: they may be inflected or uninflected. Most determiners fall in one class only, i.e. they show declension in *all* cases, cf. (10), or they *never* inflect, cf. (11). Inflected determiners exhibit the so-called “strong” declension, indicated by ‘ST’ in the examples; the corresponding inflectional “strong” ending is printed in bold-face.⁷ Uninflected determiners are marked by ‘∅’.⁸

(10) a. *der* / *des* / *dem* / *den* *Wein*
 the[NOM,ST] / [GEN,ST] / [DAT,ST] / [ACC,ST] wine

b. *jener* / *jenes* / *jenem* / *jenen* *Wein*
 that[NOM,ST] / [GEN,ST] / [DAT,ST] / [ACC,ST] wine

(11) *solcherlei Wein*
 such[∅] wine

⁷Traditional analysis assume that “weak” determiners exist as well, see fn. 11.

Note that a considerable number of quantifiers have an inflected as well as an uninflected variant, cf. (i). I consider these as two different lemmas, in contrast to, e.g., Pafel (1994). That is, I assume that declension type is an inherent property of determiners. However, the quantifier analysis argued for in this paper is compatible with the two-variant assumption as well.

(i) a. *mancher* / *manches* / *manchem* / *manchen* *Wein*
 some[NOM,ST] / [GEN,ST] / [DAT,ST] / [ACC,ST] wine

b. *manch* *Wein*
 some[∅] wine

⁸In the examples in this section, the head noun *Wein* ‘wine’ actually ought to be inflected in the genitive case: *Weins*. For ease of reading, I disregard this difference.

An exception are the indefinite article and possessive determiners: depending on case (and gender), they inflect or remain uninflected; compare the uninflected forms *ein*, *mein* in nominative singular with the other, inflected, cases (12).

- (12) a. *ein / eines / einem / einen Wein*
 a[∅] / [GEN,ST] / [DAT,ST] / [ACC,ST] wine
- b. *mein / meines / meinem / meinen Wein*
 my[∅] / [GEN,ST] / [DAT,ST] / [ACC,ST] wine

The following table presents an overview of the three declension classes of determiners and their inflectional properties. All plural forms (column ‘Pl’) behave uniformly, whereas in the singular, case and gender matters for the “mixed” class (rows ‘Nom, Gen, Dat, Acc’ and columns ‘Masc, Neut, Fem’).

| Class | Example | Sg | Pl |
|-------------|--------------------------|----|----|
| inflected | <i>der</i> ‘the’ | ST | ST |
| uninflected | <i>solcherlei</i> ‘such’ | ∅ | ∅ |

| | | | Masc | Neut | Fem | Pl |
|---------|---|-----|------|------|-----|----|
| “mixed” | <i>ein</i> ‘a’, <i>kein</i> ‘no’, <i>mein</i> ‘my’, etc. | Nom | ∅ | ∅ | ST | ST |
| | | Gen | ST | ST | ST | ST |
| | | Dat | ST | ST | ST | ST |
| | | Acc | ST | ∅ | ST | ST |

Adjectives Similarly to determiners, attributive adjectives can be inflected or uninflected and, like for determiners, it is an inherent feature of adjectives what declension type they belong to.

However, in contrast to inflected determiners, which are always strong, inflected adjectives may be “strong” or “weak”.⁹ The declension (strong/weak) of an inflected adjective depends on the declension of its determiner, i.e., adjectival declension is an *agreement* phenomenon.

The tables below present all strong and weak adjectival endings. As can be seen from the tables, the endings *-er*, *-es*, and *-em* are clear indicators for strong declension; *-e* and *-en*, while predominantly weak, are ambiguous.¹⁰

| | | Masc Sg | Neut Sg | Fem Sg | Pl |
|---------|-----|------------|------------|------------|------------|
| Strong: | Nom | -er | -es | <i>-e</i> | <i>-e</i> |
| | Gen | <i>-en</i> | <i>-en</i> | -er | -er |
| | Dat | -em | -em | -er | <i>-en</i> |
| | Acc | <i>-en</i> | -es | <i>-e</i> | <i>-e</i> |
| Weak: | Nom | <i>-e</i> | <i>-e</i> | <i>-e</i> | <i>-en</i> |
| | Gen | <i>-en</i> | <i>-en</i> | <i>-en</i> | <i>-en</i> |
| | Dat | <i>-en</i> | <i>-en</i> | <i>-en</i> | <i>-en</i> |
| | Acc | <i>-en</i> | <i>-e</i> | <i>-e</i> | <i>-en</i> |

⁹Strong declension is also called “pronominal” declension, since it is similar to the declension of pronouns. Weak declension was restricted to nouns in older stages of the language, hence it is sometimes called “nominal” declension.

¹⁰In the plural, all genders exhibit identical inflection. Except for genitive singular, strong determiner endings are identical to strong adjectival endings.

Adjectival declension depends on the declension of a preceding determiner in the following way:

- If preceded by an inflected (= strong) determiner, the adjective comes in its so-called “weak” form, cf. (13).

- (13) a. *der süße Wein*
the[ST] sweet[WK] wine ([NOM])
b. *einem süßen Wein*
a[ST] sweet[WK] wine ([DAT])

Multiple, successive adjectives show identical declension, cf. (14).

- (14) a. *der süße rote Wein*
the[ST] sweet[WK] red[WK] wine ([NOM])
b. *einem süßen roten Wein*
a[ST] sweet[WK] red[WK] wine ([DAT])

- If preceded by a non-inflected determiner or if no determiner is present, the adjective itself exhibits strong declension, cf. (15) and (16), respectively. (Note the similarity between the inflectional ending of the strong determiner in (13a) and the strong adjectives in (15) and (16).)

- (15) a. *solcherlei süßer Wein*
such[∅] sweet[ST] wine ([NOM])
b. *ein süßer roter Wein*
a[∅] sweet[ST] red[ST] wine ([NOM])

- (16) *süßer Wein*
sweet[ST] wine ([NOM])

- Uninflected adjectives like *lila* ‘purple’ never inflect and are compatible with any declension type, cf. (17). They do not yield data relevant to our purposes and can be safely ignored.

- (17) a. *der süße lila Wein*
the[ST] sweet[WK] purple[∅] wine
b. *ein süßer lila Wein*
a[∅] sweet[ST] purple[∅] wine
c. *süßer lila Wein*
sweet[ST] purple[∅] wine

The following generalization emerges from the above data: In a DP, the feature “strong” is represented (i) on the head D if present and if inflected, (ii) on attributive adjectives otherwise (similarly assumed, e.g., by Bhatt (1990); Olsen (1991, ch. 9.4)). One important conclusion is: *Determiners and adjectives show complementary declension.*¹¹

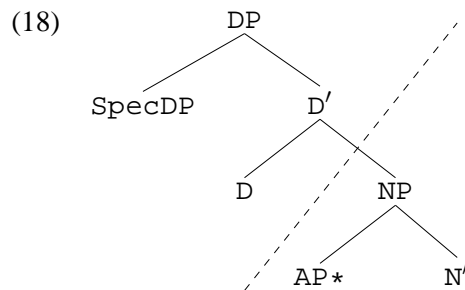
¹¹Consequently, attributive adjectives that follow the indefinite article or possessive determiners show a mixed declension: strong declension after (uninflected) *ein* ‘a[∅]’ (= MASC/NEUT,SG,NOM) and weak declension in all other cases, i.e. after *eines*

The table lists all possible combinations of declensions as predicted by the generalization:

| Determiner | Adjective |
|---------------------------------|-------------------------|
| strong | weak (or uninflected) |
| uninflected or no determiner | strong (or uninflected) |

4 A New Criterion: Declension

The generalization presented in the previous section implies that a German DP can be partitioned according to declension, as shown in the tree in (18): the part above the dotted line belongs to the domain of D, the part below that line comprises adjectives and the head noun; the parts can be formally (i.e., by surface properties) identified by complementary declension.



We now have a straightforward solution to our initial question as to how to identify the categorial status of quantifiers: by looking at their declensional properties. That is, we need to determine whether a quantifier parallels the declension of canonical determiners (such as the definite article) or whether it parallels the declension of ordinary adjectives. This is done by testing for the declension of a following attributive adjective. If the adjective shows the same declension as the quantifier in question (e.g., both show strong declension), then the quantifier is a quantifying adjective. Otherwise, if the adjective shows complementary declension, the quantifier must be a determiner.

The table summarizes the potential combinations of declension and the categorial status of the quantifier that emerges from the combinations.¹²

| Quant candidate | Adjective | C-str class |
|-----------------|-------------------------|---------------|
| strong | weak (or uninflected) | Dquant |
| strong | strong (or uninflected) | Aquant |
| uninflected | strong (or uninflected) | Dquant/Aquant |

¹²‘a[MASC/NEUT,SG,GEN,ST]’, *einem*, etc. Traditionally, this declension pattern is regarded as a declension type of its own, called “mixed declension” (see, e.g., Drosdowski (1995, p. 279) or Müller (1999, ch. 7.2)).

Authors who do not assume a mixed declension type fall in two classes: Some assume that (uninflected) *ein* (and *kein*, *mein*, etc.) are weak determiners (Pollard and Sag, 1994, ch. 2.2); others analyze *ein* as uninflected (Netter 1994) (as we shall do in our analysis). The first approach has the drawback that *ein* constitutes the only instance of a weak determiner, whereas within the second approach, *ein* behaves like any uninflected determiner.

¹²For uninflected quantifier candidates, nothing can be derived from this test: there are uninflected determiners as well as uninflected adjectives in German. Ordinary inflected adjectives that follow them have to exhibit strong declension in either case.

Applying this test, e.g., to corpus data of the quantifier *mehrere* ‘some’, reveals that *mehrere* behaves like an ordinary adjective in that it shows the same declension as a following adjective, cf. (19). Hence, *mehrere* is classified as a quantifying adjective and therefore analyzed as occupying an adjectival position, which I call the *Aquant* position.¹³

- (19) *mehrere strittige Punkte*
 some[ST] contestable[ST] points

In contrast, for the quantifier *alle* ‘all’ the test reveals that *alle* behaves like the canonical determiner *die* ‘the’: *alle* and the following adjective exhibit complementary declension, cf. (20). Hence, *alle* is classified as a determiner, occupying the *Dquant* position.

- (20) *alle politischen Parteien*
 all[ST] political[WK] parties

5 Analysis

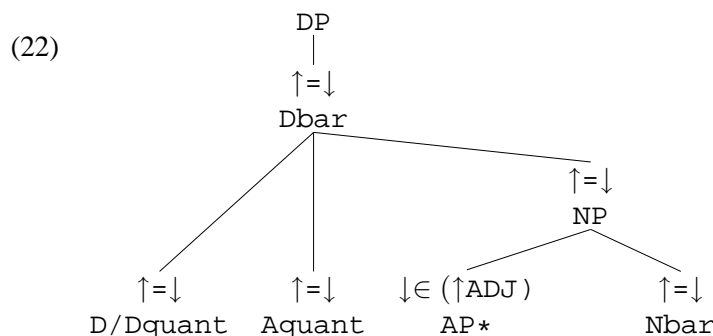
Having introduced the criteria as to how to determine the categorial status of quantifiers, I now present my c- and f-structure analysis of quantifiers in German.

Despite the variance in inflection, it seems sensible to represent quantifiers uniformly in f-structure, e.g., to facilitate subsequent semantic processing. That is, in my analysis a quantifier in the *Aquant* position—although inflecting like an ordinary adjective—functions as a specifier (contrary to ordinary adjectives). Hence, the c-structure distinction *Dquant* vs. *Aquant* does not correspond to an f-structure distinction.

Example lexicon entries for the canonical determiner *der* ‘the’, the *Dquant* determiner *alle* ‘all’, and the *Aquant* determiner *mehrere* ‘some’ are sketched out in (21).

- (21) *der* D (↑ SPEC DET PRED) = ‘die’
alle Dquant (↑ SPEC QUANT PRED) = ‘alle’
mehrere Aquant (↑ SPEC QUANT PRED) = ‘mehrere’

The schematic tree in (22), enriched by f-structure annotations, shows a slightly simplified version of my analysis.



¹³Remember that according to the “traditional” criteria, *mehrere* probably has to be classified as an article word, cf. examples (2) and (3) above.

Contrary to expectation, the c-structure position of Aquant—being an adjective according to its declensional behaviour—is not within the NP, in contrast to the position of ordinary adjectives. Instead, Aquant is dominated by DP, like determiners. There are two reasons for this: (i) Quantificational adjectives always precede all other adjectives; this is directly modeled by putting Aquant in the higher DP projection. (ii) More importantly, quantificational adjectives can be interrogative, cf. (23).

(23) *wieviele deutsche Aussiedler*
 how_many[ST] German[ST] emigrants

Treating *wieviele* as a quantifying adjective within NP would be in contrast to the generalization we otherwise observe: that the type of a DP is determined by elements of the D projection, never by some element within NP.¹⁴

Agreement with regard to declension is implemented by a feature DECL, which is projected by inflected expressions of category D, Dquant, Aquant, and A:

- D/Dquant vs. Aquant/A project incompatible feature values: (\uparrow DECL) = ST-DET vs. (\uparrow DECL) = ST-ADJ. This guarantees complementary declension of D/Dquant vs. Aquant/A.
- Weak Aquant/A introduce a constraining equation: (\uparrow DECL) =_c ST-DET. This has the desired effect that they may only occur after strong D/Dquant.
- Uninflected D, Dquant, Aquant, and A do not introduce any constraints on DECL, since they are compatible with any declension.

Outlines of example f-structures for (19) and (20) are displayed in (24) and (25), respectively.

(24) $\left[\begin{array}{ll} \text{PRED} & \text{'Punkt'} \\ \text{SPEC} & \left[\text{QUANT} \left[\text{PRED 'mehrere'} \right] \right] \\ \text{ADJUNCT} & \left\{ \left[\text{PRED 'strittig'} \right] \right\} \\ \text{DECL} & \text{st-adj} \\ \text{GEND} & \text{masc} \\ \text{NUM} & \text{pl} \\ \text{CASE} & \text{nom} \end{array} \right]$

(25) $\left[\begin{array}{ll} \text{PRED} & \text{'Partei'} \\ \text{SPEC} & \left[\text{QUANT} \left[\text{PRED 'alle'} \right] \right] \\ \text{ADJUNCT} & \left\{ \left[\text{PRED 'politisch'} \right] \right\} \\ \text{DECL} & \text{st-det} \\ \text{GEND} & \text{fem} \\ \text{NUM} & \text{pl} \\ \text{CASE} & \text{nom} \end{array} \right]$

¹⁴Note that *wieviele* actually consists of two components: *wie* 'how' and *viele* 'many'. One could argue that the interrogative part *wie* is attached outside of NP while *viele* remains within the NP projection. However, *welche* 'which', which is not composed of such transparent components, can also be used as an interrogative Aquant.

6 Ambiguous and Multiple Quantifiers

In this section, I address two further aspects of quantifiers: (i) many quantifiers are ambiguous with regard to their categorial status; (ii) multiple quantifiers do occur in German.

6.1 Ambiguous quantifiers

In the preceding section, a clear line was drawn between determiners on one side and adjectives (including quantifying adjectives) on the other, based on inflectional properties. However, the borderline is not always that clear. Many quantifiers exhibit idiosyncratic declension.

Traditional grammars note that after certain quantifiers the declension of attributive adjectives varies. For example, quantifiers preceding weak adjectives (hence determiners, according to our analysis) comprise: *solche* ‘such’, *irgendwelche* ‘any’, and *manche* ‘some’. But some of these expressions also tolerate strong adjectives (e.g. *irgendwelche*); some even prefer strong adjectives but only in plural forms (e.g. *manche*), etc.¹⁵

To get a clearer view of the data, I performed a corpus analysis on the Frankfurter Rundschau Corpus.¹⁶ The tables below summarize the results I got from the FR corpus for a selection of quantifiers. The tables show the frequency of unambiguous instances for quantifiers; the first table lists expressions with predominantly determiner declension, the second lists expressions with predominantly adjectival declension.¹⁷

| | Relative Frequency | Absolute Frequency |
|-----------------------|--------------------|--------------------|
| <i>die</i> ‘the’ | 99.9 % | 90,230 |
| <i>jede</i> ‘each’ | 99.8 % | 2,087 |
| <i>diese</i> ‘this’ | 99.7 % | 4,324 |
| <i>jene</i> ‘that’ | 98.9 % | 369 |
| <i>welche</i> ‘which’ | 96.7 % | 91 |
| <i>alle</i> ‘all’ | 95.8 % | 1,781 |
| <i>wenige</i> ‘few’ | 92.5 % | 721 |
| <i>manche</i> ‘some’ | 79.3 % | 119 |

Quantifiers with predominantly determiner declension (D/Dquant)

¹⁵Traditional grammars typically devote several sections to the problem of such idiosyncratic inflectional properties. Here is an example:

“[So wie nach dem definiten Artikel], aber mit bestimmten Einschränkungen werden die Adjektive flektiert nach den Artikelwörtern *mancher* (Plural überwiegend wie nach Nullartikel, *irgendwelcher* (durchgehend auch wie nach Nullartikel möglich), *solcher* (gelegentlich wie nach Nullartikel, nicht aber im Sg.Nom. und Akk. aller Genera und Gen.Mask. und Neutr.), *welcher* und *aller* (selten auch wie nach Nullartikel).” (Helbig and Buscha, 1993, p. 301)

Free translation: “After the following quantifiers, attributive adjectives show weak declension (with certain restrictions, listed in parentheses): *manche* ‘some’ (in plural predominantly strong), *irgendwelche* ‘any’ (strong declension equally possible), *solche* ‘such’ (sometimes strong, but not in [SG,NOM/ACC] and [MASC/NEUT,GEN]), *welche* ‘which’ and *alle* ‘all’ (rarely strong).”

¹⁶The FR corpus comprises about 40 million tokens and is delivered by the European Corpus Initiative, URL: <http://www.elsnet.org/resources/eciCorpus.html>.

¹⁷Due to case syncretism, only a subset of the corpus instances of a quantifier followed by an attributive adjective provide unambiguous evidence for determiner vs. adjectival declension. Only quantifiers with more than 50 unambiguous instances in the corpus were taken into account.

| | Relative Frequency | Absolute Frequency |
|--------------------------|--------------------|--------------------|
| <i>mehrere</i> ‘some’ | 98.0 % | 50 |
| <i>einige</i> ‘some’ | 92.1 % | 129 |
| <i>andere</i> ‘other’ | 91.2 % | 249 |
| <i>viele</i> ‘much/many’ | 88.4 % | 169 |
| <i>solche</i> ‘such’ | 60.4 % | 81 |

Quantifiers with predominantly adjectival declension (Aquant)

Below we list some of the “counterexamples”, i.e. examples exhibiting the unusual, more marked declension. The examples in (26) are at the margins of ungrammaticality; the examples in (27) are quite acceptable.

- (26) a. (*) *bei jedem mißglücktem Dribbling*
on each[ST] bad[ST] dribbling
- b. (*) *vor diesem wirtschaftlichem Hintergrund*
against this[ST] economic[ST] background
- c. (*) *mit jenem spektakulärem Triumph*
with that[ST] spectacular[ST] triumph
- d. (?) *laut mehrerer ärztlichen Atteste*
according_to several[ST] medical[WK] certificates
- (27) a. *einiges verschämte Kichern*
some[ST] bashful[WK] giggling
- b. (?) *anderer hessischen Jugendzentren*
other[ST] Hessian[WK] youth_centres

According to my analysis, (many of) the above quantifiers are ambiguous with respect to their categorial status. Hence, they are classified as both Dquant and Aquant in their lexicon entries. To encode idiosyncratic preferences, I use OT marks (Frank et al., 2001). An example entry is given in (28) for the quantifier *viele* ‘many’, which predominantly is an Aquant.

- (28) *viele* Dquant (↑ SPEC QUANT PRED) = ‘viele’
 Aquant (↑ SPEC QUANT PRED) = ‘viele’
 PreferVieleAsAquant ∈ o:.*

6.2 Multiple quantifiers

Finally, the criterion proposed in sec. 4 reveals that multiple quantifiers do occur in German (also assumed by Bhatt (1990, p. 204ff) and Pafel (1994)). In DPs such as (29), both *alle* ‘all’ and *die* ‘the’ need to be classified as determiners due to their inflectional behaviour. Further examples are given in (30).

- (29) *alle die schönen Definitionen*
all[ST] the[ST] nice[WK] definitions

- (30) a. *alle unsere schönen Sprüche*
all[ST] our[ST] pretty[WK] sayings
- b. *manch einem wissenschaftlichen Assistenten*
some[∅] (a[ST]) research[WK] assistant

Only certain (probably semantically restricted) combinations of multiple determiners or determiners plus quantifying adjectives are grammatical in German. To avoid massive overgeneration, only quantifiers that are lexically assigned a specific category, Dpre (“predeterminers”), may precede other determiners in my implementation (i.e. the class of predeterminers is restricted c-structurally). In contrast, there is no restriction on multiple Aquants.

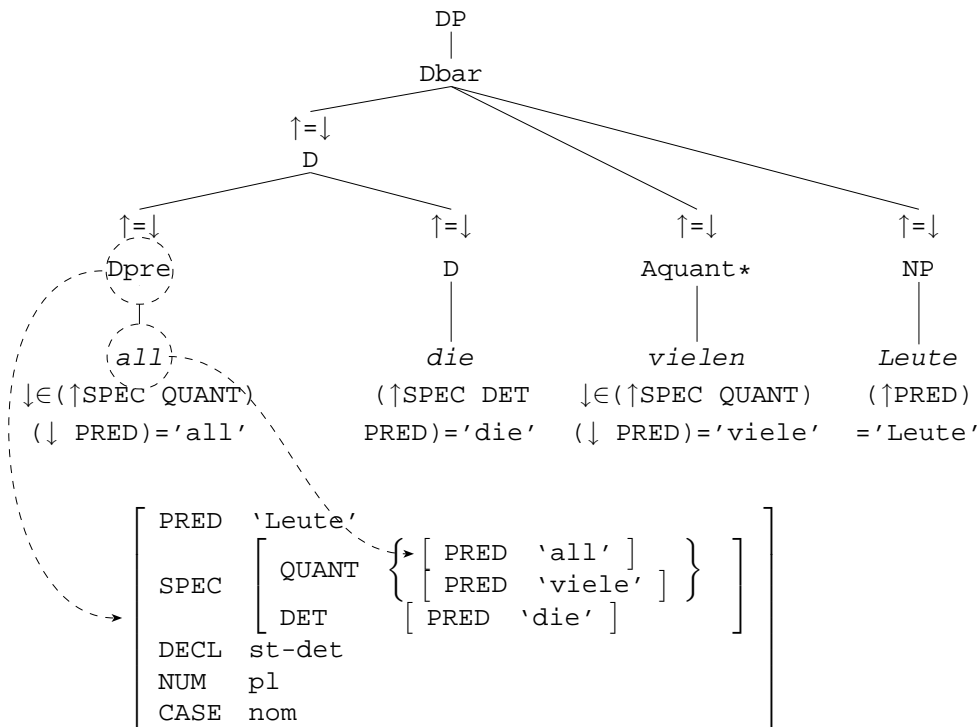
In f-structure, indefinite quantifiers project a set-valued feature QUANT, similar to the set-valued feature ADJUNCT, which is projected by (multiple) adjectives.

In contrast to the class of indefinite quantifiers, which can be iterated within in a DP, the definite and indefinite articles as well as other types of “quantifiers”, such as demonstratives, interrogatives or possessives (see the examples in (1)), can occur only once within a DP. These quantifiers project specific, single-valued features such as DET, DEM, INT, or POSS in my analysis.

(32) displays an annotated c-structure analysis and the corresponding f-structure of the example in (31), featuring three quantifiers. Exemplary ϕ -projections are shown for the terminal node *all* and its mother node Dpre.

- (31) *all die vielen Leute*
all[∅] the[ST] many[WK] people
‘all these numerous people’

(32)



7 Conclusion and Open Questions

In this paper, I have argued for a formal, non-semantic criterion for distinguishing between determiner-like and adjectival quantifiers (and related expressions). I propose to determine the categorial status of quantifiers by *declension*: the quantifier either parallels the declension of an attributive adjective and is thus classified as a quantifying adjective. Or else, they show complementary declension, and thus the quantifier is classified as a determiner. The criterion also reveals that ambiguous and multiple quantifiers do occur in German.

In my implementation, I assume the category Dpre for predeterminers, D for canonical determiners, and Dquant and Aquant for determiner-like and adjectival quantifiers, respectively. These categories are dominated by DP and function as f-structure heads. Most of the quantifiers in German are ambiguous and are assigned both Dquant and Aquant in their lexicon entries. Idiosyncratic preferences are encoded by OT marks.

While the implementation presented here allows us to analyze ambiguous quantifiers, reasons for the observed ambiguous nature have still to be found. The rule of thumb that determiners have less descriptive content than adjectives does not carry over to Dquant vs. Aquant preferences of individual quantifiers. For instance, the descriptive content of the predominantly-Dquant quantifier *wenige* ‘few’ seems very similar to the predominantly-Aquant quantifier *viele* ‘many/much’.

What my implementation does not account for is the fact that the idiosyncratic variance depends on case and number. For instance, *solche* ‘such’ sometimes inflects like a determiner but not in the cases of [SG,NOM/ACC] and [MACS/NEUT,GEN] (cf. fn. 15). Obviously the implementation does not model the variance in such detail. However, the factors that play a role in the observed variance are not yet understood; possibly phonetic factors are involved.

References

- Christa Bhatt. *Die syntaktische Struktur der Nominalphrase im Deutschen*, volume 38 of *Studien zur deutschen Grammatik*. Tübingen: Narr, 1990.
- Kersti Börjars. Clitics, affixes, and parallel correspondences. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG98 Conference*, Brisbane, Australia, 1998. CSLI Online Proceedings. <http://csli-publications.stanford.edu/LFG/3>.
- Kersti Börjars, John Payne, and Erika Chisarik. On the justification for functional categories in LFG. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG99 Conference*, Manchester, UK, 1999. CSLI Online Proceedings. <http://csli-publications.stanford.edu/LFG/4>.
- Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. The Parallel Grammar Project. In *Proceedings of the COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7, Taipei, Taiwan, 2002.
- Stefanie Dipper. *Implementing and Documenting Large-Scale Grammars — German LFG*, volume 9(1) of *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)*. IMS, University of Stuttgart, 2003.
- Günther Drosdowski, editor. *DUDEN. Grammatik der deutschen Gegenwartssprache*. Mannheim et al.: Dudenverlag, 5 edition, 1995.

- Peter Eisenberg. *Grundriss der deutschen Grammatik*, volume 2 Der Satz. J.B.Metzler, 1999.
- Anette Frank, Tracy Holloway King, Jonas Kuhn, and John Maxwell. Optimality theory style constraint ranking in large-scale LFG grammars. In Peter Sells, editor, *Formal and Empirical Issues in Optimality Theoretic Syntax*, Studies in Constraint-Based Lexicalism, pages 367–397. Stanford, CA: CSLI Publications, 2001.
- Gerhard Helbig and Joachim Buscha. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Leipzig et al.: Langenscheidt/Enzyklopädie, 15 edition, 1993.
- Stefan Müller. *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche*. Number 34 in *Linguistische Arbeiten*. Tübingen: Niemeyer, 1999.
- Klaus Netter. Towards a theory of functional heads: German nominal phrases. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, number 46 in CSLI Lecture Notes, pages 297–340. Stanford, CA: CSLI, 1994.
- Susan Olsen. Die deutsche Nominalphrase als “Determinansphrase”. In Susan Olsen and Gisbert Fanselow, editors, *DET, COMP und INFL. Zur Syntax funktionaler Kategorien und grammatischer Funktionen*, volume 263 of *Linguistische Arbeiten*, pages 35–56. Tübingen: Niemeyer, 1991.
- Jürgen Pafel. Zur syntaktischen Struktur nominaler Quantoren. *Zeitschrift für Sprachwissenschaft*, 13(2): 236–275, 1994.
- Carl Pollard and Ivan Sag. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. Stanford, CA: CSLI/Chicago, IL: Chicago University Press, 1994.

DIRECT OBJECT CLITIC DOUBLING IN OT-LFG: A NEW LOOK AT RIOPLATENSE SPANISH

Bruno Estigarribia

Stanford University

Proceedings of the LFG05 conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

Spanish expresses the direct object argument of transitive clauses as a Direct Object Clitic, as a lexical or independent pronominal NP, or both (*Direct Object Clitic Doubling*). The latter structure presents an obvious puzzle to theories that assume one form or another of functional uniqueness. Although much research has been devoted to the structural representation and semantics of DOCLD, a rather natural question has been left uninvestigated in the linguistic literature: what drives these different types of expression?

In this paper I analyze the Rioplatense dialect (henceforth RSp), which is generally described as allowing CLD more freely than other dialects (including Standard Peninsular Spanish). Using the apparatus of OT, I investigate the relation between discourse structure, cross-linguistic markedness hierarchies, and formal expression of direct objects. My examples come from a corpus of four texts that I reference in the appendix, augmented by examples from the Internet which are in the public domain. Where no source is cited, the example is a constructed one, usually a modified example from the corpus.¹

1 Introduction

Direct Object clauses in Spanish can be classified in three groups². There are structures in which the direct object argument of a transitive verb is expressed by a lexical NP:

- | | |
|--|---|
| (1) (a) <Beto-42> <i>Aldo pidió un mate</i> Aldo asked a mate Aldo ordered a “mate”. | (1) (b) <PerInv-14> <i>De inmediato miró fijamente a sus compañeros</i> Immediately looked.at-he fixedly A his-PL partners He immediately looked fixedly at his partners. |
|--|---|

In other cases, the direct object argument is instantiated by an allotactic (Haiman 1985) direct object clitic (a *special clitic*, Zwicky 1977):³

- | | |
|--|--|
| (2) (a) <Beto-42> <i>Roque lo miró fijo</i> Roque CL he-looked-at fixed Roque looked at him fixedly. | (2) (b) <Lig-45> <i>Las debés tener</i> CL you-must have You sure have them. |
|--|--|

Finally, some structures have a direct object clitic that is coreferential with a lexical NP in the same clause:

- | | |
|--|---|
| (3) (a) <Lig-110> <i>Yo <u>las</u> tenía guardadas <u>las</u> cartas⁴</i> I CL had stored the letters I had the letters stored. | (3) (b) <Beto-50> <i>¿<u>La</u> vas a llamar <u>a</u> Marta?</i> CL you-go to call A Marta Are you going to call Marta? |
|--|---|

This paper focuses on Rioplatense Spanish, the dialect of the area around the Río de la Plata, including cities like Buenos Aires and Rosario, in Argentina, and Montevideo, in Uruguay.⁵ This dialect is generally described as allowing CLD more freely than Standard Peninsular Spanish, since in the former only inanimate direct objects can be doubled. The data were extracted from a corpus of seven conversations published in Ligatto (1996). These involve 15 participants, from 10 to 60 years of age, and two

¹ Special thanks to Joan Bresnan, T. Florian Jaeger, Peter Sells, Nigel Vincent, and the audience and organizers of LFG05 at Universitetet i Bergen, Norway.

² I will not address clitic doubling of indirect objects here.

³ The direct object clitics will be glossed “CL”, without indication of person, number or gender. Refer to the appendix for a table of DOCLs. Indirect Object Clitics, when they appear, will be glossed “IOCL”.

⁴ In all the examples in this paper, underlining signals the clitic and its associated constituent, boldface signals the clitic; when necessary, small capitals will mark focal stress. The glosses will be as transparent and non-technical as possible; however, the technical gloss A will be used for the animacy marker *a* which is obligatory before animate direct objects and is homonymous with the IO marker *a* and the preposition *a*.

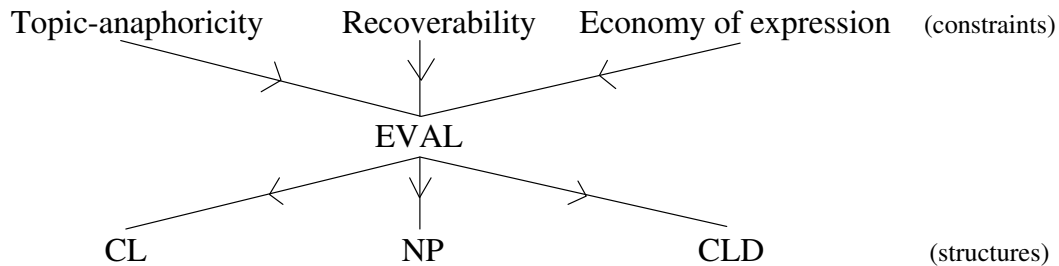
⁵ This dialect is also called Porteño Spanish or River Plate Spanish in the literature. It is spoken in a wider area than Buenos Aires, hence my choice of “Rioplatense”.

interviewers, all middle-class, born and raised in Buenos Aires. With the obvious exception of the ten-year old, all completed high-school at least. One participant was born in North-Eastern Argentina, but had lived in Buenos Aires since her adolescence. Data from a participant born in Spain were excluded. From the Ligatto corpus are also two short excerpts from a 1980's political talk-show, *Tiempo Nuevo*, involving two journalists and three politicians.

I supplemented this corpus with two short stories by Argentinian writer Roberto Fontanarrosa (1995), “Periodismo Investigativo” and “Beto”, which contain fictional dialogues in the vernacular of Rosario. The characters in these stories are middle-class men and women in their forties and fifties, acquaintances and work colleagues in the first, old friends in the second. The origin of the examples is given between angled brackets, followed by the page number in the original text. The few constructed examples I used are marked as such (all ungrammatical examples are constructed).

Much research has tried to establish a structural representation for Direct Object Clitic Doubling structures that does not violate theoretical tenets. In LFG, for instance, Andrews (1990) proposed two lexical entries: one for stand-alone clitics with an obligatory PRED, and one for doubling clitics with an optional PRED and an ANIM + specification (which restricts doubling to animates and thus works for Peninsular Spanish but not for RSp).⁶ Also, researchers have tried to find factors that explain DOCLD's restricted distribution vis-à-vis the quasi-obligatory clitic doubling of indirect objects (see Estigarribia, forthcoming). However, these analyses are ultimately unsatisfactory because they do not address the question of why CLD exists and what functions it serves, especially considering the assumption, implicit in the name “clitic doubling”, that the clitic is a pleonastic element.⁷ My working assumption here is that, in each structure, no lexical node is superfluous. Therefore, in CLD-clauses, both the presence of the DOCL and of an independent NP instantiating the same argument have to be justified. Note that this is consistent with LFG's principle of economy of expression. Bresnan (2001:147) suggests that the clitic's presence in the c-structure as a terminal node “still bears some kind of information not available elsewhere.”⁸

My proposal is that three factors “conspire” to give the range of possible RSp transitive structures: Topic-anaphoricity (Bresnan 2001), associated with pronominal expression; Recoverability, associated with lexical expression; and Economy of expression, which prefers structures with as few lexical nodes as possible. This hypothesis is clearly in the spirit of Optimality Theory: a CLD construction would be the optimal result of conflicting pressures to give an argument a certain type of expression. CLD obtains when expression of both a clitic and a NP is the optimal response to independent constraints on expression of either argument.



This analysis thus predicts under which conditions speakers will use a CLD structure. Furthermore, I will show that this approach can also explain two well-known phenomena: the so-called “obligatory” doubling of personal pronouns and the effect of animacy on DOCLD. Finally, the factorial typology predicts the existence of six types of languages. I will begin with my view of the input to the OT GEN module.

⁶ In derivational frameworks, the clitic and the lexical NP are assumed to originate in a “big DP” with a single theta-role, and Case is assigned to the “split” parts of the DP through different mechanisms (Belletti 2005).

⁷ Belletti (2005:31) speculates on a possible answer to this question, tentatively suggesting that “the clitic ultimately [contributes] to Case licensing of the noun phrase in Topic position.” We’ll see that this suggestion cannot work for Rioplatense, since topicality is not a necessary condition for CLD (section 6).

⁸ Bresnan also suggests that the clitic may be voided of nonredundant information, and in that case it would not contribute a separate node to the c-structure. This may indeed be the case for IO clitics, and would explain why clitic doubling of indirect objects is not restricted to referential arguments and is quasi-obligatory. In either case, economy-of-expression (in the classical LFG sense) is satisfied (see section 2 below).

2 The input

Two caveats are necessary here. First, there is evidence that, unlike IOCLs, DOCLs in RSp cannot appear in CLD structures with non-referential arguments (in the sense of DuBois 1980: idioms, attributive uses of NPs, conflated objects, etc.). For this reason, in what follows I will consider the input to have referential arguments only. In the non-referential cases, CLD is *prima facie* not possible.⁹ Second, although recoverability and topic-anaphoricity arguably also affect expression of subjects (e.g. pro-drop), predicates (e.g. ellipsis), or argument selection in general, I will not be concerned with them here. I will state all constraints in terms of direct objects, although more general constraints could conceivably have the same effects, and at the same time model the ellipsis phenomena mentioned above.

I will take the input to be an LFG f-structure augmented by a Saliency List (*SaL*: Buchwald et al. 2002) that encodes the discourse status of referents and is updated every time an utterance is produced, thus effectively operationalizing topichood. The most salient referent at a given point occupies the first position in SaL (SaL_1) and the remaining referents are ordered from most to least salient also. But whereas Buchwald et al.'s SaL is simply a linearly ordered list of referents, the version I will use here is a partial ordering of discourse referents, represented by a full f-structure.¹⁰ It will become clear later that SaL's being a partial order is important to model cases where one or more referents are equally salient (i.e. occupy the same position in SaL) and can thus be in competition for topichood.

| | | |
|--------------------------------------|--------------------------------------|--------------------------------------|
| Most salient | > | Least salient |
| SaL_1 | SaL_2 | ... SaL_n |
| PRED ₁ < p ₁ > | PRED ₂ < p ₂ > | PRED _n < p _n > |
| GEN ₁ g ₁ | GEN ₂ g ₂ | GEN _n g _n |
| NUM n ₁ | NUM n ₂ | NUM n _n |
| ANIM +/- | ANIM +/- | ANIM +/- |
| DEF +/- | DEF +/- | DEF +/- |
| ... | ... | ... |

The fact that some marking of discourse status of the referents in the input is needed is consistent with remarks made by Sells (2003:93): "it seems that the INPUT must be a predicate-argument structure with all relevant (semantic) features of the arguments specified, plus an indication of target scope for any potentially scopal elements, and probably a similar indication of Information Structure status (e.g., Topic and Focus)." Also, Kuhn (2003:132) proposes that some representation of discourse context ("pragmatic clues", p. 63) is necessary in the input, especially for the "actual language production task from an underlying message."¹¹

A direct object (in fact, any argument) can bear three possible relations to SaL that are of relevance here:

- 1) the DO's f-structure is subsumed by the only f-structure in SaL_1 (OBJ= SaL_1 , which entails OBJ \in SaL_1)
- 2) the DO's f-structure is subsumed by an f-structure in SaL_1 but at least one other f-structure occupies that position (OBJ \neq SaL_1 but OBJ \in SaL_1)
- 3) the DO's f-structure is not subsumed in SaL_1 (OBJ \neq SaL_1 and OBJ \notin SaL_1)

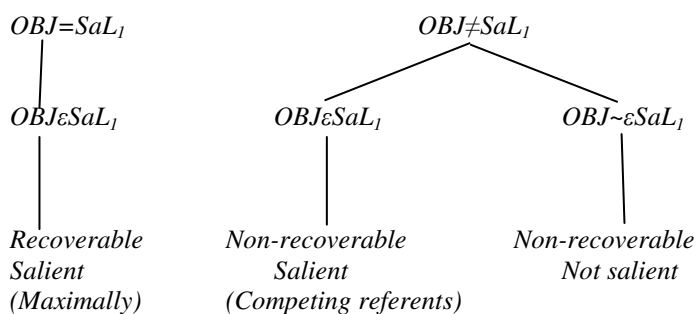
The relation OBJ \in SaL_1 (when the DO "belongs" to the set defined by the first position of SaL, that is, the DO "is in" SaL_1) is an operationalization of *salience*: any argument in that position is a salient argument. The relation OBJ= SaL_1 (when the DO is the only element of SaL_1) can be thought of as operationalizing *recoverability*. We will see that formally separating these two notions that are often confused pays off when it comes to modeling CLD.

⁹ In fact, no data from the corpus contradict this claim, although this alone does not constitute proof.

¹⁰ A straightforward extension of this proposal can be made where SaL also contains for each referent the predications that are true of it and have been introduced in discourse. This would model cases where a referent is salient but an NP is used to introduce a speaker's new point of view (see Estigarribia, forthcoming, section 3.4).

¹¹ Kuhn then goes on to say that this "actual production task ... is not of primary interest under the perspective of linguistic theory". I see absolutely no basis for such a rejection.

Figure 1: Three possible relations of OBJ to SaL



I will turn now to the relation between topic-anaphoricity, recoverability, economy-of-expression and argument expression, and how to model them in OT to derive the distribution of DOCLD. In the tableaux that follow, Clitic-only structures will be represented as CL, NP-only structures as NP, and Clitic-doubling structures as CLD to render constraint evaluation easier, but of course c-structure representations are what are being evaluated.¹²

3 The constraints

3.1 Expression of DOCLs and topic-anaphoricity

According to Bresnan (2001), reduced pronominals are crosslinguistically specialized for topic-anaphoric uses. That is, when a reduced form is available, it is the form that will express topics. That DO clitics are topic-anaphoric in Spanish is shown by the following examples (these examples closely parallel those offered by Bresnan and Mchombo (1987) to demonstrate topic-anaphoricity of reduced forms in Chicheŵa):

Discourse topics:

(4) (a)

La hiena se comió al león. Habíendose_{lo} comido, se fue a San Francisco.

The hyena ate the lion. Having-REFL-CL eaten went to San Francisco

The hyena ate the lion. Having eaten it (the lion), he went to S.F.

(b)

*La hiena se comió al león. *Habíendose comido a él, se fue a San Francisco.*

The hyena ate the lion. Having-REFL eaten A him went to San Francisco

The hyena ate the lion. Having eaten him, he went to S.F.

One could claim that the ungrammaticality of the example above is due to the fact that personal pronouns can never appear without a clitic in object position¹³, but inanimate pronouns (which can appear without a doubling clitic) show the same pattern as animate pronouns:

(5) (a)

La hiena se comió el arroz. Habíendose_{lo} comido, se fue a San Francisco.

The hyena ate the rice. Having-REFL-CL eaten went to San Francisco

The hyena ate the rice. Having eaten it (the rice), he went to S.F.

(b)

La hiena se comió el arroz. #Habíendose comido eso, se fue a San Francisco.

The hyena ate the rice. Having-REFL eaten it/that went to San Francisco

The hyena ate the rice. Having eaten it/that (something other than the rice), he went to S.F.

¹² The reader can find these representations in many grammars of Spanish.

¹³ This situation is commonly known as “obligatory” CLD of personal pronouns. See below section 5.

Dislocated topics:

(6) (a)

Este arroz, la hiena se lo comió.

This rice, the hyena REFL CL ate

This rice, the hyena ate it.

(b)

**Este arroz, la hiena se comió eso.*

This rice, the hyena REFL ate it/that

(Intended) This rice, the hyena ate it.

Resumptive relativization:

(7) (a) <Lig-116>

Generalmente viste casas viejas que las arreglan...

Generally you-saw houses old that CL they-repair

In general, you saw old houses that people repair...

(b)

**Generalmente viste casas viejas que esas/a ellas arreglan.*

Generally you-saw houses old that those/A them they-repair

(Intended) In general, you saw old houses that people repair...

I take the preceding examples to be evidence in favor of the following constraint¹⁴:

OBJ_εSaL₁(cl): Assign one violation if DO in SaL₁ (salient) but not expressed by a clitic.¹⁵

That is, salient referents prefer clitic expression. Arguably, this constraint belongs to a family of constraints that harmonically align salient referents with the hierarchy of pronominals:

OBJ_εSaL₁(∅) >> OBJ_εSaL₁(aff) >> OBJ_εSaL₁(cl) >> OBJ_εSaL₁(weak) >> OBJ_εSaL₁(free)

Since RSp does not have DO verbal inflection, null anaphora¹⁶ or weak pronouns, all candidates will violate those constraints, and we can leave them out of the picture for modeling DOCLD.

3.2 Expression of DO NPs and recoverability

Lexical NPs differ from pronominals in having semantic content that allows them to:

- Introduce new referents and also new predications about referents
- “Point” to already established referents that satisfy the NPs’ lexical description

Pronominals can also “point” to referents that are accessible enough if they match the pronominal’s features. But if the referent is not immediately recoverable in the context of utterance (possibly through competition with other referents), then a lexical description (hence, a lexical NP) is needed. Having an operational definition of *recoverability* is crucial here (see Barbosa et al. 1998, Kuhn 2003, and Pesetsky 1998 for some problems in defining recoverability in an OT framework, and Buchwald et al. 2002 for a bidirectional approach): a DO is *recoverable* if OBJ=SaL₁, that is, it is the only element of SaL₁. If this is not the case, a form with lexical content will be needed:

OBJ≠SaL₁(NP): Assign one violation for every OBJ not identical with SaL₁ and not expressed by an NP

¹⁴ In this paper, constraints are verbally stated in the format recommended by McCarthy (2002, p.40).

¹⁵ Again, bear in mind that I am restricting myself to DOs here, but such a constraint could be applied to any GF, indeed to any argument of the predicate.

¹⁶ But see Masullo (2003) for a claim that null objects do exist in Rioplatense (they are very common in Andean Spanish and Basque Spanish).

Pesetsky (1998) and Kuhn (2003) assume that a recoverability constraint should be inviolable. Bresnan (2001) states that because of learnability considerations the input must be fully recoverable from the output. Clearly, recoverability constraints have a special status in OT. It is not even clear whether it should be a (rerankable) constraint at all, unless a language could show extreme unfaithfulness *systematically*. I believe that the problem lies there: allowing reranking in the factorial typology will yield *systematically* unrecoverable languages. But we want to allow reranking to minimize stipulations about the internal structure of the universal space of constraints. Unfortunately, I cannot tackle this issue here. I will assume a form of recoverability that is undominated without discussing whether this is a violation of the Methodological Principle of OT that states that explanation should be done by constraint interaction. This assumption will obviously affect the factorial typology, but I think it is a rather reasonable one.¹⁷

3.3 Economy of expression

Economy-of-expression constraints have been proposed under several guises in the OT literature. The version I will adopt here is the following:

***STRUC:** Assign one violation for each c-structure node (Aissen 2003).

Let's see now how these three constraints are ranked in RSp and how they interact to derive the basic distribution of CL-, NP-, and CLD-structures.

4 Constraint ranking and evaluation

In Rioplatense, our three constraints are ranked as follows: **OBJ≠SaL₁(NP) >> OBJεSaL₁(cl)>> *STRUC**. Tableau 1 summarizes evaluation of the three principal types of DO: Recoverable hence salient (OBJ=SaL₁), salient but not recoverable (OBJεSaL₁ but OBJ≠SaL₁), and not salient hence not recoverable (OBJ~εSaL₁). The letters 'n' and 'k' stand for the number of nodes in each structure, the relevant fact being that a CLD-structure will always violate *STRUC more times than either CL- or NP-structures.

Tableau 1: General distribution of transitive clause types

| INPUT | CANDIDATES | (Recoverability) | (Topic-Anaphoricity) | (Economy) |
|---|------------|---------------------------|---------------------------|-------------|
| | | OBJ≠SaL ₁ (NP) | OBJεSaL ₁ (cl) | *STRUC |
| OBJ=SaL ₁ (hence OBJεSaL ₁) | CL ☺ | | | *n + 1 |
| | NP | | *! | *n + k |
| | CLD | | | *n + k + 1! |
| OBJ≠SaL ₁ OBJεSaL ₁ ¹ | CL | *! | | *n + 1 |
| | NP | | *! | *n + k |
| | CLD ☺ | | | *n + k + 1 |
| OBJ~εSaL ₁ (hence OBJ≠SaL ₁) | CL | *! | | *n + 1 |
| | NP ☺ | | | *n + k |
| | CLD | | | *n + k + 1! |

We see that the basic distribution of direct transitive structures is as follows:

- CL-structures are optimal if the DO is salient and recoverable (maximally salient)
- NP-structures are optimal if the DO is not salient
- CLD-structures are optimal if the DO is salient but not recoverable (competition with other referents)

These generalizations are borne out by naturally occurring data.

¹⁷ A different approach, involving bidirectional optimization, is possible, but I will not explore it here. See Buchwald et al. 2002 and Kuhn 2003.

a) Maximally salient DOs

A clitic suffices (CL-structure) if the DO is the most salient referent

(8) <Beto-46>

Pero, uno se va con el problema. No lo dejás acá.

but one leaves with the problem no CL3SgM you.leave here

But you take the problem with you [when you go on vacation]. You don't leave it here.

In the above example, *el problema* 'the problem' is the only referential argument in the previous sentence (*uno* being a generic 'you'), which introduces it as (maximally) salient in the context.¹⁸

b) Non-salient referents

A NP suffices (NP-structure) if the DO is not a salient referent

(9) <Beto-42>

El Negro Moreira... dejó un par de cortados. Aldo le pidió un mate.

"Black" Moreira left a couple of espressos.with.milk. Aldo IOCL ordered a mate

"Black" Moreira left a couple of espressos [on the table]. Aldo ordered a mate [from him].

c) Salient but non-recoverable referents

CLD is the optimal candidate when the DO is salient but other equally salient referents exist.

(10) <Lig-115>

es mucho más fácil montar un jardín que-que un hospital o un consultorio-

is much more easy mount a garden than a hospital or a consultancy

It is much easier to start a kindergarten than a hospital or a private practice –

el jardín lo podés hacer poner con-con nada casi

the garden CL you-can make put with nothing almost

you can set up a kindergarten with almost no investment.

In this case, three salient referents are in competition, *un jardín* 'a kindergarten', *un hospital* 'a hospital', and *un consultorio* 'a private practice'. The Clitic Left-Dislocated structure (CLLD: Cinque 1900, Escobar 1997), a case of CLD, is used to pick out the intended referent.

We can see why it is crucial to clearly differentiate recoverability and salience as two independent notions in this model. Every salient argument will require a reduced form of expression, in this case a DO clitic, while every non-recoverable argument will require lexical content, that is, a NP or at least a free pronoun.¹⁹

Although the basic pattern is captured here, the actual distribution of DOCLD in discourse is more complicated. Morphological and semantic properties of the DO have an impact on whether a DOCLD-structure is optimal or dispreferred. I will now show that an independently grounded extension of this model successfully captures two well-known phenomena: "obligatory" doubling of personal pronouns and a seeming animacy restriction on DOCLD.

5 Personal pronouns and markedness

The literature on CLD considers it obligatory with personal pronouns because of the following contrasts:

¹⁸ Also the general topic of the exchange is how to avoid a certain problem one of the participants has, which contributes to the salient status of the referent 'the problem'. I am assuming here that some mechanism for calculation of salience is available, for instance, Centering Theory (Grosz, Joshi, and Weinstein 1995).

¹⁹ Remember that free pronouns are "both distributionally and prosodically indistinguishable from a full NP/DP" (Vincent 2001). The distribution of DOCLD with free pronouns will be derived by other constraints in the next section.

(11) (a) <PerInv-16>

*Sí, dejáme a mí que yo los conozco/ *Sí, dejá a mí que yo los conozco*
Yes leave-CL A me that I CL know Yes leave A me that I CL know
Leave it to me because I know them.

(12) (b) <Lig-124>

*Yo *(te) pregunto a vos*
I *(CL) ask A you
I am asking you.

(13) (c) <Lig-82>

*Yo *(la) veo a ella vender es bárbara yo admirada*
I *CL see A her sell is great I admiring
I see her sell... she's great, I'm all admiration.

(14) (d) <Lig-148>

**(me) llamó a mí Cápura*
CL1Sg he.called A me Cápura
Cápura called [me]_{FOC}.

However, CLD-structures are not obligatory, since CL-structures are allowed:

(15) <Beto-50>

Ya te estaba extrañando
Already CL2Sg was missing
I missed you already.

Moreover, inanimate pronouns can appear without an accompanying clitic²⁰:

(16) (a) <Lig-81>

no soy yo en ese momento cuando estoy haciendo eso
no am I I that moment when I-am doing it/that
I'm not myself at the time when I am doing that.

(16) (b) <Lig-100>

el producto de la ignorancia es lo que favorece eso
the product of the ignorance is what favors it/that
It is the product of ignorance that favors that.

What needs to be accounted for is precisely the prohibition of NP-structures, when the NP is a free personal pronoun (that is, free and animate), not obligatoriness of CLD. This has been modeled in classical LFG as a morphological blocking effect (Andrews 1990). Now, the fact that only personal pronouns (which I interpret as [+anim, +pro]) present a morphological blocking effect indicates that there is an interaction with animacy that needs to be captured. What is the difference RSp makes in the treatment of animate and inanimate DOs?

In all varieties of Spanish the case of inanimate objects is indicated by the presence or absence of the preposition/marker *a*:

(17) (constructed example)

Le sacaste un botón a tu camisa
You took.out a button DAT your shirt.
You took a button off your shirt.

²⁰ These inanimate pronouns are homophonous with demonstratives. They could be argued to be demonstratives but they fill the inanimate gap in the paradigm of pronouns, and therefore, I take them to be inanimate pronouns, in accordance with prescriptive grammars of Spanish.

Note that *un botón* and *tu camisa* are both inanimate, but while the former is a DO and has no marking (except post-verbal position), the latter is an IO and is marked with *a*. However, Spanish requires all **referential animate** DOs to be marked with *a* too, and hence, case is not differentially marked for animates²¹ (example repeated from above):

(1) (b) < PerInv-14>
De inmediato miró fijamente a sus compañeros
 Immediately looked.at-he fixedly A his-PL partners
He immediately looked fixedly at his partners.

Since the marker *a* is indistinguishable from the IO marker, in (1b) case is not unambiguously marked (locally). Avoidance of ambiguity was argued, for instance, by Donohue (1999) for Fore and is usually assumed in the processing literature to play a role in language production.²² Our constraint to avoid local ambiguity is:

***?CASE:** Assign one violation for every argument ambiguously marked for case.

Importantly, this constraint has an effect only in the case of animate DOs (remember that inanimate DOs are unambiguously marked by zero marking). But what is crucial here is that DO clitics, being distinct from IO clitics, can disambiguate marking in this case (García 1975). Therefore, structures with object clitics never violate ***?CASE**.

As Aissen (2003) notes, DOs are more marked the more definite they are, and therefore the following markedness scale (in the form of avoid constraints) applies:

***OBJ/PRO >> *OBJ/PN >> *OBJ/DEF >> *OBJ/SPEC >> *OBJ/NSPEC** (Aissen 2003, p. 445)

We can conjoin the ***?CASE** requirement with the definiteness hierarchy. The constraints that are the product of this conjunction are ranked above each separate constraint (local conjunction in the DO domain; see Smolensky 1995, McCarthy 2002 for discussion.):

***OBJ/PRO&*?CASE >> *OBJ/PN&*?CASE >> *OBJ/DEF NP&*?CASE >> *OBJ/SPEC NP&*?CASE >> *OBJ/NONSPEC NP&*?CASE**

Note the parallels with Differential Object Marking (Aissen 2003). To recapitulate, unambiguous case marking is satisfied by either an inanimate DO or an animate DO that is copresent with a coreferential clitic. For ease of presentation, let's collapse the non-pronominals into the category NP. The ranking for RSp is:

OBJ≠SaL₁(NP) >> OBJ=SaL₁(cl), *OBJ/PRO&*?CASE >> *STRUC, *OBJ/NP&*?CASE

where constraints within the same stratum are separated by commas and allowed to rerank. Tableau 2 below shows that the distribution of CLD/CL-structures and prohibition of NP-structures with animate pronouns follows without direct stipulation of a constraint that requires doubling of free personal pronouns.

The cases where the DO is salient (OBJεSaL₁) behave exactly as with a non personal pronoun argument, yielding a CL- or a CLD-structure. The evaluation differs for non-salient DOs: whereas in the general case (see tableau 1 above) a non-salient DO would enforce a NP-structure, in the case of personal pronouns such a situation requires a CLD-structure to avoid the local ambiguity of having marking of the DO by *a*. That is why NP-structures are not found with personal pronouns in Spanish. The crucial factor is that the constraint towards unambiguous case manifestation is satisfied by two independent means: it is satisfied by DO clitics, but it is also satisfied by any inanimate argument (which do not take *a* when DOs and take *a* when IOs). Hence, if we have an inanimate pronoun that is not salient, we revert to the general case where NP-structures are optimal (examples repeated from above):

²¹ Animate DOs in some cases may not be marked (nonspecific ones, for instance).

²² However, see Wasow (2002) for a sceptical view.

(16) (a) <Lig-81>
no soy yo en ese momento cuando estoy haciendo eso
 no am I I that moment when I-am doing it/that
I'm not myself at the time when I am doing that.

(16) (b) <Lig-100>
el producto de la ignorancia es lo que favorece eso
 the product of the ignorance is what favors it/that
It is the product of ignorance that favors that.

To summarize, we haven't explicitly built a constraint that requires doubling personal pronouns, but we have derived its effects from a constraint on case marking and from cross-linguistic, independently grounded markedness constraints.

Tableau 2: Distribution of transitive clauses with free personal pronouns as DOs

| INPUT (all +Anim, +Pro) | CAND | (Recoverability) | (TopAnaph) | (AmbPro) | (AmbNP) | (Economy) |
|---|-------|------------------------------|------------------------------|---------------|----------------|-------------------------|
| | | OBJ≠SaL ₁ (NP) | OBJεSaL ₁ (cl) | *O/P & *?C | *O/NP & *?C | *STRUC |
| OBJ=SaL ₁ | CL ☹ | | | | | *n + 1 |
| | NP | | *! | * | | *n + k |
| | CLD | | | | | *n + k + 1 _! |
| OBJ≠SaL ₁ OBJεSaL ₁ | CL | *! | | | | *n + 1 |
| | NP | | *! | * | | *n + k |
| | CLD ☹ | | | | | *n + k + 1 |
| OBJ≠SaL ₁ OBJ~εSaL ₁ | CL | *! | | | | *n + 1 |
| | NP | | | *! | | *n + k |
| | CLD ☹ | | | | | *n + k + 1 |

6 Animacy effect

Peninsular Spanish disallows DOCLD with inanimate DOs. Although this is not the case for RSp (contra Andrews 1990, Jaeggli 1986, Roberge 1990; see example (3a) above), some researchers claim that animates are doubled more frequently, or that inanimate DOCLD is less general in this dialect (Barrenechea and Orecchia 1977, Gutiérrez-Rexach 2000, Suñer 1988)²³, therefore acknowledging the presence of an *animacy effect*. However, the precise nature of this effect has never been stated. The OT model I propose derives a precise generalization: **animates, but not inanimates, can be doubled in contexts where the DO is not salient** (see Estigarribia, forthcoming).

[CONTEXT: B asks A for a prepaid phone card. "Marta" is B's girlfriend, not the topic of the exchange.]

(18) <Beto-50>

- A: ¿La vas a llamar a Marta?
 CL3SgF you.go to call A Marta
Are you going to call [Marta]_{FOC}?

- B: No querido... La voy a llamar a estamina de la que
 No dear CL3SgF I.go to call A this girl of CL that

hablábamos anoche

we.talked last.night

No my dear, I'm gonna call [this girl we were talking about last night]_{FOC}.

²³ Colantoni (2002) is the only author I know of that claims that inanimate DOCLD is more frequent than animate DOCLD.

The predictive strength and empirical adequacy of the model is demonstrated by the fact that the constraints already proposed account for this effect without further stipulation:

Tableau 3: Animacy effect on CLD

| INPUT | CANDIDATES | (Recoverability) | (TopAnaph) | (AmbPro) | (AmbNP) | (Economy) |
|---|--------------|---------------------------|---------------------------|--------------|---------------|----------------------|
| | | OBJ≠SaL ₁ (NP) | OBJεSaL ₁ (cl) | *O/P & ?C | *O/NP & ?C | *STRUC |
| OBJ=SaL ₁ +Anim | CL ☹ | | | | | * ⁿ⁺¹ |
| | NP | | *! | | * | * ^{n+k} |
| | CLD | | | | | * ^{n+k+1} ! |
| OBJ≠SaL ₁ OBJεSaL ₁ +Anim | CL | *! | | | | * ⁿ⁺¹ |
| | NP | | *! | | * | * ^{n+k} |
| | CLD ☹ | | | | | * ^{n+k+1} ! |
| OBJ≠SaL₁ OBJ~εSaL₁ +Anim | CL | *! | | | | * ⁿ⁺¹ |
| | NP | | | | *! | * ^{n+k} |
| | CLD ☹ | | | | | * ^{n+k+1} ! |
| OBJ=SaL ₁ -Anim | CL ☹ | | | | | * ⁿ⁺¹ |
| | NP | | *! | | | * ^{n+k} |
| | CLD | | | | | * ^{n+k+1} ! |
| OBJ≠SaL ₁ OBJεSaL ₁ -Anim | CL | *! | | | | * ⁿ⁺¹ |
| | NP | | *! | | | * ^{n+k} |
| | CLD ☹ | | | | | * ^{n+k+1} ! |
| OBJ≠SaL₁ OBJ~εSaL₁ -Anim | CL | *! | | | | * ⁿ⁺¹ |
| | NP ☹ | | | | | * ^{n+k} |
| | CLD | | | | | * ^{n+k+1} ! |

The boldfaced cases are those in which the DO is not salient, and we see that animates and inanimates behave differently, the former preferring CLD-structures and the latter, NP-structures. Moreover, DOCLD with animates is optional with non-salient referents (as shown by the example below), and this is captured in the model by reranking constraints within the lowest stratum (see tableau 4):

(19) <Beto-48>
 ¿Te enganaste a la Sonia en lo del Pitu?²⁴
 REFL you.hooked A the Sonia at.Pitu's
You picked up Sonia at Pitu's?

Tableau 4: Optionality of CLD with animates

| INPUT | CAND | (Recoverability) | (TopAnaph) | (AmbPro) | (AmbNP) | (Economy) |
|--|----------|---------------------------|---------------------------|--------------|---------------|----------------------|
| | | OBJ≠SaL ₁ (NP) | OBJεSaL ₁ (cl) | *O/P & ?C | *O/NP & ?C | *STRUC |
| OBJ≠SaL ₁ OBJ~εSaL ₁ +Anim | CL | *! | | | | * ⁿ⁺¹ |
| | NP | | | | *! | * ^{n+k} |
| | CLD ☹ | | | | | * ^{n+k+1} ! |

²⁴ This is an all-focus sentence.

| | | (Recoverability) | (TopAnaph) | (AmbPro) | (Economy) | (AmbNP) |
|-----------------------|------|---------------------------|---------------------------|--------------|----------------------|---------------|
| INPUT | CAND | OBJ≠SaL ₁ (NP) | OBJεSaL ₁ (cl) | *O/P & ?C | *STRUC | *O/NP & ?C |
| OBJ≠SaL ₁ | CL | *! | | | * ⁿ⁺¹ | |
| OBJ~εSaL ₁ | NP ⊙ | | | | * ^{n+k} | *! |
| +Anim | CLD | | | | * ^{n+k+1} ! | |

7 Final ranking and synthesis of results

I checked the rankings obtained with the Gradual Learning Algorithm built into OTSoft (Hayes, Tesar and Zuraw 2003).²⁵ I assumed that the recoverability constraint was undominated (see discussion in 3.2) and I also assumed the cross-linguistic dominance relations derived from the conjunction of the overt case requirement with the object markedness hierarchy. I modeled free variation for the animate non-recoverable inputs (OBJ≠SaL₁, OBJ~εSaL₁, +Anim) by assigning winning frequencies of 0.5 to both the NP- and CLD-structure candidates. The stochastic grammar found was:

OBJ≠SaL₁(NP) = 142
 *OBJ/PRO & ?CASE = 122
 OBJεSaL₁(cl) = 108
 OBJ/NP&?CASE = 96.27
 *STRUC = 95.73

The grammar correctly predicted all winning candidates and modeled free variation in the animate non-recoverable case as output frequencies of 0.42 for the NP-structure and 0.581 for the CLD-structure, very close to the theoretical 0.5 frequencies. The average error per candidate was less than 0.3%. In what follows, I will use the ranking in stratal form, and abbreviated constraint names:

OBJ≠SaL₁(NP) >> *OBJ/PRO & ?CASE >> OBJεSaL₁(cl) >> *OBJ/NP&*?CASE, *STRUC
 that is
 Rec >> AmbPro >> TopAnaph >> AmbNP, ECON

I will also use a binary feature vector representation for the inputs for readability: [±sal, ±rec, ±anim, ±pro].

7.1 Optimization of the different cases

A) Salient recoverable DOs (maximally salient DOs = OBJ=SaL₁) : Clitic structure is optimal.

| [+sal, +rec, +anim, +pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|----------------------|
| CL ⊙ | | | | | * ⁿ⁺¹ |
| NP | | *! | * | | * ^{n+k} |
| CLD | | | | | * ^{n+k+1} ! |

| [+sal, +rec, +anim, -pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|----------------------|
| CL ⊙ | | | | | * ⁿ⁺¹ |
| NP | | | *! | * | * ^{n+k} |
| CLD | | | | | * ^{n+k+1} ! |

²⁵ Initial rankings: all constraints = 100; 50000 learning trials; initial plasticity = 2; final plasticity = 0.002; a priori rankings differ by 20; grammar tested 2000 times.

| [+sal, +rec, -anim, +pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|-------------|
| CL ☹ | | | | | *n + 1 |
| NP | | | *! | | *n + k |
| CLD | | | | | *n + k + 1! |

| [+sal, +rec, -anim, -pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|-------------|
| CL ☹ | | | | | *n + 1 |
| NP | | | *! | | *n + k |
| CLD | | | | | *n + k + 1! |

This result is optimal for every possible ranking of the constraints: the construction is optimal in every possible language.

B) Salient non-recoverable DOs (OBJ≠SaL₁, OBJεSaL₁): CLD-structure is optimal.

| [+sal, -rec, +anim, +pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|------------|
| CL | *! | | | | *n + 1 |
| NP | | *! | * | | *n + k |
| CLD ☹ | | | | | *n + k + 1 |

| [+sal, -rec, +anim, -pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|------------|
| CL | *! | | | | *n + 1 |
| NP | | | *! | * | *n + k |
| CLD ☹ | | | | | *n + k + 1 |

| [+sal, -rec, -anim, +pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|------------|
| CL | *! | | | | *n + 1 |
| NP | | | *! | | *n + k |
| CLD ☹ | | | | | *n + k + 1 |

| [+sal, -rec, -anim, -pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|------------|
| CL | *! | | | | *n + 1 |
| NP | | | *! | | *n + k |
| CLD ☹ | | | | | *n + k + 1 |

Reranking in this case may lead to other optimal candidates in other languages (see section 8 on Factorial Typology).

C) Non-salient, non-recoverable DOs (OBJ≠SaL₁ OBJ~εSaL₁):

a) CLD is optimal with personal pronouns.

| [-sal, -rec, +anim, +pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|------------|
| CL | *! | | | | *n + 1 |
| NP | | *! | | | *n + k |
| CLD ☹ | | | | | *n + k + 1 |

Reranking may lead to a different outcome in different languages.

b) CLD is optional with non-pronominal animates (NP construction is also possible).

| [-sal, -rec, +anim, -pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|--------------------|
| CL | *! | | | | * ⁿ⁺¹ |
| NP ☹ | | | | * | * ^{n+k} |
| CLD ☹ | | | | | * ^{n+k+1} |

Reranking may lead to a different outcome in different languages.

c) NP construction is optimal with inanimates.

| [-sal, -rec, -anim, +pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|----------------------|
| CL | *! | | | | * ⁿ⁺¹ |
| NP ☹ | | | | | * ^{n+k} |
| CLD | | | | | * ^{n+k+1} ! |

| [-sal, -rec, -anim, -pro] | REC | AmbPro | TopAnaph | AmbNP | ECON |
|------------------------------|-----|--------|----------|-------|----------------------|
| CL | *! | | | | * ⁿ⁺¹ |
| NP ☹ | | | | | * ^{n+k} |
| CLD | | | | | * ^{n+k+1} ! |

This configuration obtains for every possible language.

As we see in the analysis above, the cases where either CL or NP structures (undoubled structures) are optimal cannot yield a different result under reranking. Note that it may seem that this is a strange situation, at first glance. But the constraint enforcing reduced expression of topics is formulated exclusively in terms of clitics. If we included the whole hierarchy of reduced pronominal forms, the winning CL candidates would violate higher ranked constraints requesting zero and affixal inflection, and lower ranked ones requesting weak and independent pronominals. Therefore, no candidate is completely unmarked and faithful (an undesirable situation from the theoretical point of view).

8 Factorial typology

8.1 Language Space and Variation Space

With 5 constraints, the number of logically possible grammars is 120. Using OTSoft, there were 6 different output patterns, represented in Table 1 below.

Therefore, the factorial typology predicts that:

- All recoverable DOs are expressed by CL structures in every possible language.
- All the inanimate non-salient DOs are expressed by NP structures in every possible language.
- The differing outputs correspond to salient non-recoverable DOs and to non-salient animate DOs.

Note that the output correctly predicts that the implicational hierarchies of definiteness and animacy (and possibly topicality, if operationalized through salience) will be respected cross-linguistically:

- If [-pro] allows CLD, then, *ceteris paribus* [+pro] allows CLD;
- If [-anim] allows CLD, then, *ceteris paribus* [+anim] allows CLD;
- If [-sal] allows CLD, then, *ceteris paribus* [+sal] allows CLD.

Table 1: Factorial typology for direct object clitic doubling

| Language Type | | | | I Rioplatense | II Rioplatense (optionally) | III Bulgarian | IV Peninsular Spanish | V Kichaga | VI French, Italian |
|-----------------|-------------|-------|------|---------------|-----------------------------|---------------|-----------------------|-----------|--------------------|
| | | | | | | | | | |
| Recoverable | Salient | +Anim | +Pro | CL | | | | | |
| | | | -Pro | | | | | | |
| | | -Anim | +Pro | | | | | | |
| | | | -Pro | | | | | | |
| Non-Recoverable | Salient | +Anim | +Pro | CLD | NP | NP | NP | NP | |
| | | | -Pro | | | | | | |
| | | -Anim | +Pro | | | | | | |
| | | | -Pro | | | | | | |
| | Non-salient | +Anim | +Pro | | | | | | |
| | | | -Pro | | | | | | |
| | | -Anim | +Pro | | | | | | |
| | | | -Pro | | | | | | |

8. 2 Possible rankings and possible languages

Language 1

REC >> TopAnaph, AmbPro >> ECON >> AmbNP

Language 2

REC >> TopAnaph, AmbPro >> AmbNP >> ECON

Rioplatense Spanish behaves optionally like Language 1 or Language 2, since CLD/NP are both possible for [-sal, -rec, +anim, -pro]. In that case, given that variation here is modeled as reranking within a stratum, we get the ranking given in section 7 above, and repeated here for convenience:

REC >> TopAnaph, AmbPro >> AmbNP, ECON

Note that, since the economy-of-expression constraint *STRUC is ranked very low, this language has a very widespread CLD pattern.

Language 3

REC >> TopAnaph >> ECON >> AmbPro >> AmbNP

Language 3 is a language that expresses non-salient DOs as NP-structures, salient recoverable DOs as CL, and salient non-recoverable ones as CLD. Bulgarian (Jaeger and Gerassimova 2002) is possibly such a language. Since *STRUC is ranked above the unambiguous marking markedness constraints, in such a language unambiguous case marking constraints are unnecessary, and salience/topicality is the major factor that drives CLD.

Language 4

REC >> AmbPro >> AmbNP >> ECON >> TopAnaph

Language 4 only has CLD with animates. This is what is usually claimed about Peninsular Spanish. In such a language salience/topicality would play no role (“emergence of the unmarked” effects aside) in clitic expression. Unambiguous marking of cross-linguistically marked DOs is the most relevant factor here, and so the pattern of CLD reflects animacy very accurately.

Language 5

REC >> AmbPro >> ECON >> TopAnaph, AmbNP

Language 5 has CLD only for non-recoverable personal pronouns (both salient and not). Kichaga as reported by Bresnan and Moshi (1990) could be such a language. AmbNP plays no role in this case.

Language 6

REC >> ECON >> TopAnaph, AmbPro >> AmbNP

Language 6 has no CLD. Economy-of-expression is ranked very high, above all the other markedness constraints, dominated only by recoverability. In this language, unambiguous marking of objects does not play a role either.

Since it is usually claimed that French and Italian do not have CLD, this could be the ranking underlying them. However, French and Italian do not fit this picture since they do have CLLD,²⁶ that is Clitic Left-Dislocated structures, which I analyze as a particular case of CLD (Belletti 2005, Estigarribia forthcoming). However, they do not have CLD of in-situ direct objects and this suggests that other constraints may be needed for the model to make more accurate cross-linguistic predictions.

This typology allows us to make interesting observations about these possible languages. In all of them, non-recoverability of information is a major divide between possible expressions of arguments (but this result was somewhat built into the analysis; see section 9, Conclusions). A language of type 6 only differentiates recoverable and non-recoverable arguments. The other five types make additional distinctions.

Only in one of them (Language 3), the presence of a clitic unambiguously encodes salience/topicality, although this is a prevalent hypothesis in the literature for Spanish CLD. In fact, only in two language types (3 and 6), the presence of a clitic encodes one only factor: recoverability in 6, salience/topicality in 3. In the remaining 4 types, the presence of a clitic will be determined by a more complex combination of animacy, pronominality, and discourse structure.

9 Conclusions

This paper showed the route towards a complete cross-linguistic OT analysis of Clitic Doubling, based on data from Rioplatense Spanish. The dimensions of animacy, definiteness and discourse structure were shown to interact to determine the possible patterns of CLD in the world's languages. The proposal advanced here was related to functional and typological results on pronominal expression and markedness results with respect to Differential Object Marking. This, I believe, gives this contribution a solid grounding in ongoing OT research.

Also, DO clitics have been shown not to be associated with one particular function, but rather as providing means for expressing independent requirements: as salience/topicality markers or as overt markers of case in surface-ambiguous predicate structures. The crucial point on which the analysis hinges is that recoverability and salience are factors that can be teased apart from one another. Even though *recoverability* as defined here entails *salience*, the converse is not true, and DOCLD emerges precisely when *salience* obtains but *recoverability* doesn't. This situation has not been fully exploited in previous analyses.

I chose not to deal directly with structural considerations, and therefore the role of structural descriptions (like LFG) was minimized. Far from being a weakness of the analysis, I consider this to be one of its strengths. The results exposed are maximally independent from theory-internal reasoning.

Several problems with current formalizations of OT syntax were spotted. The most salient is the issue of how to deal with ellipsis and recoverability in such an approach. Methodologically speaking, it would be desirable that such effects were a result of the interaction of violable constraints. However, so far such a solution has proven elusive, and some sort of inviolable recoverability principle still needs to be assumed. Bidirectional optimization may offer a solution to this quandary (Buchwald et al. 2002, Kuhn 2003), but this is a field that needs to be explored in more depth.

²⁶ My thanks to Nigel Vincent for this observation.

Insofar as the input is concerned, the addition of a Saliency List with a range of possible positions (degrees of saliency) may be useful to model also secondary topic phenomena, of the sort discussed by Dalrymple and Nikolaeva (2005).

A potential quirk in the analysis is the received knowledge that CLD is a cline towards grammatical agreement. If this is the case, then it should be possible to get CLD across the whole spectrum of possibilities. But, as we have seen, inanimate non-salient DOs are expressed by NP-only structures in every possible language. One way of explaining this and making the analysis more powerful is to include the markedness scales (definiteness and animacy) in full. That way, economy of expression would be allowed to interact in finer-grained ways with requirements for overt case marking.

APPENDIX

Direct Object Clitics in Rioplatense Spanish

| | Masculine | Feminine |
|-------------|-----------|----------|
| 1 Sg | me | |
| 2 Sg | te | |
| 3 Sg | lo | la |
| 1 Pl | nos | |
| 2 Pl | los | las |
| 3 Pl | los | las |

Corpus references

In the examples, the texts are referred to by the abbreviations between angle brackets, followed by the page number in the original.

<PerInv> Fontanarrosa, Roberto. 1995. “Periodismo Investigativo”, in *La mesa de los galanes y otros cuentos*. Buenos Aires: Ediciones de la flor. 7-16.

<Beto> Fontanarrosa, Roberto. 1995. “Beto”, in *La mesa de los galanes y otros cuentos*. Buenos Aires: Ediciones de la flor. 37-50.

[*La mesa de los galanes* is a book of comic short stories reproducing very accurately the vernacular dialogues of the rioplatense area where RSp is spoken].

<Lig> Ligatto, Dolores. 1996. *Matériau pour l'étude de l'espagnol parlé : la variante argentine*. Limoges: Pulim. [Transcription and translation into French of several corpora of naturally occurring conversation and television interviews from the 1980s, all of them representing RSp]

<Quino> Quino (Joaquín Salvador Lavado). 1997. *Toda Mafalda*. Buenos Aires: Ediciones de la flor. [Comic strips of the 60s, 70s and 80s]

References

- Aissen, Judith. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* 21, 435-83.
- Andrews, Avery. 1990. Unification and Morphological Blocking. *Natural Language and Linguistic Theory* 8: 507-557.
- Barbosa, Pilar, Danny Fox, Paul Hagstrom, Martha McGinnis and David Pesetsky (eds.). 1998. Introduction. In *Is the Best Good Enough? Optimality and Competition in Syntax*. Cambridge, MA – London: MIT.
- Barrenechea, Ana María and Teresa Orecchia. 1977. La duplicación de objetos directos e indirectos en el español hablado en Buenos Aires. *Estudios sobre el español hablado en las principales ciudades de América*, ed. by Lope Blanch, Juan M., 351-381. México: Universidad Autónoma de México.
- Belletti, Adriana. 2005. Extended doubling and the VP periphery. *Probus* 17. 1-35.
- Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

- Bresnan, Joan. 2001. The Emergence of the Unmarked Pronoun. In *Optimality-theoretic Syntax*, edited by Geraldine Legendre, Jane Grimshaw, and Sten Vikner, 113 - 142. Cambridge, MA: The MIT Press.
- Bresnan, Joan and Sam A. Mchombo. 1987. Topic, Pronoun and Agreement in Chicheŵa. *Language* 63. 741-782.
- Bresnan, Joan and Lioba Moshi. 1990. Object asymmetries in comparative Bantu syntax. *Linguistic Inquiry* 21. 147-185.
- Buchwald, Adam, Oren Schwartz, Amanda Seidl, and Paul Smolensky. 2002. Recoverability Optimality Theory: Discourse Anaphora and Bidirectional Optimization. In Johan Bos, Mary Ellen Foster and Colin Matheson (eds.): *Proceedings of the sixth workshop on the semantics and pragmatics of dialogue (EDILOG 2002)*. 37-44.
- Cinque, Guglielmo. 1990. *Types of A'-Dependencies*. Cambridge: MIT Press.
- Colantoni, Laura. 2002. Clitic Doubling, Null Objects and Clitic Climbing in the Spanish of Corrientes, in Gutierrez-Rexach, J. (ed.). *From Words to Discourse: Trends in Spanish Semantics and Pragmatics*. Oxford: Elsevier.
- Dalrymple, Mary, and Irina Nikolaeva. 2005. Agreement and discourse function. Paper presented at LFG05, Universitetet i Bergen.
- Donohue, Cathryn. 1999. Optimizing Fore case and word order. Stanford University, ms.
- DuBois, John W. 1980. Beyond Definiteness: The Trace of Identity in Discourse. In Wallace L. Chafe (ed.), *The Pear Stories. Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Series Advances in Discourse Processes, Vol III. Norwood, NJ: Ablex.
- Escobar, Linda. 1997. Clitic left dislocation and other relatives. In Elena Anagnostopoulou, Henk van Riemsdijk and Frans Zwarts, *Materials on Left Dislocation*. Amsterdam: John Benjamins. 233-273.
- Estigarribia, Bruno. Forthcoming. Why clitic doubling? A functional analysis for Rioplatense Spanish. *Selected proceedings of the 8th Hispanic Linguistics Symposium and 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*. Somerville: Cascadilla Press.
- García, Erica. 1975. *The Role of Theory in Linguistic Analysis: The Spanish Pronoun System*. Amsterdam: North Holland.
- Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics* 21. 203-225.
- Gutiérrez-Rexach, Javier. 2000. The Formal Semantics of Clitic Doubling. *Journal of Semantics* 16. 315-380
- Haiman, John. 1985. *Natural Syntax: Iconicity and Erosion*. Cambridge: Cambridge University Press.
- Hayes, Bruce, Bruce Tesar, and Kie Zuraw (2003) "OTSoft 2.1," software package, <http://www.linguistics.ucla.edu/people/hayes/otsoft/>.
- Jaeger, T. Florian and Veronica A. Gerassimova. 2002. Bulgarian Word Order and the Role of the Direct Object Clitic in LFG. In *Proceedings of the LFG02 Conference*, Stanford, CA: CSLI.
- Jaeggli, Osvaldo A. 1986. Three Issues in the Theory of Clitics: Case, Doubled NPs, and Extraction. In Hagit Borer (ed.), *Syntax and Semantics 19: The Syntax of Pronominal Clitics*. New York: Academic Press. 15-42.
- Kuhn, Jonas. 2003. *Optimality-Theoretic Syntax – A Declarative Approach*. Stanford, CA: CSLI.
- Masullo, Pascual. 2003. Cliticless definite object drop in River Plate Spanish, paper presented at LSRL XXXIII, Indiana University.
- McCarthy, John J. 2002. *A Thematic Guide to Optimality Theory*. Cambridge, UK: Cambridge University Press.
- Pesetsky, David. 1998. In Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis and David Pesetsky (eds.), *Is the Best Good Enough? Optimality and Competition in Syntax*. Cambridge, MA – London: MIT.
- Roberge, Yves. 1990. *The Syntactic Recoverability of Null Arguments*. Kingston – Montreal: McGill-Queen's University Press.
- Sells, Peter. The INPUT and Faithfulness in OT Syntax. 2003. In Jennifer Spenader, Anders Eriksson and Oesten Dahl (ed.) *Proceedings of the workshop on Variation within Optimality Theory*, University of Stockholm, 92-101.
- Smolensky, Paul. 1995. On the Internal Structure of the Constraint Component Con of UG. Manuscript available at the Rutgers Optimality Archive <http://roa.rutgers.edu/index.php3>.
- Suñer, Margarita. 1988. The role of agreement in clitic-doubled constructions. *Natural Language and Linguistic Theory* 6: 391-434.

- Vincent, Nigel. 2001. Competition and correspondence in syntactic change: null arguments in Latin and Romance. In *Diachronic Syntax*, ed. by Susan Pintzuk, George Tsoulas, and Anthony Warner. Oxford: Oxford University Press. 25-50.
- Wasow, Tom. 2002. *Postverbal Behavior*. Stanford: CSLI.
- Zwicky, Arnold 1977. *On clitics*. Bloomington: Indiana University Linguistics Club.

OPEN ARGUMENT FUNCTIONS

Yehuda N. Falk
The Hebrew University of Jerusalem

Proceedings of the LFG05 Conference
University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications
<http://csli-publications.stanford.edu/>

Abstract

This paper proposes a new approach to open argument functions in LFG. As opposed to both the earliest approach, under which open argument functions were individuated by category (VCOMP, NCOMP, PCOMP, ACOMP), and the conventional approach, under which there is a single category-neutral open argument function (XCOMP), this paper proposes that there are three open argument functions, XOBJ_θ, XOBL_θ, and XCOMP, each of which is canonically associated with one or more c-structure categories, much like their closed counterparts.

1. The Problem

The LFG theory of grammatical functions has, since the outset, made a distinction between open and closed functions, both in the realm of argument functions and of non-argument (adjunct) functions. This paper¹ is an attempt to reexamine the inventory of open argument functions in light of problems with the conventional theory.

Open argument functions are the ones used in functional control constructions: raising and some cases of complement equi. In the earliest work in LFG (such as Kaplan & Bresnan 1982, Bresnan 1982a, Grimshaw 1982, and Andrews 1982), open argument functions were differentiated by category: nominally-headed open arguments were assigned the function NCOMP, adjectival open arguments ACOMP, prepositional open arguments PCOMP, and verbal/clausal ones VCOMP. However, encoding category directly in grammatical functions represents a mixing of levels; category is a c-structure property. Therefore, what became the standard view is that taken by Bresnan (1982b), under which there is a single category-neutral open argument function XCOMP.

The postulation of a single category-neutral XCOMP function makes sense from an architectural perspective. However, it carries with it an expectation: XCOMPs should be able to be any category. As Bresnan (1982b) states,

Nothing in our theory requires that [the complement to a raising verb] be a VP (as opposed to another phrasal category which is a maximal projection). Hence, in our theory, raising should be possible with phrasal complements other than VP; in fact, both *consider* and *seem* also allow phrasal complements of categories other than VP... (Bresnan, 1982b, 376-7)

Of course, the LFG formalism allows reference to the category of an XCOMP across the ϕ^{-1} correspondence (Falk 2001), and some verbs might be expected to have such a lexical specification, but such lexical properties would be idiosyncratic. The normal case would be verbs like *consider* and *seem*, ones which allow their XCOMPs to be any category.

This expectation is not met. The vast majority of XCOMP-taking verbs only allow verb-based XCOMPs (*to* infinitives, bare infinitives, or participles)². The following is a partial list of verbs which only take clausal XCOMPs. Some of these verbs have other uses as well; the intended one here is Raising (to either subject or object).

¹This paper was presented at the LFG 05 conference in Bergen, Norway. I would like to thank Alex Alsina, Aaron Broadwell, Mary Dalrymple, Tracy King, Helge Lødrup, and K.P. Mohanan for comments.

²We will not discuss the selection of morphological form of verb-based XCOMPs.

- (1) affirm
 allege
 begin
 cease
 come (as in *After a while, he came to understand Minimalism.*)
 concede
 continue
 demonstrate
 determine
 expect
 guess
 happen
 know
 need
 proceed
 resume
 show
 state
 suppose
 tend

This calls for a rethinking of the nature of open functions.

To understand the nature of open argument functions it is useful to consider the closed argument functions. The closed argument functions are neither individuated on the basis of category nor category-neutral. The functions OBJ/OBJ_θ, OBL_θ, and COMP are canonically associated with specific categories. While there is some cross-linguistic variation, the object functions are typically associated with nominal phrases (NP, DP, KP), the oblique functions with PP, and the COMP function with subordinate clauses (CP, IP, S). The lexical forms in (2) have the categorial consequences in (3).

- (2) a. ‘express ⟨(↑ SUBJ) (↑ OBJ)⟩’
 b. ‘think ⟨(↑ SUBJ) (↑ COMP)/(↑ OBL_{about})⟩’
 c. ‘threaten⟨(↑SUBJ) (↑COMP)⟩’
- (3) a. The non-SUBJ argument of *express* is nominal
 b. The non-SUBJ argument of *think* is either a clause or a PP
 c. The non-SUBJ argument of *threaten* is a clause.

We propose that the open functions mirror the closed functions: specifically, there is an open object function (which we assume for reasons to be discussed later is the restricted XOBJ_θ), an open oblique function (XOBL_θ), and an open complement function (XCOMP). A few limited cases still, as we will see below, require c-selection, but this is the exception. In the general case, selection of open arguments will be entirely parallel to selection of closed arguments, with no need for an additional layer of c-selection mysteriously limited to open functions.

2. The Functions

2.1. XCOMP

In agreement with the LFG tradition, we consider the primary open argument function

to be XCOMP, i.e. the open equivalent of the COMP function, which expresses propositional arguments. However, we take the position that, just as the COMP function is expressed structurally by verbal/clausal categories (CP, IP, S), XCOMP can only be expressed by verbal/clausal categories (CP³, VP). There is thus no need for the verbs in (1) to have any special lexical marking preventing them from taking NP, PP or AP complements.

- (4) a. 'expect <((↑ SUBJ) (↑ XCOMP)) (↑ OBJ)'
 b. We expect this conference [_{CP} to be interesting] (CP can express XCOMP)
 c. *We expect this conference [_{AP} interesting] (AP cannot express XCOMP)

It is natural that the normal realization of open arguments is as verbal/clausal elements. Open arguments are propositional in nature; the normal (closed) realization of propositional arguments is by the COMP function, realized as a verbal/clausal category. In fact, many of these verbs also take closed COMPs.

2.2. XOBJ_θ

Most of the verbs that allow non-clausal open complements take primarily APs, usually as an alternative to clausal complements. For some, this is the only option:⁴

- (5) a. The lexicalist lecturer proved the transformationalist [_{CP} to be crazy]. CP
 b. The lexicalist lecturer proved the transformationalist [_{AP} crazy]. AP
 c. *The lexicalist lecturer proved the transformationalist [_{DP} a madman]. *DP/NP
 d. *The lexicalist lecturer proved the transformationalist [_{PP} out of his mind]. *PP

Generally, DPs (or NPs) are also possible.

- (6) a. The lexicalist doctor declared the transformationalist [_{CP} to be crazy]. CP
 b. The lexicalist doctor declared the transformationalist [_{AP} crazy]. AP
 c. The lexicalist doctor declared the transformationalist [_{DP} a madman]. DP/NP
 d. ?The lexicalist doctor declared the transformationalist [_{PP} out of his mind]. ?PP

With some verbs, there is dialectal variation.

- (7) a. The transformationalist seems [_{CP} to be crazy]. CP
 b. The transformationalist seems [_{AP} crazy]. AP
 c. #The transformationalist seems [_{DP} a madman]. #DP/NP
 d. ?The transformationalist seems [_{PP} out of his mind]. ?PP

A very few verbs exclude clausal/verbal arguments.

- (8) a. *The transformationalist stayed [_{CP} {to be / being} crazy]. *CP/VP

³We assume the analysis of *to* infinitives as CP (Falk 2001).

⁴Some of these allow PPs with adjectival meaning and others do not. For those that do, there is a certain degree of idiolectal variation, and the PPs often have more of a colloquial feel than APs and DPs.

- | | | |
|----|---|--------|
| b. | The transformationalist stayed [AP crazy]. | AP |
| c. | *The transformationalist stayed [DP a madman]. | *DP/NP |
| d. | *The transformationalist stayed [PP out of his mind]. | *PP |

In closed arguments, a distinction is made between core and non-core arguments. Core arguments (SUBJ, OBJ, OBJ_θ) are those which are canonically realized as nominals, without the mediation of prepositions (or semantic Case). The nominal categories are NP (/DP/KP) and AP. The difference between these categories is their functional potential: NP is referential and AP is predicative/attributive. It thus is to be expected that a closed core function would be expressed by NP and an open core function by AP. In fact, closed core arguments are canonically realized as NP, not as AP. Following this line of thinking, we take the basically-AP open function to be a core function. The frequent (and dialectally variable) extension to NP reinforces the insight that the arguments in question are core (nominal) arguments.⁵

The next question is which core function this open function is the equivalent of. The closed core functions are classified into the object (OBJ, OBJ_θ) and non-object (SUBJ) functions, and into non-restricted (SUBJ, OBJ) and restricted (OBJ_θ) functions. It seems relatively clear that the grammatical function in question is an object function: it does not have any subject-like properties, and, as with the object functions, arguments of adjectives cannot be mapped to the open core function. The choice between a non-restricted XOBJ and a restricted XOBJ_θ is less clear, but it appears to be restricted. Non-restricted functions can express non-thematic arguments (such as expletives), but non-thematic open functions do not exist (nor is it clear what it would mean to have a non-thematic open function). Characterizing the open core function as restricted (or, alternatively, disallowing open non-restricted functions) provides a formal syntactic expression of this impossibility—it states that open arguments do not have the functional capacity to be non-thematic. It also accounts for the existence of only one core open function, and, in particular, the nonexistence of XSUBJ.

The core/nominal open function is thus XOBJ_θ (apparently XOBJ_{Theme}). We take the normal case to be allowing realization as either AP or NP; the relatively unusual case of verbs which exclude NP involves lexical specification of category across the ϕ^{-1} projection through the use of the CAT function.⁶ Most verbs that select XOBJ_θ also allow XCOMP as an alternative.

- (9) a. $(\uparrow \text{PRED}) = \text{'declare } \langle (\uparrow \text{SUBJ}) (\uparrow \text{XOBJ}_{\text{Theme}}) / (\uparrow \text{XCOMP}) \rangle (\uparrow \text{OBJ}) \text{'}$
- b. $(\uparrow \text{PRED}) = \text{'prove } \langle (\uparrow \text{SUBJ}) (\uparrow \text{XOBJ}_{\text{Theme}}) / (\uparrow \text{XCOMP}) \rangle (\uparrow \text{OBJ}) \text{'}$
 $\text{NP} \notin \text{CAT } (\uparrow \text{XOBJ}_{\theta})$
- c. $(\uparrow \text{PRED}) = \text{'stay } \langle (\uparrow \text{SUBJ}) (\uparrow \text{XOBJ}_{\text{Theme}}) \rangle \text{'}$

⁵As the examples show, PPs with adjective-like meanings are also sometimes possible. While PP is not a canonical realization of a core function, it is a possible realization in some languages.

⁶The value of the CAT of a grammatical function is the set of category labels in its c-structure equivalent.

d. f-structure of (7b)

| | | |
|-----------------------|--|--|
| SUBJ | [DEF + PRED 'transformationalist'] | |
| TENSE | PRES | |
| PRED | 'seem <<(\uparrow XOBJ _{Theme})>> (\uparrow SUBJ)' | |
| XOBJ _{Theme} | [SUBJ PRED 'crazy <<(\uparrow SUBJ)>>'] | |
| | | |

2.3. XOBJ₀

There are two constructions with open arguments which have been problematic for LFG analysis. They are exemplified in (10).

- (10) a. The transformationalist strikes me [as {crazy | a madman | out of his mind | being crazy}].
 b. They prevented the transformationalist [from corrupting young minds].

(10a) exemplifies the construction with perception verbs like *strike*, *impress*, *remember*, *reveal*, which require their open complements to be headed by *as*, which we hypothesize is a preposition. LFG analyses, when they relate to this kind of sentence at all, typically ignore the *as*, and consider these verbs to select category-insensitive XCOMP. (10b) exemplifies the construction with negative causation verbs like *prevent*, *dissuade*, *deter*, *prohibit*, in which the complement is a PP headed by *as* with a gerund complement. Since the verbs in question are raising verbs, the complements must be open, but it is not clear how the combination of preposition and gerund can be analyzed as an open complement under the standard analysis.

Since both of these constructions involve the use of a preposition which explicitly marks the thematic role of the complement, we propose that these two constructions involve open oblique complement functions: XOBJ_{Stimulus} and XOBJ_{Neg} (or XOBJ_{Source}).

- (11) a. 'strike <<(\uparrow OBJ) (\uparrow XOBJ_{Stim})>> (\uparrow SUBJ)'

b.

| | | |
|----------------------|---|--|
| SUBJ | [DEF + PRED 'transformationalist'] | |
| TENSE | PRES | |
| PRED | 'strike <<(\uparrow OBJ)(\uparrow XOBJ _{Stim})>>(\uparrow SUBJ)' | |
| OBJ | [PRED 'PRO' PRES 1 NUM SG] | |
| XOBJ _{Stim} | [PCASE XOBJ _{Stim} SUBJ PRED 'crazy <<(\uparrow SUBJ)>>'] | |
| | | |

- (12) a. 'prevent <<(\uparrow SUBJ) (\uparrow XOBJ_{Neg})>> (\uparrow OBJ)'

b.

| | | | | | | | | | | | |
|---------------------|---|------|--|------|---|-------|---------------------|------|----|-----|------------------|
| SUBJ | <table border="1"> <tr><td>PRED</td><td>'PRO'</td></tr> <tr><td>PERS</td><td>3</td></tr> <tr><td>NUM</td><td>PL</td></tr> </table> | PRED | 'PRO' | PERS | 3 | NUM | PL | | | | |
| PRED | 'PRO' | | | | | | | | | | |
| PERS | 3 | | | | | | | | | | |
| NUM | PL | | | | | | | | | | |
| TENSE | PAST | | | | | | | | | | |
| PRED | 'prevent $\langle(\uparrow \text{SUBJ})(\uparrow \text{XOBL}_{\text{Neg}})\rangle(\uparrow \text{OBJ})$ ' | | | | | | | | | | |
| OBJ | <table border="1"> <tr><td>DEF</td><td>+</td></tr> <tr><td>PRED</td><td>'transformationalist'</td></tr> <tr><td>PCASE</td><td>XOBL_{Neg}</td></tr> <tr><td>SUBJ</td><td></td></tr> </table> | DEF | + | PRED | 'transformationalist' | PCASE | XOBL _{Neg} | SUBJ | | | |
| DEF | + | | | | | | | | | | |
| PRED | 'transformationalist' | | | | | | | | | | |
| PCASE | XOBL _{Neg} | | | | | | | | | | |
| SUBJ | | | | | | | | | | | |
| XOBL _{Neg} | <table border="1"> <tr><td>PRED</td><td>'corrupt $\langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle$'</td></tr> <tr><td>OBJ</td><td> <table border="1"> <tr><td>PRED</td><td>'mind'</td></tr> <tr><td>NUM</td><td>PL</td></tr> <tr><td>ADJ</td><td>{[PRED 'young']}</td></tr> </table> </td></tr> </table> | PRED | 'corrupt $\langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle$ ' | OBJ | <table border="1"> <tr><td>PRED</td><td>'mind'</td></tr> <tr><td>NUM</td><td>PL</td></tr> <tr><td>ADJ</td><td>{[PRED 'young']}</td></tr> </table> | PRED | 'mind' | NUM | PL | ADJ | {[PRED 'young']} |
| PRED | 'corrupt $\langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle$ ' | | | | | | | | | | |
| OBJ | <table border="1"> <tr><td>PRED</td><td>'mind'</td></tr> <tr><td>NUM</td><td>PL</td></tr> <tr><td>ADJ</td><td>{[PRED 'young']}</td></tr> </table> | PRED | 'mind' | NUM | PL | ADJ | {[PRED 'young']} | | | | |
| PRED | 'mind' | | | | | | | | | | |
| NUM | PL | | | | | | | | | | |
| ADJ | {[PRED 'young']} | | | | | | | | | | |

3. LMT

3.1. Features

We now turn to the place of the proposed new grammatical functions, XOBJ_θ and XOBL_θ, in the LMT taxonomy of grammatical functions.

Standard LMT decomposes the closed argument functions into the features [\pm restricted] and [\pm objective], features which form the basis of the mapping of arguments to the syntax.

(13)

| | $[-r]$ | $[+r]$ | |
|--------|---------------------|------------------|---------------------------|
| $[-o]$ | SUBJ | OBL _θ | non-Themes/Patients |
| $[+o]$ | OBJ | OBJ _θ | secondary Themes/Patients |
| | Themes/ Patients | | |

The mapping to the fully specified grammatical functions (for all but the most prominent argument) then adds positive values for remaining features.

This taxonomy does not include the closed complement function COMP, nor does it cover open functions. As for the latter, we add a feature [\pm saturated], where open functions are [$-s$] and closed functions are [$+s$].⁷ Based on our discussion above, the [$-r$] functions SUBJ and OBJ do not have open equivalents; we therefore hypothesize that the feature combination [$-r, -s$] is disallowed.

⁷K. P. Mohanan (personal communication) suggests a slightly different treatment. Under the theory of open functions proposed here, every [$+r$] function has an open version; under standard assumptions, ADJ also does (XADJ). Mohanan suggests treating the open and closed functions as not being distinct grammatical functions, but simply distinguished by whether or not they are controlled. Adopting this suggestion would require a different approach to the lexical government of functional control.

(14)

| | | | | |
|------|------|------------------|-------------------|------|
| | | [-r] | [+r] | |
| | | [+s] | | [-s] |
| [-o] | SUBJ | OBL _θ | XOBL _θ | |
| [+o] | OBJ | OBJ _θ | XOBJ _θ | |

Falk (2001) adds the feature $[\pm c]$ for COMP, but does not discuss the status of COMP with respect to the features $[\pm o]$ and $[\pm r]$. Our proposal is that COMP is $[+r]$ and unspecified with respect to $[\pm o]$. The analysis of (X)COMP as $[+r]$ is originally due to Zaenen & Engdahl (1994), who consider the restriction to propositional arguments to be on a par with thematic restrictions. In the present context, the existence of the open XCOMP requires the analysis, since $[-r, -s]$ is disallowed. COMP's $[o]$ feature is less clear; we note, however, that XCOMP can alternate with both XOBJ_θ and XOBL_θ, and COMP can alternate with either OBJ, OBJ_θ, or OBL_θ,⁸ suggesting that the $[+c]$ functions are neutral between $[+o]$ and $[-o]$.

(15)

| | | | | |
|------|------|------|------------------|-------------------|
| | | [-r] | [+r] | |
| | | [+s] | | [-s] |
| [-c] | [-o] | SUBJ | OBL _θ | XOBL _θ |
| | [+o] | OBJ | OBJ _θ | XOBJ _θ |
| [+c] | [±o] | | COMP | XCOMP |

3.2. Mapping Principles

The standard LMT mapping principles are:

- (16) a. Mapping from thematic roles to a-structure (“intrinsic classification”)
 Themes and Patients are $[-r]$
 “Secondary” Themes and Patients are $[+o]$
 Non-Theme/Patients are $[-o]$

⁸A different perspective on the alternation between COMP and OBJ is offered by Dalrymple & Lødrup (2000). They suggest that clausal complements that alternate with nominal complements bear the OBJ function rather than the COMP function. Given an appropriate LMT feature decomposition, such an analysis is unnecessary. An interesting question that is raised by COMP/OBJ alternations, however, is whether COMP should be analyzed as neutral with respect to $[\pm r]$ as well. We take a different approach below.

b. Mapping from a-structure to f-structure (“default classification”)

$\hat{\theta}$
 | maps to SUBJ
 [-o]

[-r] optionally maps to SUBJ

Otherwise, ‘+’ values of unspecified features are added

We will leave the a-structure–f-structure mapping unchanged, but arguments with propositional content (both closed and open) need to be added to the mapping to a-structure.

We begin by noting that the unmarked mapping for arguments with propositional content, regardless of thematic role, is COMP, i.e. [+r, +c, +s, ±o], while non-propositional arguments are invariably mapped to [-c] functions. We take it that, in the unmarked case, the Theme/Patient mapping to [-r] is restricted to non-propositional arguments, and that the [±c] feature distinguishes propositional from non-propositional arguments.

- (17) Non-propositional Themes and Patients are [-r]
 “Secondary” Themes and Patients are [+o]
 Non-propositional arguments are [-c]
 Non-Theme/Patients are [-o]

A verb with a propositional argument will not be specified at a-structure for the feature [c], and the ‘+’ value will be filled in at f-structure.

(18) *intend*

| thematic roles | Agent | Prop. Theme |
|--|--------------|------------------------------|
| nprop T/P sec T/P nonprop non T/P | [-c] [-o] | |
| added features | [-r] [+s] | [+r] [+o] [+c] [+s] |
| GFs | SUBJ | COMP |

Verbs like *express*, which map propositional arguments to OBJ, are lexically specified exceptions.⁹

⁹The existence of exceptions to the mapping principles, although never (to my knowledge) discussed in the literature, has to be permitted by LMT to account for verbs which, for example, map a non-Theme/Patient argument to OBJ (such as *enter*). There is no reason to see this as a problem: the LMT features express syntactic properties of arguments while the mapping principles relate the features to semantic properties.

(19) *express*

| | | |
|----------------------|------------------|------------------|
| thematic roles | Agent | Prop. Theme |
| idiosyncratic | | map as non-prop. |
| nprop T/P sec T/P | | $[-r]$ |
| nonprop non T/P | $[-c]$ $[-o]$ | $[-c]$ |
| added features | $[-r]$ $+[s]$ | $+[o]$ $+[s]$ |
| GFs | SUBJ | OBJ |

Many verbs that take propositional arguments map them as both propositional and non-propositional arguments.

(20) a. *say*

| | | | |
|----------------------|------------------|----------------------------------|------------------|
| thematic roles | Agent | Prop. Theme | |
| idiosyncratic | | map as prop or non-prop | |
| nprop T/P sec T/P | | | $[-r]$ |
| nonprop non T/P | $[-c]$ $[-o]$ | | $[-c]$ |
| added features | $[-r]$ $+[s]$ | $+[r]$ $+c$ $+o$ $+[s]$ | $+[o]$ $+[s]$ |
| GFs | SUBJ | COMP | OBJ |

b. *tell*

| | | | | |
|----------------------|------------------|------------------|----------------------------|------------------|
| thematic roles | Agent | Recipient | Prop. Theme | |
| idiosyncratic | | | map as prop or non-prop | |
| nprop T/P sec T/P | | $[-r]$ | $+[o]$ | $+[o]$ |
| nonprop non T/P | $[-c]$ $[-o]$ | $[-c]$ | | $[-c]$ |
| added features | $[-r]$ $+[s]$ | $+[o]$ $+[s]$ | $+[c]$ $+[r]$ $+[s]$ | $+[r]$ $+[s]$ |
| GFs | SUBJ | OBJ | COMP | OBJ _θ |

c. *hope*

| | | | |
|--|------------------|----------------------------|------------------|
| thematic roles | Agent | Prop. Ben? | |
| idiosyncratic | | map as prop or non-prop | |
| nprop T/P sec T/P nonprop non T/P | $[-c]$ $[-o]$ | $[-o]$ | $[-c]$ $[-o]$ |
| added features | $[-r]$ $[-s]$ | $[+c]$ $[+r]$ $[+s]$ | $[+r]$ $[+s]$ |
| GFs | SUBJ | COMP | OBL _θ |

The ability of a predicate to take an open complement is a lexically idiosyncratic property. The well-known distinction between *probable* and *likely* illustrates this. We hypothesize, as a first approximation, that propositional arguments are optionally mapped as $[-s]$.

(22) *likely*

| | | | |
|--|------------------|--------------------------------------|----------------------------|
| thematic roles | | Prop. Theme | |
| idiosyncratic | non-θ | | |
| nprop T/P sec T/P nonprop non T/P prop | $[-c]$ $[-o]$ | | $[-s]$ |
| added features | $[-r]$ $[-s]$ | $[+s]$ $[+r]$ $[+o]$ $[+c]$ | $[+r]$ $[+o]$ $[+c]$ |
| GFs | SUBJ | COMP | XCOMP |

Predicates which allow only open argument functions are lexically marked as requiring the feature $[-s]$.

(23) *tend*

| thematic roles | | Prop. Theme |
|--|------------------|------------------------|
| idiosyncratic | non- θ | $[-s]$ |
| nprop T/P sec T/P nprop non T/P prop | $[-c]$ $[-o]$ | |
| added features | $[-r]$ $+[s]$ | $+[r]$ $+o$ $+c$ |
| GFs | SUBJ | XCOMP |

Similarly, predicates which only allow closed complements are lexically specified to require $+[s]$.

(24) *probable*

| thematic roles | | Prop. Theme |
|--|------------------|------------------------|
| idiosyncratic | non- θ | $+[s]$ |
| nprop T/P sec T/P nonprop non T/P prop | $[-c]$ $[-o]$ | |
| added features | $[-r]$ $+[s]$ | $+[r]$ $+o$ $+c$ |
| GFs | SUBJ | COMP |

This picture needs to be complicated a little for Raising-to-Object verbs. In case of a $[-s]$ mapping, they also require a non-thematic $[-r]$ argument.

(25) *believe*

| thematic roles | Agent | Prop. Theme | | | |
|--|--|--|--|--|--|
| idiosyncratic | | map as prop or non-prop [-s] ⇒ non-θ [-r] | | | |
| nprop T/P sec T/P nonprop non T/P prop | [-c] [-o] | | [-r] [-c] | | [-r] [-c] |
| added features | $\begin{bmatrix} -r \\ +s \end{bmatrix}$ | $\begin{bmatrix} +r \\ +c \\ +o \\ +s \end{bmatrix}$ | $\begin{bmatrix} +o \\ +s \end{bmatrix}$ | $\begin{bmatrix} +r \\ +c \\ +o \end{bmatrix}$ | $\begin{bmatrix} +o \\ +s \end{bmatrix}$ |
| GFs | SUBJ | COMP | OBJ + XCOMP | | OBJ |

All three mappings for *believe* are thus derived: with an OBJ, with a COMP, and the Raising-to-Object variant.

The version of LMT that we are proposing thus has as a consequence that the normal realization of propositional complements is as a verbal/clausal element: either a closed COMP or an open XCOMP. Deviations from this require lexical specification, and are thus taken to be exceptional. The specific deviation required for $XOBJ_\theta$ and $XOBL_\theta$ is a lexical specification allowing [-c]. A mapping of a propositional argument as [-c] requires a concomitant [-s], since either clausal or open (predicative) realization is required to express propositional content. In most cases, the [-c] specification is optional, and COMP and XCOMP mappings are also possible.

(26) *seem*

| thematic roles | | Prop. Theme | | |
|--|--|---------------|--|--|
| idiosyncratic | non-θ | [-c] | | |
| nprop T/P sec T/P nonprop non T/P prop | [-c] [-o] | [+o] | [+o] | [+o] |
| added features | $\begin{bmatrix} -r \\ +s \end{bmatrix}$ | [+r] | $\begin{bmatrix} +c \\ +r \end{bmatrix}$ | $\begin{bmatrix} +c \\ +r \end{bmatrix}$ |
| GFs | SUBJ | $XOBJ_\theta$ | COMP | XCOMP |

(27) *recognize*

| thematic roles | Agent | Prop. Stimulus | | | |
|--|--------------|------------------------|--------------|------------------|--------------|
| idiosyncratic | | [-s] ⇒ non-θ [-r] [-c] | | | |
| nprop T/P sec T/P nonprop non T/P prop | [-c] [-o] | [-o] | [-r] [-c] | [-o] [-s] | [-o] [-s] |
| added features | [-r] [+s] | [+c] [+r] [+s] | [+o] [+s] | [+c] [+r] | [+r] |
| GFS | SUBJ | COMP | OBJ + XCOMP | OBL _θ | |

To summarize, the following is the full set of LMT principles:

- (28) Non-propositional Themes and Patients are [-r]
 “Secondary” Themes and Patients are [+o]
 Non-propositional arguments are [-c]
 Non-Theme/Patients are [-o]
 Propositional arguments are [-s], obligatorily if they are [-c], optionally otherwise

4. Further thoughts

4.1. PREDLINK

It has been proposed (Butt, King, Niño, & Segond 1999) that some ostensibly open complement arguments are actually closed arguments with a function they dub PREDLINK. As they put it,

As NPs, APs, and especially PPs do not generally have an overt subject, we believe the representation of the relationship between the noun and the thing predicated of it should either be encoded at the level of argument structure, or of semantic structure. Note that if one does implement the controlled subject analysis, it becomes necessary to provide two subcategorization frames for each of these categories: one without a SUBJ argument for simple NPs such as *a cat* in *A cat ate my food*, and one for predicatively used NPs such as *a cat* in *Harry is a cat...* The PREDLINK analysis avoids these difficulties by positing a grammatical function PREDLINK.... As PREDLINK is a closed category, there is no control equation between the SUBJ and the PREDLINK and hence no need for NPs, APs, and PPs to have subject arguments. (Butt et al. 1999: 70)

A study reanalyzing open functions must consider the nature of PREDLINK.

Dalrymple, Dyvik, & King (2004) compare the open and closed analysis of predicative complements of copular verbs. From their study, it is clear that a closed PREDLINK analysis is not correct for all of the cases that Butt et al. intended it for. Predicate adjectives which agree with the NPs of which they are predicated and those which are themselves Raising

predicates need to be analyzed as having SUBJ arguments of their own, and therefore are open (controlled) arguments. On the other hand, as Dalrymple et al. point out, a functional control analysis is not possible for predicative complements which have their own subjects, such as finite subordinate clauses.

- (29) a. The problem is that the hamster will eat the cat.
 b. The issue is whether closed predicative argument functions exist.

The conclusion they reach is thus that both PREDLINK and XCOMP analyses are correct, albeit for different cases.

Two additional considerations must be added, which cast doubt on the existence of PREDLINK as a distinct grammatical function. The first, internal to the current proposal, is that there is no plausible feature decomposition available. The only gap in the LMT feature chart (15) is $[-r, \pm o, +c, -s]$, but a $[-r]$ designation for PREDLINK seems unlikely. More generally, there is something very strange about a grammatical function which serves as an argument for a very restricted class of verbs, perhaps a single verb, *be*. (*Remain* may also be a “PREDLINK”-taking verb.)

The most straightforward alternative to PREDLINK is to analyze *be* as selecting COMP in addition to open complements.

- (30)
$$\left[\begin{array}{l} \text{SUBJ} \left[\begin{array}{l} \text{DEF} \quad + \\ \text{PRED} \quad \text{'problem'}$$

This analysis is preferable to the one invoking a PREDLINK function.¹⁰

4.2. Other Languages

This paper has dealt with a relatively detailed study of open arguments in English. In this section, we will speculate on the situation in other languages.

The mapping of propositional arguments to open functions differs from language to language, suggesting that the LMT principles must be allowed to vary. For example, in Hebrew XCOMPs are much rarer as arguments of verbs than in English. The verb *nire* ‘seem’, for example, takes closed COMP and open XOBJ₀, but not open XCOMP.

¹⁰Another case for which XCOMP seems inappropriate is NPs that cannot be predicative, such as pronouns and proper names.

- (i) a. The teacher is Sara.
 b. The new president is him.

It is not clear how to analyze these cases, but they may involve a closed OBJ. These are in some sense more marginal cases, with a flavor of being less acceptable than other constructions with ‘be’.

- (31) a. Nire še mištاتفey ha- kenes nehenim me ha- harcaot.
seems that participants the- conference enjoy.PRES from the- lectures
'It seems that the conference participants enjoy the lectures.'
- b. *Mištاتفey ha- kenes nirim lehanot me ha- harcaot.
participants the- conference seem enjoy.INF from the- lectures
'The conference participants seem to enjoy the lectures.'
- c. Mištاتفey ha- kenes nirim smexim.
participants the- conference seem happy
'The conference participants seem happy.'

Lødrup (2002) discusses the frequent use of prepositions (particularly *til* 'to') in introducing open complements in Norwegian.

- (32) a. Han ser ut til å sove.
he seems PREP INF sleep
'He seems to sleep.'
- b. Vi fikk ham til å sove.
we made him PREP INF sleep
'We made him sleep.'

He links this to the observation by Dalrymple and Lødrup (2000) that complement clauses in Norwegian generally appear to be OBJ rather than COMP; the COMP function is used in a very limited number of cases. As Lødrup puts it, the canonical function for embedded clauses in Norwegian, both finite and infinitive, is OBJ rather than (X)COMP. It is more natural for a PP to be mapped to XCOMP than it is for an infinitive, and the result is the use of the preposition. From the perspective of the theory proposed here, Lødrup's account is basically correct but can be improved. In a language without COMP, we would not expect XCOMP; such a language would be one in which there are no [+c] functions, or in which the [$\pm c$] feature is not used. Norwegian is apparently not entirely lacking in COMP, but it is highly marked; the same degree of markedness would be expected for XCOMP. As a result, the only open argument functions that are normally available are XOBJ₀ (realized as AP) and XOBL₀ (realized as PP). The use of the dummy preposition allows the XOBL₀ realization of the arguments of raising verbs.

The more fine-grained approach that we are taking to open argument functions, in particular the positing of an open object function, provides a new insight into the nature of open arguments in Balinese. In Balinese (Arka 1998), some open arguments are capable of being the "pivot" (Arka's GF-SUBJ) and others are not.

- (33) a. [Naar ubad ento] tegarang tiang
ACTORVOICE.eat medicine that OBJECTVOICE.try 1
'Taking the medicine is what I tried.'
- b. *[Ngelah umah luung] ane edot ia.
ACTORVOICE.own house good REL OBJECTVOICE.want 3
'Having a good house is what (s)he wants.'

Following Falk (2000, to appear), we take PIV to be an overlay function: not the argument function “subject” (which the references cited call $\widehat{\text{GF}}$). Unlike Arka & Simpson (1998), we therefore do not take this to be evidence for a function XSUBJ. However, the fact that some open complements can be designated as PIV through the use of “object voice” while others cannot suggests that some are objects and others are not.

- (34) a. ‘try $\langle(\uparrow \text{SUBJ}) (\uparrow \text{XOBJ}_\theta)\rangle$ ’
 b. ‘want $\langle(\uparrow \text{SUBJ}) (\uparrow \text{XCOMP})\rangle$ ’

The object status of open complements such as the complement of ‘try’ is made even clearer by the ability of an applicative suffix to convert the complement of a verb like ‘want’ into a potential pivot.

- (35) [Ngelah umah luung] ane edot- in= a.
 ACTORVOICE.own house good REL OBJECTVOICE.want- APPL= 3
 ‘Having a good house is what (s)he wants.’

Applicative morphology typically converts non-objects into objects, so the most straightforward analysis of this applicative suffix is that it converts XCOMP into XOBJ_θ. The distinction between core (or “term”) and non-core (“non-term”) complements made by Arka, which is rather mysterious under the XCOMP-only view of open argument functions, thus receives a natural analysis.

5. Conclusion

We have shown that the traditional LFG view, under which there is a single open argument function XCOMP, incorrectly creates the expectation that the normal situation is for all categories to be equally possible as open arguments. We have argued that in addition to XCOMP, the grammatical functions XOBJ_θ and XOBJ_L_θ need to be recognized, and we have suggested extensions to the LMT mapping principles. We have seen in a preliminary way that phenomena in Norwegian and Balinese can be profitably analyzed in terms of the richer inventory of open argument functions. Further cross-linguistic study of open arguments will doubtless reveal additional phenomena of this kind.

The necessity of distinguishing different kinds of open arguments provides an indirect argument for the LFG conception of argument mapping as being mediated by an articulated set of grammatical functions. This contrasts with the view of argument mapping taken by other theoretical frameworks.

References

- Andrews, Avery (1982) “The Representation of Case in Modern Icelandic.” in Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 427–503.
- Arka, I Wayan (1998) *From Morphosyntax to Pragmatics in Balinese: A Lexical-Functional Approach*. Ph. D. dissertation, University of Sydney.
- Arka, I Wayan, and Jane Simpson (1998) “Control and Complex Arguments in Balinese.” in Miriam Butt and Tracy Holloway King, ed., *Proceedings of the LFG98 Conference, The University of Queensland, Brisbane*. On-line: CSLI Publications.
<http://csli-publications.stanford.edu/LFG/3/lfg98.html>

- Bresnan, Joan (1982a) “The Passive in Lexical Theory.” in Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 3–86.
- Bresnan, Joan (1982b) “Control and Complementation.” in Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 282–390.
- Butt, Miriam, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond (1999) *A Grammar-Writer’s Cookbook*. Stanford, Calif.: CSLI Publications.
- Dalrymple, Mary, Helge Dyvik, and Tracy Holloway King (2004) “Copular Complements: Closed or Open?.” in Miriam Butt and Tracy Holloway King, ed., *Proceedings of the LFG 04 Conference, University of Canterbury*. On-line: CSLI Publications. 188–198.
<http://csli-publications.stanford.edu/LFG/9/lfg04.html>
- Dalrymple, Mary, and Helge Lødrup (2000) “The Grammatical Functions of Complement Clauses.” in Miriam Butt and Tracy Holloway King, ed., *Proceedings of the LFG 00 Conference, University of California, Berkeley*. 104–121.
<http://csli-publications.stanford.edu/LFG/5/lfg00.html>
- Falk, Yehuda N. (2000) “Pivots and the Theory of Grammatical Functions.” in Miriam Butt and Tracy Holloway King, ed., *Proceedings of the LFG00 Conference, University of California, Berkeley*. On-line: CSLI Publications. 122–138.
<http://csli-publications.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. (to appear) *Explaining Subjects and Their Properties* [tentative title]. Cambridge, Mass.: Cambridge University Press.
- Grimshaw, Jane (1982) “On the Lexical Representation of Romance Reflexive Clitics.” in Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 87–148.
- 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.
- Lødrup, Helge (2002) “Infinitival Complements in Norwegian and the Form-Function Relation.” in Miriam Butt and Tracy Holloway King, ed., *Proceedings of the LFG 02 Conference, National Technical University of Athens*. On-line: CSLI Publications. 274–291.
<http://csli-publications.stanford.edu/LFG/7/lfg02.html>
- Zaenen, Annie, and Elisabet Engdahl (1994) “Descriptive and Theoretical Syntax in the Lexicon.” in B.T.S. Atkins and A. Zampolli, ed., *Computational Approaches to the Lexicon*. Oxford: Oxford University Press. 181–212.

CORPUS-BASED LEARNING OF OT CONSTRAINT RANKINGS FOR LARGE-SCALE LFG GRAMMARS

Martin Forst

University of Stuttgart
Institute for NLP

Jonas Kuhn

Saarland University
Computational Linguistics

Christian Rohrer

University of Stuttgart
Institute for NLP

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

We discuss a two-stage disambiguation technique for linguistically precise broad-coverage grammars: the pre-filter of the first stage is triggered by linguistic configurations (“optimality marks”) specified by the grammar writer; the second stage is a log-linear probability model trained on corpus data. This set-up is used in the Parallel Grammar (ParGram) project, developing Lexical Functional Grammars for various languages. The present paper is the first study exploring how the pre-filter can be empirically tuned by learning a relative ranking of the optimality marks from corpus data, identifying problematic marks and relaxing the filter in various ways.

1 Background

1.1 Linguistically precise grammars in NLP

In recent years, parsing based on large manually developed grammars that are directly informed by linguistic theory has made significant progress towards broad-coverage application. Efficient processing platforms for parsing and generation are available (in our context in particular the XLE system (Kaplan et al. 2002) for Lexical-Functional Grammar – LFG); advanced profiling techniques and tool support for grammar development are available (Oepen and Carroll 2000, King et al. 2004); effective fallback strategies have been established to achieve robustness while still taking advantage of the high depth of analysis (Riezler et al. 2002). The high initial cost of theory-driven manual grammar development pays off when grammars for new languages are added to a family of grammars, as syntactic theory often gives clear indications as to which parts of the rule system will carry over from an existing grammar and which parts have to be rewritten; the most effective methodology thus relies on multilingual grammar development based on clear cross-linguistic grammar writing conventions, as, e.g., practiced in the LFG-based Parallel Grammar (ParGram) project (Butt et al. 2002, 1997) and the Grammar Matrix approach (Bender et al. 2002) in the Head-driven Phrase Structure Grammar (HPSG) framework. The resulting grammars are particularly suited for application contexts requiring great depth of analysis (like language understanding systems with a reasoning component) and/or reversible grammars for parsing/generation (like high-quality machine translation or computer-aided language learning).

1.2 Disambiguation: a two-stage approach

Theory-driven grammar development typically leads to grammars that overgenerate only mildly, since lexical subcategorization information is taken into account and the grammatical constructions can be restricted by rich feature constraints. In other words, most of the parses that a grammar assigns to a string are linguistically justified. Nevertheless, due to the underspecified nature of natural language, ambiguity rates for non-trivial sentences are considerable: most disambiguation decisions cannot be made on strictly grammatical grounds, but involve some contextual or world knowledge. To ensure portability of the grammars across domains, hard-wiring such non-grammatical decisions in the grammar code is generally avoided. This justifies the need for sophisticated disambiguation techniques to complement the linguistic grammar in parsing.

Contrary to the situation in grammar writing, the most effective way of building a disambiguator is to exploit empirical, corpus-driven techniques. For constraint-based formalisms like LFG and HPSG, the use of log-linear probability models applied on fully or partially labeled training corpora has been established as a powerful, general machine learning technique (Johnson et al. 1999, Riezler et al. 2002, Kaplan et al. 2004, Toutanova et al. 2002). The log-linear models are typically trained using a large, schematically constructed set of learning features that check for structural and lexical configurations and co-occurrences in the linguistic representations. Trimming the features of the log-linear model is an

engineering task separated quite clearly from the grammar writing task. Hence, there is a clear conceptual split between grammar writing, driven by linguistic theory, and disambiguator development work, involving advanced machine learning techniques. However, in the ParGram project, it has proven highly productive to assume an intermediate stage between the two components – a linguistically motivated pre-disambiguation filter. Motivation comes from both ends:

(1) *Linguistic motivation*: Beyond the strict grammatical rules and principles that are encoded in the grammar proper, there is a considerable number of soft principles that the grammar writer is well aware of when working on a linguistic construction. For instance, many “rare” constructions should only be appealed to in analysis when there is no “canonical” analysis of a string.¹

Translating soft constraints into carefully conditioned hard grammatical constraints in order to keep up the two-way conceptual split often yields unintelligible, error-prone rules; it also goes against the idea of a theory-driven grammar development as it hides a clear intuitive explanation in a technically complicated rule. Leaving the soft constraints entirely out of the grammar writing picture, hoping that some constellation of the learning features will pick them up, is unsatisfactory, too. At the point when the linguist is working on a construction, s/he is best aware of the linguistically salient interactions, and it takes almost no extra work to encode the soft constraint explicitly.

Therefore the XLE grammar development platform employed in the ParGram project has been integrated with a soft constraint mechanism inspired by the strict constraint ranking system of Optimality Theory (OT) (Frank et al. 2001). The mechanism is conceptually quite simple: for particular structural configurations in the linguistic representation, an *optimality mark* or OT constraint can be introduced (e.g., for the occurrence of a topicalized object). Each optimality mark is assigned a polarity, i.e., defining it as preferred or dispreferred. Furthermore, all marks used in a grammar can be ordered in a relative ranking (where several marks can be given the same rank position). When the parser is applied, optimality mark instances are collected in a multiset and can then be used as a filter on the readings produced by the system. Following the ranking order, each mark will filter out readings that have fewer instances than the reading with the maximal instances (for preference marks) or more instances than the reading with the fewest instances (for dispreference marks). The readings of a sentence that pass all marks and are still left in the end are called “optimal”, the readings that are filtered out are called “suboptimal”.²

Experience showed that without a (potentially temporary) filtering mechanism for uncommon constructions, grammar writing would be considerably harder (King et al. 2004).

(2) *Technical motivation*: The log-linear models applied in empirical training of the disambiguator are a discriminative technique, involving the computation of the gold standard analysis as well as all alternative solutions (Johnson et al. 1999); furthermore, parameter estimation is an iterative process that passes the training data multiple times. Hence, to keep the process tractable on medium-size to large training corpora, the set of competing analyses has to be limited. Using only the analyses that pass a linguistically motivated prefilter is a very desirable set-up.

¹Awareness of such linguistic interactions goes back to Panini’s work, and in recent years, ways of including soft-constraint mechanisms in formal grammar formalisms have been explored, particularly in the framework of Optimality Theory (Prince and Smolensky 1993) or in probabilistic grammar models (Manning 2003).

²For example, the parser may assign four readings to a sentence. Reading one has the multiset $\{ C_1, C_1, C_2 \}$ of optimality marks, reading two has the multiset $\{ C_2, C_2, C_3 \}$, reading three $\{ C_1, C_1, C_2 \}$, and reading four $\{ C_1, C_1, C_3, C_3 \}$. Let us furthermore assume that the marks are ranked $C_1 \gg C_2 \gg C_3$, and that all have positive polarity, i.e., they are preference marks. In evaluation, C_1 is considered first. Readings one, three, and four have two marks of C_1 each, whereas reading two has none. Therefore reading two is filtered out at this step. For the remaining readings, C_2 is considered next. Reading four doesn’t include any C_2 marks, so it is filtered out. For the final mark C_3 , there is no difference between the remaining readings (one and three – both have no C_3 mark). Hence, there are two optimal readings: readings one and three.

2 Methodology

In this section, we discuss our experimental methodology at a conceptual level, a more detailed description and the results follow in section 3.

2.1 Trimming the linguistic pre-filter

In past work on the ParGram grammars, both the introduction of OT marks and the specification of their relative ranking was done manually. This is problematic since the various marks affect phenomena that were integrated into the grammar at different development stages, and often the appropriate relative ranking can only be determined empirically. Moreover, the question of whether or not a particular mark should be active in the two-stage filter architecture we described is also hard to answer in isolation. (But note that the structural specification and the polarity of the candidate marks *are* aspects about which the grammar writer can make an informed decision.)³

This paper is the first systematic study applying empirical methods in order (i) to determine the ranking of the OT marks, and (ii) to decide which marks should be left out of the first disambiguation stage. It has been part of this study to explore measures for the quality of a particular specification of the pre-filter. We present results for the German grammar from the family of ParGram grammars. Some of the results of our experiments are surprising and provide some interesting insights in the workings of the two-stage filter architecture.

While the technical results we report in this paper make reference to project-specific details of our system architecture, we believe that many of the higher-level observations will carry over very well to other projects involving a linguistically motivated core module that is applied in a broader context of empirically tuned system components.

2.2 Measuring the quality of the pre-filter

The fact that the component we are interested in here is a pre-filter in the context of a two-stage system has special consequences for quality assessment. It is not necessary that the pre-filter remove *all* incorrect readings – since it is followed up by a sophisticated second disambiguator. On the other hand, it is

³Let us consider the following sentences as examples that illustrate the way dispreference and preference marks work, but that also show that they sometimes prove problematic:

- (1) Der Journalist stellt ihn ihr gegenüber.
The journalist put him her opposite.
'The journalist confronted him with her.'
- (2) Weil er ihr gegenüber arrogant war, verliert er Sympathie.
Because he her opposite arrogant was, loses he sympathy.
'Since he was arrogant towards her, he is losing sympathy.'
- (3) Weil er Frau Merkel gegenüber Schwachsinn erzählte, verliert er Sympathie.
Because he Ms Merkel opposite nonsense told, loses he sympathy.
'Since he told nonsense to Ms Merkel, he is losing sympathy.'

In examples 1 and 3, ambiguity arises due to the fact that *gegenüber* can be a separable verb particle, a preposition, a postposition or an adverb. This ambiguity can be (at least partly) resolved by the OT marks *VerbParticle*, which is a preference mark, and *Ppost*, which is a dispreference mark. In example 1, *VerbParticle* correctly filters out the readings where *gegenüber* is analyzed as an adverb or postposition, keeping only the parse where it is analyzed as a separable verb particle as optimal. In example 2, *gegenüber* is analyzed as a postposition. This analysis survives the filtering by *VerbParticle* and *Ppost* because no alternative analysis of *gegenüber* is available. In example 3, *gegenüber* is wrongly analyzed as a preposition, because *Ppost* makes its intended analysis as a postposition suboptimal. Realistically, the fine-tuning between instances parallel to 2 vs. 3 can only be done taking corpus frequencies into account.

highly undesirable if the pre-filter accidentally removes the correct reading, since this would make a data point unusable for the second stage. One might describe this as a task in which recall is of greatest importance, and precision should be traded off for recall; but in order to do justice to the special setting, we will call the relevant measures “filter fidelity” and “filter efficiency”. “Filter fidelity” is defined as the proportion of sentences for which the OT mark ranking under consideration keeps the correct reading among the optimal reading(s). The intuition behind “filter efficiency”, on the other hand, is to measure the proportion of readings among all incorrect readings of a sentence which are discarded by the OT mark ranking as suboptimal. Concretely, we calculate it as the quotient of the number of readings discarded by the filter divided by the total number of readings minus one.⁴ Filter fidelity is our main criterion and it should be as close as possible to 100%, but filter efficiency does have a certain importance as well, of course, since filtering a maximum of bad readings while losing a minimum of good readings is the goal of this whole enterprise. As a combined quality measure, we therefore provide a weighted F-score where filter fidelity is weighted more strongly than filter efficiency.⁵

2.3 Corpus-based learning of a ranking

Our experiments start out with the manually specified OT mark ranking in the German grammar. An obvious technique to try out is to learn a ranking automatically from corpus data for which the correct reading has been labeled. The filter quality with the learned ranking can then be compared against the manual ranking and a uniform ranking (giving all marks the same rank).

For corpus-based learning of the OT ranking, one could in theory apply the classical Constraint Demotion Algorithm from the OT literature (Tesar and Smolensky 1998); however, due to the variation in the data the algorithm might not converge. Therefore we transform the classical discrete constraint ranking into a continuous numerical ranking for the purpose of learning. This allows us to apply robust learning algorithms like the Gradual Learning Algorithm (GLA) proposed by Boersma (1998), which is related to the perceptron algorithm. In learning, the system’s current numerical ranking (with some noise added to determine each constraint’s particular rank) is used in order to disambiguate a sentence from the training data. When the predicted solution does not match the gold standard analysis, all constraints ranked too low are promoted by a small increment (controlled by the so-called plasticity parameter); all constraints ranked too high are demoted. The noise added in application has the effect that constraints with a similar ranking can “swap” their relative rank, which leads to variation in the data, as it is often observed. This variant of OT is thus often called Stochastic OT.

2.4 Augmenting the set of OT marks

We also performed an additional experiment besides learning a ranking for just the OT marks specified by the grammar writers: we explored how pre-filter quality is affected if we systematically augment the existing set of OT marks to ensure that for common disambiguation decisions, sufficiently fine-grained distinctions in the OT marks are available. It is conceivable that for certain decisions, the OT mark set is too “sparse” to produce a reliable result, whereas a richer OT mark set might behave in a more balanced way. This is because in stochastic OT, competing marks may form clusters in the numerical ranking, and the addition of new constraints may have the effect of making such a cluster more stable.⁶

⁴At the LFG Conference in Bergen we presented figures that were based on a slightly different definition of filter efficiency, namely the quotient of the number of readings discarded by the filter divided by the total number of readings. Since this initial definition prevents filter efficiency from taking 1.0 as a value and it is highly dependent on the total number of readings for a given sentence, our new definition is more appropriate.

⁵The exact definition is $F_\beta = (1 + \beta^2) \frac{FE \times FF}{FE + \beta^2 FF}$, β being set to 0.5.

⁶To anticipate the experimental results however, we could not observe the effect of getting a more relaxed filter by providing a larger set of interacting OT marks.

In a pilot study, we thus established OT tableaux containing the OT marks employed in the German ParGram LFG and ran the GLA on these. This allowed us to identify OT marks which were reranked particularly often and/or which were regularly both demoted and promoted. Two such marks were *ObjInVorfeld*⁷ and *LabelP*.⁸ After inspection of a certain number of sentences where these OT marks made the correct reading(s) suboptimal, we introduced new, more fine-grained OT marks such as *ObjPersPronoun* (which disprefers the interpretation of personal pronouns as objects) and *SubjIndef* (which disprefers the interpretation of indefinite noun phrases as subjects), hoping these would allow to make the correct reading(s) optimal for more sentences.

In order to be able to control whether this is effectively the case, we established two sets of tableaux: the first one, henceforth the “all marks” set, contains both the 59 original and the 54 additional, newly introduced OT marks; the second one, henceforth the “original marks” set, contains only the original marks. Both sets were in turn split up into a training and a test set, so that we can examine how well rankings learned from the training sets generalize to unseen data.

We then ran the GLA on the training portions of both the “all marks” set and the “original marks” set. For training, we used a “traditional” GLA setting, i.e. a setting where the effective numerical rank of an OT mark diverts from its grammatical rank within a normal distribution due to added noise and where marks making wrong predictions are demoted or promoted on the numerical scale by a constant called plasticity (cf. Boersma (1998)).

In order to evaluate the resulting rankings, we used a variant of the GLA without any noise intervening at evaluation time. This allowed us to evaluate the resulting numerical rankings as if they were strict relative rankings, which is the type of ranking used in XLE. Moreover, the application of this variant of the GLA to the data allows us to identify marks that, even with an “optimal” OT mark ranking, cause correct readings to be evaluated as suboptimal. In this sense, it is not only a tool for the evaluation of OT mark rankings as a whole, but it can also be used to evaluate how reliable single OT marks are.

2.5 Relaxing the filter

An important additional step in our experiments (based both on the “original marks” and based on “all marks”) was the attempt to modify an existing set of marks and ranking in order to increase filter fidelity – without decreasing filter effectiveness too much. Besides learning a more adequate ranking, this could be achieved in the following ways: (1) deactivating certain OT marks, such that their filtering effect is removed, and (2) grouping together constraints with a very similar rank. For step (1), it is important to identify appropriate marks for deactivation. In the pre-filter scenario, marks that are typically involved in “highly contingent disambiguation decisions” (i.e., decisions that may turn out one way or the other) should be excluded from the set, since they will eliminate the correct solution in relatively many cases. To identify marks for deactivation we explored two strategies: (a) inspecting the results obtained with the GLA variant without noise given a ranking obtained through a training run and to deactivate the marks that caused wrong predictions and (b) automatically deactivating a certain proportion of marks being associated with ranks at the lower side of the numerical scale.

For step (2) – grouping of similarly ranked constraints – we used various threshold values representing the minimal distance that two marks have to be away from each other in order to be attributed to distinct groups with distinct ranks.⁹

⁷*ObjInVorfeld* disprefers the interpretation of case-ambiguous noun phrases in the vorfeld, i.e. the position in front of the finite verb in verb-second clauses, as objects in sentences such as [NP-SUBJ Hans] *sieht Maria* ‘John sees Mary’ vs. [NP-OBJ Hans] *sieht Maria* ‘John, Mary sees’.

⁸*LabelP* disprefers the interpretation of a noun phrase as a close apposition to another noun phrase in sentences such as *Hans stellt* [NP-OBJ *das Auktionshaus Ebay*] *vor* ‘Hans presents the auction house Ebay’ vs. *Hans stellt* [NP-OBJ *das Auktionshaus*] [NP-OBJ θ *Ebay*] *vor* ‘Hans presents the auction house to Ebay’.

⁹Such a grouping can “relax” the pre-filter in the following way: assume two dispreference marks C_1, C_2 end up with

3 Experiments

3.1 Data

For our experiments, we parsed the 40,020 sentences of the Release 1 of the TiGer Corpus¹⁰ with a variant of the German ParGram LFG (Dipper 2003) into which we had integrated the new, more fine-grained OT marks and in which the evaluation of almost all OT marks had been deactivated.¹¹ Out of 40,020 TIGER sentences, 23,962 received a full parse.¹² The resulting f-structure charts (packed f-structure representations) were matched against the f-structure charts previously derived from the TiGer graph annotation (Forst 2003a,b); OT mark profiles corresponding to the f-structure charts produced by the grammar were established, the TiGer-compatible readings being marked as target winners in them. Since the matching of the grammar output against the TiGer-derived representations is very time-intensive when the number of different analyses contained in the two f-structure charts involved (or at least in one of them) is very high, we had to limit the maximum number of matches performed between individual analyses to 10,000.¹³ By doing this, we obtained 6,418 OT mark profiles, associated with the sentences for which a proper subset of all parses is compatible to the TiGer-derived f-structure charts. (The granularity of the TiGer annotation is not sufficient to always determine one single parse as the correct one.) Sentences for which all analyses were compatible with the TiGer-derived representations had to be discarded, since there would be nothing to be learned from OT tableaux associated with sentences of this kind.

An example of an OT mark profile obtained this way is the one associated with the following sentence:

- (4) Anlaß für all das gab aus Schweizer Sicht das neue österreichische Abfallwirtschaftsgesetz.
rise for all that gave from Swiss view the new Austrian waste management law
'From the Swiss point of view, it was the new Austrian waste management law that gave rise to all that.'

| reading | AdvAttach | Obj | ObjCommon | ObjDef | ObjNoSpec | ... |
|---------|-----------|-----|-----------|--------|-----------|-----|
| A1 | 0 | 1 | 2 | 1 | 1 | ... |
| A2 | 0 | 1 | 2 | 0 | 2 | ... |
| A3 | 0 | 1 | 2 | 0 | 2 | ... |
| A4 | 0 | 1 | 2 | 1 | 1 | ... |
| A5-B1 | 1 | 1 | 1 | 0 | 1 | ... |
| ☞ A5-B2 | 0 | 1 | 1 | 0 | 1 | ... |
| A6-B1 | 1 | 1 | 1 | 1 | 0 | ... |
| A6-B2 | 0 | 1 | 1 | 1 | 0 | ... |

Table 1: Sample OT mark profile for sentence (4)

similar, but distinct, decreasing ranks. Without grouping, all readings with a C_1 mark would be filtered out, independent of their C_2 marking, whereas after grouping a reading with a C_1 and no C_2 mark is treated like one with a C_2 and no C_1 mark.

¹⁰<http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/>

¹¹The OT mark *GuessedMassNoun* was kept active, as its deactivation would have led to such an enormous increase in the numbers of readings produced by the grammar that the matching mentioned below would not have been feasible for most sentences.

¹²The grammar version employed was not chosen for its coverage but for its adherence to ParGram f-structure decisions which are reflected in the TiGer-derived representations. Moreover, the newly introduced OT marks caused a slight slow-down of the grammar, which caused additional timeouts wrt. other grammar versions.

¹³This means that sentences for which the product of the number of analyses in the grammar output and the number of analyses in the TiGer-derived representations was greater than 10,000 were discarded.

The directly resulting 6,418 OT mark profiles, which correspond to the “all marks” set, were randomly split up into a trainings set of 5,755 and a test set of 663. Then we created the “original marks” set, by replacing all values in the columns of the newly introduced OT marks by zero, and split it up into a trainings set of 5,755 profiles and a test set of 663 along the same lines as the “all marks” set.

3.2 Training and first results

The 5,755 OT mark profiles of both the “all marks” trainings set and the “original marks” training set were input to an implementation of the GLA that allows for multiple target winners. The learning was performed with a plasticity of 0.2 and in 10 iterations over the whole training set, each datum being considered 5 times. The result of these training runs were two different numerical OT mark rankings, one for the “all marks” set and another one for the “original marks” set.

The results of these two training runs are summarized in table 2.

| ranking employed | original marks | | | all marks | | |
|-------------------------|-----------------|-------------------|------------------|-----------------|-------------------|------------------|
| | filter fidelity | filter efficiency | weighted F-score | filter fidelity | filter efficiency | weighted F-score |
| after 10 iterations GLA | 81.9 | 84.9 | 82.5 | 78.3 | 87.2 | 79.9 |
| uniform ranking | 82.7 | 83.3 | 82.8 | 77.2 | 84.1 | 78.5 |
| original manual ranking | 80.5 | 84.8 | 81.3 | | | |

Table 2: Results of GLA learning on test sets

As can be seen from the figures, the ranking of the OT marks does not play a major role. Although the automatically learned ranking performs better than the manually determined ranking originally used in the German ParGram LFG, both in terms of filter fidelity (81.9% vs. 80.5%) and filter efficiency (84.9% vs. 84.8%), the improvement from the latter to the former is very slight. Also, we have to state that, for the “original marks” set, the automatically learned ranking performs worse than a uniform ranking, i.e. a ranking where all marks are equally strong, in terms of filter fidelity (81.9% vs. 82.7%), even if filter efficiency is better (84.9% vs. 83.3%). The weighted F-score we employ confirms this picture (82.8% vs. 82.5%).

Comparing the results for the “original marks” set and the “all marks” set, the observation is that although the additional marks allow for a better filter efficiency, they have a negative effect on filter fidelity. This result is a bit disappointing, because initially, we had hoped to improve both filter efficiency and filter fidelity by providing the new marks. At the same time, it is not all that surprising, since the more OT marks are used for disambiguation, the more difficult it is, of course, to maximise filter fidelity.

As to the rankings’ ability to generalize from the training data to the unseen test data, we can see in table 3 that both the figures themselves and the patterns observed above are comparable between the training sets and the test sets.

3.3 Relaxing the filter

Given that the filter fidelity we achieved with the learned ranking hardly exceeded 80%, we thought of ways of relaxing the OT filter in order to increase this value. At the same time, filter efficiency was not supposed to be affected too badly.

Inspecting (and deactivating) “problematic” OT marks: The first approach we took was to inspect the OT marks that, even with the automatically learned ranking, caused correct readings to be evaluated

| ranking employed | original marks | | | all marks | | |
|-------------------------|-----------------|-------------------|------------------|-----------------|-------------------|------------------|
| | filter fidelity | filter efficiency | weighted F-score | filter fidelity | filter efficiency | weighted F-score |
| after 10 iterations GLA | 80.2 | 85.4 | 81.2 | 78.0 | 87.3 | 79.7 |
| uniform ranking | 81.8 | 82.7 | 82.0 | 76.7 | 83.7 | 78.0 |
| original manual ranking | 79.6 | 85.2 | 80.7 | | | |

Table 3: Results of GLA learning on training sets

as suboptimal. Examples of these were, as in the pilot study mentioned in 2.4, *ObjInVorfeld* and *LabelP*. Apparently, even the newly introduced OT marks did not allow us to counterbalance them in cases where they caused wrong predictions, which leads us to the opinion that these OT marks, instead of being evaluated in the pre-filter step, should be integrated into the log-linear model as properties. As such, they can contribute to choosing the correct reading in the final disambiguation step, where, moreover, they can interact with other properties, such as the ones that weigh subcategorization frames.

Another category of OT marks that still made wrong predictions were robustness OT marks such as *AdvAttach* and *MassInPl*. The purpose of these OT marks is mainly to disprefer fall back rules that are implemented for cases where lexical information is lacking and, as a consequence, they interact tightly with this kind of information. Due to missing or erroneous information in the lexicons, it can happen that they make wrong predictions, although they are fairly reliable in all other cases. We deactivated most of these OT marks, i.e. those which caused relatively many wrong predictions, but they can potentially be reactivated once the lexicons they interact with have been improved.

As a reaction to the inspection of the “problematic” OT marks, a variant of the data was created where these marks are deactivated. We henceforth call this set of data the “unproblematic marks” set.

“Translating” the numerical rankings into strict rankings: For use in XLE, the numerical rankings obtained from GLA learning have to be “translated” into strict rankings in which marks may be grouped as equally strong. The easiest way of doing this is, of course, to have one group for each distinct numerical ranking. However, this may not be the most appropriate method of “translating” a numerical ranking into a strict ranking, because it completely ignores the information contained in the distance between two rankings. A possible alternative is to group all OT marks whose ranking have a distance smaller than a given threshold. This way, the number of groups of equally ranked OT marks is reduced, which should allow for better generalisation, and, more importantly, some of the information contained in the distance between rankings is taken into account. We experimented with groupings of this kind with thresholds 2.0 and 5.0.

The resulting rankings were then applied to both the “original marks” data set and the “unproblematic marks” set. The results are shown in table 4.

Just as in our first results (cf. subsection 3.2), we observe that the ranking has only a little influence on the results. Nevertheless, filter fidelity can be improved slightly by grouping marks whose ranks are not very distant, without filter efficiency being affected considerably. Taking the figures from the training data into account (which are not displayed here), the conclusion could be that a grouping with a threshold value of 5.0 performs best, since it basically achieves the same filter fidelity as the uniform ranking, while allowing for a slightly higher filter efficiency.

More importantly, table 4 shows that the deactivation of “problematic” marks can increase the filter fidelity considerably. We achieve a filter fidelity of about 96%, while still discarding more than 62% of the readings as suboptimal.¹⁴ This set-up also yields the highest weighted F-score of all our experiments:

¹⁴This can arguably be considered an underestimation, because the effect of the OT mark *GuessedMassNoun*, mentioned in

| ranking employed | original marks | | | unproblematic marks | | |
|----------------------------|-----------------|-------------------|------------------|---------------------|-------------------|------------------|
| | filter fidelity | filter efficiency | weighted F-score | filter fidelity | filter efficiency | weighted F-score |
| after 10 iterations GLA | 81.9 | 84.9 | 82.5 | 95.9 | 62.2 | 86.5 |
| grouped with threshold 2.0 | 82.4 | 84.8 | 82.9 | 96.2 | 62.1 | 86.7 |
| grouped with threshold 5.0 | 82.5 | 84.7 | 82.9 | 96.2 | 62.1 | 86.7 |
| uniform ranking | 82.7 | 83.3 | 82.8 | 96.1 | 60.3 | 85.9 |

Table 4: Results of disambiguation with “original marks” and “unproblematic marks”, marks being grouped according to different methods

86.7%.

Deactivating portions of the OT marks according to their ranks: An alternative approach to deactivating “unreliable” OT marks we experimented with was to discard a certain proportion of the marks corresponding to the ranks at the lower end of the numerical scale. We ran this experiment for both the “original marks” set and the “all marks” set, deactivating the lower 50% of the OT marks. The resulting variants of the data are henceforth called “upper 50% original” and “upper 50% all” respectively.

Tables 5 and 6 show the effect of discarding the lower 50% of the two OT mark sets. (The ranking used is the one obtained after 10 iterations of the GLA.)

| original marks | | | upper 50% original | | |
|-----------------|-------------------|------------------|--------------------|-------------------|------------------|
| filter fidelity | filter efficiency | weighted F-score | filter fidelity | filter efficiency | weighted F-score |
| 81.9 | 84.9 | 82.5 | 99.5 | 41.3 | 77.6 |

Table 5: Results of disambiguation with “original marks” and “upper 50% original”

| all marks | | | upper 50% all | | |
|-----------------|-------------------|------------------|-----------------|-------------------|------------------|
| filter fidelity | filter efficiency | weighted F-score | filter fidelity | filter efficiency | weighted F-score |
| 78.3 | 87.2 | 79.9 | 97.0 | 50.5 | 81.9 |

Table 6: Results of disambiguation with “all marks” and “upper 50% all”

Filter fidelity is greatly improved by this strategy, but unfortunately, there is a high price to be paid in terms of filter efficiency. In both settings, it drops to 50% or even less. Given that the filter fidelity for the “upper 50% all” set is comparable to the filter fidelity for the “unproblematic marks” set, but that the filter efficiency for it is considerably lower than for the “unproblematic marks” set, we conclude that it is a better strategy to identify problematic marks and then deactivate them than just to deactivate a certain proportion of the lower ranked marks.

subsection 3.1 as well, is not taken into account here, although it cuts down the number of readings considerably.

4 Discussion and conclusions

We presented a sequence of experiments exploring ways of empirical tuning for the first stage of a disambiguation architecture for linguistic grammars. This pre-filter is triggered by configurations that the grammar writer specifies as OT marks and uses a relative ranking among the marks. A somewhat surprising result is that training the constraint ranking on corpus data does not lead to a noticeable improvement over the use of a uniform ranking. However, it is very effective for identifying and deactivating marks that tend to exclude the correct readings in some cases. Both results are of course directly related to the somewhat unusual application context of the disambiguation routine as a pre-filter: if it were used as the only filter, one should certainly rely on a learned ranking to maximize filter effectiveness, and the OT marks that are problematic in the pre-filter scenario might well play an important role. For the given two-stage scenario however, our systematic empirical exploration showed that filter fidelity can be maximized most effectively by removing unreliable marks.

In future work, we plan to explore in more detail the technique of deactivating certain marks from the ranking automatically, among other things by combining this technique with the approach of grouping similarly ranked constraints together. Moreover, we will evaluate the effect of pre-filter variants on the training of the log-linear model used as the second disambiguation stage.

References

- Bender, Emily M., Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe (eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pp. 8–14, Taipei, Taiwan.
- Boersma, Paul. 1998. *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam.
- 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.
- Butt, Miriam, Tracy Holloway King, Maria-Eugenia Niño, and Frederique Segond. 1997. *A Grammar-Writer's Cookbook*. CSLI Publications.
- Dipper, Stefanie. 2003. *Implementing and Documenting Large-scale Grammars – German LFG*. PhD thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), Volume 9, Number 1.
- Forst, Martin. 2003a. Treebank Conversion – Creating a German f-structure bank from the TIGER Corpus. In *Proceedings of the LFG03 Conference*, Saratoga Springs. CSLI Publications.
- Forst, Martin. 2003b. Treebank Conversion – Establishing a test suite for a broad-coverage LFG from the the TIGER Treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC '03)*, Budapest.
- Frank, Anette, Tracy H. King, Jonas Kuhn, and John Maxwell. 2001. Optimality Theory style constraint ranking in large-scale LFG grammars. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality-theoretic Syntax*, pp. 367–397. Stanford: CSLI Publications.

- Johnson, Mark, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler. 1999. Estimators for stochastic “unification-based” grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL’99)*, College Park, MD, pp. 535–541.
- Kaplan, Ronald M., Tracy H. King, and John T. Maxwell. 2002. Adapting existing grammars. The XLE approach. In *Proceedings of the COLING-2002 Workshop on Grammar Engineering and Evaluation*, Taipei, Taiwan.
- Kaplan, Ronald M., Stefan Riezler, Tracy King, John T. Maxwell III, Alexander Vasserman, and Richard Crouch. 2004. Speed and Accuracy in Shallow and Deep Stochastic Parsing. In *Proceedings of the Human Language Technology Conference and the 4th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL’04)*, Boston.
- King, Tracy Holloway, Stefanie Dipper, Annette Frank, Jonas Kuhn, and John Maxwell. 2004. Ambiguity management in grammar writing. *Research on Language and Computation* 2:259–280.
- Manning, Christopher D. 2003. Probabilistic syntax. In Rens Bod, Jennifer Hay, and Stefanie Jannedy (eds.), *Probabilistic Linguistics*, pp. 289–341. Cambridge, MA: MIT Press.
- Oepen, Stephan, and John Carroll. 2000. Parser engineering and performance profiling. *Natural Language Engineering* 6:81–97.
- Prince, Alan, and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical Report Technical Report 2, Rutgers University Center for Cognitive Science.
- Riezler, Stefan, Dick Crouch, Ron Kaplan, Tracy King, John Maxwell, 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 Association for Computational Linguistics (ACL’02)*, Pennsylvania, Philadelphia.
- Tesar, Bruce B., and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29: 229–268.
- Toutanova, Kristian, Christopher D. Manning, Stuart M. Shieber, Dan Flickinger, and Stephan Oepen. 2002. Parse disambiguation for a rich HPSG grammar. In *First Workshop on Treebanks and Linguistic Theories (TLT2002)*, pp. 253–263.

ON PARENTHETICALS (IN GERMAN)

Christian Fortmann
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

Proceedings of the LFG05 Conference
University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005
CSLI Publications
<http://csli-publications.stanford.edu/>

Abstract. Optional constituents of a clause which, according to a couple of criteria, are commonly called parentheticals show quite diverging properties with respect to structurally determined aspects of constituency and interpretation like constituent placement, scope or bound variable reading of pronominals. One type of parenthetical string may form regular constituents of a clause, if the string is not parenthetically marked, and shows the same facts about interpretation like other regular constituents. The other type may not. Strings of the former type have to be represented as parts of their host at both levels of syntactic representation: c-structure and f-structure. Parentheticals of the latter type must be treated in a different way, since they exhibit properties usually attributed to strings which are not constituents of the host. I will propose an analysis of this type which rests on the integration of the parenthetical string into the c-structure of the host but its separate non-integrated representation at f-structure.*

1 (Something like) a Definition

Parentheticals are usually characterized by a couple of quite heterogeneous empirical properties. Separation by intonational breaks from the surrounding constituents is the principle characteristic of this type of constructions (in written text typographical means like dashes and parentheses function as surrogates). Parentheticals are optional; this means that obligatory constituents of a clause cannot be parenthetically separated from their co-constituents. Parentheticals express a comment by the speaker on the content of the hosting clause. Furthermore, it is often assumed that parentheticals are not fully integrated into the syntactic structure of the hosting clause. This diversity of qualifications points to the difficulties of giving a concise conclusive explanation of this type of construction. On the other hand, parentheticals occur quite frequently in corpora.¹ They thus deserve some attention.

In the following, I will try to shed some light on the question to what extent a genuine syntactic account of parenthetical constructions is possible, as well as appropriate. Although the empirical base of this study is restricted to German, some of the theoretical results may be helpful to account for parenthetical constructions in other languages as well. But some caution required. Parentheticals may exhibit rather language specific properties with respect to their internal structure, as well as to their distribution.

After a short survey of different instances of parenthetical constructions, I will concentrate on sentential parentheticals in order to account for a certain type of parenthetical construction, which I will refer to as *syntactic parentheticals*.

2 Parentheticals in German – Category and Distribution

If we adopt the prosodic and pragmatic definition as a means of detecting parentheticals in a clause, it seems to be hopeless to also provide for a consistent *syntactic* characterization of all the strings which are identified.

To begin with, there is no restriction with respect to syntactic category. Any type of maximal category may be inserted parenthetically into (a constituent of) a clause (cf. (1) examples extracted from the TIGER-corpus).² The same holds for the determination of grammatical func-

* I wish to thank Judith Berman, Martin Forst, Hans Kamp, Christian Rohrer, and the participants of the *Generative Grammatik des Südens* meeting 2005 in Tübingen and of the LFG05 conference in Bergen for hints, discussion and comments.

¹ About 4 – 5% of the sentences in the TIGER-Corpus, which comprises 40.000 sentences of German newspaper text contain a string which is tagged as a parenthetical.

² There is no example of a parenthetical formed by an adjective phrase (AP) in (1). This is due to the fact that these parentheticals are all clause level in the sentences they are excerpted from. Parentheticals formed by attributive APs are common within NPs.

tion, at least in principle. Adjuncts and adjectival modifiers are appropriate candidates par excellence (cf. (2a-c), (2e)), but even governable grammatical functions may provide a parenthetical. Of course, this is only possible if the respective function is optional. Mostly this is not the case, but there are exceptions (cf. (2d), (2f))

- | | | |
|-----|---|-----------------|
| (1) | a. ... (etliche sind allerdings noch unbesetzt) ... <i>some are however still not filled</i> | CP/V2-clause |
| | b. ... – das sind Zinszahlungen über den Bankschalter in bar – ... <i>that are payments of interest across the bank counter in cash</i> | CP/V2-clause |
| | c. ... (seien sie struktureller oder anderer Art) ... <i>be they structural or other kind</i> | CP/V1-clause |
| | d. ... die italienische Kriegsmarine, auf Flottenbesuch in Venedig, ... <i>the italian navy on fleet visit in Venice</i> | PP |
| | e. ... 1987, nach Marta Feuchtwangers Tod, ... <i>1987 after Marta Feuchtwangers death</i> | PP |
| | f. ... – unter ihnen viele Obstbauern – ... <i>among them many fruit farmers</i> | PP ³ |
| | g. ... (davon 80 Prozent mit öffentlichen Kunden) ... <i>that-of 80 percent with public customers</i> | PP |
| | h. ..., weltweit ein Spitzenreiter, ... <i>worldwide a leader</i> | NP |
| | i. ..., gleichsam ein Nono aus Fernost, ... <i>virtually a Nono from far east</i> | NP |
| | j. ... , jetzt zumeist als Genossenschaft oder GmbH geführt, ... <i>now mostly as cooperative or Ltd. led</i> | VP |
| | k. ..., von Klose selbst bei hitzigen Bundestagsdebatten bloß als Halt für den rechten Oberarm eingesetzt, ... <i>by Klose even in heated parliamentary debates merely as holder for the right upper arm employed</i> | VP |
| (2) | a. Theo hat die Tür – mit einem Dietrich – geöffnet <i>Theo has the door – with a pass key – opened</i> | PP adjunct |
| | b. Theo ist – ohne anzuklopfen – eingetreten <i>Theo is – without knocking – entered</i> | VP adjunct |
| | c. Theo ist – als es zu regnen begann – eilig nach Hause gelaufen <i>Theo is – when it to rain began – quickly to home run</i> | CP adjunct |
| | d. Theo hat reichlich – insbesondere Burgunder – getrunken <i>Theo has plenty – in particular Burgundy wine – drunk</i> | NP argument |
| | e. einen – gewiß vermeidbaren – Fehler habe ich gestern gemacht <i>a – certainly avoidable – mistake have I yesterday made</i> | AP modifier |
| | f. ein – auf seine Verfehlungen – stolzes Individuum <i>a – on his offences – proud individual</i> | PP argument |

Parenthetical placement is also quite liberal in German. A parenthetical may be inserted at clause level (cf. (2a-d)) or at constituent level (cf. (2e/f)). There is only one condition that has to be observed: The parenthetical must be preceded by at least one constituent of the hosting clause, it need not be an immediate one.

³ The categorization of the parenthetical as a PP in this and the following example may be problematic, since the parenthetical consists of a PP and a NP which functions as the subject of a predicate formed by the PP.

But even if an empirical scenario like the one sketched in (1) and (2) may cast doubt on an attempt to give an overall syntactic characterization of what is conceived of as a *parenthetical*, it is reasonable to consider whether every string of terminal elements that may form a parenthetical may also occur as a regular constituent of a clause or as its regular subconstituent in a non-parenthetical context. From this point of view some non-trivial questions arise, since there are actually certain constructions in German which prohibit a non-parenthetical use of a parenthetical string. Before going into details of an analysis, the notion of a *regular constituent* must be clarified. This clarification amounts to both conditions of structural representation and some empirical means by which a non-regular constituent can be detected.

3 Regular and Non-Regular Constituency

A *regular constituent* is a string of elements with properties, which are reasonably attributed to conditions of syntactic structure representation.

The first criterion that comes to mind is *inclusion*. A string, for instance, which is preceded and followed by a constituent of a clause is also a constituent of that clause. This, of course, is only a sufficient but not a necessary condition, since it does not hold for the first and the last constituent.

In a language like German, which permits some variation in constituent placement at clause level, the option of variable placement of a string is a second criterion of constituency. In declarative main clauses, the position in front of the finite verb (the so called *Vorfeld*) must be filled by – exactly – one constituent. Otherwise the clause is restricted to interrogative mood. Apart from some phonologically motivated exceptions, any constituent that occurs in a position following the finite verb can also fill the position before. Hence, in (3) the two phrases *Theo* and *die Tür* are qualified as constituents of the clause.

- (3) a. Theo **hat** die Tür geöffnet
 Theo has the door opened
 b. die Tür **hat** Theo geöffnet
 the door has Theo opened

Capability of placement in the pre-finite position may thus count as a second characteristic of a regular constituent of a clause.

On the other hand, certain conditions on interpretation are determined by syntactic structure, namely scope relations between elements or phrases and the bound variable reading of a pronoun in the context of a quantified NP.

For certain scope relations to hold between two strings, these strings must be represented as part of the overall syntactic structure of a clause. The specific content of the relation is determined by the structural relation between the scope sensitive elements within the hosting structure and by their lexical content. So for instance, the scope of the negative element *nicht* in (4) comprises the NP *die Tür* only if it precedes the latter, as in (4a) but not in (4b).

- (4) a. Theo hat **nicht** die Tür geöffnet Neg > OBJ
 Theo has not the door opened
 b. Theo hat die Tür **nicht** geöffnet *Neg > OBJ
 Theo has the door not opened

Likewise scope relations between different quantified NPs as well as a bound variable reading of a pronoun are determined by syntactic structure.

Constituent placement, scope of negation and the availability of a bound variable reading of a pronoun will be used in the following sections as a means to determine the degree of syntactic integration into a host clause. A given string is regarded as a *regular constituent* if it matches all the relevant empirical conditions. A *non-regular constituent*, on the other hand, is a string which only partially meets them. Finally, a *non-constituent* is one that does not match any.

If this tripartite differentiation has substantive content, the question arises of how to structurally represent *non-regular constituents*. The two other cases are unproblematic.

In a multi-level representation of syntactic structure as postulated in Lexical Functional Grammar, the aforementioned aspects of syntactic structure are modelled at the two levels of representation: c-structure and f-structure.⁴

A *regular constituent* of a clause may be conceived of as a string that is part of the c-structure as well as the f-structure representation of its hosting clause. A *non-regular constituent*, on the other hand, is a string which is only part of the c-structure of the host but not of its f-structure. In the following, I restrict myself to clause level phenomena of regular and non-regular constituency. An extension to constituent level calls for some further refinements, but is nevertheless possible.

4 Regular Constituent Parenthetical

We may now consider the examples from section 2 once again. As a matter of fact, all parentheticals are surrounded by constituents of their host – setting apart for the moment cases in which they immediately follow a clause. Due to inclusion, they are at least non-regular constituents. In a large number of cases, the parenthetical is also a regular constituent of the hosting clause. This holds for the examples (2a) – (2d) which exhibit all of the relevant properties. (5) shows placement in the *Vorfeld*.

- (5) *Occurrence in pre-finite-position*⁵
- a. *mit einem Dietrich hat* Theo die Tür geöffnet
with a pass key has Theo the door opened
 - b. *ohne anzuklopfen ist* Theo eingetreten
without to knock is Theo entered
 - c. *als es zu regnen begann ist* Theo eilig nach Hause gelaufen
when it to rain began is Theo quickly to home run
 - d. *insbesondere Burgunder hat* Theo reichlich getrunken
in particular burgundy wine has Theo plenty drunk

The facts about scope of negation and variable binding are illustrated by the PP-example (2a); the other cases pattern alike.

⁴ If more than the two fundamental levels – CS and FS – are postulated, some of the phenomena may be accounted for with reference to other levels of representation.

⁵ The fact that a given string of lexical elements which otherwise can be inserted as a parenthetical may occupy the pre-finite position in a clause (i.e. the SpecCP position) does not mean that a parenthetical reading is also possible in the latter case. It seems to be a general condition on parentheticals that at least one constituent of the host – not necessarily an immediate one – precedes them. A parenthetical reading, then, emerges as an option for optional constituents in a suitable environment.

- (6) *Scope of Negation*
 a. Theo hat **nicht** – mit einem Dietrich – die Tür geöffnet Neg>Adjunct, Object
Theo has not – with a pass key – the door opened
 b. Theo hat – mit einem Dietrich – **nicht** die Tür geöffnet Adjunct, Neg>Object
Theo has – with a pass key – not the door opened
- (7) *Variable Binding*
 jeder/niemand_i hat – mit seinem_i Dietrich – die Tür geöffnet
everyone/no one has – with his pass key – the door opened

With respect to the syntactic representation of these examples, we have to assume that the parenthetical phrase is not only part of the c-structure of the matrix clause, but that it is also mapped onto an f-structure that serves as the value of an attribute of the host's f-structure. In the case of (2a) – (2c) this is an adjunct. Parenthetical optional objects as in (2d) are integrated into the f-structure of the governing predicate anyway, since they are subject to the coherence condition.

Now, if the c- to f-structure mapping of this type of parentheticals does not differ from their non-parenthetical counterparts, both would remain indistinguishable – an unsatisfactory result in the face of the peculiarity of interpretation and pragmatic use of the parenthetical. But this can be remedied if we make the plausible assumption that parenthetical marking – by means of prosody or by typographic marks – is expressed by a suitable attribute at f-structure so that the relevant information is available for further processing.

5 Sentential Parentheticals

Sentential parentheticals make up the next type to be considered. At first, we have to state that all three variants of possible clause structure in German may also form licit parentheticals. Verb-second and verb-first clauses, which exhibit the canonical structure of an independent main clause, and verb-final clauses, the canonical exponent of subordinated clauses, may be parenthetically inserted into a hosting clause. (8) and (9) show a verb-second and a verb-first clause.

- (8) Theo hat – er **ist** Klempner – die Heizung im Handumdrehen repariert
Theo has – he is plumber – the heating in a jiffy fixed
- (9) Theo hat – **ist** er Klempner? – die Heizung im Handumdrehen repariert
Theo has – is he plumber? – the heating in a jiffy fixed

In both cases shown in (8) and (9), c- to f-structure mapping of the parenthetical string on its own brings about a complete and coherent f-structure. The parenthetical clause, therefore, may be used independently. The interpretation of the parenthetical likewise parallels that of the non-parenthetical clause. The parenthetical's mood is declarative in (8) and interrogative in (9). This has to be borne in mind because of cases that do not fit into this pattern and to which we will return immediately.

Verb-final clauses that function as adjuncts permit parenthetical insertion into a clause, too.

- (10) Theo hat – als es plötzlich kalt **wurde** – die Heizung im Handumdrehen repariert
Theo has – when it suddenly cold got – the heating in a jiffy fixed

In principle, a clausal parenthetical may be headed by any verbal predicate. Nevertheless, certain peculiar properties of the construction emerge if the predicate governs a COMP-function. The facts to be observed are independent of clause structure with respect to the position of the finite verb. Verba dicendi may illustrate the case.

- (11) Theo hat – er sagte **daß es kalt würde** – die Heizung im Handumdrehen repariert
Theo has – he said that it cold would get – the heating in a jiffy fixed

In (11) – a verb-second clause – the COMP-function governed by the parenthetical verb *sagte* is provided by a finite complement clause. Instead of a clause, a pronominal may occur as in (12).

- (12) a. Theo hat – er sagte **es** mir später – die Heizung im Handumdrehen repariert
Theo has – he told it me later – the heating in a jiffy fixed
 b. Theo hat – so sagte er **es** mir später – die Heizung im Handumdrehen repariert
Theo has – so told he it me later – the heating in a jiffy fixed

In constructions like (12) the interpretation of the pronominal is strictly restricted to coreference with the hosting clause. This restriction may be due to pragmatic reasons. Since a deictic interpretation is hard to construe, the pronominal must corefer with some suitable antecedent. In (12) the host is the next available candidate. This restriction on interpretation does not preclude the parenthetical clause from independent use as long as the discourse provides for an antecedent of the pronoun.

In (12) the interpretation of the host clause as an argument of the parenthetical's verbal predicate is syntactically mediated by the pronominal complement *es* of the parenthetical verb. We finally have to consider a third case of clausal parentheticals with more or less the same interpretation as (12). This kind of construction lacks an overt c-structure representation of the verbs complement within the parenthetical string. Instead, the host clause by itself seems to provide the complement directly. This type results from erasing the pronouns in (12).

- (13) a. Theo hat – **sagte** er mir später – die Heizung im Handumdrehen repariert
Theo has – told he me later – the heating in a jiffy fixed
 b. Theo hat – so **sagte** er mir später – die Heizung im Handumdrehen repariert
Theo has – so told he me later – the heating in a jiffy fixed

The mood of the parenthetical clause is declarative in both cases. In the case of (13b) this follows from the verb-second structure of the parenthetical. But the same holds true of (13a) despite the fact that this parenthetical has the shape of a verb-first clause which otherwise determines interrogative mood.⁶

As mentioned before, the peculiar relation between the host and the parenthetical's predicate illustrated in (12) and (13) does not depend on the internal clause structure of the parenthetical. It also obtains in verb-final clauses as in (14).

- (14) a. Theo hat – wie er mir später **sagte** – die Heizung im Handumdrehen repariert
Theo has – as he me later told – the heating in a jiffy fixed
 b. Theo hat – wie er *es* mir später **sagte** – die Heizung im Handumdrehen repariert
Theo has – as he it me later told – the heating in a jiffy fixed

⁶ Since the principle interest of this work is to elucidate the syntactic relation of the parenthetical to its host, I will not further investigate the structural conditions that determine declarative mood in verb-first clauses like that in (13a).

For convenience, clausal parentheticals which contain all constituents that are necessary for projecting a locally complete f-Structure will furthermore be called *internally complete*, those which do not will be called *internally non-complete*.

After this survey of possible instances of clausal parentheticals, let us come back to the question of their status with respect to syntactic integration into the host. As before, we will apply occurrence in pre-finite position, scope of negation and variable binding as tests. In all cases to be examined, the criterion of *inclusion* is trivially satisfied, because the parenthetical string can be preceded and followed by integral parts of the host (cf. (8) – (14)).

5.1 Non-Regular Constituent Parenthetical – Internally Complete

Verb-second parenthetical

Verb-second parentheticals do not pass any of the three tests. They are excluded from pre-finite position in a clause (cf. (15)), they are exempt from the scope of negation (16) and a quantifier in the host cannot bind a variable within the parenthetical (17).

(15) *Occurrence in pre-finite position*

- a. Theo **hat** – *ein Klempner war nicht zu erreichen* – den Rohrbruch selbst repariert
Theo has – a plumber was not to reach – the pipe burst self fixed
- b. **ein Klempner war nicht zu erreichen* **hat** Theo den Rohrbruch selbst repariert
a plumber was not to reach has Theo the pipe burst self fixed

(16) *Scope of negation*

- a. Theo hat **nicht** – er ist kein Klempner – den Rohrbruch selbst repariert parenth>Neg
Theo has not – he is no plumber – the pipe burst self fixed
- b. Theo hat – er ist kein Klempner – **nicht** den Rohrbruch selbst repariert parenth>Neg
Theo has – he is no plumber – not the pipe burst self fixed
- c. Theo hat – er ist kein Klempner – den Rohrbruch **nicht** selbst repariert parenth>Neg
Theo has – he is no plumber – the pipe burst not self fixed

(17) *Variable binding*

- (fast) jeder/keiner_i wird – er_i ist Klempner – nur aus Vergnügen arbeiten⁷
(nearly) everyone/no one will – he is plumber – only for fun work

⁷ There are some exceptions from the ban on variable binding. So for instance (i) is grammatical with a bound variable reading of the pronominal.

- (i) (fast) jeder/keiner_i wird – er_i mag Klempner sein oder nicht – nur aus Vergnügen arbeiten
(nearly) everyone/no one will – he may plumber be or not – only for fun work

Variable binding in cases like (i) depends on specific conditions. Subjunctive mood of the verb within the parenthetical or the occurrence of the modal verb *mögen* 'may' makes it possible. But this is not specific to a parenthetical construction with a verb second parenthetical. Apart from the parenthetical construction, variable binding also becomes possible in the case of two successive clauses.

- (ii) Keiner wird nur aus Vergnügen arbeiten. Er mag Klempner sein oder nicht
no one will only for fun work. He may plumber be or not

Variable binding in cases like (i) and (ii), which may be considered instances of modal subordination, calls for an explanation independent of the string's status as a parenthetical.

From these facts, we can conclude that the parenthetical string of a verb-second clause cannot be a *regular constituent*. Since, on the other hand, the parenthetical string can be included in the host (cf. (8)), it has the status of a *non-regular constituent*.

Verb-first parenthetical

Internally complete verb-first parentheticals pattern like the verb-second cases. They are excluded from the pre-finite position in a clause.

(18) *Occurrence in pre-finite position*

- a. Theo – *hat er denn überhaupt das nötige Werkzeug?* – **will** den Rohrbruch
Theo – has he_{D-PART} at all the necessary tool? – wants the pipe burst
 selbst reparieren
self fix
- b. **hat er/Theo denn überhaupt das nötige Werkzeug?* **will** Theo/er
has he/Theo_{D-PART} at all the necessary tool? wants Theo/he
 den Rohrbruch selbst reparieren
the pipe burst self fix

A negative element in the host does not have scope over the parenthetical clause.

(19) *Scope of negation*

- a. Theo will **nicht** – habe ich recht? – den Rohrbruch selbst reparieren
recht>Neg/*Neg>recht
Theo wants not – have I right? – the pipe burst self fix
- b. Theo will – habe ich recht? – **nicht** den Rohrbruch selbst reparieren recht>Neg
Theo wants – have I right? – not the pipe burst self fix

A bound variable reading of a pronoun within the parenthetical is not available for binding by a quantifier within the host.

(20) *Variable binding*

- (fast) jeder/keiner_i wird – ist er*_i Klempner? – nur aus Vergnügen arbeiten
(nearly) everyone/no one will – is he plumber – only for fun work

Hence, verb-first parentheticals are also *non-regular constituents* of the hosting clause.

5.2 Non-Regular Constituent Parenthetical – Internally Non-Complete

Verb-second and verb-first parentheticals that are internally non-complete in the sense that they do not enclose a constituent which is mapped onto the COMP function of the verbal predicate exhibit the same restrictions with respect to placement, scope and variable binding as their complete counterparts discussed in section 5.1. The relevant facts are shown in (21) – (23).

Verb-second parenthetical

- (21) *Occurrence in pre-finite position*
- a. Theo **hat** – so sagte man mir – den Rohrbruch im Handumdrehen repariert
Theo has – so said one me – the pipe burst in a jiffy fixed
- b. *so sagte man mir **hat** Theo den Rohrbruch im Handumdrehen repariert
so said one me has Theo the pipe burst in a jiffy fixed
- (22) *Scope of negation*
- a. Theo hat **nicht** – so sagt man – selbst den Rohrbruch repariert sagt>Neg/*Neg>sagt
Theo has not – so says one – self the pipe burst fixed
- b. Theo hat – so sagt man – **nicht** selbst den Rohrbruch repariert sagt>Neg
Theo has – so says one – not self the pipe burst fixed
- (23) *Variable binding*
- a. *jeder/keiner_i wird – so sagt er_i – nur aus Vergnügen arbeiten
everyone/no one will – so says he – only for fun work
- b. *jeder/keiner_i wird – so sagt man ihm_i – nur aus Vergnügen arbeiten
everyone/no one will – so tells one him – only for fun work

Verb-first parenthetical

- (24) *Occurrence in pre-finite position*
- a. Theo **hat** – sagte man mir – den Rohrbruch im Handumdrehen repariert
Theo has – said one me – the pipe burst in a jiffy fixed
- b. *sagte man mir **hat** Theo den Rohrbruch im Handumdrehen repariert
said one me has Theo the pipe burst in a jiffy fixed
- (25) *Scope of negation*
- a. Theo hat **nicht** – sagt man – selbst den Rohrbruch repariert sagt>Neg/*Neg>sagt
Theo has not – says one – self the pipe burst fixed
- b. Theo hat – sagt man – **nicht** selbst den Rohrbruch repariert sagt>Neg
Theo has – says one – not self the pipe burst fixed
- (26) *Variable binding*
- a. *jeder/keiner_i wird – sagt er_i – nur aus Vergnügen arbeiten
everyone/no one will – says he – only for fun work
- b. *jeder/keiner_i wird – sagt man ihm_i – nur aus Vergnügen arbeiten
everyone/no one will – tells one him – only for fun work

In sum, we have to state that any occurrence of a verb-second or a verb-first parenthetical is only partially integrated into the syntactic structure of the hosting clause. Parentheticals with main clause structure thus have to be considered non-regular constituents.

5.3 Regular Constituent Parenthetical – Internally Complete

Parentheticals formed by ordinary adverbial verb-final clauses have already been addressed in section 2 at least partially. Apart from inclusion in the matrix clause, verb-final clauses may occupy the pre-finite position in main clauses (cf. (27)). They show scope interaction with nega-

tion in the matrix clause (cf. (28)) and the interpretation of a pronoun as a variable bound by a quantified NP in the matrix clause is also possible (cf. (29)).

(27) *Occurrence in pre-finite position*

- a. Theo **kam** – *als es regnete* – mit einem großen Schirm
Theo came – *when it rained* – with a big umbrella
- b. *als es regnete* **kam** Theo mit einem großen Schirm
when it rained came Theo with a big umbrella

(28) *Scope of negation*

- a. Theo kam **nicht** – *als es regnete* – mit einem großen Schirm
Neg>regnete/regnete>Neg
Theo came not – *when it rained* – with a big umbrella
- b. Theo kam – *als es regnete* – **nicht** mit einem großen Schirm regnete>Neg
Theo came – *when it rained* – not with a big umbrella

(29) *Variable binding*

- a. jeder_i wird – wenn er_i etwas wissen will – fragen
everyone will – *if he something know wants* – ask
- b. keiner_i wird – wenn man ihn_i fragt – antworten
no one will – *if one him asks* – answer

Verb-final parenthetical clauses, therefore, are regular constituents of a clause.

5.4 Regular Constituent Parenthetical – Internally Non-Complete

The second type of verb-final parenthetical clause is instantiated by the *wie*-parenthetical which corresponds to the *as*-parenthetical in English. With respect to the syntactic realization of the verb's complement function, this construction is on a par with the verb-second and verb-first parentheticals discussed in section 5.2. It crucially differs from the latter in all aspects concerning syntactic integration into the host. The string formed by the *wie*-parenthetical may occur in pre-finite position (cf. (30)) – a parenthetical reading is not available in this case. A negative element in the matrix clause may get scope over them (cf. (31)) and variable binding is possible (cf. (32)).

(30) *Occurrence in pre-finite position*

- a. Theo **hat** – *wie Paul sagt* – den Rohrbruch im Handumdrehen repariert
Theo has – *as Paul says* – the pipe burst in a jiffy fixed
- b. *wie Paul sagt* **hat** Theo den Rohrbruch im Handumdrehen repariert
as Paul says has Theo the pipe burst in a jiffy fixed

(31) *Scope of negation*

- a. Theo hat **nicht** – wie man mir sagte – selbst den Rohrbruch repariert
Neg>sagte/sagt>Neg
Theo has not – *as one me said* – self the pipe burst fixed
presupposition: someone said: Theo has fixed the pipe burst by himself
- b. Theo hat – wie man mir sagte – **nicht** selbst den Rohrbruch repariert sagte>Neg
Theo has – *as one me said* – not self the pipe burst fixed
presupposition: someone said: Theo has not fixed the pipe burst by himself

(32) *Variable binding*

- a. jeder/keiner_i wird – wie er_i sagen mag – nur aus Vergnügen arbeiten
everyone/no one will – as he say may – only for fun work
- b. jeder/keiner_i wird – wie man ihm_i sagen mag – nur aus Vergnügen arbeiten
everyone/no one will – as one him tell may – only for fun work

wie-parentheticals, like other verb final parenthetical clauses, form regular constituents of the matrix clause.

5.5 Parenthetical and Non-Parenthetical Clause Sequences

As mentioned in footnote 5, a parenthetical string must be preceded by at least one constituent of the hosting clause. In the cases discussed so far, there is also some constituent of the host which follows it. Instead of being encapsulated in a clause, a parenthetical string may also follow its host immediately. The prosodic as well as the pragmatic properties of the core case of the parenthetical construction may be conserved. Furthermore, a sequence of those two clauses may be uttered by two different speakers in a discourse. In the latter case it is unreasonable to assume that both clauses are part of one and the same syntactic representation. On the other hand, the parenthetical construction and the corresponding sequence of host and parenthetical clauses pattern alike with respect to scope of negation and variable binding. Scope of negation is restricted to the preceding clause.

(33) *scope of negation*

- A: Theo hat nicht den Rohrbruch selbst repariert
Theo has not the pipe burst self fixed
- B: er hatte das Werkzeug vergessen
he had the tool forgotten
- B': so sagt er
so says he

A pronoun in the second clause does not allow a bound variable reading.

(34) *variable binding*

- A: jeder/keiner_i wird nur aus Vergnügen arbeiten
everyone/no one will only for fun work
- B: *er_i ist Klempner
he is plumber

(33) and (34) show the relation that regularly holds between independent clauses. These parallels between the parenthetical construction and a sequence of clauses is a further indicator of the syntactic independence of the parenthetical.

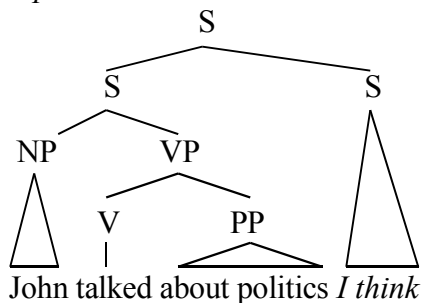
As a result of the preceding sections, we can state that parentheticals differentiate into two classes. One type subsumes strings of terminals which otherwise permit a non parenthetical use and which show the properties of other regular constituents of a clause. The other type subsumes strings – verb second and verb first clauses, as investigated so far – which cannot form a regular constituent of a clause. Parentheticals of the latter type may be conceived of as *syntactic parentheticals* as distinguished from mere prosodic parentheticals.⁸

⁸ The examples of verb-second and verb-first clauses discussed so far represent the core cases of syntactic parentheticals. Another quite frequent type is (i)

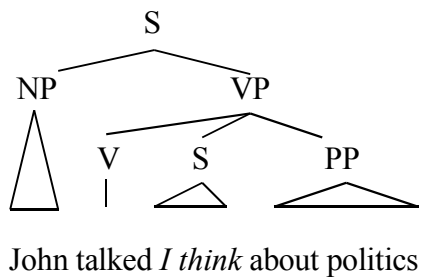
6 Previous Analyses

Current analyses of parenthetical constructions face a fundamental dilemma. Obvious aspects of syntactic disintegration like those illustrated in the previous sections have to be captured within a completely integrated structural description of these facts, since the parenthetical and its host are equally subordinated to one root node in the phrase structure representation of the whole construction. (cf. Ross (1973), Emonds (1976), (1979), McCawley (1981), Potts (2002), Stowell (2003)). Interpretative effects like the escape from scope and the blocking of a bound variable reading can be attributed to adjunction of the parenthetical string at the level of syntactic representation that feeds the interpretation of the whole construction. The superficial inclusion of the parenthetical in the host, then, is explained by movement of either the parenthetical into the host (cf. Ross (1973)) or of parts of the host into a position following that of the adjoined parenthetical (cf. Emonds (1979)).

(35) *Input*



(36) *Output: Ross (1973)*



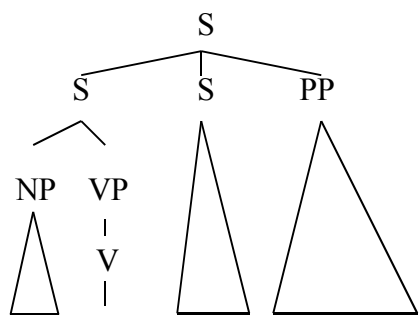
-
- (i) heute hat – so ein Nachbar – Theo die Heizung repariert
today has – so a neighbour – Theo the heating fixed

The parenthetical which lacks a verbal predicate must contain the adverb *so* as its initial constituent and it has roughly the same interpretation as (ii)

- (ii) heute hat – so sagte ein Nachbar – Theo die Heizung repariert
today has – so said a neighbour – Theo the heating fixed

This type of parenthetical raises a couple of intricate questions which cannot be pursued here.

(37) Output: *Emonds (1979)*



John talked *I think* about politics

In a third variant of analysis McCawley (1982) assumes that all constituents are base generated in place. By allowing crossing edges in the tree, the parenthetical is not dominated by the S-node that immediately dominates the host, but is adjoined to it.

While it might be possible to capture the interpretative properties of the parenthetical construction by adjunction of the parenthetical string, the restrictions on its distribution in German do not result from these analyses. There is no plausible motivation for the assumption that a string which is adjoined to a clause should be banned from movement into the pre-finite position of that clause if, at the same time, movement into a clause internal position is postulated to be possible and even necessary.

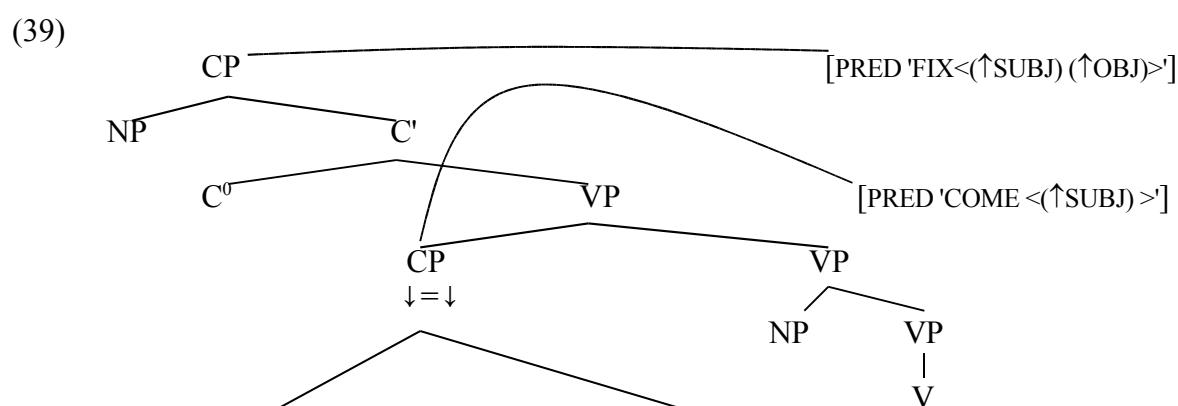
If, on the other hand, the parenthetical string is fixed in its adjoined position immediately dominated by the root node and the superficial constituent order is attributed to movement of some material from within the host, several well established general restrictions on movement cannot be maintained. Since the parenthetical may occur between the first constituent and the finite verb in main clauses, rightward movement of an intermediate C'-projection would have to be assumed. And even non-constituent movement would have to be considered because parenthetical insertion is possible at the right edge of nearly any constituent of a clause in German.

7 An LFG-Account

A satisfactory analysis of the parenthetical construction calls for a solution of the dilemma just sketched. One way out of it would be the dissociation of the syntactic representation of the parenthetical and its host in the case of a syntactic parenthetical. Of course, such a dissociation cannot be absolute, since the parenthetical is enclosed into the string of terminal elements forming the host. A separated structural representation of the parenthetical and its host must be achieved under the condition of their integration in a terminal string. Espinal (1994) offers an account in this spirit.⁹ Lexical-Functional Grammar with its co-representation of syntactic structure, on the other hand, provides a natural means to cope with this task. The level of c-structure can serve as the locus of integration of the parenthetical into its host, whereas at the level of f-structure the representations of the respective clauses are separated. The empirical facts that give rise to a disjunction of the syntactic representation of the parenthetical and its host (scope and variable binding) can be modelled in terms of f-structure configurations (cf. Bresnan (2001), Dalrymple (2001)). The question, then, is how to project two independent non-integrated f-structures from an integrated c-structure.

⁹ According to her analysis, the host and the parenthetical form two different phrase structure trees which are not dominated by a unique root node. This account is applied to any occurrence of a parenthetically marked string. It is not restricted to cases of non-regular constituents in the sense developed above.

Every c-structure node is mapped onto an f-structure which, in the first place, bears no relation to the f-structure of any other node within that c-structure. An integrated f-structure emerges from two fundamental operations, namely unification of two f-structures and function application by which an f-structure is introduced as the value of an attribute of another f-structure. Both operations are mediated by functional annotation of either a trivial equation $\uparrow = \downarrow$ or an equation of the form $(\uparrow GF) = \downarrow$. In the case of a syntactic parenthetical, neither annotation is suitable. Since by convention the annotation of a trivial equation is presumed if a c-structure node lacks an explicit annotation, we need an explicit annotation of the parenthetical node. An equation of the form $\downarrow = \downarrow$ is convenient in this case. Like any other node, the parenthetical node is mapped onto an f-structure. As a consequence of the annotation, this f-structure is not unified with any other f-structure (with the exception of the parenthetical's head) nor is it introduced as the value of any attribute. (39) illustrates the case of an internally complete verb-second parenthetical.¹⁰



Theo hat – der Klempner war nicht gekommen – die Heizung repariert
Theo has – the plumber had not come – the heating fixed

The interpretation of the syntactic parenthetical, namely its exclusion from the scope of a negative element in the host and the unavailability of a bound variable reading of a pronoun inside the parenthetical, now follow in a straightforward way.

7.1 Internally Complete Parenthetical

Scope of negation can be modelled by means of F-precedence.¹¹ The negative element must F-precede a constituent in order to get scope over it. F-precedence is defined as follows (cf. Bresnan (2001)):

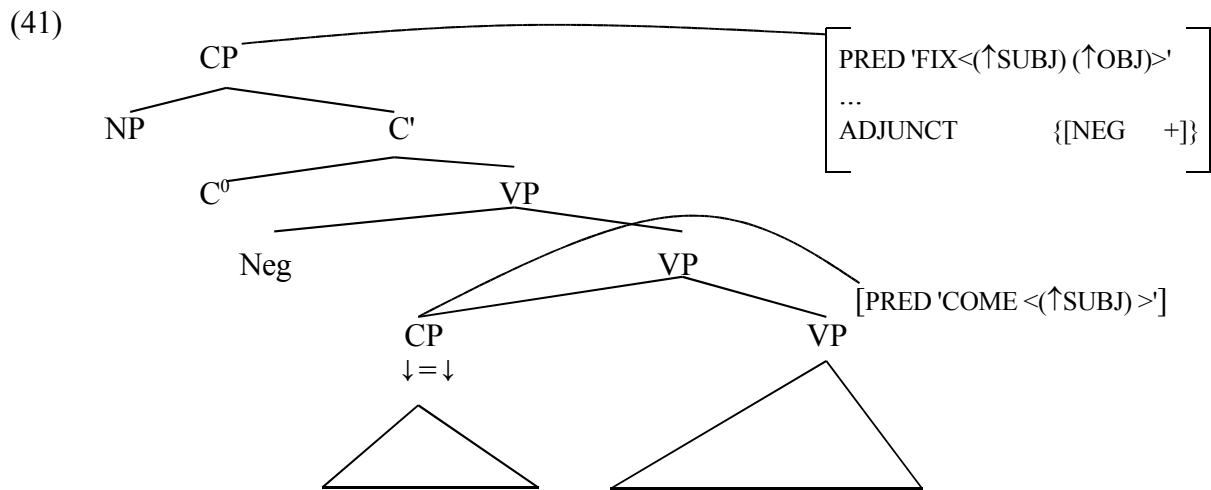
(40) *F-precedence*

Given a correspondence mapping ϕ between a CS and its FS, and given two subsidiary f-structures α and β , α F-precedes β if the rightmost node in $\phi^{-1}(\alpha)$ precedes the rightmost node in $\phi^{-1}(\beta)$.

¹⁰ The c-structure of the host is modelled in the line of Berman (2003). A verb-second clause is represented by a CP functional category. The complement of its C-head is formed by a (recursive embedding of) VP. The parenthetical verb-second clause is adjoined to VP.

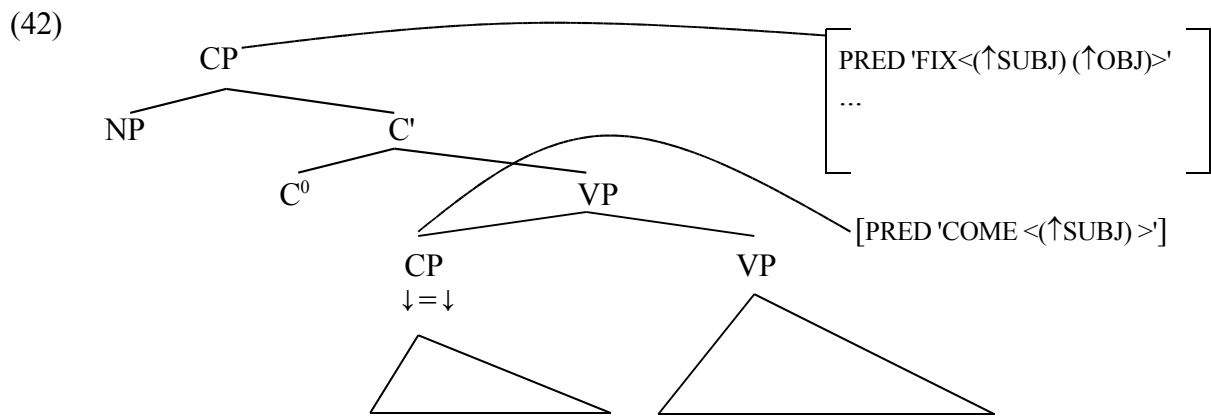
¹¹ The position of the negation relative to other constituents affects its scope; cf. the examples in (31).

For an F-precedence relation to hold between two f-structures, it is necessary that an f-structure is given which contains both of them. But there is none in the case of a syntactic parenthetical (cf. (41)). Hence, the parenthetical is excluded from the scope of the negation.



Theo hat nicht – er kam zu spät – den Rohrbruch selbst repariert
Theo has not – he came too late – the pipe burst self fixed

A pronominal is capable of a bound variable reading if it is enclosed in the domain of a potential binder. The domain of a binder is defined as the minimal f-structure containing the binder (cf. Bresnan (2001)). Since there is no f-structure which includes both the f-structures corresponding to the QNP and to the pronominal, the required structural relation does not hold.



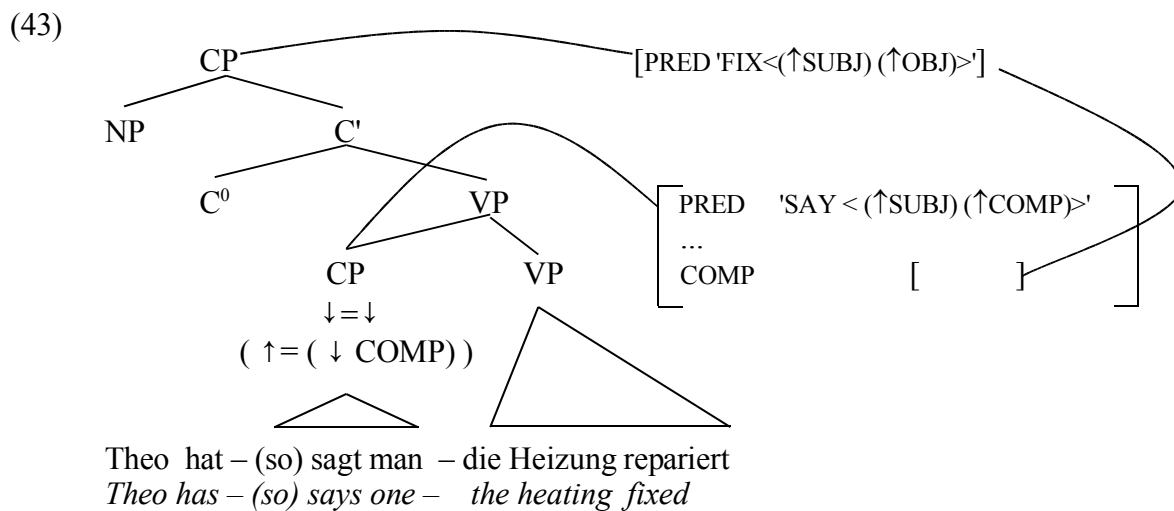
jeder/keiner_i wird – er_i ist kein Klempner – nur aus Vergnügen arbeiten
everyone/no one will – he is no plumber – only for fun work

The separation of the structural representations of the host and the parenthetical clause at the level of f-structure yields an account of the interpretation of the internally complete syntactic parenthetical.

7.2 Internally Non-Complete Parenthetical

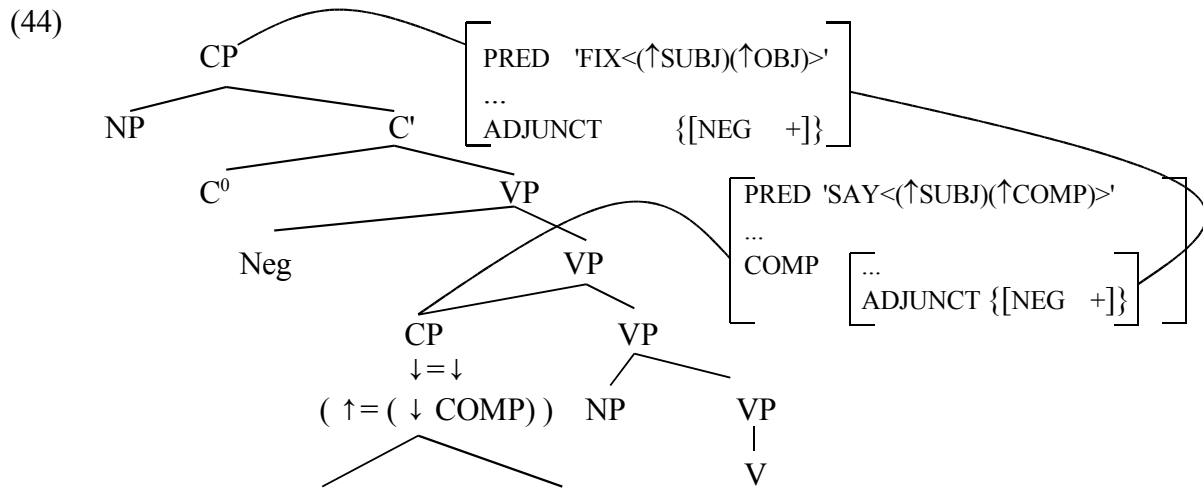
In the case of internally non-complete syntactic parentheticals, the fact has to be captured that the host is interpreted as the complement of the parenthetical's verb. In the ordinary case of complementation of a clause, a definition of the form $(\uparrow \text{COMP}) = \downarrow$ is annotated to the CP-node of the complement. This annotation is, of course, inappropriate for the parenthetical. On the one hand, it would contradict the annotation of the equation $\downarrow = \downarrow$, since the f-structure corresponding to the parenthetical ends up as the value of some COMP attribute hence integrated into an f-structure. On the other hand, the resulting f-structure would violate the completeness and coherence conditions, as the host's predicate does not govern a COMP function whereas the parenthetical's predicate does.

In a first approximation, we may optionally annotate the node that dominates the parenthetical string with $\uparrow = (\downarrow \text{COMP})$, as in (43), in order to meet the interpretation of the construction.



The restrictions on scope of negation and variable binding also follow from this mode of representation.

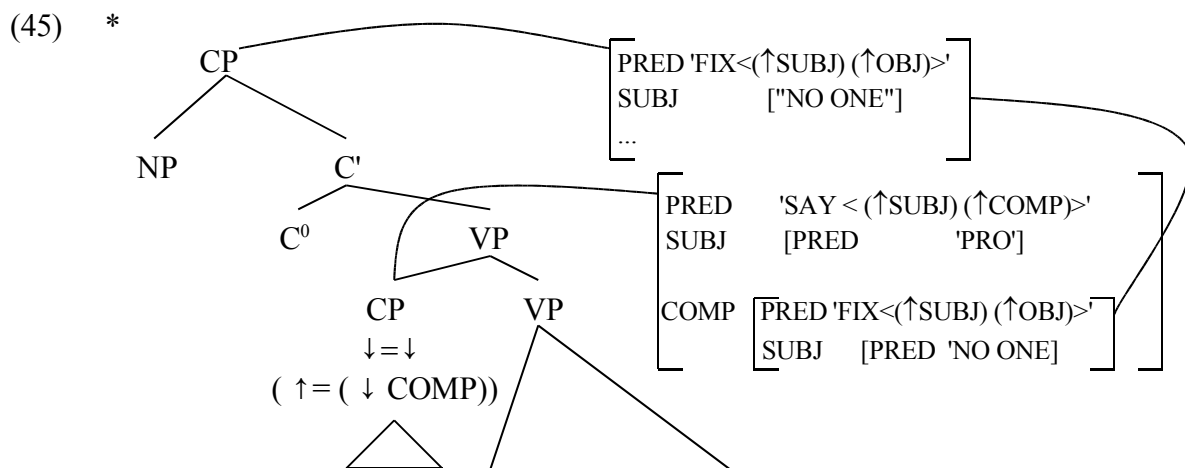
In the case of negation, the f-structure value of the COMP function contains an f-structure corresponding to the negative element.



Theo hat nicht – (so) sagt man – die Heizung repariert
Theo has not – (so) says one – the heating fixed

Since at c-structure the negation precedes the parenthetical the f-structure corresponding to the negation F-precedes the f-structure of the parenthetical. But F-precedence is not sufficient for the negation to get scope over the parenthetical. In general, the scope of a negative element which is enclosed in a complement clause cannot extend to the matrix clause.¹² Since the f-structure of the negation is embedded in the f-structure value of the parenthetical's COMP function, the parenthetical is excluded from its scope despite F-precedence.

Binding of a pronoun within the parenthetical by a quantified NP within the host is prohibited because the pronoun is not included in the domain of the quantifier if the f-structure of the host is copied into the parenthetical's f-structure as a value of the COMP function. The SUBJ within COMP does not outrank any function outside of COMP.



jeder/keiner_i hat – sagt er_i – die Heizung repariert
every-/no one has – says he – the heating fixed

From the proposed annotation of $\uparrow = (\downarrow \text{COMP})$ to the parenthetical node, it follows that the entire f-structure of the host is copied into the parenthetical's f-structure. This account is appro-

¹² A negative element within a complement clause that fills the *Vorfeld* position or that is left dislocated does not get scope over the matrix clause.

appropriate for the examples of internally non-complete syntactic parentheticals given so far. But, possibly, it is insufficient if we take into consideration some further instances of parenthetical constructions. So for instance, multiple embedding of parentheticals is possible as in (46).

- (46) heute hat Theo – sagt Paul – die Heizung – sagt Fritz – im Handumdrehen repariert
today has Theo – says Paul – the heating – says Fred – in a jiffy fixed

(46) is a possible statement in a context of identical statements by *Paul* and *Fred* respectively. But (46) is also licit as a summary of statements that are not completely identical. If *Paul* and *Fred*, referring to the same event, utter the sentences in (47), sentence (46) is a possible résumé of both.

- (47) a. Fritz: heute hat jemand im Handumdrehen die Heizung repariert
Fred: today has someone in a jiffy the heating fixed
 b. Paul: heute hat Theo irgendetwas im Handumdrehen repariert
Paul: today has Theo something in a jiffy fixed

The proposed annotation of the parenthetical, instead, would give rise to an interpretation that presupposes identical statements by *Paul* and *Fred* and hence is not compatible with a discourse like (47). This defect may be avoided if the complement of the parenthetical's predicate is represented by an empty pronominal, restricting the annotation of the parenthetical node to the equation $\downarrow = \downarrow$. The interpretation of the complement, then, is a matter of anaphora resolution.

7.3 Integrity of the Parenthetical string

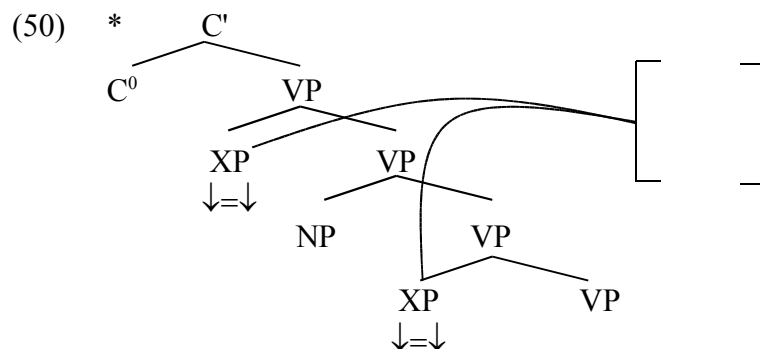
As illustrated by (46), multiple occurrences of syntactic parentheticals within one host is possible. Besides the case in (46), recursive embedding is equally possible.

- (48) Theo – er ist – sagen einige – in solchen Dingen geschickt – hat gestern
Theo – he is – say some – in such things skilled – has yesterday
 den Rohrbruch im Handumdrehen repariert
the pipe burst in a jiffy fixed

In the parenthetical construction, the host is a discontinuous constituent by definition. However, parts of the host cannot be interspersed among constituents of the parenthetical. The parenthetical cannot be discontinuous relative to its host, as shown in (49).

- (49) *Theo – er ist – hat gestern – ein versierter Klempner – im Handumdrehen
Theo – he is – has yesterday – a skilled plumber – in a jiffy
 den Rohrbruch repariert
the pipe burst fixed

Espinal (1991), who deals with this fact, explicitly stipulates a specific condition to exclude ungrammatical cases as in (49). In the analysis elaborated above, no stipulation is necessary. As a consequence of the annotation, no path into the host can be defined. That means that the f-structure of a given parenthetical node cannot be unified with any other f-structure. Hence, a c- to f-structure mapping as in (50) does not emerge.



A structure like (50), however, would be necessary to integrate the scattered parts of a discontinuous parenthetical within a complete and coherent f-structure.

8 Summary

Due to the distribution of the parenthetical string and due to the syntax-based conditions on its interpretation, two types of parenthetical constructions can be differentiated in German. Syntactically integrated parentheticals are strings which otherwise may form regular constituents of a clause. They are part of the c- and f-structure representation of the containing clause. Syntactically non-integrated parentheticals (*syntactic parentheticals*) are formed by strings which do not permit a non-parenthetical use and which diverge from regular constituents with respect to placement and interpretation. Syntactic parentheticals share the c-structure representation with their host, but they are mapped onto an independent f-structure that is not part of the f-structure of the host.

References

- Altmann, P., (1981) *Formen der 'Herausstellung' im Deutschen*. Tübingen: Niemeyer.
- Berman, J., (2003) *Clausal Syntax of German*. Stanford, California: CSLI.
- Emonds, J., (1976) *A Transformational Approach to English Syntax*. Academic Press.
- Emonds, J., (1979) Appositive Relatives Have no Properties. *Linguistic Inquiry* 10: 211–243.
- Espinal, M.T. (1991) The Representation of Disjunct Constituents. *Language* 67(4): 726–762.
- Grewendorf, G., (1988) *Aspekte der Deutschen Syntax*. Tübingen: Narr.
- Haider, H., (1993) ECP-Etüden. *Linguistische Berichte* 145: 185–203.
- McCawley, J., (1982) Parentheticals and Discontinuous Constituent Structure. *Linguistic Inquiry* 13(1): 91–106.
- Potts, C., (2002) The Syntax and Semantics of *As*-Parentheticals. *Natural Language and Linguistic Theory* 20(3): 623–689.
- Reis, M., (1995) *wer glaubst du hat recht?* on So-called Extractions from Verb-Second Clauses. *Sprache & Pragmatik* 36: 27–83.
- Ross, J., (1973) Slifting, in: M. Gross e.a. (eds.) *The Formal Analysis of Natural Languages*. 133–169.
- Stowell, T., (2003) Parenthetical (Phrase) Structure. Ms.

STRONG DOMAIN VARIATION AND TREEBANK-INDUCED LFG RESOURCES

John Judge, Michael Burke, Aoife Cahill, Ruth O'Donovan, Josef van Genabith, and Andy Way
National Centre for Language Technology and School of Computing,
Dublin City University, Dublin, Ireland

Proceedings of the LFG05 Conference
University of Bergen
Miriam Butt and Tracy Holloway King (Editors)
2005
CSLI Publications
<http://csli-publications.stanford.edu/>

Abstract

In this paper we present a number of experiments to test the portability of existing treebank-induced LFG resources. We test the LFG parsing resources of Cahill et al. (2004) on the ATIS corpus which represents a considerably different domain to the Penn-II Treebank Wall Street Journal sections, from which the resources were induced. This testing shows an underperformance at both c- and f-structure level as a result of the domain variation. We show that in order to adapt the LFG resources of Cahill et al. (2004) to this new domain, all that is necessary is to retrain the c-structure parser on data from the new domain.

1 Introduction

Probabilistic, treebank-based parsing resources (Collins, 1999; Charniak, 2000; Bikel, 2002) are of high quality and can be rapidly induced from appropriate treebank material. However, treebank- and machine learning-based grammatical resources reflect the characteristics of the training data. They generally underperform on test data substantially different from the training data. In this paper we investigate the effects of strong domain variation on the treebank-induced, “deep”, probabilistic Lexical-Functional Grammar resources of Cahill et al. (2004) and show how these resources can be adapted to handle strong domain variation. In our experiments, we use the Penn-II treebank (Marcus et al., 1994) Wall Street Journal (WSJ) newspaper sections and the ATIS (Hemphill et al., 1990) transcribed spoken language airline reservation resource. The Penn-II WSJ vs. ATIS domain change results in a markedly stronger drop in performance, both on the trees and the f-structures, for the Penn-II trained LFG resources of Cahill et al. (2004), compared to the drop observed by Gildea (2001) for the Penn-II WSJ vs. Brown domain variation experiments with Collins’s (1997) parser.

This poses a research question: is the observed performance drop of the LFG resources of Cahill et al. (2004) due to the decrease in quality of c-structure parsing, or is it a lack of coverage of the f-structure annotation algorithm (ibid.), or both? We report on experiments which answer this question. The main, and surprising, result is that, while the Penn-II trained c-structure component of Cahill et al. (2004) requires retraining, the f-structure annotation algorithm (originally designed for Penn-II WSJ data) requires no changes or extensions. The linguistic information encoded in the f-structure annotation algorithm is already complete with respect to strong domain variation as exemplified between the Penn-II WSJ and ATIS corpora. This is a surprising result as Penn-II WSJ data represents a markedly different text domain to that of ATIS, as discussed in Section 3. A possible explanation is that, compared to c-structure, f-structure is a more abstract and “normalised” level of representation in the LFG architecture, less affected by domain variation than c-structure.

Section 2 gives a brief outline of related work on treebank induced resources. In Section 3, we compare and contrast the ATIS corpus with the WSJ sections from the Penn-II Treebank. We outline our baseline experiments and present the results in Section 4. We analyse the results, investigate the underperformance and present experiments to improve performance in Sections 5 and 6. We investigate retraining the c-structure parser with appropriate data. In a CCG-style

experiment with the retrained parser we achieve a c-structure labelled f-score of 86.07 and an f-structure all grammatical functions f-score of 88.11. This constitutes an improvement of over 14% on c-structure parsing, and over 7% on f-structure annotation compared to unadapted parsing and annotation with the same system. In some additional experiments we parameterise the amount of WSJ material in the parser’s training set. We then measure the effect of adding punctuation to the ATIS test set and assess the question/non-question performance of the parser and annotation algorithm and perform a back-testing experiment with the retrained resources.

2 Background Work and Motivation

Wide coverage parsers are now being used for question analysis in open-domain question answering (QA) systems as described in Pasca and Harabagiu (2001) for example. In ongoing work we are investigating the use of the LFG annotation algorithm of Cahill et al. (2004) with Bikel’s (2002) parser to analyse TREC¹ question material into f-structures to develop a question tree- and f-structure bank resource for developing QA systems.

2.1 Previous Work

Domain variation and its effects on “shallow”² probabilistic parser performance has been investigated by Gildea (2001). For example, training on the Penn-II Treebank WSJ sections and parsing Brown corpus text resulted in a drop in labelled bracketing f-score for trees of 5.7% compared to parsing the WSJ. This shows the negative effect of domain variation on parser performance even when the test data is not substantially different from the training data (both the Penn II and Brown corpora consist primarily of written texts of American English, the main difference is the considerably more varied nature of the text in the Brown corpus). Gildea also shows how to resolve this problem by adding appropriate data to the training corpus, but notes that a large amount of additional data makes little impact if it is not matched to the test material.

Clark et al. (2004) have worked specifically with question parsing to generate dependencies for QA with Penn-II treebank based Combinatory Categorical Grammars (CCG’s). In their work they focus on “what” questions taken from the TRECQA dataset. Their solution is to retrain the lexical annotation component (the supertagger) of the parser rather than the whole parser. They evaluate accuracy at the lexical category level. In their work the supertagger’s accuracy improves over 13% with retraining on appropriate data. This gives a good indication of what can be achieved by retraining resources for questions.

Burke et al. (2004), Cahill et al. (2004), and O’Donovan et al. (2004) present a substantial body of work on automatically producing LFG resources from treebanks. However, to date no previous

¹<http://www.trec.nist.gov>

²A “shallow” grammar defines a language as a set of strings and may associate syntactic representations with strings. A “deep” grammar (in addition) associates strings with information/meaning representations, usually in the form of predicate-argument structures, dependency relations or logical forms. In order to construct accurate and complete “meaning” representations, deep grammars usually resolve long-distance dependencies.

research has been carried out to test the effect of domain variance on the treebank-induced LFG parsing resources of Cahill et al. (2004). Given that the resources are induced from the Penn-II Treebank, the expectation is that performance will suffer in a similar way as the experiments of Gildea with Collins' (1997) parser showed. In Section 4, we present experiments to test this hypothesis on the ATIS corpus, which contains transcribed spoken language with a significant proportion of question material and constitutes an instance of strong domain variation.

3 Corpus Description

3.1 ATIS

The Air Travel Information System (ATIS) corpus (Hemphill et al., 1990) is a transcription of spoken dialog with an automated air travel information system. ATIS represents a different style of language from the Wall Street Journal texts of the Penn-II Treebank: a significant proportion of the sentences in ATIS are questions, imperatives and non-sentential utterances, which are generally shorter than those in the WSJ sections of Penn-II and the transcription does not contain punctuation marks.

1. Are there any flights arriving after eleven a.m
2. Show me the T W A flight
3. I need a flight from Los Angeles to Charlotte today
4. Flights from Los Angeles to Pittsburgh
5. On Tuesday arriving before five p.m
6. What flights from Philadelphia to Atlanta

Figure 1: Example ATIS utterances

Figure 1 illustrates typical ATIS corpus data including both question (1) and non-question sentences (2,3), as well as sub-sentential (4,5) and incomplete utterances (6). Note also, that punctuation has not been added.

3.2 Penn-II WSJ vs. ATIS

| | ATIS | Penn-II WSJ |
|-------------------------|--|-----------------------|
| Words | 4000 words | 1 Million words |
| Sentences | 578 sentences | 50,000 sentences |
| Average sentence length | 7 words | 21 words |
| Source | Transcription of spoken dialog | WSJ Newspaper text |
| #Questions | 213 Direct questions | 233 Direct questions |
| Sentence type | Interrogatives, imperatives, and fragments | Declarative sentences |
| Inter-Word Punctuation | None | Punctuated |

Table 1: Corpus statistics compared

Both Penn-II WSJ and ATIS are POS- and parse-annotated corpora (ie. treebanks) following the same general annotation guidelines (Bies et al., 1995). Despite these similarities, the two treebanks exhibit strong differences as regards size, domain, phrase type distribution and punctuation.

Table 1 shows a comparison of the Penn-II WSJ sections and the ATIS corpus. The most striking difference between the Penn-II Treebank WSJ sections and the ATIS is the difference in size between the two corpora: the WSJ sections of the Penn-II Treebank with 50,000 sentences are over eighty times the size of ATIS with only 578 sentences. Another important difference between the two is in the average sentence length, those in ATIS tend to be much shorter than the WSJ, with an average length of 7 words, compared to 21 words in the WSJ. Figure 2 plots the number of sentences against the sentence length for the ATIS corpus and Section 23 of the WSJ section of the Penn-II treebank illustrating the difference in sentence length distribution between the corpora.

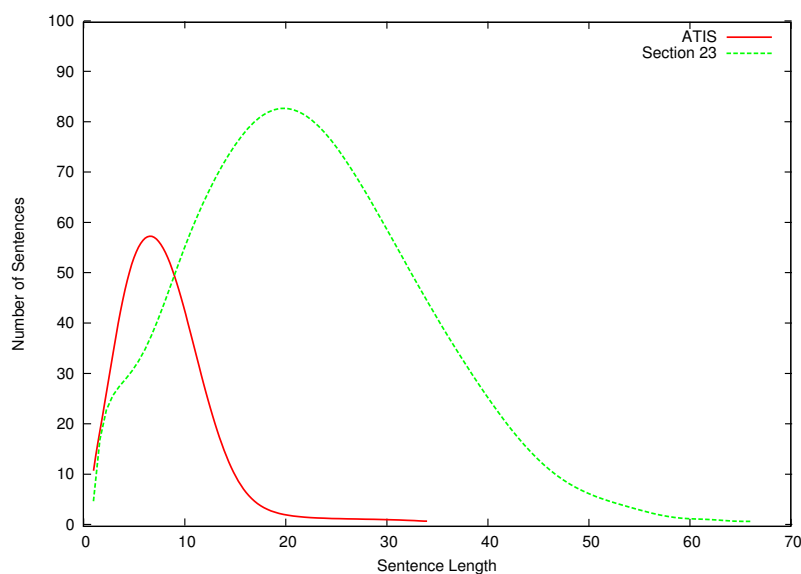


Figure 2: Sentence length distributions ATIS vs WSJ Section 23

The graph shows how significantly larger a single section of the Penn-II Treebank WSJ sections is than ATIS. It also shows the broader distribution of data over the sentence lengths in the section of the Penn-II Treebank, which has a much wider spread over the sentence lengths. Section 23 has a mean sentence length of 21 words with a standard deviation of 8.6, while ATIS has a mean sentence length of 7 words with a standard deviation of 2.9.

The source of text for the two corpora also highlights some important differences. The source for the ATIS corpus is spoken dialogue which tends to be more casual and brief (Figure 1) than the longer, more complex structures found in the Penn-II Treebank (Figure 3). Also the nature of the air travel information system results in the ATIS corpus containing sentences of a predominantly interrogative nature. Of the 578 sentences in the ATIS corpus, 213 are questions, accounting for over 36% of the entire corpus. Comparatively, the WSJ has very few interrogative sentences or questions, only 233 over the entire WSJ sections (accounting for less than a half of a percent of the corpus). In addition, many of these are embedded or rhetorical questions (Figure 4 (3)), which unlike those in the ATIS do not seek information. None of the 233 questions in the WSJ sections are to be found in section 23 of the treebank, which is the standard testing section for parser evaluation. Therefore, none of the evaluations carried out on this section reflect the quality of parsing/annotation of question data.

1. Shares of UAL, the parent of United Airlines, were extremely active all day Friday, reacting to news and rumors about the proposed \$6.79 billion buy-out of the airline by an employee-management group.
2. Ports of Call Inc. reached agreements to sell its remaining seven aircraft to buyers that weren't disclosed.
3. As a group, stock funds held 10.2% of assets in cash as of August, the latest figures available from the Investment Company Institute.

Figure 3: Example Penn-II Treebank WSJ sentences

1. For example, what exactly did the CIA tell Major Girolodi and his fellow coup plotters about U.S. laws and executive orders on assassinations?
2. Who'd have thought that the next group of tough guys carrying around reputations like this would be school superintendents?
3. What is the way forward?
4. But if rational science and economics have nothing to do with the new environment initiative, what is going on?

Figure 4: Example Penn-II Treebank WSJ questions

4 Preliminary Experiments and Results

4.1 Baseline Resources

This section describes our baseline experiments to determine the portability of the resources of Cahill et al. (2004) to a new domain, the ATIS corpus.

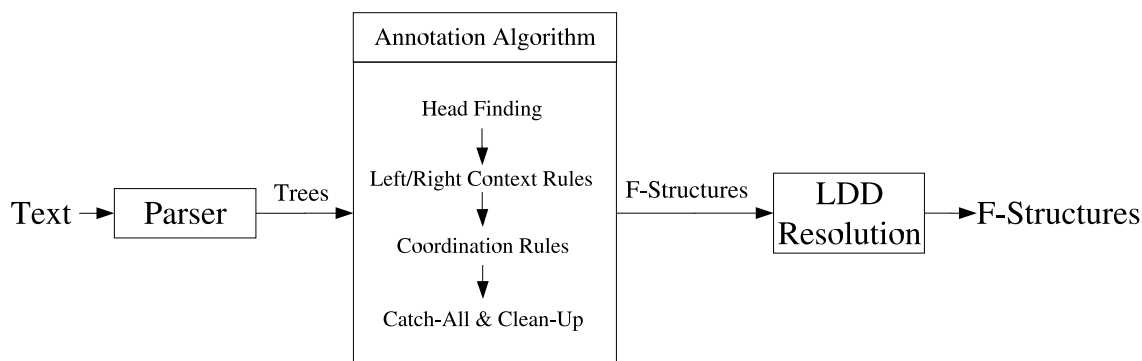


Figure 5: Pipeline Architecture

We use the pipeline model of Cahill et al. (2004) (Figure. 5) to generate f-structures from raw text. The c-structure parser used is that of Bikel (2002) which emulates Collins’ (1999) model 2 parser. The grammar used by the parser is trained on sections 2-21 of the Penn-II Treebank. The f-structure annotation algorithm (also developed on Penn-II WSJ material) is modular, taking c-structure trees and automatically adding LFG f-structure equations to each node in the tree. A modified version of Magerman’s (1994) scheme is used for determining the head of each subtree. The first module of the algorithm (Left-Right Context Rules) assigns annotations to the tree nodes based on whether they occur to the left or right of the head. Since the analysis of co-ordination in the Penn-II Treebank is very flat, co-ordination is treated separately in order to keep the left-right context rules concise. In the “Catch-All and Clean-Up” module of the algorithm, overgeneralisations made by the previous modules are corrected. The three modules generate “proto” f-structures which are then passed to a post-annotation long distance dependency (LDD) resolution module, which resolves long distance dependencies and outputs the final “proper” f-structures which we evaluate.

4.2 Evaluation

We use the pipeline architecture shown in Figure 5 to generate c- and f-structures from raw strings taken from the ATIS corpus. We evaluate both the c-structure trees outputted by the parser using PARSEVAL metrics (Black et al., 1991), and the LDD-resolved f-structures output by the annotation algorithm using the triple encoding and evaluation software of Crouch et al. (2002). The parser output is evaluated against the parse trees in the ATIS corpus, and the f-structures are evaluated against a hand crafted gold standard of f-structures for 100 sentences randomly selected from the ATIS corpus. We also perform a CCG-style (Hockenmaier, 2003) evaluation whereby we generate f-structures for the entire ATIS corpus from the original ATIS treebank trees and evaluate f-structures generated from the parser output against these 578 pseudo gold standard f-structures.

4.3 Results

(a)

| 100 Gold Standard | | Precision | Recall | F-Score |
|-----------------------------|------------|-----------|--------|---------|
| Trees (labelled bracketing) | | 73.77 | 67.05 | 70.25 |
| F-Structures | All GFs | 82.17 | 67.41 | 74.06 |
| | Preds-only | 70.33 | 56.97 | 62.95 |

(b)

| 578 ATIS | | Precision | Recall | F-Score |
|-----------------------------|------------|-----------|--------|---------|
| Trees (labelled bracketing) | | 75.49 | 67.77 | 71.42 |
| F-Structures | All GFs | 81.23 | 80.29 | 80.76 |
| | Preds-only | 69.27 | 67.02 | 68.13 |

(c)

| DCU 105 | | Precision | Recall | F-Score |
|-----------------------------|------------|-----------|--------|---------|
| Trees (labelled bracketing) | | 86.56 | 85.59 | 86.07 |
| F-Structures | All GFs | 83.45 | 78.95 | 81.14 |
| | Preds-Only | 76.32 | 72.0 | 74.10 |

(c)

Table 2: Results for baseline experiments

Table 2 gives the results for the two evaluations described above. Table 2 (a) shows the evaluation against the 100 sentence ATIS hand-crafted f-structure gold standard. Compared to the most recent results for the Penn-II WSJ section 23 based DCU 105³ evaluation in Table 2(c), the treebank-based LFG parsing resources of Cahill et al. (2004) show a significant drop in both the tree- and f-structure-based analysis scores for the ATIS material. The c-structures output by the parser have an f-score around 16% less than in the in-domain (section 23) evaluation for the same parser/grammar combination (Bikel trained on sections 02-21 of the Penn-II Treebank). Likewise the f-structure evaluation has suffered, with the preds-only f-score over 11% lower than on in-domain data.

³<http://nclt.dcu.ie/gold105.txt>

| Dependency | Precision | Recall | F-Score |
|-----------------|-----------------|-----------------|-----------|
| adjunct | 159/258=62 | 159/353=49 | 55 |
| comp | 0/5=0 | 0/3=0 | 0 |
| coord | 15/23=65 | 15/24=62 | 64 |
| det | 56/64=88 | 56/70=80 | 84 |
| focus | 9/9=100 | 9/33=27 | 43 |
| obj | 172/206=83 | 172/216=80 | 82 |
| obj2 | 17/18=94 | 17/18=94 | 94 |
| obl | 1/2=50 | 1/12=8 | 14 |
| obl2 | 0/0=0 | 0/5=0 | 0 |
| poss | 1/1=100 | 1/1=100 | 100 |
| quant | 2/16=12 | 2/6=33 | 18 |
| relmod | 9/13=69 | 9/16=56 | 62 |
| subj | 10/27=37 | 10/17=59 | 54 |
| topicrel | 10/27=37 | 10/17=59 | 45 |
| xcomp | 23/33=70 | 23/46=50 | 58 |

Table 3: Annotation results for selected features

Table 3 shows a more detailed analysis of the f-structure evaluation in Table 2(a) for selected features. The table shows that in particular for features such as **focus** and **topicrel**, which are important to analyse correctly in questions, the performance is quite low. This indicates that, as it stands, the Penn-II treebank-based LFG parsing system is not well suited to analysing questions and performance has suffered substantially as a result of the change in domain.

We have seen that by changing the domain from WSJ text to ATIS, the overall performance for c-structure analysis and f-structure analysis has dropped significantly. The strong domain variance between ATIS and WSJ data has affected both shallow (c-structure trees) and deep (f-structure dependencies) analyses and is more pronounced than was observed in earlier work by Gildea (2001).⁴

5 Why the Performance Drop?

The drop in performance can be attributed to the domain variance, but the question remains which module in the pipeline parsing architecture in Figure 5 (c-structure parser, f-structure annotation algorithm or LDD resolution) is underperforming due to the change in domain, or is it a combination? We can narrow the possibilities down to two of the three modules shown in Figure 5.⁵ Either the c-structure parser is underperforming and consequently the annotation algorithm is unable to generate sufficiently good f-structures from the bad c-structures, or the annotation algorithm is incomplete with respect to the domain variance.

⁴Gildea's work focused on c-structure parsing as opposed to full LFG f-structures.

⁵Testing on the long distance dependency resolution module showed that problems with LDD resolution were directly related to bad c-structure parsing.

The results in Table 2 have shown that the c-structure parser performance has dropped by almost 16% as a result of the domain variance. Previous work has shown that parser performance can be boosted through retraining with appropriate data (Gildea, 2001; Clark et al., 2004). We carry out an experiment to try and boost the question domain performance of Bikel’s parser by retraining a grammar with appropriate material from the ATIS corpus.

6 Retraining Experiments and Results

6.1 Retraining (WSJ + ATIS)

In order to improve the performance of the c-structure parser on ATIS sentences we create a new training set from which to extract a grammar for the parser. This new, larger, training set consists of sections 02-21 of the Penn-II Treebank WSJ (the original training data) and 90% of the ATIS corpus. We then train the parser on this new training set, and repeat the parsing and annotation experiments outlined in Section 4. C-structures for each of the 578 ATIS sentences are generated by retraining a grammar and parsing using a 10-fold cross-validation experiment with a 90%:10% training:test split over the ATIS corpus, and adding the 90% ATIS split to sections 02-21 of the Penn-II Treebank WSJ for training. The parser output c-structures are then passed to the f-structure annotation algorithm and LDD-resolution and the f-structures evaluated as before.

(a)

| 100 Gold Standard | | Precision | Recall | F-Score | Diff |
|-----------------------------|------------|-----------|--------|---------|--------|
| Trees (labelled bracketing) | | 88.03 | 78.78 | 83.14 | +12.89 |
| F-Structures | All GFs | 88.04 | 79.10 | 83.33 | +9.27 |
| | Preds-only | 80.17 | 73.66 | 76.77 | +13.82 |

(b)

| 578 ATIS | | Precision | Recall | F-Score | Diff |
|-----------------------------|------------|-----------|--------|---------|--------|
| Trees (labelled bracketing) | | 80.66 | 92.26 | 86.07 | +14.65 |
| F-Structures | All GFs | 87.27 | 88.97 | 88.11 | +7.35 |
| | Preds-only | 80.21 | 80.81 | 80.51 | +12.38 |

Table 4: Results for experiments with retrained grammar for 10-fold cross validation

Tables 4 (a) and (b) give the results of evaluating c-structures and f-structures generated with Bikel’s parser retrained as described above. Evaluating against the 100-sentence ATIS gold standard, the c-structure f-score has increased by almost 13% to 83.14. The quality of the f-structures has also increased with an improvement of almost 14% in the preds-only f-score, to 76.77. The performance over the whole corpus, in a CCG-style experiment against automatically generated

f-structures for the original 578 treebank trees, has increased correspondingly, with the c-structure f-score increasing over 14% to 86.07, and a preds only evaluation of the f-structures gaining over 12% to achieve an f-score of 80.51.

| Dependency | Precision | Recall | F-Score | Diff |
|-----------------|------------------|-----------------|-----------|------------|
| adjunct | 229/292=78 | 229/324=71 | 74 | +19 |
| comp | 0/4=0 | 0/3=0 | 0 | - |
| coord | 16/24=67 | 16/24=67 | 67 | +3 |
| det | 67/66=92 | 61/70=87 | 90 | +6 |
| focus | 23/23=100 | 23/33=70 | 82 | +39 |
| obj | 193/223=87 | 193/216=89 | 88 | +6 |
| obj2 | 17/17=100 | 17/18=94 | 97 | +3 |
| obl | 1/1=100 | 1/12=8 | 15 | +1 |
| obl2 | 0/0=0 | 0/5=0 | 0 | - |
| poss | 1/1=100 | 1/1=100 | 100 | - |
| quant | 2/16=12 | 2/6=33 | 18 | - |
| relmod | 14/19=74 | 14/16=88 | 80 | +18 |
| subj | 75/89=84 | 75/133=56 | 68 | +14 |
| topicrel | 14/19=74 | 14/17=82 | 78 | +33 |
| xcomp | 25/30=83 | 25/46=54 | 66 | +12 |

Table 5: Annotation results for selected features

Table 5 shows a more detailed analysis of the evaluations in Table 4(a) for a number of features. Compared to Table 3 the table shows that the retraining has had no negative effect on any of the features. The majority of features have improved in terms of both precision and recall. Of those features which benefited from the retraining, two features have gained significantly more than the others, **focus** and **topicrel**. These are two features which are important for analysing questions correctly.

Our experiments so far indicate that the annotation algorithm of Cahill et al. (2004), Burke et al. (2004), and O’Donovan et al. (2004) is complete with respect to the strong domain variance encountered in our experiments. We have seen that in order to cope with a new domain only the c-structure parser needs to be retrained.

In order to estimate an upper bound for our experiments, we took the original ATIS treebank trees for the 100 sentences in the gold standard and automatically annotated them to produce f-structures, thereby removing the c-structure parser margin of error. We then evaluated these f-structures against the hand-crafted f-structures in the gold standard. In this evaluation the all grammatical functions f-score is 92.80 and the preds-only f-score is 89.88 (Table 6). This is a satisfactory upper bound and the results are comparable to a similar experiment on the DCU 105.

| | All GFs | Preds-only |
|---------|---------|------------|
| F-Score | 92.80 | 89.88 |

Table 6: Upper bound for gold standard trees

These results demonstrate that improving the c-structure parsing is sufficient to improve the overall performance of the annotation algorithm on sentences outside of the domain on which it was developed. This is quite a surprising result, as we did not modify the annotation algorithm of Burke et al. (2004) in any way.

6.2 Parameterisation of Penn-II WSJ Training Data

We have seen above that adding a (relatively) small amount of domain appropriate material to the training set for the c-structure parser has resulted in quite significant gains for both c-structure and f-structure analysis of ATIS sentences. Previous work by Gildea (2001) has shown that a large amount of additional data makes little impact if it is not matched to the test material. With this in mind one can wonder if, due to its relative size, the Penn-II Treebank WSJ material in the training set for the parser might constitute such a large amount of redundant additional data.

In order to test, this we conducted a number of ablation experiments using the automatically f-structure annotated 578 ATIS trees as gold standard in a CCG-style experiment, where we evaluate c-structures and f-structure parser output algorithm, while reducing the amount of Penn-II Treebank material in the parser’s training set. The graphs in Figures 6 and 7 show the effect for evaluations against the entire ATIS corpus in a series of 10-fold cross validation experiments, in which the training set for the parser consists of 90% of the ATIS corpus and a varying (randomly selected) percentage of the Penn-II Treebank.

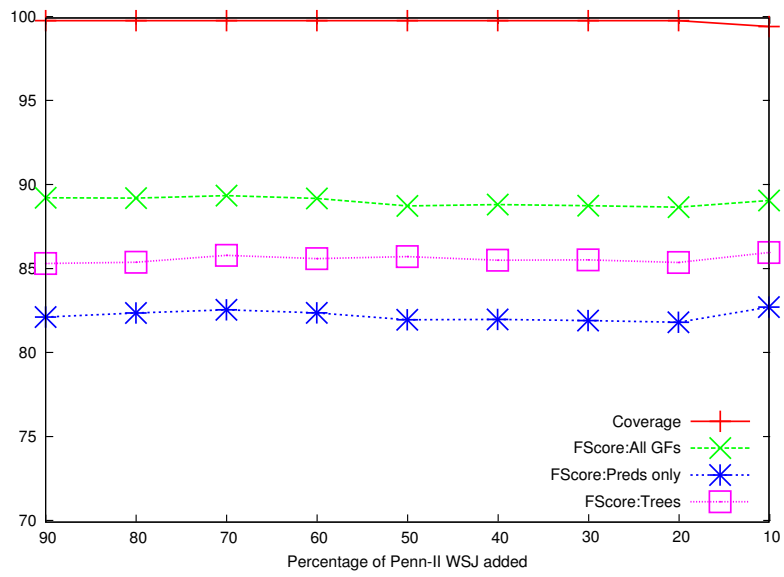


Figure 6: Reducing Penn-II Treebank content (90%-10% of sections 02-21 WSJ, CCG-style experiment)

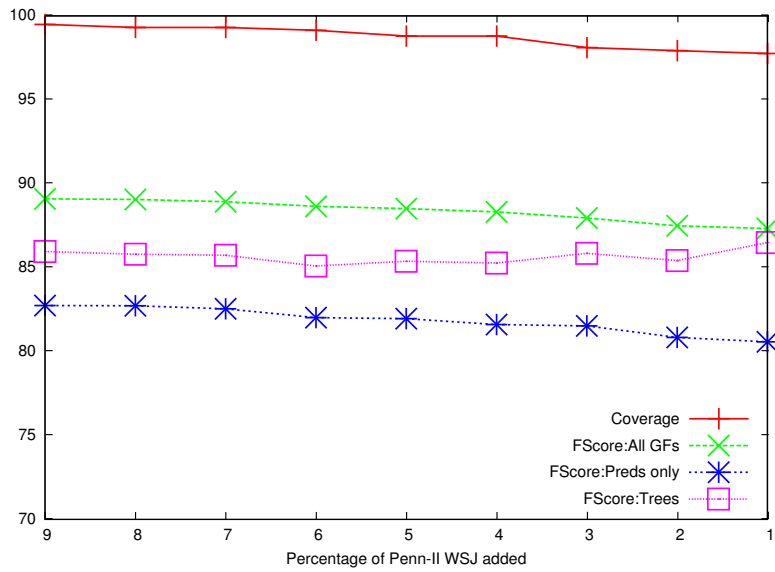


Figure 7: Reducing Penn-II Treebank content (9%-1% of sections 02-21 WSJ, CCG-style experiment)

The graphs show that reducing the amount of Penn-II Treebank WSJ material in the training set adversely affects the overall performance. Grammar coverage, c-structure parsing and f-structure annotation all suffer to varying degrees. Both c-structure and f-structure evaluations start to decline when less than 70% of the treebank is included in the training set. Grammar coverage proves to be less affected in this case: it does not decline significantly until the amount of treebank WSJ training material falls below 20%. Nevertheless, the system is capable of achieving coverage in the region of 99%, a c-structure f-score of over 85%, and f-structure f-scores of over 88% (all grammatical functions) and over 82% (preds-only), when the c-structure parser is trained on 90% of the ATIS corpus and only 10% of the Penn-II Treebank.

6.3 Punctuation

The Penn-II Treebank Wall Street Journal sections used for training the c-structure parser contains properly punctuated text. On the other hand, the ATIS strings are unpunctuated. This is another factor that could possibly explain the underperformance of the c-structure parser and (consequently) annotation algorithm in our earlier experiments, as we would expect grammars trained on Penn-II Treebank sections to perform better on punctuated text.⁶

To test this with the ATIS corpus, we added basic punctuation to each of the ATIS sentences. Each of the 213 questions had a question mark added, the remaining sentences had a fullstop added, and the sub-sentential fragments were left unpunctuated. We then reran the parsing experiments with both the baseline WSJ-only trained grammar, and also the improved WSJ and 90% ATIS trained grammar in a 10-fold cross validation experiment.

| | WSJ | | | WSJ + ATIS 90% | | |
|----------------|--------------|------------|-------|----------------|------------|-------|
| | Unpunctuated | Punctuated | Diff | Unpunctuated | Punctuated | Diff |
| Coverage | 100 | 99.83 | -0.17 | 100 | 99.83 | -0.17 |
| F-Score(Trees) | 71.42 | 71.31 | -0.11 | 86.07 | 85.36 | -0.71 |

Table 7: Parsing results for punctuated ATIS sentences

Table 7 shows the evaluation results for c-structure analysis of the 578 ATIS sentences with basic punctuation added. The table shows the coverage and f-scores for both the baseline grammar, trained on sections 02-21 of the Penn-II Treebank WSJ, and the grammar retrained with added ATIS sentences, and the difference between these scores and those for parsing the ATIS sentences without punctuation. It is interesting to note that all of the scores have decreased slightly as a result of adding punctuation, when the naive assumption, stated above, would be that the parser should perform better given that its training data is punctuated. This emphasises the effect of the domain difference between the ATIS corpus and the Penn-II Treebank.

⁶This was pointed out to us by Tracy King (p.c.).

6.4 Question vs Non-Question

The ATIS corpus contains both question and non-question data. Our 100-sentence gold standard is taken from the ATIS corpus and so comprises both question and non-question sentences. Table 8 shows the breakdown of the upper bound (established following the procedure detailed in Section 6.1) for both question and non-question sentences in the gold standard.

| | Non-question | Question |
|------------|--------------|----------|
| All GFs | 94.82 | 90.77 |
| Preds-only | 92.94 | 86.81 |

Table 8: Question and non-question f-score upper bounds

The upper bound breakdown shows a slight leaning towards a higher upper bound for non-question sentences, but the upper bound for questions is still quite high.

Table 9 gives the breakdown of the scores for question and non-question sentences in the 100 sentence gold standard parsing evaluations.

| | WSJ Trained | | WSJ + ATIS Trained | | | |
|------------|--------------|----------|--------------------|-------|----------|--------|
| | Non-Question | Question | Non-Question | | Question | |
| | F-Score | F-Score | F-Score | Diff | F-score | Diff |
| Trees | 74.75 | 61.92 | 80.55 | +5.8 | 88.35 | +26.43 |
| All GFs | 77.40 | 70.52 | 82.62 | +5.22 | 84.38 | +13.86 |
| Preds-only | 68.96 | 54.12 | 76.28 | +7.32 | 77.56 | +23.44 |

Table 9: Question and non-question scores for the annotation algorithm

The breakdown in Table 9 clearly shows the effect of both the domain variance and the re-training in the earlier experiments. The left of the table shows the breakdown for the baseline experiments before the parser was retrained. In this experiment it is clear that both the c-structure parser and the f-structure annotation algorithm are underperforming on questions as opposed to non-question sentences. The right of the table shows the same breakdown, but for the experiments with the parser retrained on both Penn-II Treebank WSJ and ATIS sentences. It is clear that this retraining has benefited both the c-structure and f-structure evaluations for the questions in particular. The c-structure tree evaluation has improved over 26% with an f-score of 88.35, likewise the f-structure evaluations have improved for evaluations of all grammatical functions and preds-only, improving by 13.86% and 23.44% respectively. It is also interesting to note that none of the scores have decreased as a result of this retraining, the results for the non-question sentences have also improved (albeit to a lesser extent).

6.5 Back-Testing the Retrained Grammar

The experiments above show that retraining the c-structure parser for the new domain has allowed us to adapt the treebank-based LFG resources to a new domain and achieve similar f-scores in c- and f-structure evaluations on data from a new domain compared to in-domain results. In order to ensure that this retraining process has not adversely affected the overall system performance, we back-test the retrained parser and annotation algorithm on sentences from the original WSJ domain (the DCU 105 gold standard). We parsed the 105 sentences with each of the 10 retrained grammars from the 10-fold cross validation experiment in Section 6.1, then evaluated both c- and f-structures against the DCU 105 gold standard. The averaged results are shown in Table 10 (a), along with the results for the grammar trained only on sections 02-21 of the Penn-II Treebank in the same evaluation (b).

| WSJ 02-21 trained | | Precision | Recall | F-Score |
|-------------------|------------|-----------|--------|---------|
| Trees | | 86.56 | 85.59 | 86.07 |
| F-Structures | All GFs | 83.45 | 78.95 | 81.14 |
| | Preds-Only | 76.32 | 72.0 | 74.10 |

(a)

| WSJ 02-21 + 90% ATIS trained | | Precision | Recall | F-Score |
|------------------------------|------------|-----------|--------|---------|
| Trees | | 87.05 | 86.10 | 86.57 |
| F-Structures | All GFs | 83.92 | 79.34 | 81.56 |
| | Preds-Only | 77.32 | 72.85 | 75.02 |

(b)

Table 10: Results for backtesting retrained grammar and baseline grammar on DCU 105

The results show that the retraining process has resulted in no loss of accuracy at either c- or f-structure level. The scores have in fact improved slightly as a result of the retraining; however the improvements, when tested, were not statistically significant (paired t-test). From this we conclude that there has been no significant negative effect on the LFG parsing resources of Cahill et al. (2004) on WSJ material as a result of retraining the c-structure grammar to adapt the treebank-based LFG resources to a new domain.

7 Conclusions

Our experiments have shown that treebank induced LFG resources underperform when the domain is varied from that of the training material. This holds for both c-structure and f-structure analyses. To adapt the treebank-based LFG resources of Cahill et al. (2004) to a new domain, all that was necessary was to retrain the c-structure parser. The f-structure annotation module is able to handle the domain variance without modification. We have also shown that the f-structure

annotation algorithm is general: given high-quality c-structure trees, it can achieve a high upper bound for f-structures in a new domain. More generally, our experiments support the claim that the f-structures generated are a more normalised linguistic representation which are less affected by domain variance than the level of c-structure representation.

In our experiments we have adapted our LFG parsing resources to a new domain with a c-structure labelled f-score of 86.07 and an f-structure all grammatical functions f-score of 88.11 in a CCG-style experiment. This constitutes an improvement of over 14% on c-structure parsing, and over 7% on f-structure annotation compared to unadapted parsing and annotation with the same system.

We plan to extend our work by developing a larger question corpus. With such a resource we will be able to parameterise the amount of question data needed in retraining the c-structure parser to reach an optimal result.

References

- Bies, A., Ferguson, M., Katz, K., and MacIntyre, R. (1995). Bracketing guidelines for Treebank II style Penn Treebank Project. Technical report, University of Pennsylvania, Philadelphia, PA.
- Bikel, D. M. (2002). Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of Human Language Technology (HLT) 2002*, pages 24–27, San Diego, CA.
- Black, E., Abney, S., Flickenger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingira, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., and Strzalkowski, T. (1991). A Procedure for Quantatively Comparing the Coverage of English Grammars. In *Proceedings of the February 1991 DARPA Speech and Natural Language Workshop*, pages 306–311, Pacific Grove, CA.
- Burke, M., Cahill, A., O’Donovan, R., van Genabith, J., and Way, A. (2004). The Evaluation of an Automatic Annotation Algorithm against the PARC 700 Dependency Bank . In *Proceedings of the Ninth International Conference on LFG*, pages 101–121, Christchurch, New Zealand.
- Cahill, A., Burke, M., O’Donovan, R., van Genabith, J., and Way, A. (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, Barcelona, Spain.
- Charniak, E. (2000). A maximum entropy inspired parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 132–139, Seattle, WA.
- Clark, S., Steedman, M., and Curran, J. R. (2004). Object-extraction and question-parsing using ccg. In Lin, D. and Wu, D., editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 111–118, Barcelona, Spain.

- Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, Philadelphia, PA.
- Crouch, R., Kaplan, R. T., King, T., and Riezler, S. (2002). A comparison of evaluation metrics for a broad coverage parser . In *Beyond PARSEVAL Workshop, Language Resources and Evaluation (LREC)*, pages 67–74, Las Palmas, Canary Islands, Spain.
- Gildea, D. (2001). Corpus variation and parser performance. In Lee, L. and Harman, D., editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 167–202, Pittsburgh, PA.
- Hemphill, C. T., Godfrey, J. J., and Doddington, G. R. (1990). The ATIS Spoken Language Systems pilot corpus. In *Proceedings of DARPA Speech and Natural Language Workshop*, pages 96–101, Hidden Valley, PA.
- Hockenmaier, J. (2003). Parsing with generative models of predicate-argument structure. In *Proceedings of 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 359–366, Sapporo, Japan.
- Magerman, D. (1994). *Natural Language Parsing as Statistical Pattern Recognition*. PhD thesis, Department of Computer Science, Stanford University, CA.
- Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1994). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- O’Donovan, R., Burke, M., Cahill, A., van Genabith, J., and Way, A. (2004). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II Treebank . In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 368–375, Barcelona, Spain.
- Pasca, M. and Harabagiu, S. M. (2001). High Performance Question/Answering. In *Research and Development in Information Retrieval*, pages 366–374, New Orleans, LA.

**THE INS AND OUTS OF THE
PARTICIPLE-ADJECTIVE CONVERSION RULE**

Anna Kibort
Surrey Morphology Group, University of Surrey, UK

Proceedings of the LFG05 Conference
University of Bergen
Miriam Butt and Tracy Holloway King (Editors)
2005
CSLI Publications
<http://csli-publications.stanford.edu/>

Abstract

Bresnan (1978:8-9, 1982:29, 2001:31) proposes that English has a general morphological process of participle-adjective conversion which enables any verbal participle to be used as an adjective. The phenomenon captured by the Participle-Adjective Conversion Rule has been used as key evidence from English that passivisation is a lexical relation change, not a syntactic transformation, and as such, it can feed lexical processes of derivational morphology. The discussion offered in this paper supports fully the lexical character of both passivisation and participial formation. However, it argues for a small but important revision to the formulation of the Conversion Rule. Specifically, the morphological derivation of participles does not engage the syntactic level of argument structure. The input to the Conversion Rule is a verb, and the output is a deverbal form (a participle) which is category-neutral between a verb and an adjective (i.e. it is both a verb and an adjective at the same time). Passivisation is also a derivation, but it is morphosyntactic: it occurs at the level of argument structure of the predicate and it is an operation on grammatical functions. A patient-oriented (derived) participle is passive only if it is used as the main verb of the passive construction.

I Introduction

This paper proposes a reformulation of the standard Participle-Adjective Conversion Rule (Bresnan 1978; 2001). It argues that the morphological derivation of participles does not engage the syntactic level of argument structure, so the Conversion Rule should simply be formulated as deriving a participle (a deverbal form) from a verb. The participle (e.g. *broken*) is a lexical form that is category-neutral between an adjective and a verb – that is, the same lexical form can function as either an adjective (*The vase appeared broken*), or a verb (*The vase got broken by the burglars*), depending on the construction in which it is used.

When used as an adjective in predicative function with a copula *be* (*The vase was broken*), the participle is frequently indistinguishable from the main verb of the passive construction accompanied by an auxiliary. This is because the subject-complement construction and the passive construction in English overlap in their surface syntactic (phrasal) expression. Furthermore, an apparently ‘passive’ participle used as an adjective in attributive function does not have to be passive (*the broken vase* ~ *the vase that has broken*; cf. *the fallen leaf* ~ *the leaf that has fallen*). This is because the morphological derivation of the participial form is independent of passivisation which is a different, morphosyntactic derivation. However, the participle in *the broken vase* is indeed ambiguous between being active and passive because the periphrastic passive construction happens to use the same derived participial form as its main verb as the form that is used in the (non-passive) resultative construction.

The fact that the same participial form can be ambiguous between being an adjective and a verb, as well as between being active and passive, means that the frequently posited distinction between ‘verbal passives’ and ‘adjectival passives’ is not very helpful, as it collapses the two distinctions into just one. The distributional and morphological diagnostic tests that have been devised to distinguish between ‘verbal passives’ and ‘adjectival passives’ do indeed help to confirm the categorial status of the participial form in some clauses, but

they are nevertheless incapable of disambiguating all clauses, and incapable of identifying the passive. Therefore, instead of supporting the distinction within the category of the passive between verbal and adjectival passives, I suggest that two different processes conspire to produce the variety and ambiguity of the forms in question: (i) the derivation of the participle, captured by the revised rule which says that participles ($V_{A/V}$) are morphologically derived from the base verb (V); and (ii) the usage of the participle in two different types of construction.

A **resultative participle**, which is semantically oriented towards the affected participant, is both an adjective and a verb and can function as either: (a) the head of the predicative complement to the main predicator, with *be* as the main verb, head of the verbal phrase, alternating with other copular verbs such as *appear*, *look*, or *seem*; or (b) the main verb of the passive construction, with *be* as an auxiliary alternating with *become* or *get*. Thus, instead of the distinction within the category of the passive between verbal and adjectival passives, the distinction that should be drawn is that between **the resultative** (a semantically restricted construction), which results in clauses in which the morphologically derived resultative participle fulfills the function of the main verb's complement, and **the passive** (a syntactically restricted construction), which results in clauses in which the resultative participle fulfills the function of the main verb.

Starting from the Participle-Adjective Conversion Rule of standard LFG, in the sections below I discuss the distribution of the resultative participle in English, analyse the behaviour of resultative participles and hypothesise the lexical rule deriving them, contrast it with the formation of the passive, and finally formalise the revised rule of participial formation from the base verb.¹

2 The Participle-Adjective Conversion Rule in LFG

The Participle-Adjective Conversion Rule was first suggested by Bresnan (1978) and defined by Bresnan (1982), with the following formulation (Bresnan 1982:23):

- | | | |
|-----|----------------------------|--|
| (i) | Morphological change: | $V_{Part} \mapsto [V_{Part}]_A$ |
| | Operation on lexical form: | $P(\dots(\text{SUBJ})\dots) \mapsto \text{STATE-OF } P(\dots(\text{SUBJ})\dots)$ |
| | Condition: | $\text{SUBJ} = \text{theme of } P$ |

The rule has played the key role in the argument for the lexical character of passivisation. It has led to the general acceptance by lexicalist syntactic frameworks of the hypothesis that passivisation is a lexical relation change which can feed further lexical processes of derivational morphology, such as adjective formation, nominalisation, or compounding. It has served as the foundation for analyses of passive/past participles in English (Bresnan 1995, Ackerman & Goldberg 1996) and has been applied in analyses of deverbal adjective formation in other languages (Hungarian: Ackerman 1992, Komlósy 1994; Modern Greek: Markantonatou 1995, Kordoni 2002).

Apart from passive participles, the rule also applies to perfect and present participles. Thus, the following derivations are postulated for English:

- (2) a. *the food that is/was eaten* \Rightarrow *the eaten food*

¹The description and analysis presented in this paper is taken from my PhD thesis on passive and passive-like constructions in English and Polish (Kibort 2004), in particular Chapter 6.

- b. *a leaf that has fallen* ⇒ *a fallen leaf*
- c. *an argument that is/was not convincing* ⇒ *an unconvincing argument*

and the categorial status of all converted adjectives can be confirmed with the help of three distributional diagnostic contexts and one morphological diagnostic test, proposed by Bresnan (1982).

The hypothesised participle-adjective conversion rule naturally accounts for the fact that the participles in both uses – the verbal and the adjectival – have the same form. Levin & Rappaport additionally remark that '[Bresnan's] rule also captures the generalization, noted by Lieber (1980), that although the passive morpheme has a number of allomorphs, the verbal and adjectival passive participles of any given verb always involve the same allomorph: *the food was eaten, the eaten food; the ballad was sung, a badly sung ballad*' (1986:629). In other words, '(...) adjectival passives show the full range of passive participle morphology that we find with passive verbs' (Bresnan 2001:31).

Beside displaying the same allomorphs, the identity of form between English verbal and adjectival participles has also been observed in passives containing a verb and a preposition (examples from Bresnan 2001:31-32):

- (3) a. *After the tornado, the fields had a marched through look.*
- b. *Each unpaid for item will be returned.*
- c. *You can ignore any recently gone over accounts.*
- d. *His was not a well-looked on profession.*
- e. *They shared an unspoken, unheard of passion for chocolates.*

and in the fact that exceptions to the adjectival passive are also exceptions to the passivisation of a prepositional verb (examples adapted from Bresnan 2001:32):

- (4) a. **a looked-like twin*
- b. **The twin is looked like by his brother.*
- (5) a. **the left-for reason*
- b. **No reason was left for.*

All this has been taken as evidence supporting the hypothesised rule converting verbal participles (passive or other) into adjectives. Arguing against postulating a separate rule of adjectival passivisation in addition to verbal passivisation, Bresnan proposes that the input to passive adjectival formation rule is the passive lexical form of the verb, as in (1) above. If there were a separate morphological rule of 'adjectival passivisation' alongside of verbal passivisation – she argues – all the morphological parallels between verbal and adjectival passives would be an unexplained accident (Bresnan 2001:31).

I suggest, however, that in order to explain the coincidence of verbal and adjectival participial forms we do not have to posit the derivation of adjectives from verbal participles. A different proposal, decentering the passive – that the analytic passive verb is one of the uses of the morphologically derived resultative participle – preserves all the above observations regarding the morphological and syntactic behaviour of the participial form in its different environments, and similarly does not require a separate rule of adjectival passivisation in addition to the rule which passivises the verb.

3 The distribution of the resultative participle in English

3.1 Actional and statal passives

We have seen that the same participle can be found in verbal and adjectival passives in English. The distinction between verbal and adjectival passives seems to correspond to another which is sometimes drawn between the so-called ‘actional’ or ‘dynamic’ passives and ‘statal’ or ‘stative’ passives (e.g. Huddleston 1984:322, Quirk et al. 1985:168). Huddleston illustrates the two kinds of passive construction in English with the following examples (respectively):

- (6) a. *The vase was broken by Tim.*
b. *The vase was already broken.*

and argues that actional passives say that a certain event took place, while statal passives attribute to their subject the property of being in the state resulting from a certain event. Specifically, in sentence (a) above the actional passive says that the breaking of the vase took place, while in sentence (b) the statal passive attributes to the vase the property of ‘being in the state resulting from the event wherein it was broken in the actional sense’ (1984:323).

However, if we remove the agent phrase (to which Huddleston refers as ‘the complement’) from sentence (a) and the modifier from sentence (b), we are left with *The vase was broken*, which can belong to either category. The same ambiguity is found in *They were married*, which can mean ‘The marriage ceremony took place’ (actional) or ‘They were husband and wife’ (statal); in *The gate was closed*, which can mean ‘The closing of the gate took place’ (actional) or ‘The gate was in a closed state, i.e. the opposite of open’ (statal); and so on (all examples from Huddleston 1984:323).

The corollary of positing any such distinction within passive participles is that the verb *be* in adjectival or statal passives is considered a main verb, head of the verbal phrase, with the participle functioning as (head of) the predicative complement. Being a complement to the main predicator, the participle can occur with other copular verbs than *be*, as in *The vase appeared/looked/seemed broken* (analogous to *The vase was/appeared/looked/seemed very valuable*) (examples from Huddleston 1984:323). On the other hand, in verbal or actional passives the verb *be* is an auxiliary, and it may alternate with other acceptable passive auxiliaries such as *become* or *get*.

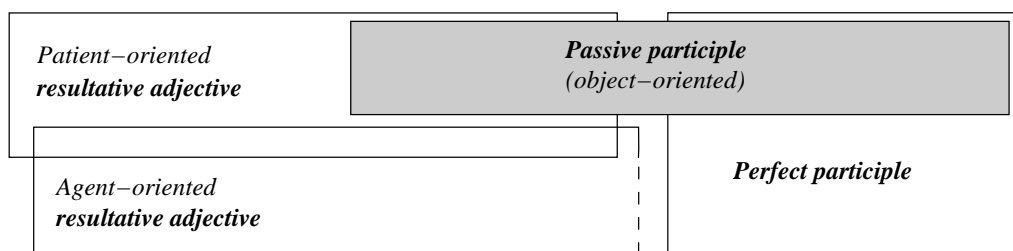
This paper argues that the proposed distinction within the passive construction between verbal/actional passives and adjectival/statal passives is, in fact, an unnecessary extension to the observation that the same participial form can be used by two constructions: the morphosyntactic passive construction and the subject-complement construction. The verb phrase in the passive should indeed be analysed as a ‘analytic verb’ comprising an auxiliary and a main verb (the participle). Subject-complement constructions, on the other hand, are made up of a copular verb and a predicative (adjectival, nominal or adverbial) complement of the subject.

However, it is not the morphological form of the verbs or the surface structure of the clause that determine whether the clause is a passive construction or a non-passive subject-complement construction. Because of the use of the participial form by both constructions, and because of the overlap of the phrasal expression between the passive and the subject-complement construction it is often impossible, as well as unnecessary, to attempt to identify the passive on the basis of the form of the verb or surface syntax. The same

participial form used by both constructions is best regarded not as a ‘passive participle’, but as a verb-derived ‘resultative participle’ which has categorial status neutral between an adjective and a verb. The notion of the ‘resultative’, which is central to this proposal, will be discussed in detail in Section 4. In the meantime, the following diagram illustrates the overlap in the use of the resultative participle as an adjective and as a verb in English:

(7) Resultative participle and the passive construction (Kibort 2004:169,412)

used attributively ----- *used predicatively* ----- *part of a verb compound*
 (with copula ‘be’) (with aux such as ‘be’/
 ‘become’, or ‘have’)



Thus, the distinction between actional/dynamic and statal/stative passives turns out to be primarily an issue of the distribution of a form (the participle). While the passive and the subject-complement construction may and do overlap in the semantic function of expressing stativity, there is no direct relation between passives and states.²

3.2 Subject-complement constructions

As argued in this paper, the similarity between the periphrastic passive and the predicative adjectival construction is not only apparent – some passive clauses indeed have the very same *phrase structure* as the predicative adjectival construction.³ Frajzyngier (1978:149) proposes that the logical structure of subject-complement constructions, which he calls ‘nominal sentences’, can be either $X=Y$ (as in: *Elizabeth II is the present Queen of England*) or $X \in Y$ (as in: *Salt is white*) (both examples originally from Suppes 1957:101). He argues that ‘be’-passives differ from other nominal sentences only in the fact that the predicate in ‘be’-passives (i.e. the participle following the copula/auxiliary) is morphologically derived from the lexical class of verbs, while in other nominal sentences it does not have to be so derived. On the other hand, ‘we might have languages, such as Semitic, in which not only verbal adjectives are derived from verbs but such nominal categories as agent, instrument,

²Frajzyngier (1978:153) argues that while subject-complement sentences are inherently stative, the passive construction can have (at least) two types of meaning: stative and non-stative. In some languages this distinction is marked morphologically, i.e. there are (at least) two different passive forms. In this case the stative passive and the subject-complement construction share the ‘stative’ interpretation, while the other passive form is designated to express the ‘non-stative’. If a language has only one passive form, this form will be ambiguous with respect to the stative/non-stative distinction and it is likely that another form may be brought in to disambiguate the construction. The present-day English ‘be’-passives are ambiguous in just this respect and Frajzyngier argues after Visser (1973:2089) that, in modern English, *get* is becoming the most important auxiliary to indicate the non-stative passive.

³For this reason, many analyses have treated the participles in all periphrastic passive clauses as adjectives. See Siewierska (1984:127,139-149) for useful summaries of these accounts.

name of action and place of action' (1978:150). Therefore, syntactically – he argues – there is no distinction between 'be'-passives and other nominal sentences.

Based on the analysis of a sample of over thirty languages chosen at random from several language families, Frajzyngier further points out that there are no languages that have 'be'-passives but do not have nominal sentences formed with a copula. Moreover, the passive form in a language will contain the equivalent of 'be' only if the nominal sentence contains 'be' (cf. the common phenomenon, as in Russian, of the absence of the copula both in the present tense of nominal sentences and in *-n/-t* resultatives, versus the presence of the copula in the past tense of nominal sentences, *-n/-t* resultatives, and participial passives).⁴

Finally, diachronic analysis shows that 'be'-passives are, generally, more recent forms than other passives or statives. 'In languages for which the *be*-passives are attested in the oldest available texts, one can claim that actually there is no distinction between *be*-passives and nominal sentences' (Frajzyngier 1978:154). The most natural explanation of the similarity between 'be'-passives and stative nominal sentences is, therefore, that the former developed from the latter, and this happened because nominal sentences with a copula presented a suitable structure for the realisation of the passive.

Haspelmath (1990:38) argues that the elements like 'be' and 'become' in Indo-European periphrastic passives were indeed initially main verbs and formed subject-complement constructions. As they entered into the passive construction which gradually grammaticalised, they became grammatical verbs (auxiliaries).⁵ When the passive as a morphosyntactic category grammaticalises, it may expand to include other auxiliaries than the copula 'be', thus becoming more distinct and independent from the predicative adjectival construction.

A morphosyntactic analysis of the passive operation allows one to specify that the difference between the passive and the predicative adjectival construction is lexical. The passive is produced as a result of a morphosyntactic operation on the argument structure of the predicate, while the resultative adjective results from the morphological derivation of an adjective from a verb. Thus, both processes are lexical derivations, but the formation of a resultative adjective does not require the application of the passive rule or constraint, nor does it require appealing to argument structure at all.

4 The resultative

4.1 Resultative participles: overview

All English participles referred to as 'past', 'perfect' and 'passive' result from the same morphological derivation. 'Past passive' (or, 'passive') participles formed from transitive verbs – as in *the solved problem* – and 'past active' (or, 'perfect'/'active unaccusative') participles formed from intransitive verbs – as in *the escaped prisoner* – are instances of the same

⁴This phenomenon also shows that the resultative participle itself does not seem to be sufficient to support the passive structure. The participial passive makes use of the semantics of the resultative participle, but it also needs a finite auxiliary, such as the copula of subject-complement clauses, to support its ability to refer to various time frames in the analogous way to the corresponding active. Therefore, it would be inaccurate to attribute the interpretation of the participial passive construction to the participle itself, rather than to the argument structure of the predicate expressed with the analytic verb form which includes an auxiliary.

⁵Furthermore auxiliaries may subsequently become affixed to the verb stem and lose their verbal status, thus turning into purely grammatical affixes, as can be demonstrated to have happened in the passive constructions of numerous (non-Indo-European) languages (Haspelmath 1990:38).

participial formation which is best understood with reference to the notion *resultative*. Nedjalkov & Jaxontov (1988:6), who undertook a crosslinguistic study of resultative constructions, define the term resultative as indicating ‘those verb forms that express a state implying a previous event’. Both the past passive deverbal participle and the perfect active deverbal participle are, thus, the same *resultative participle* which characterises its head ‘by expressing a state that results from a previous event’ Haspelmath (1994:159).

Haspelmath (1994:159-161) discusses some of the semantic restrictions on the formation of the resultative participle, which all boil down to the fact that ‘a thing cannot always be characterized by means of a state resulting from an event in which it participated’. One obvious restriction that can be posited makes use of the notion of *affectedness*: it is possible, and indeed useful, to characterise a participant by means of a resulting state only if the previous event affected or changed it somehow (cf. *the abused child, the wilted dandelion*). For this reason, resultative participles formed from transitive verbs are most commonly patient-oriented. Another restriction, also semantic in nature and deriving from the notion of affectedness, is that the verb may need to be telic to be able to form a resultative participle. This requirement is particularly relevant in the formation of those resultative participles which are agent-oriented.

In the sections below, I will first discuss the notions of ‘participle’ and ‘resultative’, and then discuss all the known restrictions on the formation of resultative participles. I will then show how the passive construction makes use of the available forms of the resultative participle and the subject-complement construction in which the resultative participle appears as an adjective in predicative function. By offering a systematic account of both the resultative and the passive, I aim to show that the participial form itself is a morphologically derived lexeme which is not passive unless it is used in a passive construction identified on the basis of morphosyntax. When referring to the isolated participial form, it is therefore more accurate to use the term ‘resultative participle’. Using the term ‘passive’ with reference to this form outside the passive construction may be supported by the fact that a large proportion of resultative participles are ‘semantically passive’. However, this functional classification is misleading and creates unnecessary problems for analysis. On the basis of the presence of the semantically passive participial form, many more constructions are classified as passive than are genuinely morphologically passive.

4.2 Adjectives and participles

It is a widely accepted fact that adjectives do not constitute a universal syntactic category. Languages which lack (or have few) distinct adjectives use verbs or nouns to express properties or qualities. Similarly, in languages regarded as having a distinct adjective class, the adjectives tend to share morphological and/or syntactic properties with nouns or with verbs.

In languages like English which have a distinct open class of adjectives, property concepts are traditionally considered to be encoded either as adjectives, adjectival nouns, or as adjectival verbs – even though none of these subcategories is, in fact, clearly identifiable or homogeneous. Adjectives tend to split up into ‘noun-like’ and ‘verb-like’, and the boundaries between adjectives on the one hand and adjectival nouns and adjectival verbs on the other appear to be extremely fuzzy. As for the adjectivals, whatever their word class status is considered to be, they are typically attached to the nominal or verbal system of the

language in question (Wetzer 1996:5-6). In languages like English the adjectivals which are derived from verbs and considered part of the verbal inflectional paradigm are traditionally referred to as participles.

The last characteristic, whose consequence is the retention of verbal valency at some level of representation other than just conceptual, is often considered necessary for a verb-derived adjective to be called a participle (Haspelmath 1994:152). It is this characteristic which distinguishes English verb-derived adjectives such as *understandable*, *reliable*, etc., which are not normally considered participles, from verb-derived adjectives such as *singing*, *smiling*, *sung*, *gathered*, etc., which are generally considered to be participles.

4.3 Verb-derived adjectives: orientation and tense

When participles are used in attributive function, they modify the head noun with which they are combined in the same way as adjectives do. I assume, after Haspelmath (1994:153, and footnote 5), that in a modifying relation, the *modifier* is relational and has a slot for its head which coincides with its referent. (In a governing relation, on the other hand, the *head* is relational, has slots for its arguments, and has a separate referent.) Furthermore, whenever the meaning of an attributive word is a concept involving more than one semantic participant, it is possible for the word to express a specific orientation towards one of the participants.

Taking Haspelmath's example of the English adjectives *dreadful* and *apprehensive*, we understand that they both involve fear which, in turn, involves the experiencer of the fear and the cause of the fear (the stimulus). Using the notion of orientation (which Haspelmath attributes to Lehmann 1984:152) we can say that *dreadful* is oriented towards its stimulus participant (i.e. the noun modified by *dreadful* is understood to be a stimulus), while *apprehensive* is oriented towards its experiencer (i.e. the noun modified by *apprehensive* is understood to be an experiencer).

Participles have a similar ability to display orientation and, moreover, any one verb can in principle produce a number of participles oriented towards any of the verb's participants. According to Nedjalkov & Jaxontov (1988:8-11) who do not use the notion of orientation but, in a similar spirit to Haspelmath's argumentation, propose a taxonomy of resultative constructions according to 'diathesis type', there are languages in which resultative participles may even be oriented towards non-core participants such as locations and beneficiaries (Nedjalkov & Jaxontov refer to them by their oblique-argument names of 'locatives' and 'datives'). However, the most frequently attested participles crosslinguistically are agent-oriented (also referred to as 'active' or 'subjective') and patient-oriented (also referred to as 'passive' or 'objective').⁶

Another widely acknowledged feature of participles, apart from their orientation, is that they can display some tense characteristics, in a similar way to finite verbs. This means that in addition to coding the particular property of the referent in terms of (or, with reference to) the event denoted by the verb, the participle can also specify the time at which the property of the referent applies relative to the time of the event. This has often been taken to mean that participles indicate tense (the location of the event in time) and has led to the

⁶Haspelmath (1994:154) further points out that, in cases such as the ones discussed in this paper – i.e. when the participial marker specifies the orientation – the participle is oriented *inherently*. However, it is also possible for participles to be inherently *unoriented* and oriented only *contextually*, as in, for example, Lezgian.

widely used labels ‘past’ and ‘present’ with reference to participles.

However, despite being traditionally called ‘past’ or ‘present’, some participles may not indicate tense at all, and instead they may, in fact, be able to refer to various time frames. An example of the participle whose time-reference is relative is the so-called ‘present’ participle in modern English (e.g. *singing*). It should be referred to as ‘contemporaneous’, since this term captures better the fact that the participle is non-finite, can be used within any time frame and interpreted accordingly.

4.4 The orientation of resultative participles

In this and the next two sections I will present an overview of what is known about the restrictions on the formation of resultative participles, based mostly on studies of English. The first restriction discussed is that of the affected participant.

In general, in transitive verbs the action usually affects the patient or theme, not the agent, and, for this reason, most transitive verbs tend not to make agent-oriented resultative participles. That is, **the sung performer* is implausible, and therefore unacceptable, even if the verb happens to be used intransitively (compare: *the sung ballad*), and the only available interpretation of *the abused teacher* is that the participle characterises the patient of the activity denoted by the verb.

However, it is inaccurate to say that resultative participles can only characterise patients. The particular semantic role fulfilled by the participant does not seem to be relevant to the formation of the resultative participle characterising that participant. Instead, what is relevant is whether the action has affected the participant – whether patient, theme, or agent – in a way that can be used to characterise it.

For example, sometimes a transitive action may be such that it affects the agent. If this is the case, it is possible to characterise the agent by means of the state resulting from the action, and resultative participles with ‘active’ orientation can be formed. Haspelmath (1994:160) cites examples of transitive agent-oriented participles from Hindi-Urdu. Polish is another language in which resultative participles of many semantically transitive verbs can be agent-oriented. These participles can be formed from telicised (as well as morphologically perfective) forms of verbs such as: *jeść* ‘eat’, *pić* ‘drink’ (and semantic derivatives of these two, e.g. *zreć* ‘devour/pig out’, *chlać* ‘tope/guzzle’), *ubrać* ‘put on’ and *zziębić* ‘cause to be cold’.

Verbs of this type encode actions which, despite involving two participants one of whom is a theme or patient, affect the agent saliently – that is, the action affects *both* the agent and the theme/patient. Therefore, as Haspelmath argues, it is not surprising that in some of these verbs the resultative participle can be *either* agent- or patient-oriented. The following examples are an illustration of this phenomenon in Polish⁷ (N.B. also the English translation of (8d)):

- (8) a. *wypita herbata* ‘(the) tea that has been drunk up’
b. *spity nektar* ‘(the) drunk nectar’
c. *ale jestem napity* ‘how full of drink I am (coll.)’

⁷Also, many derived reflexive verbs in Polish which ‘internalise’ the agent/experiencer in a transitive action can make agent-oriented resultative participles. This phenomenon is discussed in more detail in Kibort (2004), Chapter 3, Section 3.2.5.2.

- d. *spity chłopak* ‘a/the drunk boy’
- (9) a. *ubrany płaszcz* ‘a/the coat that is/was being worn’
b. *ubrany chłopak* ‘a/the boy who is/was dressed’
- (10) a. *przeziębione gardło* ‘a/the sore throat’ (lit. ‘a/the throat that has been exposed to the cold’)
b. *jestem przeziębiony* ‘I have a cold’
c. *przeziębiony chłopak* ‘a/the boy who has/had a cold’

In their typological survey of resultative constructions, Nedjalkov & Jaxontov (1988:9) treat agent-oriented resultative constructions as a separate category, calling them ‘possessive resultatives’, since in most of such constructions ‘the underlying object of the affecting action refers to a body part or possession of the underlying subject or to something in immediate contact with the latter’. They identify eight main groups of verbs that form ‘possessive resultatives’ crosslinguistically, including verbs of obtaining (‘take’, ‘receive’, ‘lose’), wearing (‘put on’, ‘wear’), ingestion (‘eat’, ‘drink’), and ‘mental ingestion’ (‘see’, ‘learn’, ‘study’) (cf. Haspelmath 1994:174, footnote 10).

Haspelmath (1994:161) further notes that agent-oriented resultative participles formed from transitive verbs had already been noted by Brugmann (1895) and Wackernagel (1920:288) with reference to the Latin ‘exceptionally active past participles’ such as *cenatus* ‘having eaten’ and *potus* ‘having drunk’.

Arguing in support of a different hypothesis, Bresnan provides more examples from English in which transitive agent-oriented resultative participles have been formed (2001:36, adapted):

- (11) a. *a confessed killer* [a killer who has confessed (his/her crime)]
b. *a recanted Chomskyan* [a Chomskyan who has recanted (his/her opinion about Chomsky)]
c. *(un)declared juniors* [juniors who have (not) declared (majors)]
d. *a practised liar* [a liar who has practised (lying)]
e. *an unbuilt architect* [an architect who has not built (buildings)]

She argues that all these verbs designate actions (verbal or other) that change one’s moral, legal, or administrative status. Resultative participles formed from these verbs are, therefore, felicitous both with patient/theme and agent orientation.

If we now look at intransitive verbs, both the semantically ‘unaccusative’ ones (having one patient participant) and the semantically ‘unergative’ ones (having one agent participant), the situation is not much different. Whether the participant of the action is semantically a patient or an agent, a resultative participle can be formed if the action has affected the participant and caused it to assume a state resulting from the action. The following are examples from English (Bresnan 1978:8 and 2001:34,35; see also Levin 1993:87):

- (12) a. *elapsed time* e. *a lapsed Catholic* i. *a stuck window*
b. *a fallen leaf* f. *a failed writer* j. *an escaped convict*
c. *the drifted snow* g. *wilted lettuce* k. *a risen Christ*
d. *a collapsed lung* h. *a grown man* l. *an undescended testicle*

Thus, all this evidence suggests that the orientation of resultative participles is ultimately determined by the semantics of the whole predicate rather than by any syntactic differences between the arguments of the verb, or even by the thematic classification of participant roles.

4.5 Semantic restrictions on resultative participles

It is clear that the formation of resultative participles is not restricted to involuntary events. Whether the change of state is involuntary or volitional, it is generally possible to form resultative participles characterising the participants which have undergone the change of state. However, a further restriction on the event has been noted: the verb expressing it has to be telic (in the sense of Vendler 1957, Dowty 1979).

Since the function of the resultative participle is to characterise an entity by means of a resulting *state*, atelic events which are not construed as resulting in any state cannot provide the semantic basis required for the formation of the resultative participle.

Haspelmath (1994:159) gives the following example. The English verbs *bloom* and *sleep*, which have single non-agentive participants, do not make resultative participles (**the bloomed dandelion*, **the slept dog*) because they are atelic. However, in languages in which atelic verbs can be telicised by a locative particle, resultative participles can, nevertheless, be formed from the derived telic variants of the verbs.

This is the case, for example, with German and Polish, in which both *bloom* and *sleep* can be telicised as in the following examples (the German ones are cited directly from Haspelmath; also compare with Polish (8)-(10)):

- (13) a. **der geblühte Löwenzahn* ‘the bloomed dandelion’
 b. *der aufgeblühte Löwenzahn* ‘the bloomed (‘blown’) dandelion’
- (14) a. **der geschlafene Hund* ‘the slept dog’
 b. *der eingeschlafene Hund* ‘the dog that has fallen asleep’
- (15) a. **kwitnięty mlec* or **kwitły mlec* ‘a/the bloomed dandelion’
 b. *rozkwitnięty mlec* or *rozkwitły mlec* ‘a/the bloomed (‘opened up’) dandelion’
- (16) a. **spany pies* or **spały pies* ‘a/the slept dog’
 b. *rozspany pies* or *ospały pies* ‘a/the dog that has been affected by too much sleep; a/the sleepy dog’

Furthermore, in a similar way to the German and Polish examples above, where verbs have been telicised by prefixation, English too can form agent- and patient-oriented resultative participles from some atelic verbs if they are accompanied by an appropriate telicising preposition or adverbial. Just as in the German and Polish examples, the English telicising elements too change slightly the meaning of the base verb:

- (17) (examples (a) and (b) adapted from Bresnan (2001:31))
 a. *After the tornado, the fields had a marched-through look.*
 b. *You can ignore any recently gone-over accounts.*
 c. *What’s the difference between a run-over snake and a run-over attorney?*⁸

⁸There are skid marks in front of the snake.

Cf. the unacceptability of: **marched fields*, **gone accounts*, **a run snake/attorney*.

- (18) ((a)-(e) adapted from Bresnan 2001:34,35)
- a. **a run slave vs a run-away slave*
 - b. **an exercised athlete vs an over-exercised athlete*
 - c. **a flown bird vs a flown-away bird*
 - d. **a flown pilot vs the most-distance-flown pilot*
 - e. **!a travelled correspondent vs a widely-travelled correspondent*
 - f. **a read person vs a well-read person*

Activities expressed with atelic verbs which lack an inherent result state can, thus, be supplied with goals, limits, or result states and provide the necessary semantic basis for the formation of the resultative participle.

Bresnan (2001:34-35) discusses a couple of other cases of English resultative participles in more detail, in an attempt to tease out semantic distinctions between the verbs that can, and the verbs that cannot form them. One of the discussed verbs is *leave*. Bresnan argues that **a recently left woman* is unacceptable because the predicate focuses on the source of motion, not on the goal or result state.⁹

The verb *grow*, on the other hand, displays the following contrast: *a grown man* is acceptable, while *?a grown tree* is problematic. Bresnan cites the following explanation by Goldberg (p.c.): ‘The former refers to a culturally recognized end-point, namely adulthood, while the latter does not since there is no culturally recognized end state of treehood.’ It is, nevertheless, possible to imagine that the latter phrase might be uttered by an expert gardener with respect to a plant whose state of ‘adulthood’ he or she is able to assess.

Finally, Bresnan discusses the phrase **a thanked person*, which she considers ill-formed ‘because there is no salient result state defined by the process of thanking’. Similarly, the phrase **untaken advantage* is unacceptable (although *untaken seats* is acceptable) because ‘complex predicates consisting of verb and noun combinations like *take advantage of* do not define a result state of the internal noun (e.g. *advantage*), which forms part of the idiom’ (2001:35).

The telicity restriction on the formation of resultative adjectives is, then, the consequence of the semantic requirement that the verb phrase must denote an event which has an end point or results in a state.

4.6 Pragmatic restrictions on resultative participles

Finally, it has been observed that the semantic condition of telicity stated above is a sufficient, but not a necessary condition for the formation of resultative adjectives.

Bresnan (2001:37) cites the following examples (from p.c. with Adele Goldberg) of resultative adjectives based on *atelic* verbs, both activities (19) and states (20):

- (19)
- | | |
|---------------------------------------|-------------------------------------|
| a. <i>long anticipated event</i> | c. <i>much talked about idea</i> |
| b. <i>much hoped for consequences</i> | d. <i>strongly backed candidate</i> |

⁹Bresnan attributes this and the following observation to Adele Goldberg (p.c.).

- (20) a. *much-loved doctor* d. *despised politician*
b. *much-feared consequence* e. *highly acclaimed actor*
c. *communally owned property* f. *well-known performer*

and remarks that most of these examples require adverbial modification to be felicitous. In fact, some examples given in (18) above can be argued to demonstrate just this point (i.e. *a widely-travelled correspondent, a well-read person, a well-prepared teacher*, etc.).

Without the appropriate adverbial modification or contrastive context, even some of the apparently most canonical – i.e. patient-oriented, transitive, and telic (due to the involvement of an appropriate theme and the location of the event in the past) – resultative adjectives seem to be problematic, cf. *?a read book, ?a drunk cup of tea, ?a built house*, in contrast with, e.g. *an unread book, a quickly/slowly drunk cup of tea, or a well/nicely built house*.

Ackerman & Goldberg (1996) explain this phenomenon by resorting to a general pragmatic condition of informativeness. Bresnan sums it up as follows: ‘The adverbial modification increases the informativeness of the attribute, and thus its acceptability. Pragmatic informativeness and the semantic result state condition are members of what may be a family of sufficient (but not necessary) conditions on the use of adjectives’ (2001:37).

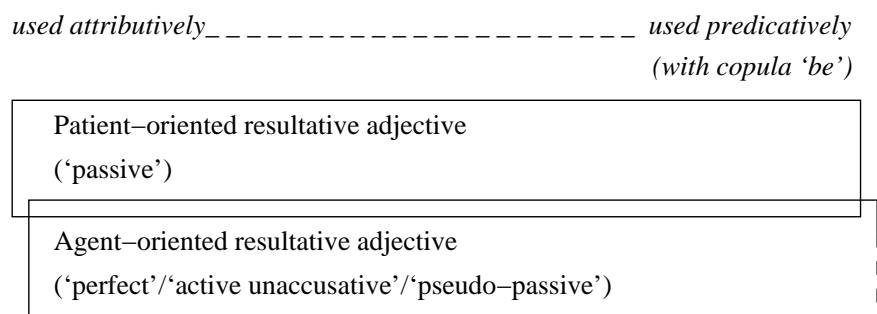
Thus, the formation of resultative adjectives is not driven by syntax, but it is driven (or, determined) by semantics. It is, however, ultimately licensed by the pragmatic requirement of ‘informational balance’. An utterance has to be non-trivial in the given context. If a resultative adjective is informationally deficient, it will not be considered acceptable, even if morphosyntax allows the formation of the resultative participle from the particular verb. The informationally-felicitous use of resultative adjectives may require adding some semantic material to the modifier to make the utterance non-trivial.

4.7 The uses of resultative participles as adjectives

In all the examples given above, I have so far concentrated on the attributive use of resultative participles. However, as modifiers, resultative participles can be used both attributively and predicatively (with the copula ‘be’). That is, just like other adjectives, most resultative participles can also appear as adjectival complements in subject-complement clauses.

The following diagram represents the uses of deverbal resultative adjectives in English:

- (21) Resultative adjectives (Kibort 2004:410)



The area of overlap between the two types of adjectives, i.e. between the patient-oriented ones and the agent-oriented ones, indicates those cases in which both types of participles can be formed from the same base verb (as discussed in the sections above). ‘Patient’ is understood here as either patient or theme, and ‘agent’ as either agent or experiencer of the situation denoted by the verb. All resultative adjectives are produced by the same morphological derivation. It is a lexical derivational process which is sensitive to the semantics of the predicate. All resultative adjectives are oriented towards the affected participant which is typically a patient, a theme, an experiencer, and occasionally an agent.

5 The passive

5.1 The passive construction: overview

Passivisation is a morphosyntactic derivation: it occurs at the level of argument structure of the predicate and it is an operation on grammatical functions. It downgrades the first argument of an unergative predicate to the status of an oblique, thus enabling the ‘promotion’ of the second argument, if there is one, to subject. It creates a new lexeme whose argument structure is different from the basic one: it is syntactically detransitivised. The alternative mapping of grammatical functions onto the arguments of the predicate provides a means to take a different perspective on truth-functionally equivalent situations (Ackerman & Moore 2001:3) and serves a useful discourse function by enabling a choice of different syntactic pivot.

Passive morphology is ‘an accidental fact’ about individual languages (Dryer 1982), because a passive predicate is recognisable by the syntactic status of its arguments (as compared with the active), not by its communicative function or by its form. In fact, the passive construction in many languages overlaps both in its communicative function and in its form with other, non-passive, constructions in the same language.¹⁰

It is to be expected that the passive can use different types of morphology both crosslinguistically and intralinguistically (see e.g. Haspelmath 1990). By far the most common strategy of forming the passive seems to be adding a passive affix to the stem of the verb, inside aspect, tense, and person markers (Dryer 1982:55, Haspelmath 1990:29). This suggests that the change in the interpretation of the predicate due to passivisation is more relevant to the meaning of the verb itself, or more internal to the predicates’ structure, than the modification brought about by a change of aspect or tense. This is consistent with our treatment of the passive as valency-changing, hence derivational, as opposed to tense which is inflectional.

Despite verb affixation being the most common way of forming passives, English does not use this method to derive its passives. Instead, it is typical of the Indo-European family of languages in using an auxiliary verb (a form of *be*, *become*, or *get*) combined with a form of the verb referred to as ‘passive participle’, as in *The window was broken by the boys from next door*. Although the participial verb form used in passive sentences is indeed passive in meaning (or, orientation), I argued in Section 4 that it did not derive this meaning

¹⁰It is widely known that the grammatical morphemes that mark the passive can have other – different but somehow related – uses, such as reflexive, reciprocal, resultative, anticausative, potential passive, fientive, reflexive-causative, deobjective, and desubjective (‘impersonal’) (see e.g. Shibatani 1985, Haspelmath 1990). In English, passive (participial) morphology is shared with the resultative.

from being part of the passive construction. Instead, it is a patient-oriented *resultative participle*. It can function as an adjective (and be used either attributively or predicatively with the copula ‘be’) and it can also be used in the passive construction as a main verb (with an accompanying auxiliary verb). In other words, resultative participles derived from unergative verbs can have an additional predicative function: they can be used as main verbs in the analytic verb form of the passive construction.

Due to this dual predicative function, the deverbal resultative elements occurring in the passive have been analysed as either adjectives (for example, by most movement-dependent syntactic accounts of the periphrastic passive) or as main verbs of analytic predicates (in traditional descriptions of English, e.g. Curme 1935:217ff; also, more recently, in Ackerman & Webelhuth 1998).¹¹

The present analysis follows the latter tradition in treating the analytic passive verb as a ‘verb complex’ comprising an auxiliary and a syntactically detransitivised main verb (the participle). In LFG, this can be understood as periphrastic exponence of the associated f-structure, with the auxiliary required because the participle is non-finite (Bresnan 2001:78).

5.2 The question of the ‘passive’ participle

The fact that the Indo-European periphrastic passive uses the resultative participle has made it problematic to identify the passive construction (see, for example, Quirk et al.’s 1985:167ff widely accepted, standard description of the ‘passive gradient’). This, in turn, has caused innumerable problems in attempts to define the passive and to account for it formally. Unless we accept that one form may be shared by two morphologically different constructions, distinguishing the passive participle from the resultative participle becomes a very difficult or even an impossible task, as the two participles are indistinguishable in some contexts (or, as I argue, because there is only one participial form, used by both constructions).

We can devise tests to establish the categorial status of the participial forms occurring in the constructions in question (this was mentioned above in Section 2). It has, however, often been assumed that all the clauses or phrases tested in this way are already passive, and so the tests have been assumed to distinguish ‘verbal *passives*’ from ‘adjectival *passives*’. Levin (1993:86-87), who provides a comprehensive list of publications which have discussed this distinction, points out, however, that ‘[t]here is some debate about whether a notion of “adjectival passive” that is distinct from “verbal passive” should be recognized’.

The widely held assumption about the ‘passiveness’ of all (or most) patient-oriented resultative participles stems from a particular understanding of the process of the morphological derivation of the deverbal adjective. In the previous sections I argued that from a large class of resultative adjectives, some (the ones which can be formed from unergative verbs and used as object-oriented) are also used by the passive construction. In some accounts (e.g. Bresnan 1978 and later publications; Grimshaw 1990, Huddleston 1984) a different hypothesis is offered. It is argued that the participial verb form which appears

¹¹According to Blevins (2001:356), ‘[t]he distributional criteria applied by post-Bloomfieldians such as Wells 1947 likewise define an extended notion of “verb” that encompasses auxiliary-verb and verb-particle combinations. This analysis survives in fact into the earliest transformational studies. Chomsky 1957 expands the category *Verb* as *Aux* + *V*, and then describes an analysis of *V* into *V_I* + *Prt* as “the most natural way of analyzing these [verb-particle] constructions” (page 39). It is only with the subsequent decision to exclude discontinuous constituents *tout court* that the status of such complex predicates became in any way problematic for generative accounts’.

in periphrastic passives may also function as a deverbal adjective. Huddleston, for example, states that in English 'in addition to the morphological process converting *-en* forms into central adjectives like *worried*, *surprised*, we have one converting *-en* forms into more marginal adjectives like *broken*' (1984:323). It may be argued that the latter hypothesis is organised passive-centrally, and that the two hypotheses are inverse of each other with respect to the passive.

As a result, in passive-centric accounts, sentences such as *The broken window was dangerous* are regarded as structurally passive by analogy with their assumed counterparts such as *The window was broken by the boys from next door*. The latter type of sentence is commonly referred to as a 'verbal passive' and taken to contain a 'verbal passive participle', while the former type of sentence is referred to as an 'adjectival passive' and taken to contain an 'adjectival passive participle'. As I showed in the previous sections, the participial forms used in both sentences are indeed morphologically identical.

Since passivisation is considered to be a lexical relation change altering the argument structure of the predicate, it should follow that adjectives derived from a passivised predicate should inherit the same altered argument structure. However, I suggest that it is, in fact, both impossible and unnecessary to determine whether the deverbal adjective is passive or not. Patients, themes or experiencers (i.e. affected participants) which provide the orientation for resultative participles do not have to be syntactically pre-specified as [-r] arguments ('underlying *objects*'), which would be the case if they were arguments of a passivised predicate. I also demonstrated above that some verbs (e.g. verbs of ingestion or wearing, in some languages) can form *both* theme-oriented and agent-oriented resultative participles using the same morphological means. The result, in both cases, is the same participle and it is unnecessary to posit that one has a passive argument structure while the other does not.

Looking now at all participles from the point of view of argument structure, the hypothesised rule which converts verbal participles into participial adjectives is assumed to operate in parallel either on passive (lexical) forms of verbs to produce 'passive' adjectival participles, or on non-passive forms of verbs to produce 'perfect' or 'present' adjectival participles. Since passivisation is assumed to be an argument-structure changing operation on the predicate, 'passive' adjectival participles derived from passive verbal stems are expected to have a passive argument structure, while the argument structures of 'perfect/past' participles are assumed to be non-passive.

Since the English 'passive' participle is identical in form with the 'perfect' participle, to distinguish between them we need to stipulate which one is (underlying) object- and which subject-oriented – that is, whether any particular morpholexical operation has been applied to the predicate prior to converting it into an adjective. This is done on the basis of the orientation that the participle displays towards a *semantic* participant.

The *eaten food* is assumed to be passive and understood as 'the food that is/was eaten' because eating food implies an agent performing the eating. Similarly, a *fallen leaf* is assumed to be non-passive ('perfect') because, on this understanding of *fall*, the phrase could not have been derived from a two-argument lexical structure corresponding to **someonesomething fell the leaf*, but instead it derives from the single-argument structure corresponding to *a leaf fell/has fallen*. However, verbs denoting actions which can be perceived as either agent-caused or spontaneous form participles which may be analysed as, simultaneously, either passive or non-passive. This can be illustrated with the following participles,

functioning as verbs or adjectives, and their potential source constructions ('counterparts'):

- | | | | |
|------|--|------|--|
| (22) | a. <i>the window was broken</i> | (23) | a. <i>the door was closed</i> |
| | b. <i>the broken window</i> | | b. <i>the closed door</i> |
| | c. <i>~ the window that was broken by the boys</i> | | c. <i>~ the door that was closed by me</i> |
| | d. <i>or ~ the window that has broken</i> | | d. <i>or ~ the door that has closed</i> |

In neither of the (b) phrases is it possible to determine whether the 'verbal' or 'adjectival' participle (or a construction of which it is part) is passive or non-passive. It is, therefore, not possible to determine which one of the hypothesised argument structures should be assigned to it.

I suggest that the formation of the so-called 'adjectival passive' is analogous to the formation of any other construction with a resultative participle in attributive or predicative function, and it does not require the application of the passive rule. There is no need to assume that adjectival passives have to be derived from a passivised verb phrase and there is no need to resort to the syntactic tier of argument structure in order to determine the orientation of the resultative participle. The resultative participle is neutral between being an adjective and a verb and can be used in both functions, including the function of the main verb of the passive construction.

Bresnan (2001:34) observes that deverbal adjectives in general denote a state derived from the semantics of the base verb. This seems to be true for all participles, whether resultative with patient or agent orientation, or contemporaneous (such as *a smiling woman*). As I argued in Section 4, all restrictions on the formation of resultative participles are semantic and pragmatic in nature, not syntactic.

To sum up, the classification of participles into passive and non-passive is misleading. If, as argued here, passivisation is a morpholexical operation on argument structure, a verb form can be called 'passive' only if its argument structure has been altered by this operation. Resultative participles (of all orientations) result from the process of morphological derivation in the lexicon and, like the verbs they are related to, they may have both argument structure and/or event structure, but their argument structure does not need to be altered when they are used as adjectives. All restrictions on the formation of resultative participles can be accounted for with recourse to semantics and pragmatics, while the primary constraint on the formation of the passive is syntactic (the predicates that passivise are syntactically unergative; Perlmutter 1978). The passive construction uses the resultative participle as the main verb of its analytic predicate and provides the only context in which a 'passive participle' can be identified as such.

5.3 The overlap of the passive and the resultative construction

Although passivisation is derivational, it is a morphosyntactic rule (or constraint) rather than a morphological derivational rule. It is both driven and determined solely by syntax. The passive operation targets the underlying subject of the predicate which is identified on the basis of its syntactic properties. If the argument structure of the predicate contains an underlying object argument, it becomes the syntactic subject of the passive clause and the situation denoted by the verb is predicated of it.

This last point captures the syntactic overlap between the passive and the resultative. In the active, the most typical affected participant, a patient or theme, is coded as an object. However, the predicative use of the resultative participle allows any type of affected participant (including the one which is an object in the active) to be coded as subject. In this way, the syntactic structure of the resultative in the form of an adjectival complement to the affected patient as subject (i.e. the predicative use of the resultative adjective) turns out to be a convenient vehicle to express the passive.

Apart from this area of the overlap, the two constructions diverge into areas exclusive to each of them. Resultative participles as adjectives can modify all sorts of subjects, including affected experiencers and affected agents, most of whom would be excluded from appearing as subjects in the passive construction either because of the unaccusativity of the predicate or because the argument bearing the agent role would be suppressed in the passive. In general, the passive can be formed of a subset of the verbs which allow the resultative. However, while the resultative participle as adjective has to modify a nominal head, the passive can be formed of intransitive predicates and, thus, the passive construction does not have to have a subject (i.e. there exist impersonal passives of intransitives, as in German or Polish). Additionally, *be* is the only verb which can accompany the participle in both constructions.

The diagram in (7) illustrated the overlap in the use of the resultative adjective and the passive participle in English and additionally showed the area of overlap between the passive and the analytic perfect tense construction, which are both driven by syntax and make use of the same derived verbal form as the semantically-driven resultative.

6 The revised rule of participial formation

Thus, we could formulate the following rule of participial formation from the base verb:

- (24) Morphological change: $V \mapsto [V_{Part}]_{A/V}$
 Operation on lexical form: (non-oriented) $P \mapsto$ semantically oriented P

Formulated as above, the rule holds for all categories of participles ('passive', 'perfect', 'present', etc.), with different *semantic* conditions on their derivation leading to their different semantic interpretations. For the resultative participle (with its particular morphology), the condition is that the derived lexical form P has to be semantically oriented towards the affected participant. For the 'present' (contemporaneous) participle (with its different morphology), the derived lexical form has to be semantically oriented towards the first participant, etc. Most importantly, the semantic orientation does not involve the syntactic notions of subject or object. Furthermore, all participles can in principle perform the function of either an adjective or a verb (A/V).

Thus, the morphological derivation of the resultative participle does not engage the syntactic level of argument structure at which the passive rule operates. The resultative derivation rule produces resultative participles which can be used attributively or predicatively, some of which are also suitable to be used by the passive construction. A patient-oriented resultative participle does not have to have been 'passivised' in order to be used as an adjective, just as the morphologically identical resultative participles with an orientation towards the first participant (agent or experiencer) are, naturally, not regarded as 'passivised'. Morphosyntactic passivisation is not required to have occurred either in the

predicative adjectival construction (with 'be') such as the one labelled 'verbal passive', or in the attributive adjectival construction such as the one labelled 'adjectival passive'.

Because of its direction, LFG's lexical rule of (verbal) participle-adjective conversion, cited at the beginning of this paper in (1), assumes that passivisation (if needed) occurs *before* the derived verb form can be used as an adjective in attributive constructions with 'passive' and 'perfect' deverbal adjectives. However, as I have shown, the construction with a resultative adjective, either in its attributive or predicative use, cannot always be unambiguously assigned a passive or non-passive argument structure, nor does it need to be always unambiguously classified as passive or non-passive. The participial formation rule in (24) does not come in the way of analysing passivisation in lexical terms as a constraint on argument structure, and predicts correctly the observations regarding the morphology and distribution of the resultative participle in its various functions.

References

- Ackerman, F. (1992). Complex predicates and morpholexical relatedness: locative alternation in Hungarian. In Sag, I. & Szabolsci, A. (eds.), *Lexical Matters*. Stanford, CA: CSLI Publications. 55–83.
- Ackerman, F. & Goldberg, A. E. (1996). Constraints on adjectival past participles. In Goldberg, A. E. (ed.), *Conceptual Structure, Discourse, and Language*. Stanford, CA: CSLI Publications. 17–30.
- Ackerman, F. & Moore, J. (2001). *A Theory of Argument Structure*. Stanford, CA: CSLI Publications.
- Ackerman, F. & Webelhuth, G. (1998). *A Theory of Predicates*. Stanford, CA: CSLI Publications.
- Blevins, J. P. (2001). Realisation-based lexicalism. *Journal of Linguistics* 37. 356–365.
- Bresnan, J. (1978). A realistic transformational grammar. In Halle, M., Bresnan, J. & Miller, G. A. (eds.), *Linguistic Theory and Psychological Reality*. Cambridge, MA: The MIT Press. 1–59.
- Bresnan, J. (1982). The passive in lexical theory. In Bresnan, J. (ed.), *The Mental Representation of Grammatical Relations*. Cambridge, MA: The MIT Press. 3–86.
- Bresnan, J. (1995). Lexicality and argument structure. Paper presented at Paris Syntax and Semantics Conference, 12 October 1995. Available online. 1–28.
- Bresnan, J. (2001). *Lexical-Functional Syntax*. Oxford: Blackwell.
- Brugmann, K. (1895). Die mit dem Suffix *-to-* gebildeten Partizipia im Verbalsystem des Lateinischen und des Umbrisch-Oskischen. *Indogermanische Forschungen* 5. 89–152.
- Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.
- Curme, G. O. (1935). *A Grammar of the English Language, Volume 1: Parts of Speech*. Boston, MA: Heath.
- Dowty, D. R. (1979). *Word Meaning and Montague Grammar*. Dordrecht: Riedel.
- Dryer, M. (1982). In defense of a universal passive. *Linguistic Analysis* 10(1). 53–60.
- Frajzyngier, Z. (1978). An analysis of be-passives. *Lingua* 46(2). 133–156.
- Grimshaw, J. (1990). *Argument Structure*. Cambridge, MA: The MIT Press.
- Haspelmath, M. (1990). The grammaticization of passive morphology. *Studies in Language* 14(1). 25–72.

- Haspelmath, M. (1994). Passive participles across languages. In Fox, B. & Hopper, P. (eds.), *Voice: Form and Function*. Amsterdam: John Benjamins. 151–177.
- Huddleston, R. (1984). *Introduction to the Grammar of English*. Cambridge: CUP.
- Kibort, A. (2004). *Passive and Passive-like Constructions in English and Polish*. Ph.D. thesis, University of Cambridge, Cambridge. Available online.
- Komlósy, A. (1994). Complements and adjuncts. In Kiefer, F. & Kiss, K. (eds.), *Syntax and Semantics 27: The Syntactic Structure of Hungarian*. San Diego, CA: Academic Press. 91–178.
- Kordoni, V. (2002). Participle-adjective formation in Modern Greek. In Butt, M. & King, T. (eds.), *Proceedings of the LFG02 Conference, National Technical University of Athens, Athens*. Stanford, CA: CSLI Publications. 220–238. Available online.
- Lehmann, C. (1984). *Der Relativsatz*. Tübingen: Gunter Narr.
- Levin, B. (1993). *English Verb Classes and Alternations*. Chicago, IL: University of Chicago Press. Part 1 also published as *English Verbal Diathesis*, Lexicon Project Working Papers 32, 1989.
- Levin, B. & Rappaport, M. (1986). The formation of adjectival passives. *Linguistic Inquiry* 17(4). 623–661.
- Lieber, R. (1980). *On the Organisation of the Lexicon*. Ph.D. thesis, MIT, Cambridge, MA. Distributed by Indiana University Linguistics Club, Bloomington.
- Markantonatou, S. (1995). Modern Greek deverbal nominals: an LMT approach. *Journal of Linguistics* 31. 267–299.
- Nedjalkov, V. P. & Jaxontov, S. J. (1988). The typology of resultative constructions. In Nedjalkov, V. P. (ed.), *Typology of Resultative Constructions*. Amsterdam: John Benjamins. 4–62.
- Perlmutter, D. M. (1978). Impersonal passives and the Unaccusative Hypothesis. *Berkeley Linguistics Society* 4. 157–189.
- Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Shibatani, M. (1985). Passives and related constructions: a prototype analysis. *Language* 61(4). 821–848.
- Siewierska, A. (1984). *The Passive. A Comparative Linguistic Analysis*. London: Croom Helm.
- Suppes, P. (1957). *Introduction to Logic*. New York: Van Nostrand.
- Vendler, Z. (1957). Verbs and times. *Philosophical Review* 56. 143–160.
- Visser, F. T. (1973). *An Historical Syntax of the English Language, Part Three, Second Half*. Leiden: Brill.
- Wackernagel, J. (1920). *Vorlesungen über Syntax. Band 1*. Basel: Emil Birkhäuser & Cie.
- Wells, R. (1947). Immediate constituents. *Language* 23. 81–117.
- Wetzer, H. (1996). *The Typology of Adjectival Predication*. Berlin: Mouton de Gruyter.

CLITICIZING LFG

Tracy Holloway King
Palo Alto Research Center

Proceedings of the LFG05 Conference
University of Bergen
Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications
<http://csli-publications.stanford.edu/>

Abstract

Why aren't *you* working on clitics?

1 Introduction

This paper represents the introduction to the workshop on LFG approaches to clitics held at LFG05. The workshop comprised: this introduction, three twenty-minute papers, and a discussion period. The first paper was by Michael Wescoat on “English nonsyllabic auxiliary contractions: An analysis in LFG with lexical sharing”. Further details can be found in Wescoat 2002 and in other works of his. The second paper was by Ana Luis and Ryo Otoguro on “Morphological and syntactic well-formedness: The case of European Portuguese proclitics”. Further details can be found in Luis 2004 and Luis and Otoguro 2004 and 2005. The third paper was by Rob O'Connor on “Clitics in LFG: Prosodic Structure and Phrasal Affixation”. Further details can be found in O'Connor 2002 and 2004.

Clitics have long been a fascinating topic for linguists because they involve intricate interactions between different grammar components, including syntax, prosody, and information structure. The two main goals of this workshop were: to present some issues and analyses of clitics in LFG so that they are accessible to the general LFG audience and to spark interest in clitics leading to further work on clitic phenomena from an LFG perspective.

LFG is an ideal theory in which to investigate clitic phenomena because its projection architecture provides a clear division between grammar components and a concrete notation with which to frame analyses within each component. However, there has been relatively little work on clitics in LFG compared to other theories. A quick look at some recent book-length publications shows a surge in the formal linguistic analysis of clitics, relatively little of which is in LFG. These works include Anderson 2005 (A-morphous morphology with Optimality Theory), Monachesi 2005 and 1999 (Head-driven Phrase Structure Grammar), Bošković 2001 (Minimalism), Gerlach and Grijzenhout (eds.) 2001, Franks and King 2000 (typology and Minimalism), and Halpern 1995 (Prosodic Inversion).

This is not to say that there is no LFG work on clitics or that the work done in other frameworks is incompatible with LFG. Instead, as the three papers presented at this workshop showed, LFG is an excellent theory in which to work on clitics: the reason for this workshop was to encourage more work in this area. In addition to the work presented at the workshop by Ana Luis, Ryo Otoguro, Rob O'Connor, and Michael Wescoat, there is LFG work on clitics that was not presented there. This includes detailed analyses of Tagalog second-position clitics by Paul Kroeger (Kroeger 1993) and of Hindi discourse clitics by Devyani Sharma (Sharma 2003). All of these works draw upon the insights of previous analyses of clitics. In some cases, LFG provides a way to account for a broader range of data than other frameworks. In other cases, the formal analysis of clitic constructions results in a clarification of how the different projections work together or in proposals for extensions of the theory.

Why should LFG focus more on the analysis of clitics? LFG's basic design involves a modularity of the grammar such that different types of information (syntactic, semantic, prosodic, etc.) are represented differently but can make reference to each other in well-defined ways. This interaction is best understood in the ϕ -mapping between c-structure and f-structure, although significant work has been done in the connection between argument-structure and f-structure and between information and discourse-structure and c- and f-structure. Despite this work, there are still gaps in our understanding of how the different components of the grammar interact with each other and the exact nature of each component and its mappings to other components.

This is where the study of clitics can provide vital clues to the structure of the formal system. The

analysis of clitics crucially involves more than one type of linguistic information. Among the most important are prosody, syntax, morphology, and information structure. Not all clitics make crucial reference to each component, but many languages contain clitics which require many or all of these and whose analysis will be complete only when all the components can be connected properly. It is these clitics that are of particular interest to most linguists working on clitics, including those who presented at this workshop. They show how LFG's projection architecture provides the necessary tools to analyze clitic constructions, and where the necessary formal devices are not yet in place, they provide proposals for this. In particular, LFG is a theory that allows access to each of these types of information, access from each type to the others, and a formal way to analyze these interactions. All of these are crucial to the analysis of clitics which is why work on clitics within LFG is vital both to the linguistic understanding of clitics and to LFG theory.

2 What is a Clitic?

Clitics are notoriously difficult to define, with linguists often resorting to the "I know one when I see one" approach to clitics. Some of the clearest discussion of the issue of defining clitics can be found in Klavans 1982 and Zwicky 1977; Anderson's 2005 work is also useful in this regard. In this section, I outline a few of the defining characteristics of clitics that are most relevant for the types of clitics that are likely to be of interest to linguists working in LFG.

The first part of the definition of a clitic is prosodic. Clitics are prosodically deficient elements in that they cannot appear on their own. For example, a clitic pronoun cannot be used as an answer to a question. Instead, a full pronoun must be used. This prosodic dependency means that clitics are often restricted to having an accented element to their left (for enclitics) or to their right (for proclitics).¹ This dependency is often represented by having the clitic be an element that prosodically subcategorizes for a prosodic word (Inkelas 1989). This is shown in (1).

- (1) a. Enclitic: [[]_w —]_w
 b. Proclitic: [— []_w]_w

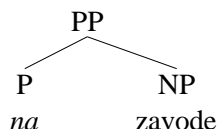
This prosodic requirement means that the clitic will form a prosodic unit with its host. This prosodic unit may or may not correspond to a syntactic unit. This frequent mismatch is particularly conducive to analysis in a theory like LFG which does not too tightly bind prosodic structure to syntactic structure.

A second characteristic of clitics is that they are (morpho)syntactically independent. This is what differentiates them from affixes which are morphologically part of the word to which they attach. An interesting theoretical proposal is that clitics are in fact a special subtype of affix: they are affixes that attach to syntactic or prosodic phrases which are larger than the prosodic word.

These two characteristics can give rise to very unusual behaviour from the viewpoint of the syntax. Zwicky 1977 captures this difference in his description of simple and special clitics. Simple clitics have a syntactic distribution similar to their non-clitic counterparts. This is seen in many languages which have proclitic prepositions. The prepositions themselves are clitics and form a prosodic word with the first prosodic word of the NP that is the object of the preposition. Syntactically, these clitics are usually considered uninteresting because their distribution is predictable: their prosodic dependency does not result in an unusual placement in the string. This can be seen with most Russian prepositions which are proclitics, as in (2).

¹Some clitics are less discriminating and can have a prosodic host either to their left or right.

- (2) a. Ona *na* zavode.
 she in factory
 ‘She is in the factory.’
- b. Prosody: [na [___]_ω]_ω ⇒ [na zavode]_ω
- c. C-structure:



In contrast to simple clitics whose syntactic distribution is similar to that of their non-clitic counterparts, special clitics have a syntactic distribution which is different from their non-clitic counterparts. Second position and verb adjacent clitics are almost always special clitics. Their position in the string is more restricted than that of their non-clitic counterparts and sometimes their non-clitic counterparts cannot appear in the position that the clitics do. The next section will discuss an example of Serbian/Croatian/Bosnian (SCB) second position clitics in detail. Languages often have both special and simple clitics. Here I focus on special clitics because they are more interesting from a syntactic perspective and hence to the majority of the LFG community.

One interesting phenomenon surrounding special clitics is that these clitics often form clusters. That is the privileged clitic position can contain a number of clitics, usually in a fixed order. These clitics can be of diverse types. For example, they can contain auxiliaries, pronominal arguments, and certain sentential adverbs, as in the examples in (3) which contain argument and auxiliary clitics together in the cluster. (Clitics are italicized in the examples.)

- (3) a. Kupila *mi* *ga* *je* jučer Vesna.
 bought me-DAT it-ACC AUX-3SG yesterday Vesna
 ‘Vesna bought it for me yesterday.’
- b. Ja *sam* *ga* *se* bojao.
 I AUX-1SG him/it-GEN refl-ACC feared
 ‘I was afraid of him/it.’

There are raging arguments in the literature as to how these clusters are formed and whether the ordering is synchronically morphologically templatic or a reflection of deeper syntactic factors. As we will see below, SCB has very complex clitic clusters in second-position.

2.1 Serbian/Croatian/Bosnian Second Position Clitics

SCB pronominal argument and auxiliary clitics occur in second position within their clause (Halpern 1995, Franks and King 2000 and references therein, and O’Connor 2002). Examples with both auxiliary and pronominal clitics were shown in (3). In (3a) the clitics follow the participial main verb *kupila*. The clitics comprise the indirect object *mi*, the direct object *ga*, and the auxiliary verb *je*. A sentential adverb and the subject follow the clitics. In (3b) the clitics follow the subject *ja* which is a full, non-clitic pronoun, and precede the participial verb *bojao*. The clitics comprise the auxiliary *sam*, the direct object *ga*, and an inherent reflexive *se*.

The clitics are in a fixed order. This order can be shown templatically as in (4). (All the auxiliaries except the third singular *je* appear initially; *je* appears finally; there are involved discussions as to why this occurs, e.g., Mišeska-Tomić 1996.)

(4) Q AUX DAT ACC GEN REFL AUX-JE

If they occur in any other order, the resulting string is ungrammatical, as in (5). This ungrammaticality does not appear to be prosodic in origin in that each of the clitics is an enclitic requiring a prosodic host to its left; this host is provided by the sentence initial constituent.

- (5) a. *Kupila *mi je ga* jučer Vesna.
b. *Kupila *ga mi je* jučer Vesna.
c. ...

SCB is largely discourse configurational: different constituents may appear in first position depending on the information structure. In each case, it is this initial constituent which hosts the clitic cluster. As seen in (6), the host can be a subject NP (6a), the main verb (6b), or an object NP (6c); other constituents such as PPs and adverbs can also act as hosts for the clitics.

- (6) a. Čovek *je* voleo Mariju.
man-NOM AUX-3SG loved Maria-ACC
'The man loved Maria.'
b. Voleo *je* Mariju čovek.
c. Mariju *je* čovek voleo.

In all the above examples, the clitics are in second position. In SCB, it is not only possible for the clitics to appear in second position, it is necessary that they do so. The clitics cannot appear in initial position, nor can they appear in third or further position in their clause.

First consider the situation in which the clitics are in initial position. Sentence word orders in which the clitic(s) are initial are ungrammatical, as in (7). The initial clitic(s) are ruled out for prosodic reasons. Since SCB clitics are enclitic, they require a prosodic host to their left. If they are in initial position, there will be no host to their left and hence the string is prosodically ill-formed. It may also be the case that there is a syntactic requirement that something appear before them, although this is difficult to detect given the prosodic requirements.

- (7) a. **Je* čovek voleo Mariju.
AUX-3SG man-NOM loved Maria-ACC
'The man loved Maria.'
b. **Je* voleo čovek Mariju.

The next situation to consider is when the clitics are in third position or even further from the beginning of the clause. Having clitics further to the right in the clause than second position is ungrammatical, as seen in (8).

- (8) a. *Čovek voleo *je* Mariju.
man-NOM loved AUX-3SG Maria-ACC
'The man loved Maria.'
b. *Čovek voleo Mariju *je*.

There are two situations in which clitics can occur further to the right. The first is the relatively uninteresting case in which the clitics are second in their clause, but the clause is subordinate, as in (9) where the clitic *ga* encliticizes to the complementizer *da*.

- (9) Marko ne zna da *ga* voli Vesna.
 Marko neg know-3SG C him-ACC loves Vesna
 ‘Marko doesn’t know that Vesna loves him.’

In this situation, the clitics always follow the complementizer which is the first element in their clause. Thus, within the relevant CP, the clitics are still in second position. The only potential difficulty is to make sure that the clitics are associated with the correct clausal domain. The second situation in which clitics occur further to the right is when there is additional material, such as left-dislocated constituents, which form a different prosodic phrase from the main clause. These are discussed more in section 3.

In the above discussion of SCB second position clitics, we have avoided an important question: What does second position mean? There are two possibilities: after the first syntactic constituent or after the first prosodic word. In most of the above examples, the first syntactic constituent (generally assumed to be the first maximal projection) also coincided with the first prosodic word because the constituents were prosodically simple. However, more complex syntactic constituents, such as complex NPs and PPs can shed light on this question since they can comprise multiple prosodic words. It turns out that both possibilities appear to occur. In (10), the first syntactic constituent is *taj čovek* which can comprise two prosodic words. The clitics can appear after the entire NP, as in (10a) or after the first prosodic word, as in (10b). Thus, upon initial inspection, second position clitic placement in SCB appears to be driven by either prosodic or syntactic considerations.

- (10) a. [_{NP} Taj čovek] *je* svirao klavir.
 that man AUX-3SG played piano
 ‘That man played the piano.’
 b. [_ω Taj] *je* čovek svirao klavir.

3 Syntax, Prosody, or Both?

In analyses of clitic placement, there are three main schools of thought: the placement is driven entirely by the prosody and other phonological factors; the placement is driven entirely by the syntax; the placement is driven by a combination of prosodic and syntactic factors. All three proposals have been made in the analysis of SCB second position clitics. The basic idea behind these is discussed here to provide an example of the types of issues that arise in the placement of special clitics. To make the discussion more concrete, I will focus on the example in (10) above in which the clitic appears either after a complex nominal subject or between a demonstrative and the head noun it is associated with.

3.1 Pure Prosodic Analyses

Purely prosodic accounts of SCB clitic placement are relatively rare. The most complete purely prosodic account of clitic placement in SCB is that of Radanović-Kocić (1988, 1996). Under these accounts, the clitic cluster is associated with a given clause, or verb heading a clause, via the syntax but the placement within the clause is derived entirely on prosodic grounds. In particular, the clitics will appear after the first phonological phrase in the prosodic domain (usually the intonational

phrase) that they belong to. The difference in placement in examples like (10) reflects a difference in the way in which the phonological phrases are composed in the clause.

- (11) syntax: *je* [_{NP} *taj* *čovек*] *svirao klavir*
 prosody 1: [_ω *taj*] *je* *čovек svirao klavir*
 prosody 2: [_ω *taj* *čovек*] *je svirao klavir*

Arguments in favor of the prosodic analysis include the straightforward explanation for why complex nominal and prepositional phrases can be split by clitic clusters and for why clitic third can be found in sentences which have heavy fronted material or “comma” intonation, as in (12). With comma intonation, the material before the pause (indicated by the # in (12a)) does not count when determining second position for the clitics.

- (12) a. [*Ove godine*] # [*taj pesnik*] *mi je napisao knjigu.*
 this year that poet me-DAT AUX-3SG wrote book
 ‘That poet wrote me a book this year.’
 b. [*ove godine*] [[[*taj pesnik*]_ω *mi ga*]_ω *napisao knjigu*]

Under the prosodic accounts, the first constituent *ove godine* forms a distinct prosodic phrase from that of the prosodic phrase that comprises the clitic domain. Hence, there is no difficulty with accounting for clitic-third placement because the clitics still form a prosodic word with the first prosodic word in their domain.

3.2 Pure Syntax Analyses

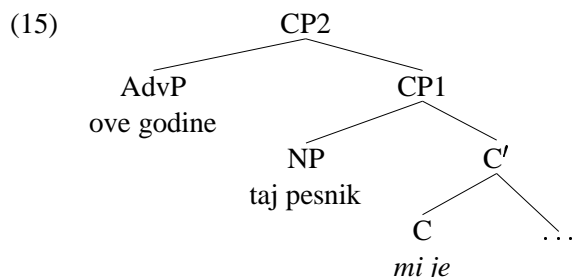
There are also purely syntactic approaches to clitic placement. Under these analyses, the syntax places the clitic cluster in the appropriate location and the prosody just determines whether the clitics are enclitic or proclitic. For the example in (10), this means that the demonstrative *taj* appears independently in initial position in the syntax, as shown in (13).

- (13) syntax/prosody: [_{NP/ω} *taj*] *je* [_{NP} ∅ *čovек*] *svirao klavir*

It turns out that in the majority of cases, the pre-clitic material can move independently of the post-clitic material in SCB since separable constituents are more common in SBC than in, for example, English. (14a) shows an example where a possessive adjective *Anina* is discontinuous from its head noun *sestra*. The demonstratives shown in the examples in this paper can similarly be split from their head nouns, as in (14b). As there are no clitics, these examples show independent evidence of split constituents in SCB.

- (14) a. *Anina dolazi sestra.*
 Ana’s come-3SG sister
 ‘Ana’s sister is coming.’
 b. *Tog Milena voli čovjeka.*
 that Milena loves man
 ‘Milena loves that man.’

However, the pure syntax approaches run into difficulties with some of the clitic third examples, where a fronted constituent appears to be in the same clause, but in a different prosodic phrase. This can occur when the clause contains left-extrapolated material, such as certain dislocated topics. This material often forms its own prosodic phrase, separate from that of the core of the clause. When it does, then the clitics follow the next constituent, as in (12). Radanović-Kocić (1988, 1996) discusses examples of this type in detail. This data is often used as evidence against syntactic accounts of second position clitic placement in SCB because the prosodic phrasing of the dislocated element is crucial for determining second position. However, the proponents of pure syntax approaches argue that material such as *ove godine* in (12) is syntactically, as well as prosodically, outside the relevant clausal domain, suggesting syntactic structures such as (15) where CP1 represents the relevant domain for clitic placement.



In addition, some syntax-only accounts have trouble with verb-initial clauses in which the verb hosts the clitics, as in (16).

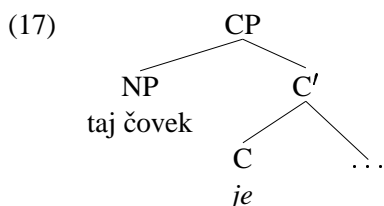
- (16) Kupila *mi ga je* jučer Vesna.
 bought me-DAT it-ACC AUX-3SG yesterday Vesna
 ‘Vesna bought it for me yesterday.’

The issue here is how to motivate the movement of the verb before the clitics. One approach is to claim that when the structure would be prosodically ill-formed because the clitics would be in initial position, then the verb can move to a position before the clitics. This involves an interesting interaction between the prosody and the syntax; another manifestation of this interaction is discussed below for the hybrid approaches. However, many proponents of the pure syntax approach believe that this foreknowledge of a prosodic failure by the syntax cannot govern syntactic movement or the realization of an alternative c-structure. If the prosodic needs of the clitics cannot motivate the verb’s preceding the clitics, which are generally assumed to be in C^0 or somewhere else high in the clause structure, then it is difficult to find other motivation for this as participles are normally quite low in the SCB phrase structure. Although “movement” of the verb is not an issue for LFG accounts, the same basic issue would arise with an LFG syntax-only account of SCB clitics: why is the verb sometimes generated high in the clause and sometimes lower?

3.3 Hybrid Analyses

Under hybrid approaches, some orders are derived syntactically and others prosodically. Halpern’s (1995) Prosodic Inversion analysis is a canonical example of a hybrid approach. The clitics occur in a particular syntactic environment, such as C^0 .² If there is non-clitic material in SpecCP or C^0 which can act as a host to the enclitics, then the structure is well-formed and no special prosodic processes are invoked. For the example in (10a), the syntactic tree would be as in (17). Since there is a full NP in SpecCP, the prosodic requirements of the enclitic are satisfied.

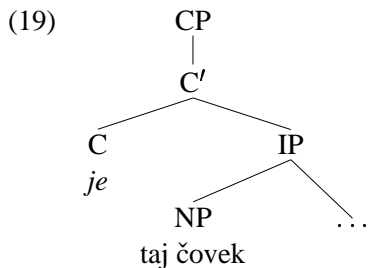
²Halpern (1995) in fact argues that they are slightly lower in the tree. Here I show the C^0 analysis for simplicity.



Note that the appearance of the NP *taj čovek* in SpecCP is independent of the prosodic requirements of the clitics. Instead, it appears in SpecCP for information structure reasons, such as topic interpretation. A similar situation arises in subordinate clauses where the complementizer in C^0 hosts the clitics, as in (18) repeated from (9).

- (18) Marko ne zna da *ga* voli Vesna.
 Marko neg know-3SG C him-ACC loves Vesna
 ‘Marko doesn’t know that Vesna loves him.’

However, if there is no complementizer in C^0 and no fronted phrase in SpecCP, then the syntactic structure of the clause leaves the clitics in initial position, as in (19).

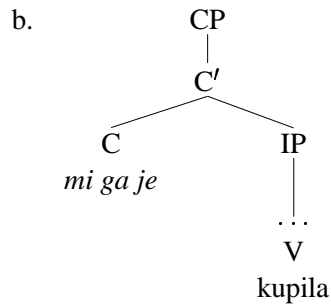


Under pure syntax accounts, such a configuration would result in an ungrammatical structure. Under Halpern’s Prosodic Inversion account, there is a prosodic process whereby the enclitics appear after the first prosodic word to their right; in (19) this is the demonstrative *taj*. This results in the surface word order seen in (10b). The prosodic aspect of this process is shown in (20).

- (20) *je* [taj]_ω čovek ...
 ↖ ↗ ↑

Prosodic Inversion is sometimes referred to as a last resort strategy because it only applies when there is no other way for the prosodic requirement of the clitics to be satisfied. This situation also arises in what would otherwise be verb initial constructions; these occur relatively frequently in SCB due to pro-drop of subjects, although the discourse configurability of the language does result in initial objects and other constituents. In general, finite verbs in SCB are assumed to be in I^0 or a related exploded Infl position while participles are in V^0 (see Bošković 2001 for detailed discussion as to verbal and clitic c-structure position in SCB; his argument is that the phrase structure of SCB clauses is much more complicated, especially as regards the position of the clitics). If the clitics are high in the c-structure in C^0 while the verb is lower in the clause in I^0 or V^0 , then clauses composed of just a verb and its clitic arguments (and possibly clitic auxiliary) will look roughly like (21b) for the sentence in (21a).

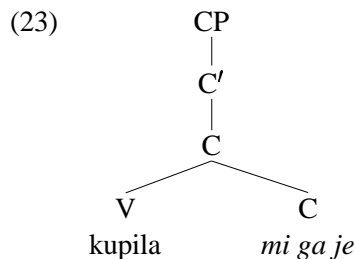
- (21) a. Kupila *mi* *ga* *je*.
 bought me-DAT it-ACC AUX-3SG
 ‘She bought it for me.’



Under the Prosodic Inversion analysis, the syntactic structure stays as in (21b) and the clitics are provided with a host in the prosody by occurring after the first prosodic word to their right, in this case the verb participle *kupila*, as in (22).

(22) *mi ga je* [*kupila*]_ω ...

Under analyses in which prosodically ill-formed structures, such as enclitic initial structures like that in (22b), can be repaired by changes to the syntactic structure, the canonical structure in (22b) is replaced by a c-structure as in (23) in which the head of the clause appears in C⁰ and provides a host for the clitics, similar to the situation in subordinate clauses with an overt complementizer, as in (18).



Hybrid accounts thus depend on a complex interaction of syntactic and prosodic factors to determine clitic placement. Most hybrid accounts assume that there is a canonical syntactic placement for the clitics in the c-structure and that when their prosodic requirements are met in that position, then there is no change to that structure. That is, the surface word order is identical to that of the leaves of the tree. However, if the clitics' prosodic requirements are not met, in the case of SCB not having a host to their left, then either the prosody must alter the order of the string, as in the Prosodic Inversion accounts, or some other syntactic structure must be found that does provide an adequate prosodic host.

Thus, we have seen that clitic placement in SCB has been analyzed as being purely prosodic, purely syntactic, or a combination of syntactic and prosodic factors. In each approach, it is necessary that the clitics have a proper prosodic host within the domain of their clause.³

³Clitic climbing provides an interesting challenge to the idea of clitic domain. In clitic climbing, the clitics that are associated with a subordinate clause verb appear on the matrix verb (see Monachesi 2005 for extensive Romance data). This often occurs cross-linguistically in causative and light verb constructions, and generally constructions involving clause union (Aissen and Perlmutter 1983). Clitic climbing is often obligatory, although some languages have optional clitic climbing in certain constructions. A SCB example of clitic climbing is shown in (i).

- (i) *Marija ju je pustila da pliva.*
 Maria it/her-ACC AUX-3SG let C swims
 'Marija let it/her swim.'

4 LFG's Projection Architecture

I hope that the brief discussion of SCB clausal clitics in the preceding sections provides an answer for the question why clitic phenomena are of interest to LFG. The analysis of clitics crucially involves more than one type of linguistic information, a strength of LFG theory. As seen in the discussion, the types of information needed for the analysis of clitics most obviously include syntax, especially c-structure, and prosody (see Butt and King 1998 on a prosodic projection for LFG).

Although it was not discussed here in detail, information structure is also important in the analysis of clitics. Information structure can influence the syntax of the clause and hence the relative placement of the clitics, and many discourse markers are themselves clitics (Sharma 2003). In addition, information and discourse structure can influence whether a clitic pronoun is chosen as opposed to a full form pronoun or even a noun phrase. Finally, information structure is crucial in determining when clitic doubling occurs. A Bulgarian example of clitic doubling is shown in (24) from Jaeger and Gerassimova (2002), who propose an LFG analysis of Bulgarian clitic doubling crucially involving information and discourse structure.

- (24) Decata *ja* običat Marija/neja.
children-DEF her-3SG love-3PL Maria/her
'The children love her.'

Also not discussed in detail in this paper, morphology is also needed for the analysis of clitics. Morphology is often assumed to be relatively unimportant for clitic placement within the clause (but see Anderson 2005 and Luis 2004 for analyses that make crucial use of morphology for clitic placement). Morphology and phonology play a more prominent role in the ordering of clitics within the cluster, and analyses differ as to whether these are entirely responsible for cluster ordering or if the order is all or partially derived from the syntax. Morphophonology is also crucial in accounting for unexpected surface forms of clitics. These peculiarities occur in the clitic systems of many languages. For example, in SCB when the cluster contains the accusative third singular clitic *je* directly adjacent to the third singular auxiliary clitic *je*, they surface as *ju je*; the *ju* form is not seen elsewhere. As another example, the SCB third singular auxiliary *je* can be dropped after the reflexive *se* and sometimes after the first and second person clitics *me* and *te*. These facts are generally assumed to follow from the morphophonological analysis of the clitic cluster.

LFG's projection architecture provides access to each of these types of information, including access from each type to the others. These interactions are formally well defined, allowing the theory to make concrete predictions about clitics in a given language. In addition, all of the components that are needed for the analysis of clitics have been independently proposed for LFG. Thus, the analysis of clitics does not require a radical reformulation of the theory or architecture of LFG. Instead, their analysis should shed light on the nature of the interactions between the different projections. The bottom line is that LFG theory is not complete without an analysis of clitics and analysis of clitics will make the overall architecture of the LFG theory and formalism clearer.

References

- Aissen, J., and D. Perlmutter. 1983. Clause Reduction in Spanish. In D. Perlmutter (ed.) *Studies in Relational Grammar 1*, 360–403. The University of Chicago Press.
- Anderson, S. 2005. *Aspects of the Theory of Clitics*. Oxford University Press.
- Bošković, Ž. 2001. *On the Nature of the Syntax-Phonology Interface*. North-Holland Elsevier.

- Butt, M., and T.H. King. 1998. Interfacing Phonology with LFG. In M. Butt and T.H. King (eds.) *Proceedings of LFG98*. CSLI Publications.
- Franks, S., and T.H. King. 2000. *A Handbook of Slavic Clitics*. Oxford University Press.
- Halpern, A. 1995. *On the Placement and Morphology of Clitics*. CSLI Publications.
- Gerlach, B., and J. Grijzenhout. 2001. *Clitics in Phonology, Morphology, and Syntax*. John Benjamins.
- Inkelas, S. 1989. *Prosodic Constituency in the Lexicon*. PhD thesis, Stanford University.
- Jaeger, T.F., and V.A. Gerassimova. 2002. Bulgarian Word order and the Role of the Direct Object Clitic in LFG. In *Proceedings of LFG02*, 197-219. Stanford: CSLI Publications.
- Klavans, J. 1982. *Some Problems in the Theory of Clitics*. Indiana University Linguistics Club.
- Kroeger, P. 1993. *Phrase Structure and Grammatical Relations in Tagalog*. CSLI Publications.
- Luis, A. 2004. *Clitics as Morphology*. PhD thesis, University of Essex.
- Luis, A. and R. Otaguro. 2004. Proclitic Contexts in European Portuguese and their Effect on Clitic Placement. In *Proceedings of LFG04*, CSLI On-line Publications.
- Luis, A. and R. Otaguro. 2005. Morphological and syntactic well-formedness: the case of European Portuguese proclitics. In *Proceedings of LFG05*, CSLI On-line Publications.
- Mišeska-Tomić, O. 1996. The Balkan Slavic Clausal Clitics. *NLTT* 14:811-872.
- Monachesi, P. 1999. *A Lexical Approach to Italian Cliticization*. CSLI Publications.
- Monachesi, P. 2005. *The Verbal Complex in Romance: A Case Study in Grammatical Interfaces*. Oxford University Press.
- O'Connor, R. 2002. Clitics and Phrasal Affixation in Constructive Morphology. *Proceedings of LFG02* 315–332. CSLI On-line Publications.
- O'Connor, R. 2004. *Information Structure in Lexical-Functional Grammar: The Discourse-Prosody Correspondence in English and Serbo-Croatian*. PhD thesis, University of Manchester.
- Radanović-Kocić, V. 1988. *The Grammar of Serbo-Croatian Clitics: A Synchronic and Dyachoric Perspective*. PhD thesis, University of Illinois, Urbana-Champaign.
- Radanović-Kocić, V. 1996. The Placement of Serbo-Croatian Clitics: A Prosodic Approach. In A. Halpern and A. Zwicky (eds.) *Approach Second: Second Position Clitics and Related Phenomena*. CSLI Publications.
- Sharma, D. 2003. Discourse Clitics and Constructive Morphology in Hindi. In M. Butt and T.H. King (eds.) *Nominals: Inside and Out*. CSLI Publications.
- Wescoat, M. 2002. *On Lexical Sharing*. PhD thesis, Stanford University.
- Zwicky, A. 1977. *On Clitics*. Indiana University Linguistics Club.

WHY GLUE A DONKEY TO AN
F-STRUCTURE WHEN YOU CAN
CONSTRAIN AND BIND IT INSTEAD

Miltiadis Kokkonidis
Computer Laboratory, University of Cambridge, UK
and
Meta Research, Athens, Greece

Proceedings of the LFG05 Conference
University of Bergen
Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications
<http://csli-publications.stanford.edu/>

Abstract

The semantic treatment of anaphora using λ -DRT with Glue which I present combines the strengths of both by assigning to each the task where it arguably fares best: Glue composes meanings and DRT deals with anaphoric resolution. Key to this approach is a simple first-order system for λ -DRT that allows LFG syntactic constraints to be transferred into the dynamic representation language. This parallels the transfer of such constraints into Glue types. Whereas approaches treating anaphora using Glue context management take advantage of this transfer, an earlier approach also leaving the treatment of anaphora to a compositional variant of DRT failed to account for syntactically-motivated anaphoric resolution constraints. On the other hand, the existing Glue context management approaches come not only at the cost of coupling context management with meaning composition, but also at the additional cost of the various remedies to the problems this uneasy cohabitation results in. The best of these approaches and the one presented here are currently very similar with respect to the range of phenomena they can correctly account for, but there are reasons to believe the latter is more scalable.

1 Introduction

DRT (Kamp, 1981; Kamp and Reyle, 1993) and its various compositional variants such as λ -DRT (Bos et al., 1994; Kohlhase et al., 1995) treat anaphors as underspecified terms to be resolved to bound variables. On the other hand, it is common practice in Glue literature (Dalrymple et al., 1999; Crouch and van Genabith, 1999; Dalrymple, 2001) to treat anaphora resolution within some system of anaphoric context management operating alongside meaning composition.

Not only is the common Glue approach more complex computationally and intuitively, but it has also proven hard to get right. The most complete treatment of this genre, that of Dalrymple (2001), resorts to imitating DRT, adapting its scoping rules and maintaining a DRT-style context of discourse referents.

If a DRT-style treatment is the aim, then in the spirit of a modular approach, especially given the close historic links of the Glue community with the LFG community, a design using a compositional variant of DRT as the meaning representation Glue has to work with makes much more sense. As a major part of LFG's success is that it allows talking separately about functional and constituent structure, recognising the importance of both and showing how they relate to each other, one would expect the same approach to be taken in semantics. Combining Glue with a system such as λ -DRT leads to a simple, modular design. Glue can be used to combine the λ -DRT expressions corresponding to the meaning contributions of the parts into an anaphorically underspecified meaning of the whole, which can then lead to different fully-specified meanings by means of DRT anaphoric resolution. The separation of meaning composi-

tion from anaphoric resolution is not a novel idea by any means. It underlies the design of compositional variants of DRT that predate the Glue attempts of dealing with both issues together.

Given the popularity of DRT within the LFG community one could reasonably expect that Glue would be used to replace other ways of combining meaning expressions given in λ -DRT (or some other DRT variant) and that anaphoric resolution would be left to DRT. That would not only be reasonable; it would also have been remarkably straightforward. The only technical challenge would be making DRT respect syntactic anaphoric constraints as expressed in LFG.

However, that was not what happened. A fascination with linear logic and the success of Glue in composing meanings lead to research trying to use it for a variety of other somewhat related tasks including anaphora resolution. Even though the work of Dalrymple et al. (1999) on using Glue for anaphora was already showing some signs that doing something like that could be problematic, the limited success that early approach had was taken as an indication that with more work more complete treatments could emerge. The modular solution based on a dynamic meaning representation language first appeared in the work of van Genabith and Crouch (1997), but Glue research remained focused on the context management approach.

Due to recent work advancing the context management approach (Dalrymple, 2001) on one hand and the present work on the other, the two approaches to the semantic treatment of anaphora have now reached the same (not very high) standards of coverage of the phenomenon. This is a welcome development as the previous *status quo* was quite unsatisfactory. Earlier all-Glue attempts had insufficiently developed context management, but could take into account syntactic anaphoric constraints. The proposal of van Genabith and Crouch (1997) got advanced anaphoric management for free by combining Glue with CDRT but failed to address the issue of enforcing LFG syntactic constraints during anaphoric resolution.

The technical contribution of the present work lies not in the straightforward combination of Glue with λ -DRT, but in showing how syntactic information can be imported into the meaning representation language, thus enabling the enforcement of syntactic constraints during anaphoric resolution. This technique is fairly generic and is tied to neither LFG, Glue, nor λ -DRT, but here it will be used to link λ -DRT discourse referents to their corresponding *f*-structures¹ in order to enforce LFG syntactic binding constraints (Dalrymple, 1993) within DRT.

Section 2 shows how LFG, Glue and λ -DRT are combined and how they deal with a simple example involving compositional ambiguity. It will be interesting to note how minimally intrusive the proposed technique is: nothing changes as far as Glue is concerned and all that is added in DRSs is a simple type for each discourse referent. Section 3 discusses how syntactic anaphoric constraints are imported into λ -DRT representations. The one change to the well-formedness rules of DRT is that they now require that the two variables appearing on the left and on the right of an equals sign have the same type. A function mapping *f*-structures to anaphoric indices is used to encode syntactic anaphoric constraints in a way that can be conveniently combined with the simple DRT system. Section 4 discusses earlier approaches to treatments of anaphora in the Glue literature. Finally, Section 5 argues for the approach presented here, claiming it is simpler and more scalable than its Glue context management counterpart.

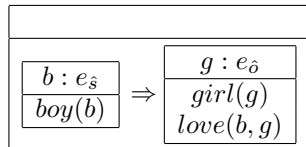
¹Not only is the technique not tied to Glue in any way, it bypasses it completely.

2 Setting the scene: LFG - Glue - λ -DRT

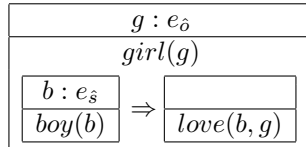
Below we have a sentence and the semantic representations corresponding to its two readings as given by our approach. Comparing these with their plain λ -DRT counterparts, the only addition is the type assignments for the discourse referents. Given that in plain λ -DRT they would all be treated as having the same entity type, the only real difference is that in the approach presented here the types of discourse referents are differentiated according to the anaphoric index associated with their corresponding f-structure. It really is that simple. However, as we will see, in this case at least, with simplicity comes power. Having given away the ending, let us see how we get there. We will start by seeing how LFG, Glue and λ -DRT combine.

Every boy loves a girl. (1)

Reading 1

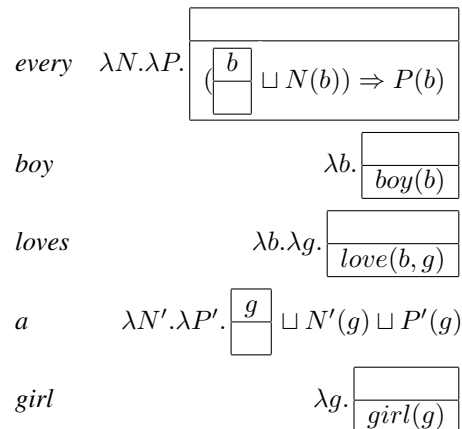


Reading 2



Our fundamental assumption will be that the meaning of a sequence of words is composed of the meanings of the words plus the meanings of certain syntactic constructs found in it (e.g. relative clauses) which we recognise as making a semantic contribution when the contributions of the words alone cannot account for its composite meaning. The question then is how we get from sequences of words to semantic representations for these sequences in a precise, systematic fashion.

Our first step will be to assign a meaning to each word of the given sentence. Using plain λ -DRT as our semantic notation, the meaning assignment for (1) is:



Using the words as shorthands for their meanings we could write the two readings of (1) as

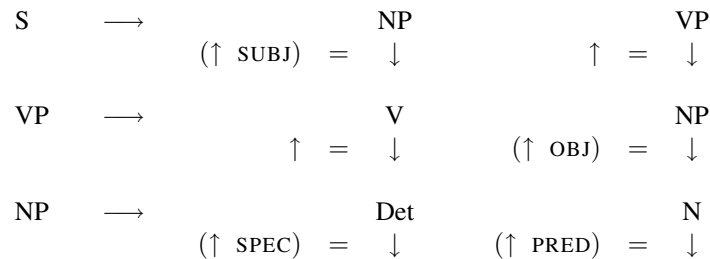
every boy $\lambda x. a \text{ girl } (\text{loves } x)^2$

and

a girl $\lambda y. \text{every boy } \lambda x. \text{loves } x y$

respectively.³ The question now is how we get these two readings while excluding other combinations that are correct DRSs but do not correspond to a possible interpretation of (1). This is where Glue enters the picture. Glue is a type system that is used for discovering the well-formed combinations of meaning contributions. Its resource sensitivity immediately rules out combinations where a meaning contribution is arbitrarily duplicated or omitted and more importantly its label sensitivity rules out combinations where meaning contributions are combined in a manner not justified by the syntactic structure of the given word sequence. Kokkonidis (2006) gives details and examples.

This brings us nicely to our next step in getting from (1) to its meaning. While there is also a long tradition of studying syntax independently of semantics, the study of semantics usually presupposes syntactic structures have been assigned to word sequences that are to subsequently be analysed semantically. We have approached the problem from the point of view of semantics. Now we will approach it from a syntactic viewpoint.



LFG (Kaplan and Bresnan, 1982; Dalrymple, 2001) views syntax not only in terms of c(onstituent) structure, but also in terms of f(unctional) structure. LFG researchers have often argued that there are generalisations that can be expressed in terms of f-structure that are not easily expressed in c-structure terms. What is particularly interesting for our purposes is how c-structure and f-structure relate to each other and how Glue and our enhanced λ -DRT draw sufficient information from them to avoid erroneous readings.

LFG constituent structure rules are expressed in a notation not very different from that used in other formalisms. Below each LFG constituent structure rule are the constraints for forming the corresponding functional structure. Up and down arrows are metavariables standing for f-structures. Their use is best understood if we consider these constraints as being evaluated on the arcs of the syntax tree resulting from the application of the constituent structure rules above them: the up arrows stand for the f-structure of the category above (the one on the left-hand side of the constituent structure rule) and the down arrows stand for the f-structure of the category below (the right-hand side category below which the constraint is written). The f-structure for a sentence is the minimal f-structure that satisfies all f-structure constraints.

²The η -equivalent expression *every boy* $\lambda x. a \text{ girl } \lambda y. \text{loves } x y$ is perhaps more familiar to some readers.

³Strictly speaking these λ -DRT expressions are not identical, but β -equivalent to the DRT readings originally given. We get the original expressions by β -reduction.

The resulting f-structure f for (1) is shown below. The f-structure for the VP is the same as that for the sentence i.e. f , the f-structure for the subject NP is $s = f$ SUBJ. Since the f-structure for the VP is f , the f-structure for the object NP is $o = f$ OBJ.

$$f : \left[\begin{array}{l} PRED \quad \text{'LOVE'} \\ SUBJ \quad s : \left[\begin{array}{l} SPEC \quad \text{'EVERY'} \\ PRED \quad \text{'BOY'} \end{array} \right] \\ OBJ \quad o : \left[\begin{array}{l} SPEC \quad \text{'A'} \\ PRED \quad \text{'GIRL'} \end{array} \right] \end{array} \right]$$

The SPEC and PRED attributes of the NPs and their values, as well as the VP (and sentence) PRED attribute and its value all come from the lexical entries. The first line of a lexical entry gives the syntactic category of the word and the f-structure constraints it comes with.

every Det $(\uparrow \text{ SPEC}) = \text{'EVERY'}$
 $every : (e_{\uparrow \text{label}} \multimap t_{\uparrow \text{label}}) \multimap ((e_{\uparrow \text{label}} \multimap \alpha) \multimap \alpha)$
 $\lambda N. \lambda P. \left[\begin{array}{l} \boxed{b : e_{\uparrow}} \\ \boxed{\quad} \sqcup N(b) \Rightarrow P(b) \end{array} \right]$

boy N $(\uparrow \text{ PRED}) = \text{'BOY'}$
 $boy : e_{\uparrow \text{label}} \multimap t_{\uparrow \text{label}}$
 $\lambda b. \left[\begin{array}{l} \boxed{\quad} \\ \boxed{boy(b)} \end{array} \right]$

loves V $(\uparrow \text{ PRED}) = \text{'LOVE'}$
 $loves : e_{(\uparrow \text{ SUBJ}) \text{label}} \multimap (e_{(\uparrow \text{ OBJ}) \text{label}} \multimap t_{\uparrow \text{label}})$
 $\lambda b. \lambda g. \left[\begin{array}{l} \boxed{\quad} \\ \boxed{love(b, g)} \end{array} \right]$

a Det $(\uparrow \text{ SPEC}) = \text{'A'}$
 $a : (e_{\uparrow \text{label}} \multimap t_{\uparrow \text{label}}) \multimap ((e_{\uparrow \text{label}} \multimap \beta) \multimap \beta)$
 $\lambda N. \lambda P. \left[\begin{array}{l} \boxed{g : e_{\uparrow}} \\ \boxed{\quad} \sqcup N(g) \sqcup P(g) \end{array} \right]$

girl N $(\uparrow \text{ PRED}) = \text{'GIRL'}$
 $girl : e_{\uparrow \text{label}} \multimap t_{\uparrow \text{label}}$
 $\lambda g. \left[\begin{array}{l} \boxed{\quad} \\ \boxed{girl(g)} \end{array} \right]$

While grammar rules may have semantic content, most do not. On the other hand most words do. The third line of each of the lexical entries contains the meaning of the word expressed in λ -DRT, this time complete with simple types for discourse referents. The second line contains a meaning placeholder (the word itself) and its compositional (Glue) type. This is the interface between syntax and semantics with respect to semantic composition. The simple types added to λ -DRT discourse referents constitute the interface between syntax and the dynamic semantics representation with respect to anaphoric resolution.

The Glue typing context Γ for a sentence is formed of the meaning placeholders and their types.⁴ What a Glue implementation does is derive all $\beta\eta$ -irreducible terms T of type t_f (where f is the f-structure for the sentence) such that $\Gamma \vdash T : t_f$. Replacing the meaning placeholders with the corresponding meaning at any time gives the composite meaning Glue has formed, although usually β -reduction needs to be applied also to produce more comprehensible, but otherwise equivalent, semantic expressions.

To recap, for each word we have its syntactic specification (line 1), its compositional specification (line 2) and its semantic specification (line 3).⁵ This presentation deviates from the standard modern presentations of Glue (Dalrymple, 2001), but only slightly. Meaning placeholders have been introduced for a number of reasons. One is to make it clear that the Glue types do not apply to the meaning expressions but only determine how the latter can combine. Another is to emphasise that Glue does not rely on knowing the details of meaning expressions it works with. A third reason is to make the structure of the Glue derivations more evident. There are also some formal reasons of minor importance. However, one can always follow tradition and use the meaning expressions with their corresponding glue types in the derivation and even perform β -reductions at intermediate steps. The end result will be the same.

The function ‘label’, written as a subscript to its argument, maps an f-structure to its label. The labels acting as arguments to the base type constructors anchor the base types to f-structures. We also have variables (lowercase Greek letters) as arguments (subscripts) to the base type constructors. They are implicitly universally qualified in prenex normal form, i.e. the actual type is obtained by adding universal quantifiers for the variables on the left hand side (say in order of appearance); so the type of *every* is really $\forall\alpha.(e_{\uparrow_{\text{label}}} \multimap t_{\uparrow_{\text{label}}}) \multimap ((e_{\uparrow_{\text{label}}} \multimap \alpha) \multimap \alpha)$.

The key to the present solution to anaphora using LFG, Glue and λ -DRT is that it uses the same kind of linking between f-structures and types that Glue uses. This linkage is essential for Glue not to compose meaning in an erroneous way; it is also essential for the treatment of anaphora being proposed to avoid erroneous anaphoric binding. With reference to the title of the paper, we should note that instead of using a lot of Glue to treat anaphora, we can thus allow our version of λ -DRT with simple discourse referent types to take care of it through its simple and elegant resolution mechanism based on variable binding while also enforcing the relevant syntactic constraints.

⁴If the same meaning placeholder name as it appears in the lexical rules appears more than once in the typing context, we can number its occurrences using subscripts to make them unique within the context. In our examples this is not necessary.

⁵For certain words and syntactic constructs it could be more practical or even necessary to break a meaning contribution into smaller and simpler ones, in which case there will be two or more compositional-semantic specification pairs. For those that do not make a semantic contribution there will not be any.

Returning to our example sentence (1), given the f-structure for it we obtain the following Glue typing context Γ :

$$\begin{aligned} \text{every} &: (e_s \multimap t_s) \multimap ((e_s \multimap \alpha) \multimap \alpha), \\ \text{boy} &: e_s \multimap t_s, \\ \text{loves} &: e_o \multimap (e_s \multimap t_f), \\ a &: (e_o \multimap t_o) \multimap ((e_o \multimap \beta) \multimap \beta), \\ \text{girl} &: e_o \multimap t_o. \end{aligned}$$

According to the Glue type-inference rules below

$$\Gamma \vdash \text{every boy } \lambda x. a \text{ girl (loves } x)$$

and

$$\Gamma \vdash a \text{ girl } \lambda y. \text{every boy } \lambda x. \text{loves } x y.$$

These are the only semantically distinct readings available for the sentence. Replacing the meaning placeholders with their corresponding meaning in the derived terms and β -reducing we get the two readings in DRT as we had originally set out to do.

$$\begin{aligned} & \frac{N : T, \Gamma, N' : T', \Gamma' \vdash E : T''}{N' : T', \Gamma, N : T, \Gamma' \vdash E : T''} \text{ (Exchange)} \\ & \frac{}{N : T \vdash N : T} \text{ (Axiom)} \\ & \frac{\Gamma, X : T \vdash E : T'}{\Gamma \vdash \lambda X. E : T \multimap T'} \text{ (}\multimap\text{Intro.)} \\ & \frac{\Gamma \vdash E : T' \multimap T \quad \Gamma' \vdash E' : T'}{\Gamma, \Gamma' \vdash E E' : T} \text{ (}\multimap\text{Elim.)} \\ & \frac{\Gamma \vdash E : \forall V. T}{\Gamma \vdash E : T[V := L]} \text{ (}\forall\text{Elim.)} \end{aligned}$$

Figure 1: First-Order Glue Inference Rules

Notes:

1. The Exchange rule is unnecessary if we regard the context as being a multiset.
2. The \forall Intro rule is not needed and has been excluded.

We have at our disposal a tripartite framework that handles syntax (LFG), meaning composition (Glue), and semantics (λ -DRT) and we have seen it at work with a simple example. Our choice of meaning expressions and Glue types (modulo f-structure labels) guarantees the well-formedness of the resulting meaning expression.⁶ The f-structure labels used as parameters in our Glue types ensure that meaning composition does not result in arbitrary semantic expressions given a multiset of meaning contributions, but all and only those readings that correspond to the given sentence or discourse. Glue pulls its weight remarkably well and has a clear, simple, yet powerful interface to LFG (and other grammar formalisms). In the following section it will be λ -DRT's turn to demonstrate the same qualities when dealing with its assigned task, anaphoric resolution.

⁶This guarantee does not cover anaphoric resolution which is an matter entirely internal to DRT in the presented approach.

3 Importing Syntactic Anaphoric Constraints

The classic DRT anaphoric resolution mechanism (Kamp, 1981; Kamp and Reyle, 1993) was based on the visibility of discourse referents. It was remarkably simple and elegant. However, in its basic form this mechanism completely disregards basic syntactic requirements such as number and gender agreement, thus potentially generating erroneous readings. There are also constraints on anaphoric resolution that are on the level of discourse. One of the strengths of DRT is in dealing with such constraints. That strength is taken full advantage of here, and so is the strength of the syntactic analysis that informs the anaphoric resolution mechanism of the constraints emanating from the syntactic form of the sentence. The latter is achieved thanks to a simple technique for importing syntactic constraints into the chosen dynamic representation language. So in neither of the two examples below will the interpretation implied by the coreference indicators be allowed, but for the first this will be thanks to DRT semantic form constraints, whereas for the second it will be thanks to the imported LFG syntactic constraints on anaphora.

*No student₁ arrived. He₁ yawned. (2)

*Every man₁ likes him₁. (3)

The most prominent feature of the classic DRT analysis of pronouns is the introduction of a new discourse referent that comes with a condition that equates it to a question mark. Informally, the question mark may be seen as a promissory note for an accessible discourse referent. Formally, we can treat it as a metavariable ranging over discourse referents. Then according to the scoping rules of DRT, ? can only be an accessible discourse referent. We return to this shortly.

$$\begin{array}{l}
 \text{himself} \quad \text{NP} \quad (\uparrow \text{ PRED}) = \text{'PRO'} \\
 \text{himself} : (e_{\uparrow \text{label}} \multimap \alpha) \multimap \alpha \\
 \lambda P. \begin{array}{|c|} \hline p : e_{\hat{\uparrow}} \\ \hline p = ? \\ \hline \text{male}(p) \\ \hline \end{array} \sqcup P(p)
 \end{array}$$

Positive and negative constraints for each pronoun are expressed in its lexical entry in terms of expressions involving inside-out functional uncertainty (Dalrymple, 1993). Such expressions determine which parts of the f-structure are the candidates allowed to act as antecedents (positive constraint) and which are disallowed (negative constraint). The antecedent of the reflexive pronoun 'himself' obeys the Minimal Complete Nucleus positive constraint; therefore its f-structure will have to satisfy the expression

$$\begin{array}{l}
 ((\quad \text{GF}^* \quad \text{GFpro} \quad \uparrow) \text{ GF}) \\
 \neg(\rightarrow \text{ SUBJ})
 \end{array}$$

where '↑' stands for the f-structure of the pronoun. To capture the above positive constraint for 'himself' in our typed λ-DRT, we add the following to its lexical entry:

$$\hat{\uparrow} \in \{ \hat{L} \mid L = ((\quad \text{GF}^* \quad \text{GFpro} \quad \uparrow) \text{ GF}) \} \\
 \neg(\rightarrow \text{ SUBJ})$$

Central to our discussion is a function $\hat{\cdot}$ from f-structures to anaphoric indices, satisfying positive and negative constraints, but otherwise assigning different indices to different f-structures. For stylistic reasons, we write $\hat{\cdot}(X)$ as \hat{X} . Coreference will be modelled as anaphoric index equality. As a pronoun can corefer, it is possible that $\hat{\cdot}$ will map two or more f-structures to the same index. As pronouns do not necessarily have to corefer (exophora), this will not necessarily be the case for all pronouns. The DRT condition $x = y$ is well formed if and only if $x : e_{\hat{X}}$ and $y : e_{\hat{Y}}$ are accessible discourse referents at the point the condition $x = y$ appears and $\hat{X} = \hat{Y}$.

The way syntactic anaphoric constraints are expressed in LFG is powerful, but cryptic. Dalrymple (2001) explains inside-out functional uncertainty, gives a brief overview of the LFG research on such constraints and links that discussion to Glue. However, the following examples should be easy to follow without a deep understanding of LFG and its way of dealing with syntactic constraints on anaphora.

For an example illustrating positive constraints we can take a sentence with a reflexive pronoun such as

John hit himself. (4)

$$f : \left[\begin{array}{l} PRED \quad \text{'HIT'} \\ SUBJ \quad s : [PRED \quad \text{'JOHN'}] \\ OBJ \quad o : [PRED \quad \text{'PRO'}] \end{array} \right]$$

constraints: $\hat{o} \in \{\hat{s}\}$

The positive constraint for $\hat{\cdot}$ simply means that $\hat{s} = \hat{o}$. So with ? standing for the subject discourse referent we can only have the following correct reading:

$$\boxed{\begin{array}{l} j : e_{\hat{s}}, h : e_{\hat{o}} \\ j = John \\ hit(j, h) \\ h = j \end{array}} \quad [\hat{o} = \hat{s}] .$$

For an example involving negative constraints we can take a mini-discourse such as the following:

An elephant saw a mouse. She frightened her. (5)

This is a rather interesting example as two readings should be available. The f-structure for the first sentence is:

$$f_1 : \left[\begin{array}{l} PRED \quad \text{'SEE'} \\ SUBJ \quad s_1 : \left[\begin{array}{l} SPEC \quad \text{'A'} \\ PRED \quad \text{'ELEPHANT'} \end{array} \right] \\ OBJ \quad o_1 : \left[\begin{array}{l} SPEC \quad \text{'A'} \\ PRED \quad \text{'MOUSE'} \end{array} \right] \end{array} \right] .$$

The f-structure for the second sentence is:

$$f_2 : \left[\begin{array}{l} PRED \quad \text{'FRIGHTEN'} \\ SUBJ \quad s_2 : [PRED \quad \text{'PRO'}] \\ OBJ \quad o_2 : [PRED \quad \text{'PRO'}] \end{array} \right]$$

constraints: $\hat{o}_2 \notin \{\hat{s}_2\}$

Before anaphoric resolution, we have two distinct question mark metavariables in our DRS.

| | |
|---|--------------------------------------|
| $e : e_{s_1}, m : e_{o_1}, s : e_{s_2}, h : e_{o_2}$ <i>elephant</i> (e) <i>see</i> (e, m) <i>mouse</i> (m) $s = ?$ <i>frighten</i> (s, h) $h = ?'$ | $[\hat{o}_2 \notin \{\hat{s}_2\}]$ |
|---|--------------------------------------|

If we have $\hat{s}_2 = \hat{s}_1$ and resolve $?$ to e , then the negative constraint on the non-reflexive pronoun 'her' in object position in the second sentence means that $\hat{o}_2 \neq \hat{s}_1$, leaving $\hat{o}_2 = \hat{o}_1$ and $?' = m$ as the only option. This gets us the first reading:

| | |
|--|--|
| $e : e_{s_1}, m : e_{o_1}, s : e_{s_2}, h : e_{o_2}$ <i>elephant</i> (e) <i>see</i> (e, m) <i>mouse</i> (m) $s = e$ <i>frighten</i> (s, h) $h = m$ | $[\hat{s}_2 = \hat{s}_1, \hat{o}_2 = \hat{o}_1, \hat{o}_2 \notin \{\hat{s}_2\}]$ |
|--|--|

If we have $\hat{s}_2 = \hat{o}_1$ and resolve $?$ to m , then the negative constraint on the non-reflexive pronoun 'her' in object position in the second sentence means that $\hat{o}_2 \neq \hat{o}_1$, leaving $\hat{o}_2 = \hat{s}_1$ and $?' = e$ as the only option. This gets us the second reading:

| | |
|--|--|
| $e : e_{s_1}, m : e_{o_1}, s : e_{s_2}, h : e_{o_2}$ <i>elephant</i> (e) <i>see</i> (e, m) <i>mouse</i> (m) $s = m$ <i>frighten</i> (s, h) $h = e$ | $[\hat{s}_2 = \hat{o}_1, \hat{o}_2 = \hat{s}_1, \hat{o}_2 \notin \{\hat{s}_2\}]$ |
|--|--|

4 Earlier work

Dalrymple et al. (1999) present the original Glue-based context management approach to anaphoric resolution. The basic idea behind that is that a pronoun makes an additional copy of the meaning of its antecedent; it does so by consuming that meaning x and producing a pair (x, x) . So the semantics of a pronoun is given by the expression $\lambda x.(x, x)$. The glue type for a pronoun found at the part of the sentence f-structure labelled Y that has an antecedent at X is $e_X \multimap e_X \otimes e_Y$. This is the resource duplication Glue treatment of anaphora.⁷

As the point of using linear, rather than, say, intuitionistic, logic in Glue was that it provides resource sensitivity,⁸ while as far as anaphora resolution is concerned a discourse referent that is in the current context can be referenced any number of times, one can immediately see a problem with trying to treat anaphora within Glue. The explicit resource duplication Glue treatment of anaphora cleverly addresses this problem, but this problem alone. Dalrymple et al. (1999) find that this approach does not work if sentence-by-sentence processing is assumed. An alternative approach, using the ! (‘of course’) linear logic modality, addresses the problem resource duplication has when sentence-by-sentence processing is assumed, but only that and at the cost of complicating Glue. Dalrymple et al. (1999) find problems with that approach too. Furthermore, neither of the two approaches takes into account that there is a difference between the anaphoric context available within a sentence and how it affects the context for other sentences. These proof-of-concept approaches address only the problems resource sensitivity causes for the treatment of anaphora within Glue.

Taking the next step in the evolution of the context management approaches, Crouch and van Genabith (1999) make a bold attempt to address intersentential anaphora issues using a para-Glue e-type anaphora context management approach. They add assignments from NP labels to e-type descriptions to the standard Glue system of the time. They also change what the final result of a derivation is in order to allow these assignments to appear alongside the meaning of a sentence at the end of the derivation. The basic idea is fairly simple. For the sentence ‘A man walks’ the result this approach gives is a pair. The first element is the meaning of the sentence $\exists x.man(x) \wedge walk(x)$ and the second is an assignment of the sentence’s subject label to the description $\lambda x.man(x) \wedge walk(x)$. A pronoun consumes a description assignment such as the above, uses it in its meaning and also produces a new one. So the result of subsequently analysing the sentence ‘He whistles’ also produces a pair. The sentence meaning element is $\exists x.man(x) \wedge walk(x) \wedge whistle(x)$ while the description assignment part assigns to the subject of this sentence the description $\lambda x.man(x) \wedge walk(x) \wedge whistle(x)$.

This is indeed as simple as it should be. Unfortunately, some of the details were omitted. Interactions between quantifier scope and context assignments complicate matters. Reinforcing the arguments against entangling meaning composition with anaphoric context management, Crouch and van Genabith (1999) also identify a problem arising “from the need to build up a collection of context assignments in addition to a single meaning assignment for the sentence”. They resort to a higher order solution to solve this. What started with a simple idea ended up being very complicated at the

⁷There is also an alternative version of this treatment that does not require \otimes to be a part of the Glue. In that version, the meaning expression for the pronoun is $\lambda x.\lambda P.P x x$ and the corresponding glue type is $e_X \multimap (e_X \multimap e_Y \multimap t_\alpha) \multimap t_\alpha$.

⁸Kokkonidis (2006) argues that although resource sensitivity is probably a desirable feature of Glue, it is not as essential as it is believed to be.

end, while only covering simple cases. Also absent from their treatment is an account of the difference between, say ‘A man walks’ and ‘No man walks’ with respect to the anaphoric context that a subsequent sentence will have available.

The same authors address this issue elsewhere (van Genabith and Crouch, 1997) by simply using CDRT and allowing it to deal with anaphora. This is the approach that is closest to the one presented here; indeed they anticipate similar work by noting that dynamic representations other than CDRT can be used as the meaning representation language in such an approach. However, they do not address the issue of imposing syntactic constraints within the dynamic representation language whereas the Glue and para-Glue approaches did.

The approach presented by Dalrymple (2001)⁹ also addresses the issue of the management of different contexts successfully, albeit at the cost of additions to standard Glue. In many ways this is a continuation of the research of Crouch and van Genabith (1999). However, a DRT-style approach is taken. In effect what this approach does is take the discourse referents universe of a DRS and stick it next to meanings derived using Glue. This may be seen as having the advantage of offering some of the benefits of DRT when other meaning representations are used. However, it does not make much sense if DRT itself is to be used as the meaning representation language. Furthermore, if one wanted to combine the characteristics of DRT with those of another representation language, an obvious solution would have been doing exactly that and using the result as the meaning representation language. Much of this is a matter of opinion and personal taste. The fact is that, historically, that was the first Glue context management approach that could control context equally well as the DRT-based approaches such that of van Genabith and Crouch (1997) and the present one. Furthermore, it was the first approach that did that and at the same time respected syntactic constraints.¹⁰ Having said that though, it does come with notational clutter and like its predecessor cannot avoid the complications caused by combining meaning composition with context management. This is evident in the proposed lexical entries for quantifiers such as ‘nobody’ and ‘somebody’ (Dalrymple, 2001). Even additional inference rules are added to help deal with the complexities this juggling with too many balls at the same time brings. However, additional rules add complexity in their own right.

5 Conclusions

The two initial approaches described in Dalrymple et al. (1999) did not require anything more than what linear logic had to offer: one required ‘ \otimes ’ but there was also a version that did not need any extension beyond the implicative fragment, and the other required ‘!’. While the fragment of Glue needed for meaning composition is first-order (Kokkonidis, 2006) the approach of Crouch and van Genabith (1999) needs genuine higher order universal quantification to deal with the context, and interestingly enough it needed that for doing something as simple as adding something to the existing context in order to construct a new one. The approach of Dalrymple (2001) is far more drastic: it not only introduces new concepts such as the context and the meaning-context combination, as well as their counterparts on the type-system side, but also new rules

⁹The approach of Dalrymple (2001) is based on joint unpublished work with Martin van den Berg, Dick Crouch, and John Lamping.

¹⁰There is a problem with enforcing negative constraints in all three Glue context management approaches discussed as using an ANT attribute does not capture the transitivity of coreference. However, this problem can be solved e.g. by using a formal device similar to ‘ \wedge ’.

for splitting meaning and context and for merging context and meaning. The evolution path seems to have taken us from essentially first-order Glue treatments that can only deal correctly with very simple cases (Dalrymple et al., 1999), to Glue plus genuine higher-order quantification (Crouch and van Genabith, 1999), to a Glue-DRT hybrid with many formal innovations (Dalrymple, 2001).

So what is the return on investment? The hybrid Glue-DRT approach of Dalrymple (2001) only covers the case of singular pronouns. How much more complication will have to be introduced to cover plural anaphora? If DRT is so good at dealing with anaphora, why not adopt it as the semantic representation and get all of it rather than trying to copy its behaviour in a hybrid Glue-DRT system?

The work of van Genabith and Crouch (1997) and the present work effortlessly tackle many problems Glue context-management approaches have found challenging by leaving them to dynamic semantic representations designed to deal with them. Combining CDRT or λ -DRT with Glue required no ingenuity whatsoever. CDRT and λ -DRT can easily work with various systems for composing meanings and Glue can work with various semantic representation languages. DRT does certain things well and Glue does other things well. They complement each other nicely. One problem not addressed by van Genabith and Crouch (1997) was that of syntactic constraints. This is addressed here.

The next step would be to provide analyses for a wider range of anaphoric phenomena. It seems that all that needs to be done if the modular approach is taken is to ensure the existing DRT analyses for this wider range of phenomena fit in well with the type system proposed. On the other hand, it seems that if one takes the approach of Dalrymple (2001) much more of DRT would have to be incorporated into the Glue-DRT hybrid system, most likely at the cost of even more complexity in order to achieve comparable results.

Complexity not only makes the system more difficult to explain, but it also hinders further development. Modular design is usually a good way of managing complexity, and in treating semantic composition and anaphoric resolution separately it certainly seems to have helped keep it to manageable levels. First-order Glue suffices for the composition of meanings. Only very simple types and a new well-formedness requirement on the equals sign were added to λ -DRT. There are no strange interactions and conflicting requirements to be dealt with. This means that one can concentrate on dealing with the phenomena, rather than problems with the formalism.

Acknowledgements

I am grateful to Mary Dalrymple for the very idea behind this project (combining λ -DRT with Glue while somehow making sure syntactic constraints on anaphoric binding are respected), as well as a number of very useful discussions throughout its duration. Most of all, I am grateful to her for her endless enthusiasm and support. Thanks also go to Tracy Holloway King and Miriam Butt for their meticulous proofreading of an earlier version and their good advice. Needless to say, I take full responsibility for any new mistakes that have, no doubt, tried to replace all those they caught. Last but not least, I thank the audience at LFG05 for their comments and questions and the local organising committee for their hospitality.

References

- Bos, J., E. Mastenbroek, S. McGlashan, S. Millies, and M. Pinkal: 1994, 'A compositional DRS-based formalism for NLP applications'. In: *Proceedings of the International Workshop on Computational Semantics*. Tilburg.
- Crouch, R. and J. van Genabith: 1999, *Context Change, Underspecification, and the Structure of Glue Language Derivations*'. In Dalrymple (1999).
- Dalrymple, M.: 1993, *The Syntax of Anaphoric Binding*, No. 36 in CSLI Lecture Notes. CA: Stanford.
- Dalrymple, M. (ed.): 1999, *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*. MIT Press.
- Dalrymple, M.: 2001, *Lexical Functional Grammar*, No. 42 in Syntax and Semantics Series. Academic Press.
- Dalrymple, M., J. Lamping, F. C. Pereira, and V. Saraswat: 1999, *Quantification, Anaphora, and Intensionality*. In Dalrymple (1999).
- Kamp, H.: 1981, 'A theory of truth and representation'. *Formal Methods in the Study of Language*.
- Kamp, H. and U. Reyle: 1993, *From Discourse to Logic*. Dordrecht: Kluwer.
- Kaplan, R. M. and J. Bresnan: 1982, 'Lexical Functional Grammar: A formal system for grammatical representation'. In: J. Bresnan (ed.): *The Mental Representation of Grammar Relations*. MIT Press, pp. 173–281.
- Kohlhase, M., S. Kuschert, and M. Pinkal: 1995, 'A type-theoretic semantics for λ -DRT'. In: *Proceedings of the Tenth Amsterdam Colloquium*.
- Kokkonidis, M.: 2006, 'First-Order Glue'. *Journal of Logic, Language and Information*. To appear.
- van Genabith, J. and R. Crouch: 1997, 'How to glue a donkey to an f-structure or porting a dynamic meaning representation language into LFG's linear logic based glue language semantics'. In: *Proceedings of the International Workshop for Computational Semantics*. Tilburg, pp. 52–65.

MORPHOLOGICAL AND
SYNTACTIC WELL-FORMEDNESS:
THE CASE OF EUROPEAN PORTUGUESE PROCLITICS

Ana R. Luís and Ryo Otaguro
(University of Coimbra/University of Essex)

Proceedings of the LFG05 Conference
University of Bergen
Miriam Butt and Tracy Holloway King (Editors)
2005
CSLI Publications

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

Abstract

European Portuguese proclitics illustrate a mismatch between inflectional status and syntactic separability which is challenging to lexicalist theories of syntax. On the one hand, they form morphologically complex clitic clusters and realise verbal properties; on the other, they may be separated from the verb by lexical items, showing no sign of being morphologically attached to it. The question then is how to account for the partly inflectional and partly phrasal behaviour of proclitic affixes in a theory of syntax that prohibits elements smaller than words from being syntactically visible. In defence of the principle of Lexical Integrity (Bresnan 2001:92), Luís&Sadler (2003) take the view that proclitic affixes may not be assigned a c-structure position. In this paper, we also endorse the view that morphology and phrase structure constitute separate levels of analysis, but explore an alternative analysis.

1. Introduction

The problem posed by pronominal proclitics in European Portuguese (EP) arises from the fact that they exhibit both inflectional and syntactic properties¹. On the one hand, they form morphologically complex clitic clusters and realise verbal properties (exactly like their enclitic counterparts); but on the other, they may be separated from the verb by lexical items, showing no sign of being morphologically attached to it. These features suggest to Luís (2004) that proclitic affixes in EP should be analysed as phrasal affixes (i.e., verbal affixes with phrasal status). However, at the level of c-structure, it is not entirely clear how phrasal affixes can be accommodated in a theory that assumes lexical integrity.

The same problem has been addressed in Luís&Sadler (2003), within LFG, who argue that proclitic affixes may not be assigned a c-structure position on the grounds that such an analysis constitutes a violation of the Principle of Lexical Integrity (Bresnan 2001:92). Luís&Sadler (2003) sketch a proposal in which the proclitic affix is represented as pronominal f-structure information associated with a phrasal V-VP node. The affix itself however does not appear in the c-structure. Because of the somewhat unconventional model of c-structure adopted in that analysis, this paper aims to explore an alternative approach. We formulate a mapping between morphology and c-structure which assigns a c-structure

¹ We are grateful to Miriam Butt, Mary Dalrymple, Ron Kaplan, Tracy H. King, Gergana Popova, Louisa Sadler and Andrew Spencer for helpful comments in the early stages of this work and throughout. Remaining errors are solely our own.

position to proclitic affixes without making the assumption that incomplete morphological strings may be represented in the syntax (Luís&Otoguro 2004, to appear).

Section 2 surveys the basic facts about the EP data. Section 3, argues that mismatch phenomena in LFG pose problems to the principle of Lexical Integrity and that enough supporting evidence has been provided in the literature to justify the search for an alternative morphology/c-structure mapping. Section 4 presents the Morphological Token analysis which assumes a revised interface between morphology and syntax. Section 5 provides a short conclusion and outlines avenues for further research.

2. Overview of the data

In this section, we survey morphological and syntactic evidence in support of the claim that proclitics in EP constitute phrasal affixes. We show that proclitics are formally and semantically exactly identical to enclitics. However, while enclitics behave like genuine verbal suffixes, proclitics display phrasal properties.

2.1 Inflectional properties

It is well-known that European Portuguese, like other Romance languages, has two types of pronominal clitics. Depending on whether clitics precede or follow the verbal host, they may be enclitic to the verb, as in (1a), or proclitic, as in (1b).

Luís (2004) shows that enclitics display a significant number of affix properties such as fusion (1a), syncretism (3a), and cluster-internal allomorphy (5a), in addition to rigid ordering and idiosyncratic co-occurrence restrictions. Proclitics show exactly the same range of cluster-internal allomorphy and rigid ordering, as the examples in (1b), (3b) and (5b) illustrate.

Illustrating these properties in more detail, portmanteau forms appear when 3rd person accusative clitics follow either 1st/2nd person singular or 3rd person plural dative clitics, as in (1). A partial inventory of opaque clitic clusters is given in (2).

- (1) a. disse-**mo** (*me-o)
said-DAT.1SG-ACC.3SG.M
's/he said it to me'
- b. ... que **mo** disse (*me-o)
... that DAT.2PL-ACC.3SG.M-said
'...that s/he said it to me'

(2)

| | | | | |
|---------|--------------|-------------|---------------|---------------|
| | 3sg.masc.acc | 3sg.fem.acc | 3pl.masc.acc | 3pl.fem.acc |
| 1sg.dat | mo (= me+o) | ma (= me+a) | mos (= me+os) | mas (= me+as) |
| 2sg.dat | to (= te+o) | ta (= te+a) | tos (= te+os) | tas (= te+as) |

Table 1

When 3rd person dative clitics co-occur with 3rd person accusative clitics, the plural features on the dative forms are neutralised giving rise to syncretism, as shown in (3), where *lho* can either mean ‘V it to him’ or ‘V it to them’. The complete set of syncretic forms is provided in (4).

(3) a. deu-**lho** (*lhe-o)

gave-DAT.3SG/PL-ACC.3SG.M

‘s/he gave it to him/them’

b. ... que **lho** deu (*lhe-o)

... that DAT.3SG/PL-ACC.3SG.M-gave

‘...that s/he gave it to him/them’

(4)

| | | | | |
|---------|--------------|-------------|--------------|-------------|
| | 3Acc.Masc.Sg | 3Acc.Fem.Sg | 3Acc.Masc.Pl | 3Acc.Fem.Pl |
| 3Dat.Sg | lho | lha | lhos | lhas |
| 3Dat.Pl | | | | |

Table 2

Cluster internally, object pronouns generally exhibit phonological alternation when 3rd accusative pronouns (*o, a, os, as* ‘him, her, them.masc, them.fem’) are preceded by a 1st/2nd person plural dative pronoun. The dative clitics loses its final consonant and an 3rd person accusative allomorph surfaces (i.e., *lo, la, los, las*).

(5) a. deu-**no-lo** (*nos-o)

gave-DAT.2PL-ACC.3SG

‘s/he gave it to us’

b. ... que **no-lo** disse

... that DAT.2.PL-ACC.3SG.M-said

‘...that s/he said it to us’

The complete inventory of clusters combining 1st/2nd person plural datives with 3rd person accusatives is shown in (6).

(6)

| | 3sg.m.acc | 3sg.f.acc | 3pl.m.acc | 3pl.f.acc |
|---------|-------------------|-------------------|---------------------|---------------------|
| 1pl.dat | (nos+o)→ no-lo | (nos+a)→ no-la | (nos+os)→ no-los | (nos+as)→ no-las |
| 2pl.dat | (vos+o)→ vo-lo | (vos+a)→ vo-la | (vos+os)→ vo-los | (vos+as)→ vo-las |

Table 3

The morphophonological changes taking place inside the cluster suggest that a morphological analysis of EP pronominals should be preferred. To capture the fact that enclitics and proclitics are formally and semantically exactly identical, Luís (2004) develops an inflectional analysis within a revised version of Paradigm Function Morphology (Stump 2001) which generates enclitics and proclitics through one and the same realisation rule (e.g., $R \{ \text{Case:Dat, Nmb:Sg, P:3} \}_{\text{def}} \langle lhe \rangle$). Such realisation rule R defines affixes as ‘ambifixal’ exponents, that is as affixes which may either attach as prefixes or as suffixes (cf. Stump 1993 on Fula). A morphological alignment function is provided which places the clitic to the left or to the right of the host.

2.2. Enclitic suffixes

Shape variations found at the boundary between verbs and enclitics suggest that enclitics constitute verbal suffixes. For example, pronominal allomorphy is found when 3rd person accusative pronouns, i.e. *-a*, *-o*, *-os*, *-as*, are preceded by verbs ending in *-r*, *-s* or *-z* (7a) or by 3rd person plural verb forms (7b). In the first context, accusative clitics surface in their *l*-form, as *-lo*, *-la*, *-los*, *-las*; in the second context they appear in their *n*-form, as *-no*, *-na*, *-nos*, *-nas*.

- (7) a. Levamo **-la** (not: levamos-a)
take -acc.3.sg.fem
‘We will take her’
- b. Os meninos levam **-nos** (not: *levam-os)
the boys take -acc.1.pl
‘The boys take us’

Enclitics also trigger phonological changes on the verb. In particular they induce word-final consonant deletion in the two following contexts: a) when *l*-initial 3rd person accusative clitics are preceded by verb forms ending in *-s*, *-z* or *-r* (7a), and b) when 1st/2nd person plural clitics, i.e. *-nos* and *-vos*, follow 1st person plural verb forms (8).

- (11) a. ... acho que ela **o** ainda não disse.
 ... think that she ACC.3SG.MASC yet not told
 ‘... I think that s/he hasn’t told it to him/her/them yet’
- b. ... embora eu saiba que **a** já tens
 ... although I know that ACC.3SG.FEM already have
 em grande dose.
 in big portion
 ‘... although I know that you already have tons of it (= patience)’

What the data shows is that the difference between enclitics and proclitics is not just a question of right/left linearisation to the host. Based on the above evidence, Luís (2004) accounts for the asymmetry between enclitics and proclitics by analysing enclitics as verbal suffixes and proclitics as phrasal affixes. This proposal elaborates on the well-known distinction between word-level affixation and phrasal-affixation, formulated originally by Klavans (1985) and developed more recently by Anderson (1992), Legendre (2000), Spencer (2000), Spencer & Luís (to appear).

In section 2.1 we alluded to the fact that enclitics and proclitics should be derived through an inflectional realisation rule *R* (cf. *lhe*, in cf. 2.1). In addition, the fact that enclitics and proclitics constitute the same exponent is accounted for by deriving both through the same realisation rule and by formulating an alignment function which positions the clitic affix either to the left or to the right of the host. We have now seen that the difference between enclitics and proclitics is not merely positional: it is not just enough to determine the direction of attachment of the clitic affix but it is also necessary to define the nature of the host the clitic affixes attaches to. Hence, in Luís (2004), the alignment function is formulated so as to allow clitics to attach to the right edge of a verbal stem (for enclitics) and to the left of a phrasal node (for proclitics). The asymmetric placement accounts for the difference in status between stem-level suffixation and phrasal affixation.

Summarising: from the point of view of morphology, EP pronominal affixes are constructed within the morphology using a realisational architecture of Paradigm Function Morphology. The assumption is that proclitic affixes are assigned the ability to select their host in the syntax. The question we will address in the following sections is how to capture the phrasal status of proclitics at the level of c-structure.

3. Lexicalism and c-structure

Even though enclitics and proclitics contribute the same f-structure information to LFG c-structure (i.e., OBJ/OBJ2), it is not clear how to incorporate phrasal affixes into a lexicalist model of syntax. The essence of the problem may be summarised as follows: on the one hand, an approach that places the proclitic

affix and its immediately adjacent host under the same terminal node is theoretically in line with lexicalist assumptions but lacks empirical support; on the other hand, an approach that assigns phrasal status to proclitic affixes, at the c-structure level, is empirically correct but in violation with lexicalist assumptions.

3.1 Lexical Integrity

LFG treats morphology and syntax as independent levels of linguistic structure. A strong division is assumed between word-internal structures, on the one hand, and structures between words, on the other, with the underlying conviction that word-formation cannot take place in the syntax.

In a lexicalist theory of grammar the role of morphology is to process morphological operations (e.g. combining a root and affixes, changing stem forms and so forth) and to create fully inflected words. In LFG, those morphological operations are completely separated from syntactic ones, as defined in the principle of Lexical Integrity:

- (12) “morphologically complete words are leaves of the c-structure tree and each leaf corresponds to one and only one c-structure node” (Bresnan 2001:92).

Hence, at the level of c-structure a terminal node can only be instantiated by a single and morphologically complete word.

The only way of adjusting pronominal proclitics to this assumption would be to analyse them either as a) verbal prefixes or as b) fully-fledged words. As prefixes they would attach to the verb and surface as part of an inflected word; as words, they would themselves constitute their own c-structure node. The problem, however, is the lack of empirical evidence supporting these analyses.

There is no data suggesting that proclitics are morphologically attached prefixes, simply because proclitics do not select the category of the word they are adjacent to. In this respect, the representation in (13) would be correct for pronominal enclitic in EP (or for enclitics and proclitics in Italian, Monachesi 1999), but not for EP proclitics:

- (13)
- $$\begin{array}{c}
 \text{VP} \\
 | \\
 \uparrow = \downarrow \\
 \text{V} \\
 | \\
 \text{vêem-nos} \\
 (\uparrow \text{OBJ PRED}) = \text{PRO} \\
 \text{'they see us'}
 \end{array}$$

Luís (2004) also makes a strong case against analysing proclitics as words, more precisely as non-projecting X^0 units. Empirically, the strong resemblance between enclitic clusters and proclitic clusters (cf. section 2) can only be insightfully captured if these sequences are effectively generated through the same inflectional mechanisms. Differentiating between clusters that are proclitic and clusters that are enclitic entails the assumption that *lho* or *se-lhe* would be analysed as sequences of affixes in enclitic position but as lexical units in proclitic position, even though they are formally, semantically and morphotactically exactly the same. In addition, if we did differentiate between lexical clusters and inflectional clusters, other problematic questions would arise about proclitic clusters, in particular: a) would the internal structure of *se-lhe* be analysed as a sequence of two function words or as an opaque unit? If proclitic clusters are regarded as sequences of function words, then how would the many co-occurrence restrictions and morphophonological idiosyncrasies be accounted for? Likewise, if proclitic clusters are treated as an opaque forms, how could one explain that the clitic *se* can co-occurs productively (and agglutinatively) with many other clitic forms, as in *se-me*, *se-lhes*, *se-nos*, etc.

Supposing that there are technical answers to all these questions, one would still need to explain, as alluded to above, why the mechanisms for the derivation of proclitic clusters must be different from those applied in the derivation of enclitic clusters, considering that clusters in either position are formally and semantically exactly identical.

These and other questions suggest to Luís (2004) that the treatment of proclitics as function words – even though technically possible – is not tenable and that clusters should be uniformly analysed as complex inflectional exponents. It would also be unsound to rule out the theoretical status of phrasal affixation solely on the grounds that it challenges Lexical Integrity. Instead, it would seem to be more correct to explore ways of solving the problem of phrasal affixation without violating the integrity of words (cf. section 4 for proposal).

3.2 Morphology-syntax mismatches in LFG

In this section, we briefly survey the analysis developed by Wescoat (2002) for the treatment of morphology-syntax mismatches. Wescoat (2002) provides evidence to support the claim that well-formed morphological words do not always correspond to one and only one terminal node. English non-syllabic auxiliaries are among the phenomena examined by Wescoat.

The claim that non-syllabic auxiliary forms are morphologically attached to the (subject) pronoun was originally formulated by Spencer (1992). Luís (1997) provides empirical evidence which shows that the auxiliary-pronoun combination does effectively behave phonologically, morphologically and syntactically like one single word. Adopting Zwicky & Pullum's criteria for affixation (Zwicky & Pullum 1983), Luís (1997) points out, among other aspects, that word-internal

phonological rules, such as vowel laxing, apply to the non-syllabic auxiliary, reducing a bimoraic unit into a monomoraic one.

(14) (Luís 1997)

He'll { /hi:l/ → /hɪl/ } go

We'll { /wi:l/ → /wɪl/ } go

You've { /ju:v/ → /juv/ } been watching tv.

Luís also shows that non-syllabic auxiliary forms trigger non-productive allomorphy on the pronominal host, as illustrated in (15).

(15) (Luís 1997)

you /yu:/ but *you're* /jɔ:/

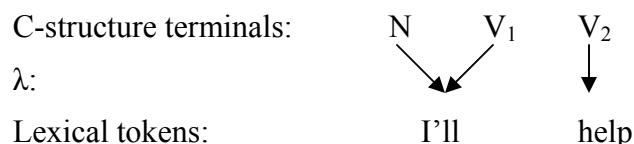
we /wi:/ but *we're* /wɜ:/ (i.e. same as 'were')

they /ðei/ but *they're* /ðɛ:/ (i.e. same as 'there')

(Other affix properties include narrow scope and high degree of selectivity).

The problem with inflected pronouns such I'll [aɪl], as Wescoat (2002) observes, is that they appear to be composed to two syntactically accessible parts. To capture this insight, 'lexical-sharing trees' are proposed which allow two or more 'terminal' nodes to share the same morphological object. The 'lexical sharing' relation is schematically illustrated below:

(16) (Wescoat 2002, p.5)



The mapping developed by Wescoat presupposes a new model of phrase-structure tree in which the Single Root Condition is not obligatory (Partee et al 1993:437-44). The analysis, thus, shows that a more complex approach to the interface between morphology and phrase-structure is necessary, one in which the well-formedness of phrase structure may have to be sacrificed. The question we would like to address now is whether the analysis may be adopted for EP phrasal affixation.

The principle of 'homomorphic lexical integrity, which requires shared nodes to be immediately adjacent, rules out any attempt at applying the analysis to EP phrasal affixes. As alluded to before, proclitic affixes do not attach morphologically to the verb and need not be adjacent to it. What this effectively means is that the proclitic-verb combination does not constitute a single word form. However, it is defined by the morphology as a well-formed inflectional

string for the following reasons: a) the sequence corresponds to a cell in the inflectional paradigm of the lexeme VER ‘see’ (Luís 2004) and b) the clitic affix realises features associated with that lexeme.

In what follows, we will try to develop an analysis which shares with Wescoat (2002) the claim that there is enough supporting evidence in favour of a revised view of the relationship between words and phrase structure.

4. Proposal

In this section, we present the Morphological Token analysis. This analysis, which is broadly outlined in Luís&Otoguro (2004) and in Luís&Otoguro (to appear), assumes that morphological well-formedness and integrity are defined solely in the morphology, through morphology-internal principles, and that morphological strings cannot be inserted directly into c-structure. Additional structure mediates between the level of morphology and the level of c-structure.

4.1 Analysis

At the interface between morphology and c-structure, we put morphological tokens in correspondence with syntactic atoms.

- (17) a. *Morphological token*: each morphological token corresponds to a well-formed stem-affix string that are defined by morphology-internal principles.
 b. *Syntactic atom*: syntactic atoms are leaves on c-structure trees; each leaf corresponds to one and only one terminal node; the insertion of syntactic atoms into c-structure is subject to standard phrase structure constraints, such as linearisation, immediate dominance, and instantiation.

The mapping between morphological tokens and syntactic atoms, as shown in (18), takes as input morphological tokens and delivers labelled syntactic atoms. In the labelling function given below, the variables y and z stand for the affixes and H represents the verbal base:

$$(18) \quad [x-H-y] \Rightarrow x_{CL} H-y_I$$

These minor alterations help us formulate the core idea of our analysis: we prohibit morphological strings from being inserted directly into phrase structure and define the ‘integrity’ of words as a condition over morphological tokens. Under this view, complete morphological strings constitute morphological tokens

which are mapped onto c-structure. Only complete strings will be properly mapped. To make our proposal clearer, we will show how the analysis works.

a) Morphological well-formedness

Within Generalised Paradigm Function Morphology (GPFM) (Luís&Spencer 2005, Spencer ms.), the well-formedness of each stem-affix string is determined as follows: the Paradigm Function *PF* takes the pair $\langle \text{VER}, \sigma \rangle$ (i.e., the lexeme VER and a set of morphosyntactic features σ associated with the lexeme) and delivers two complete stem-affix combinations: $v\hat{e} \langle me$ and $me \langle v\hat{e}$. Each inflectional string is the well-formed realisation of a pair $\langle \text{VER}, \sigma \rangle$.

(19) PF analysis ($v\hat{e}$ -*me*/ *me* $v\hat{e}$ ‘sees me’)

- a. $PF(\text{VER}, \sigma) =_{\text{def}}$
 - i. $S(\text{VER}, \sigma) = v\hat{e}$
 - ii. $R \dots = me$
 - iii. $L = v\hat{e} \langle me$

- b. $PF(\text{VER}, \sigma) =_{\text{def}}$
 - i. $S(\text{VER}, \sigma) = v\hat{e}$
 - ii. $R \dots = me$
 - iii. $L = me \langle v\hat{e}$

Clarifying in more detail the Paradigm Function *PF* in (19), we note that the *PF* defines a) the selection of the stem *S*, b) the realisation of the affix *R* and c) the linearisation of the affix with respect to the stem *L*. Both PFs yield the same stem $v\hat{e}$ and the same exponent *me*. Only the linearisation differs: the affix follows the stem in (19a) and precedes it in (19b) (see Luís&Otoguro 2004 for an analysis of the morphosyntactic contexts triggering preverbal positioning).

Adopting Generalised Paradigm Function Morphology (Luís&Spencer 2005, Spencer ms), our morphological analysis factors out the realisation of affixes from their linearization, allowing us to capture the idea that the same affix may be subject to different linearization constraints.

Finally, the *PF* delivers the complete morphological strings $me \langle v\hat{e}$ and $v\hat{e} \langle me$ which constitute two distinct morphological tokens.

b) At the morphology/c-structure interface

The correspondence between morphological tokens and c-structure nodes is mediated through the algorithm in (20) which takes as input morphological

tokens and delivers labelled syntactic atoms that are inserted into c-structure as instantiations of terminal nodes².

The algorithm may be formalised as in (20), where y and z are the affixes and H represents the verbal base. The morphological token is represented in square brackets, on left side of the arrow. The syntactic atoms, which appear on the right side of the arrow³.

$$(20) \quad [x-H-y] \Rightarrow x_{CL} H-y_I$$

In (21), the mapping function has been applied to the morphological tokens derived in (19).

$$(21) \quad \begin{array}{l} \text{a. } [me, v\hat{e}] \Rightarrow me_{CL} v\hat{e}_I \\ \text{b. } [v\hat{e}, me] \Rightarrow v\hat{e}-me_I \end{array}$$

In (19a), a single morphological token corresponds to two syntactic atoms, me_{CL} $v\hat{e}_I$. This mismatch, we claim, is what separates phrasal affixation from simple affixation at the level of c-structure. In most cases, a single morphological token corresponds to a single syntactic atom, thus in (19b) no mismatch is found and the correspondence is one-to-one. In other words, in simple affixation, one stem-affix string will be inserted under one single terminal.

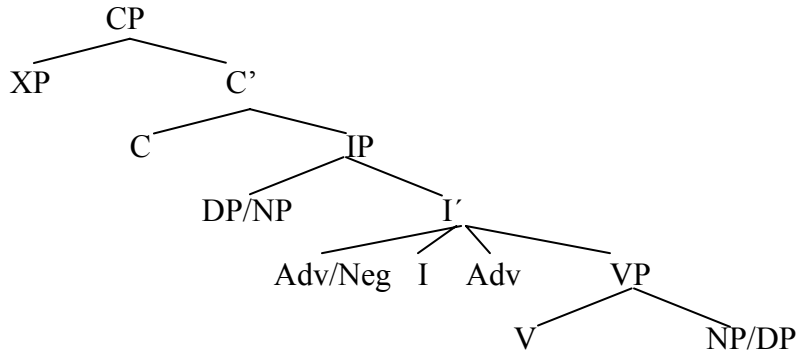
c) The c-structure

The insertion of syntactic atoms into phrase-structure is regulated by standard phrase structure principles (e.g., immediate dominance, linearisation and instantiation) in combination with PS rules. By the phrase structure in (23), proposed in Luís&Otoguro (2004) for EP, the example in (24a) has the c-structure representation in (24b).

² This process is similar to tokenisation in XLE (e.g., Kaplan&Newman 1997, Butt et al. 1999, Kaplan et al. 2004).

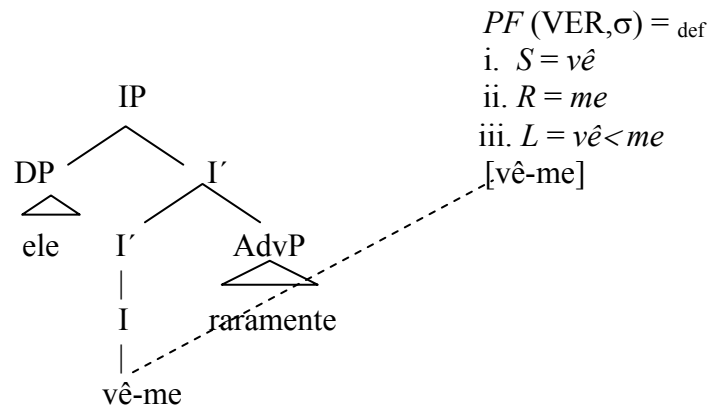
³ We position the finite verb in an I position, following the proposal in Luís&Otoguro (2004).

(23) (Luis&Otoguro 2004)



(24) a. O João vê- me raramente.
 the J. sees- ACC.1SG rarely
 'John sees me rarely'

b.



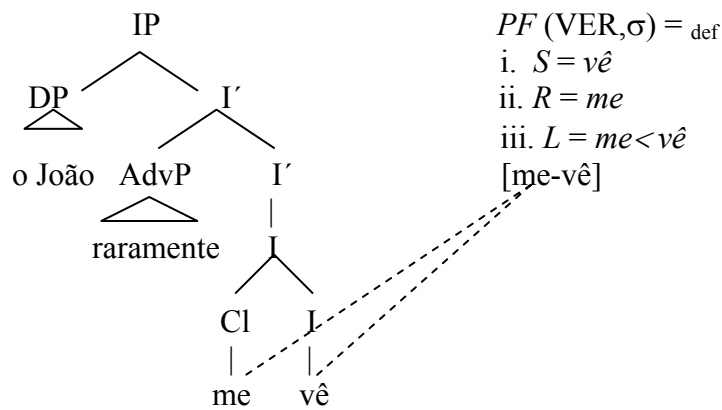
In (24b), the sequence *vê-me* is analysed as a single syntactic atom and, as predicted by the mapping in (21), it is positioned under one single c-structure node.

The mismatch is illustrated in (26), where a proclitic-verb combination is represented at the level of c-structure. Given the analysis in (21a), which associates the stem-affix sequence *me-vê* to two syntactic atoms, the sequence *me-vê* appears under two separate nodes. The correct insertion of the proclitic and the verb under Cl and I', respectively, is defined by the annotated phrase-structure rule in (25). The combination of (23) with (25) yields the c-structure representations in (26b).

(25) I → Cl Adv* I
 ↑ = ↓ ↓ ∈ (↑ ADJ) ↑ = ↓

(26) a. O João raramente me vê
 the J. rarely ACC.1SG sees
 'John rarely sees me '

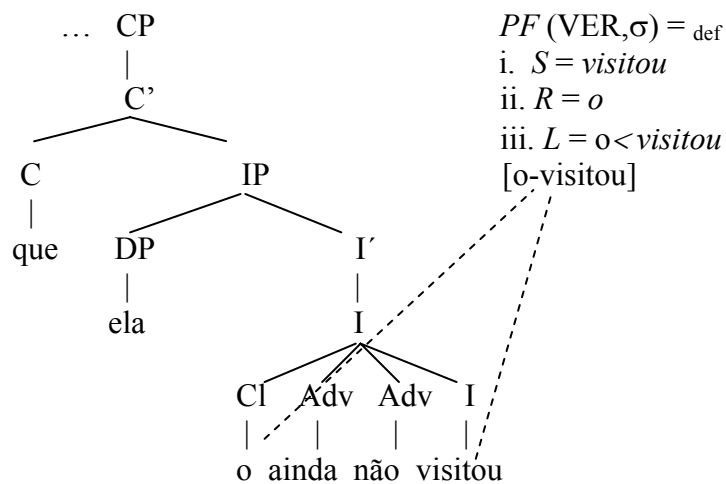
b.



In (27b), the proclitic is followed by interpolated adverbial particles: These are adjoined to I, following the proposal in Luís (2004). Again, by the annotated phrase-structure rule in (25), we represent (27a) as in (27b).

(27) a. Eu sei que ela o ainda não visitou.
 I know that she ACC.3SG.MASC yet not visited
 'I know that she has no visited him yet'

b.



4.2 Summary

We may summarise the assumptions underlying the Morphological Token analysis as follows:

- a) morphological strings are not allowed to be inserted directly into c-structure;
- b) the morphological well-formedness of words is defined in the morphology through morphology-internal principles such as the Paradigm Function which regulates the inflectional paradigm of a given language;
- c) at the interface between morphology and c-structure, a labelling algorithm takes as input morphological tokens and delivers labelled syntactic atoms;
- d) in c-structure, each syntactic atom instantiates a single syntactic terminal node, by general principles of phrase structure and PS rules.

The key goal of the analysis is to allow a single morphological token (i.e., stem-affix combination) to be mapped onto one or more syntactic atoms without incurring any violation of lexical integrity. In terms of the theoretical features of the analysis, we point out that our revised view of the interface between morphology and c-structure requires no changes in the formal model of c-structure trees, nor in the nature of the f-structure to c-structure mapping.

5. Conclusion and avenues for further research

In this paper, we have been concerned with the c-structure representation of proclitic affixes (i.e., phrasal affixes). What the above discussion has revealed is that it is possible to represent phrasal affixes without violating the integrity of words. Our claim is based on the view that ‘integrity’ is defined as a condition on morphological tokens (i.e., complete and well-formed stem-affix sequences defined through morphology-internal principles), rather than as a condition on the mapping between words and c-structure terminals.

The mapping we propose between morphological tokens and syntactic atoms finds theoretical support in the parallel linguistic structures of LFG grammar. Also, by assuming that each level is defined by its own set of well-formedness conditions, our proposal is in full harmony with the division of labour between morphology and syntax, one of the building blocks of lexicalist grammars.

In future research, we examine the scopal behaviour of proclitics in light of the c-structure representation provided in this paper. As alluded to in section 2, proclitics can take wide scope. Thus, any phrase structure representation should also accommodate these coordination properties. Also, further work will be

necessary to determine the different mismatch phenomena that our mapping theory can allow⁴.

References

- Anderson, Stephen R. (1992) *A-Morphous Morphology*. Cambridge Studies in Linguistics 62. Cambridge: CUP.
- Bresnan, Joan (2001) *Lexical Functional Syntax*. Oxford: Blackwell.
- Butt, Miriam et al. (1999) *A Grammar Writer's Cookbook*. Stanford, CA: CSLI Publications.
- Kaplan, Ronald & Paula Newman (1997) Lexical resource reconciliation in the Xerox Linguistic Environment. In D. Estival, et al. *Proceedings of the ACL Workshop on Computational Environments for Grammar Development and Linguistic Engineering*, 54-61. Chicago: Chicago University Press.
- Kaplan, Ronald et al. (2004) Integrating finite-state technology with deep LFG grammars. In *Proceedings of the Workshop on Combining Shallow and Deep Processing for NLP*. ESSLI 2004.
- Klavans, Judith (1985) The Independence of Syntax and Phonology in Cliticisation. *Language*, 61, pp. 95-120.
- Legendre, Géraldine (2000a) Positioning Romanian Verbal Clitics at PF: an Optimality-Theoretic Analysis. In: Birgit Gerlach and Janet Grijzenhout (eds.) *Clitics in Phonology, Morphology, and Syntax*. John Benjamins Publishing Company: Amsterdam/Philadelphia, pp. 219-254.
- Miller, Philp and Ivan Sag (1997) French Clitic Movement without Clitics or Movement. *Natural Language and Linguistic Theory* 15 (3), pp. 573-639.
- Monachesi, Paola (1999) *A Lexical Approach to Italian Cliticization*. CSLI, Stanford.
- Luís, Ana (2004) *Clitics as Morphology*. PhD thesis, University of Essex.
- Luís, Ana (1997) *English Reduced Auxiliaries*. Ma. Diss. University of Essex.
- Luís, Ana and Ryo Otaguro (to appear) Morphological tokens and phrase structure: towards a treatment of morphology-syntax mismatches. In Ryo Otaguro, G. Popova and A. Spencer (eds.) *Essex Research Reports in Linguistics*. University of Essex.
- Luís, Ana and Ryo Otaguro (2004) Proclitic contexts and their effect on clitic placement. In Miriam Butt and Tracy Holloway King (eds.) *Proceedings of the LFG04 Conference*. Stanford, CA: CSLI Publications.

⁴ In Luís&Otaguro (to appear) we apply the Morphological Token analysis to Japanese lexical compounds which illustrates one other type of morphology-syntax mismatch. In the case of Japanese, we argue that two morphological tokens are treated as one syntactic atom at the level of c-structure. The mapping thus is from many to one, rather than from one to many as in EP.

- Luís, Ana and Louisa Sadler (2003) Object Clitic and marked Morphology. In: Beyssade, C., et al. (eds.) *Empirical Issues in Formal Syntax and Semantics 4*. Paris: Presses de l'Université de Paris-Sorbonne, pp. 133-153.
- Luís, Ana and Andrew Spencer (2005) Paradigm Function Account of 'mesoclitisis' in European Portuguese. In: G. Booij and J. van Marle, eds. *Yearbook of Morphology*. Kluwer: Dordrecht. Miller, P. and I. Sag (1997) French Clitic Movement without Clitics or Movement. *Natural Language and Linguistic Theory* 15 (3), pp. 573-639.
- Luís, Ana and Andrew Spencer (to appear) Udi clitics: a Generalized Paradigm Function Approach. In Ryo Otoguro, G. Popova and A. Spencer (eds.) *Essex Research Reports in Linguistics*. University of Essex.
- Spencer, Andrew (1992) *Morphological Theory*. Oxford: Basil Blackwell.
- Spencer, Andrew (2000) Verbal clitics in Bulgarian: A Paradigm-Function approach. In: Birgit Gerlach and Janet Grijzenhout (eds.) *Clitics in Phonology, Morphology and Syntax*. Amsterdam/Philadelphia: John Benjamins, pp. 355-386.
- Spencer, Andrew (ms.) Extended Paradigm Function Morphology. Unpublished manuscript, University of Essex. (Available at <http://privateweb.essex.ac/~spena/papers/epfm.pdf>)
- Stump, Gregory T. (1993) Position classes and morphological theory. In: Booij, G. et. al. (eds.) *Yearbook of Morphology 1992*. Kluwer: Dordrecht, pp. 129-180.
- Stump, Gregory T. (2001) *Inflectional Morphology. A Theory of Paradigm Structure*. Cambridge: Cambridge University Press.
- Wescoat, Michael (2002) *On Lexical Sharing*. PhD thesis, Stanford University.

Ana Luís
 University of Coimbra, Portugal
aluis@fl.uc.pt

Ryo Otoguro
 University of Essex, UK
rotogu@essex.ac.uk

PARTITIONING DISCOURSE INFORMATION: A CASE OF CHICHEWA SPLIT CONSTITUENTS

Sam Mchombo, Yukiko Morimoto & Caroline Féry
UC Berkeley, ZAS Berlin & Universität Potsdam

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

Chicheŵa is said to display mixed properties of configurability such as the existence of VP on the one hand and discontinuous constituents (DCs) on the other. In the present work we examine the discourse and syntactic properties of DCs, and show that DCs in Chicheŵa arise naturally from the discourse configurational nature of the language. We argue that the fronted DCs in Chicheŵa are contrastive topics that appear in a left-dislocated external topic position, and the remaining part of the split NP in the right-dislocated topic position. We develop an analysis that integrates the discourse information of split constituents into the parallel architecture of LFG by assuming a direct mapping between c-structure and i(nformation) structure.

1 Introduction

LFG traditionally encodes discourse information such as topic and focus in f-structure via annotations on c-structure, recognizing them as grammaticized discourse functions (e.g. Bresnan and Mchombo 1987, Alsagoff 1992, King 1995).¹ In the last decade, with increasing interest in the role of discourse information in core syntactic phenomena like word order, proposals have been made to posit an independent projection i(nformation)-structure in addition to the traditional parallel LFG architecture (e.g. Choi 1997, 1999, 2001, King 1997, Cook 2001, King and Zaenen 2004).

King (1997), for example, examines Russian predicate focus, and points out that the traditional treatment of focus as part of the f-structure information has worked adequately for arguments and adjuncts, but it fails to capture the correct scoping of predicate focus: the standard annotations on the focused predicate would include more materials (e.g. selected arguments) in focus than intended. One solution to this problem is to posit an independent i-structure, which is projected off the c-structure, and to separate the i-structure information from the argument structure information.

One of the questions King leaves open for future research, which we wish to take up in this paper, is whether the f-structure should encode any discourse information at all. As King points out, topic and (perhaps to a lesser extent) focus are more syntacticized in some languages (cf. Bresnan and Mchombo 1987 for Bantu; Alsagoff 1992 for Malay) than others. These languages therefore suggest that some i-structure roles are represented in f-structure.

In this paper, we examine split constituents in Chicheŵa, in which parts of an argument (the head, its modifying adjective and demonstrative) have different discourse (topical) roles (= i-structure property), yet the head-marking morphology (= f-structure property) cross-references the argument as a whole as a topic. That is, on the one hand, the morphology indicates that there is one (f-structure) topic, but on the other hand, the c-structure positions these split parts of an argument occupy encode different i-structure roles associated with them. This aspect of the Chicheŵa split construction therefore makes it a curious empirical domain in which to further explore the question raised above—more precisely the question of how much and what type of discourse information should be represented in f-structure. The present work should also serve to illustrate once again the significant role of discourse information in syntactic phenomena that are as fundamental as constituency, and how the LFG parallel architecture is well suited to capture the partitioning of discourse information across multiple levels of representation.

¹We are grateful to Laura Downing and Al Mtenje on the discussion on Bantu tones and information structure, and to Peter Sells for valuable feedback on the formal aspects of the paper. We also thank Mary Dalrymple for her help with technical details of LFG, the audience at LFG-05 for useful questions and comments, and Tracy King and Miriam Butt for editorial comments. The intonational part of this paper is part of the project on nominal and prepositional discontinuous phrases at the Institute for Linguistics in Potsdam, financed by the DFG and conducted by the third author in collaboration with Gisbert Fanselow and Martin Haspelmath. We are solely responsible for all remaining errors or misrepresentations.

The discussion in the rest of the paper proceeds as follows. In section 2 we present data on split constituents in Chicheŵa, focusing on four observations about the construction pertinent to our discussion, and establish the discourse basis of our analysis. Section 3 provides phonological evidence for the discourse properties of split constituents and our syntactic analysis to follow. The analysis is presented in section 4, which highlights the c- to f-structure mapping for the functional identification of split constituents with the f-structure TOPIC on the one hand, and c- to i-structure mapping for the configurational identification of i-structure topics. Our analysis follows earlier proposals by King (1997), Choi (1999, 2001), and Cook (2001), who also assume that the i-structure projects off the c-structure. The final section summarizes the results.

2 Chicheŵa Split Constituents

In this section we present the following four properties of Chicheŵa split constituents: (i) a left-edge constraint, (ii) obligatoriness of a topic anaphoric pronoun on the verb corresponding to the split argument, (iii) fixed ordering of two (or more) contiguous elements, and (iv) splitting of complex possessive NPs. The first two properties are particularly relevant in our analysis. The properties noted in (iii) and (iv) should also fall out of the proposed syntactic analysis.²

To begin with, we show in (1) an example of a complex NP with rich noun class concord. In (1a), the complex NP *these foolish hunters* all agree in the noun class of the head (class 2). In (1b) the head noun *mikango* ‘lion’ is class 4, and the modifiers also must agree.

- (1) a. Njúchí izi zi-ná-lúm-á álenje awa ópúsa.
 10.bees 10.these 10-PST-bite-fv 2.hunter 2.these 2.foolish
 ‘These bees bit these foolish hunters.’
- b. Mikángó i-tátu i-ná-gúmúl-á makólá ónse a-náyi.
 4lions 4three 4PST-pull.down-fv 6corrals all 6four
 ‘Three lions pulled down all the four corrals.’

Although parts of these complex NPs typically occur together with the head noun,³ it is possible, though restricted, to split these nominal constituents. Example (2a) shows the canonical NP structure in Chicheŵa. As shown, it exhibits a strict head-initial structure with the Head-Demonstrative-Adjective order. The examples in (2b-f) show various patterns of discontinuity of that NP (boldfaced).

²Throughout the paper, we will be focusing on split object NPs, even though subjects can also be discontinuous. Because of the unambiguous status of the object marker (OM) as an anaphoric pronoun rather than agreement in Chicheŵa (Bresnan and Mchombo 1987), we can restrict the range of possible alternative analyses. Also, as opposed to the subject, the canonical position of the object (SVO) makes it easier to see when part of it is fronted.

³The integrity of the complex NPs in (1) can be shown by their occurrence in displaced positions such as passive, topicalization, and cleft (see Kathol and Rhodes 2000 for relevant observations).

- (i) a. Álenje awa ópúsa a-ná-lúm-ídw-á ndí njúchí izi.
 2.hunter 2.these 2.foolish 2-PST-bite-PASS-fv by 10.bees 10.these
 ‘These foolish hunters were bitten by these bees.’
- b. Ndi makólá ónse anáyi améné mikángó itátu íná-gúmúl-á.
 COP 6.corrals all 6.four 6.replo 4.lion 4.three 4-PST-pull.down-fv
 ‘It was all the four corrals that the three lions pulled down.’

- (2) a. Njúchíí izi zi-ná-lúm-á **A**enje awa ópɔsa. ... [H D A]
 10.bees 10.these 10-PST-bite-fv 2.hunter 2.these 2.foolish
 ‘These bees bit these foolish hunters.’
- b. awa njúchíí izi zi-ná-wá-lúm-a **A**enje ópɔsa. D ... [H A]
- c. **A**enje njúchíí izi zi-ná-wá-lúm-a awa ópɔsa. H ... [D A]
- d. **A**enje awa njúchíí izi zi-ná-wá-lúm-a ópɔsa. [H D] ... A
- e. awa ópɔsa njúchíí izi zi-ná-wá-lúm-a **A**enje. [D A] ... H
- f. **A**enje ópɔsa njúchíí izi zi-ná-wá-lúm-a awa. [H A] ... D

(i) **Left-edge constraint:** The first observation to note here is that discontinuous constituents (DCs) in Chicheŵa must occur clause-initially, as shown in all the discontinuous examples in (2b–f). The clause-initial DC *awa* ‘these’ in (2b), for instance, cannot be placed elsewhere, as demonstrated in (3). In (3a,b), *awa* is placed clause-medially—immediately pre- and post-verbally. In (3c), *awa* is placed in clause-final position.

- (3) a. *njúchíí izi awa zi-ná-wá-lúm-a **A**enje ópɔsa.
 b. *njúchíí izi zi-ná-wá-lúm-a awa **A**enje ópɔsa.
 c. *njúchíí izi zi-ná-wá-lúm-a **A**enje ópɔsa awa.

(ii) **OM requirement:** The second observation is that all the instances of discontinuity of the object NP above are accompanied by the presence of the OM (*wa* in (2)) that is coreferential with the whole NP, regardless of which part of the object NP (head or modifier) is discontinuous. Without the OM the examples are ungrammatical:

- (7') b. *awa njúchíí izi zi-ná-∅-lúm-a **A**enje ópɔsa.
 c. ***A**enje njúchíí izi zi-ná-∅-lúm-a awa ópɔsa.
 d. ***A**enje awa njúchíí izi zi-ná-∅-lúm-a ópɔsa.
 e. *awa ópɔsa njúchíí izi zi-ná-∅-lúm-a **A**enje.
 f. ***A**enje ópɔsa njúchíí izi zi-ná-∅-lúm-a awa.

The presence of the OM is crucial in that those NPs that cannot be cross-referenced by the corresponding OM (or SM) cannot be discontinuous. For example, an instrumental phrase like *ndí makÆæ awa óbantha* ‘with these blunt hoes’ in (4) in a non-applicative construction cannot be discontinuous.

- (4) a. Mikángó yókálamba i-ná-zí-gúmúl-a ndí mak~~Æ~~æawa ób~~æ~~antha nkhókwe.
4.lion 4.aged 4SM-PST-10OM-demolish-fv with 6.hoe 6.these 6.blunt 10.granary
'The aged lions pulled down the granaries with these blunt hoes.'
- b. *Awa ób~~æ~~antha mikángó yókálamba i-na-zí-gúmúl-a
6.these 6.blunt 4.lion 4.aged 4-PST-10OM-demolish-fv
ndí m~~Æ~~æ~~æ~~u nkhókwe.
with 6.hoe 10.granary

Chicheŵa exhibits object asymmetry (cf. Bresnan and Moshi 1990, Alsina and Mchombo 1993, Ngonyani 1998). In an applicative construction, only the applied object has the properties associated with the primary object. For example in (5), only the beneficiary object *mikángó yókálamba* 'aged lions', introduced by the applicative and not the theme object *mak~~Æ~~æawa ób~~æ~~antha* 'these blunt hoes', can be in anaphoric relation with the incorporated pronominal object.

- (5) a. Anyání a-na-í-gúl-íl-á makású awa óbúntha mikángó yókálamba.
2.baboon 2-PST-4OM-buy-APPL-fv 6.hoe 6.these 6.blunt 4.lion 4.aged
'The baboons bought (for) them these blunt hoes, (for) the aged lions.'
- b. *Anyání a-na-wa-gúl-íl-á mikángó yókálamba makású awa óbúntha.
2.baboon 2-PST-6OM-buy-APPL-fv 4.lion 4.aged 6.hoe 6.these 6.blunt
[Intended as:] 'The baboons bought them for the aged lions, these blunt hoes.'

The examples in (6) show that only the applied beneficiary, and not the theme object, can be discontinuous.

- (6) a. Yók~~Æ~~amba anyání a-na-í-gúl-íl-á makású awa óbúntha mik~~Æ~~gó.
4.aged 2.baboon 2-PST-4OM-buy-APPL-fv 6.hoe 6.these 6.blunt 4.lion
'The baboons bought the aged lions these blunt hoes.'
- b. *Awa ób~~æ~~antha anyání a-na-wa-gúl-íl-á mak~~Æ~~æ
6.these 6.blunt baboon 2-PST-6OM-buy-APPL-fv 6.hoe
mikángó yókálamba.
4.lion 4.aged

Similarly, the oblique agent in a passive sentence cannot be cross-referenced by an OM and hence resists discontinuity, as shown in (7).

- (7) a. Mikángó i-na-ph-édw-á ndí alenje awa ó-dzí-kónd-a.
4.lion 4-PST-kill-PASS-fv by 2.hunter 2.these 2-REFL-love-fv
'The lions were killed by these selfish (self-loving) hunters.'
- b. *Ó-dzí-kónd-a mikángó i-na-ph-édw-á ndí alenje awa
- c. *Awa mikángó i-na-ph-édw-á ndí alenje ó-dzí-kónd-a

As expected by the obligatory presence of the topic-anaphoric OM with a discontinuous object NP, the DCs receive topic interpretation. More precisely, our preliminary inquiry into discourse contexts of various instances of DCs suggests that the fronted element is often a contrastive topic equivalent to a left-dislocated topic, rather than simply given information, or a continuing topic. Given the analysis of the Chicheŵa OM as a topic-anaphoric pronoun, the fact that the OM is required when part of the object NP is discontinuous shows that at least the fronted discontinuous part of the NP must be outside the minimal clausal domain.

The first observation that DCs appear in the clause-peripheral position seems to be true for a majority of languages that allow such split NP construction (cf. Baker (1996) for polysynthetic languages; Dahlstrom (1987) for Algonquian languages in particular). Given that in many languages, clause-initial position is reserved for discourse-related elements such as topic and focus, the observation in (i) lends itself well to another aspect noted in (ii) that fronted DCs receive topic interpretation. In fact, we will show that “topicalizability” is a precondition for any constituent to be discontinuous (at least in Bantu). As argued by Bresnan and Mchombo (1987), the Chicheŵa object marker is employed only as a pronominal argument anaphoric to a floating topic outside the minimal clause nucleus (S/IP), never as grammatical agreement to a non-topical (clause-internal) NP. The observation in (ii) is therefore confirmed by the morphosyntax as well. In previous generative studies of DCs (e.g. Jelinek 1984, Speas 1990, Baker 1996), however, relatively little attention is given to the discourse function of DCs.

There is nonetheless some important work that recognizes the role of information structure in split constituents in general: Reinholtz (1999), for example, argues that clause-initial DCs in Swampy Cree has the discourse function of Focus, and that more generally, the Swampy Cree split NP construction has “all of the hallmarks of *wh*-movement in so-called configurational languages” (p.202) in that “. . . both movement types show the ability to span several clauses, a limited application in relative clauses or embedded questions, and an inability to move any material out of adverbial constituents” (p.218). Reinholtz therefore argues that DCs arise as a result of *wh*-movement.

Fanselow (2001) examines split XP constructions in general, such as a split VP as in (8) and a split DP as in (9) in German.

- (8) **Keine Bücher** hat er [___ gelesen].
 no books has he read

- (9) **Schrecklicher Morde an Studenten** ist er **vieler** beschuldigt worden.
 horrible murders at students is he many accused been
 ‘He has been accused of many horrible murders of students.’

Fanselow argues that such split XP constructions are generally associated with a particular pragmatic structure: “in a split construction, the right part of XP must be focal, while the lefthand part may be a (link-)topic or a second focus” (p.85). Although the precise pragmatic nature of the fronted elements still deserves further discussion, these studies nonetheless suggest that discourse-pragmatic functions of split constructions must be part of any analysis.

(iii) Fixed ordering of contiguous elements: Two other observations are relevant for our analysis of the syntax of Chicheŵa DCs. First, regardless of the position, the ordering of *contiguous* elements is fixed—H(ead) > D(emonstrative) > A(djective)—as shown by the contrast between (2) and (10).

- (10) a. *Njúchíí izi zi-ná-lúm-á **awa** ~~A~~enje óposa. *... [D H A]
 b. ***awa** njúchíí izi zi-ná-wá-lúm-a óposa ~~A~~enje. *D ... [A H]
 c. *~~A~~enje njúchíí izi zi-ná-wá-lúm-a óposa **awa**. *H ... [A D]
 d. ***awa** ~~A~~enje njúchíí izi zi-ná-wá-lúm-a óposa. *[D H] ... A
 e. *óposa **awa** njúchíí izi zi-ná-wá-lúm-a ~~A~~enje. *[A D] ... H
 f. *óposa ~~A~~enje njúchíí izi zi-ná-wá-lúm-a **awa**. *[A H] ... D

The ordering restriction on the fronted elements suggests that they form a single constituent. This need not always be the case, however. For example when the subject NP is left-dislocated, it can come between the two parts of the object DCs, as in (11). In such cases, these discontinuous parts of the object NPs may come in any order, each forming a separate constituent: as shown in (11), the canonical head-modifier ordering *mik*~~A~~ngo (*lion*) *ó-k*~~A~~amb-a (*aged*) is not maintained.

- (11) **Yó-k**~~A~~amb-a anyaní **mik**~~A~~ngo a-na-í-gúl-íl-á makású awa ó-búnth-a.
 4.aged 2.baboons 4.lion 2-PST-4-buy-APPL-fv 6.hoes 6.these 6-blunt-fv
 ‘The aged lions_j, the baboons_i, they_i bought them_j these blunt hoes.’

(iv) **Splitting of complex possessive NPs:** The second additional observation concerns DCs involving complex possessive NPs. As shown by example (12), a possessive NP can be split in Chicheŵa.

- (12) a. Anyaní á mísala a-ku-pwány-a **chipanda** ~~ch~~A **kazit**~~A~~Ø
 2.baboon 2ASSOC 4.madness 2-PRES-smash-fv 7.calabash 7ASSOC 1.spy
 ‘The mad baboons are smashing the calabash of the spy.’
 b. **Chipanda** anyaní á mísala a-ku-chí-pwány-a ~~ch~~A **kazit**~~A~~Ø
 ‘The calabash, the mad baboons are smashing (it) of the spy’
 c. ~~Ch~~A **kazit**~~A~~Ø anyaní á mísala a-ku-chí-pwány-a **chipanda**.
 ‘Of the spy, the mad baboons are smashing (it) the calabash’

However, as soon as we add another layer of possessive NP, splitting becomes more constrained. Consider the examples in (13). Example (13a) is a non-discontinuous example. The element in question, the object possessive NP, is in boldface. In (13b) we front the head noun of the possessive NP, and the result is ungrammatical.⁴ In (13c) we front a possessor *a mfumu* ‘of the chief’. Again the example is rendered ungrammatical. Example (13d), on the other hand, shows that it is possible to front the entire possessor and leave the head noun postverbally.

⁴Note that the example (13b) would be good if there is no OM. In this case, however, we only get the appositive interpretation of the fronted element. The absence of the corresponding OM thus suggests that nothing is out of the basic clause, and that the sentence-initial element is added on to the sentence as an appositive. We return to this contrast between (13b) and the appositive reading without an OM when we discuss the information structure of the non-fronted elements.

- (13) a. Anyaní a-na-mphwanya **chipanda chÆ alenje a mfumu.**
 2.baboons 2-PAST-smash 7.calabash 7ASSOC 2.hunter 2.ASSOC 1.chief
 ‘The baboons smashed the calabash of the hunters of the chief.’
- b. ***Chipanda_i** anyaní a-na-chi-mphwanya ____i **chÆ alenje a mfumu.**
 7.calabash 2.baboons 2-PAST-7-smash 7.ASSOC 2.hunter 2.ASSOC 1.chief
 ‘The calabash, the baboons smashed of the hunters of the chief.’
- c. ***A mfumu_i** anyaní a-na-chi-mphwanya **chipanda chÆ alenje ____i.**
 2.ASSOC 1.chief 2.baboons 2-PAST-7-smash 7.calabash 7.ASSOC 2.hunter
 ‘Of the chief, the baboons smashed the calabash of the hunters.’
- d. **ChÆ alenje a mfumu_i** anyaní a-na-chi-mphwanya **chipanda ____i.**
 7.ASSOC 2.hunter 2.ASSOC 1.chief 2.baboons 2-PAST-7-smash 7.calabash
 ‘Of the hunters of the chief, the baboons smashed the calabash.’

At this point, we leave these facts simply as an additional observation about complex possessive NPs. In the analysis to follow, we suggest that the constraint that bans the examples in (13b,c) must be formulated in terms of the information structure and heaviness of the parts of the NP that remain postverbally rather than the syntax of complex possessive NPs.

3 Discourse Functions and Syntactic Position of DCs

Based on the basic properties observed earlier that (i) DCs in Chicheŵa must occur clause-initially; and (ii) clause-initial DCs receive topic interpretation and require an anaphoric pronoun on the verb corresponding to the whole NP, we analyze the split constituents as instances of left-dislocation, in which the dislocated element is outside the minimal clause nucleus and receives contrastive topic interpretation. The analysis is consistent with the fact that every instance of DCs requires the OM on the verb and the analysis given by Bresnan and Mchombo (1987) that the OM in Chicheŵa is reserved only for topic-anaphoricity.

Furthermore, the fact that every instance of a discontinuous object NP requires the corresponding OM suggests that no part of the object NP remains inside the VP, given the topic-anaphoric analysis of the OM proposed by Bresnan and Mchombo (1987). This means that the remaining postverbal part of the object NP must be right-dislocated. This assumption is in line with the presumed discourse function of this part of the DC: it is old, non-prominent information. The discourse functions and their structural correlates we wish to explore are supported by cross-linguistic studies of left- and right-dislocated elements and by phonological evidence.

3.1 Cross-Linguistic Functional Evidence

According to C. Lee (1999a,b) while TOPIC is prototypically given, presupposed, and anchored in speech situation, CONTRASTIVE TOPIC has a focal part in contrast with the rest of the parts, and the speaker has the alternatives in contrast or contrast set in mind. While topic can be unaccented, contrastive topic shows a prominent intonation pattern cross-linguistically.

In Chicheŵa, the contrastive part of a topic constituent appears in the left-dislocated position, resulting in a split construction. For example for the split example in (2d), repeated here in (14), the most

likely context is where there are two sets of foolish people in prior discourse—these foolish hunters and those foolish fishermen. *Aenje awa* ‘these hunters’ is then contrasted with ‘those fishermen’ in the example. The ‘foolish’ part of the NP is old, non-contrastive information, and remains postverbal. We return to this point shortly.

- (14) **Aenje awa** njúchíí izi zi-ná-wá-lúm-a **ópøsa**.
 2.hunter 2.these 10 bees 10.these 10-PST-2-bite-fv 2.foolish
 ‘These bees bit these foolish hunters.’

Additional data show that “topic-hood” is in fact a pre-condition for a constituent to be discontinuous. For example, Chicheŵa has a number of verb-object idioms, in which the object is formally non-referential, as in example (15a). Non-referential NPs can never be topics, and, as such, they cannot be discontinuous, as demonstrated in (15b,c).

- (15) a. Nd-a-gwil-a mwendo wáko.
 1SG-PREF-grab-fv 3.leg 3.your
 (lit.) ‘I have grabbed (your) leg.’ = ‘I apologize.’
 b. *Wáko nd-a-gwil-a mwendo.
 c. *Mwendo nd-a-gwil-a mwendo

Similarly *wh*-phrases, which are inherent focused, cannot be fronted:

- (16) a. Mikango u-na-gumula **nyumba ya yani?**
 lion sm-past-destroy house of who
 ‘Whose house did the lions destroy?’
 b. ***ya yani** mikango u-na-gumula **nyumba?**

Crosslinguistically, these types of discourse topic seem to be associated with the syntactic positions just noted.⁵ For example, regarding the left-peripheral topic, in verb-initial languages, D. Payne (1990, 1992) identifies the preverbal position to be what she refers to as the “pragmatically marked” (PM) position. The PM information is non-presupposed asserted new information, contrastive information (i.e. focus) as we as given, discourse-prominent information (topic). Payne shows that in strongly verb-initial languages, these pragmatically marked constituents, either focus or topic, appear sentence-initially. Cooreman (1992:244) essentially makes the same observation: the non-verb initial order in the canonically verb-initial language Chamorro is commonly found when “the thematic unity of the [narrative] is disrupted”, such as change of events, or when the paragraph theme is temporarily suspended. Cooreman’s description of these sentence-initial elements in Chamorro is comparable to Aissen’s (1992) description of the external topic—the new or contrastive topic. Subsequent work on verb-initial languages makes similar observations about the discourse function of the sentence-initial position (e.g. Harold (1995:50) for Biblical Hebrew).

In SVO languages, new or contrastive topic also appears at the left-periphery in a dislocated position. Birner and Ward (1998:256–257) show that among the various syntactic constructions that encode

⁵The discussion the following cross-linguistic studies is based on the fuller review of the cited literature in Morimoto (2000, chapter 2).

different types of discourse referents in English (e.g. inversion, *by*-phrase passive, topicalization, existential, left-dislocation, right-dislocation), new or contrastive topic (hearer-new or discourse-new in Birner and Ward's taxonomy) is expressed in the left-dislocated position. In another SVO language Tok Pisin, a creole language in Papua New Guinea, Sankoff (1993) provides an example showing that (what we would call) a new/contrastive topic appears in a left-dislocated position followed by an anaphoric pronoun.⁶

In SOV languages, where scrambling and case marking are common typological features, contrastive topics may not always appear in a left-dislocated position. They are nonetheless morphologically and prosodically clearly marked, according to C. Lee (1999a,b). In Korean, for example, even though topics with the topic marker *-(n)un* can scramble, the canonical position of these topics seems to be clause-initial (Choi 1999). In German, contrastive elements (topic or focus) appear in the left-peripheral position (e.g. SpecCP for Choi 1999, Berman 2000).

As for the right-dislocated topic, it is observed for a number of languages that the right-dislocated position is reserved for afterthought or discourse-old information—e.g. Takami (1995) for Japanese and English, Birner and Ward (1998) for English, Sells (1998) for Japanese, Kimenyi (1980) for Kinyarwanda; see also Morimoto (2000, chapters 4–5), who discusses the afterthought function of right-dislocated elements in Bantu languages.

These crosslinguistic studies on left- and right-topics collectively tell us that there is a robust tendency that these types of topics are structurally defined. As shown below, our preliminary findings on phonological phrasing of these left- and right-topics indicate that they each form their own phonological phrase (also shown by Downing, et al. 2005, as cited below). These observations about the structural correlates at the syntactic and phonological level together suggest a grammatical architecture in which there is a flow of information, or mapping, (at least) between discourse or information structure ('i-structure') and c-structure on the one hand, and i-structure and prosodic structure on the other.

3.2 Phonological Evidence

Our preliminary investigation of the prosodic structure of split constructions in Chichewa also corroborates the preceding observations regarding the discourse status and the proposed syntactic positions. In order to test the prosodic phrasing, we elicited spoken utterances from the second author, Sam Mchombo (native speaker of Chichewa). The results of our experiment are also supplemented by those of Downing, Mtenje, and Pompino-Marschall (2005), who investigated phonological phrasing with respect to focus.⁷

Kanerva's (1990) study of focus and phrasing in Nkhotakota Chichewa showed that in a canonical, discourse neutral SVO sentence, the subject forms its own phonological phrase (henceforth p-phrase) separate from the VP, and the verb and object form one p-phrase together (see also Bresnan and Kanerva

⁶An example of a new/contrastive topic from Sankoff (1993:121) is given below. The dislocated topic is in small caps, and the anaphoric pronoun is underlined.

(i) kakaruk na pik wonem samting i-stap. Na OLGETA MAN IA ol i-poret long guria na ol i-go
 chicken and pig what something stay and all people DET 3pl afraid of earthquake and 3pl go
 pinis.
 complete
 '(Only) chickens and pigs and whatever were there. But ALL THE PEOPLE, they were afraid of the earthquake and they had all left.'

⁷The investigation of phonological phrasing of relevant utterances is only preliminary, and we have not yet tested all the relevant utterances with split constituents. Nonetheless, the sampling we obtained so far conforms to results reported by Kanerva (1990) and Downing et al. (2005) on phonological phrasing of discourse-prominent elements in Chichewa, and we therefore take our sampling to represent reliable evidence.

(1989), Downing et al. (2005:15, ex.(19a))). Kanerva (1989, 1990) discusses several phonological rules, summarized in (17), that are sensitive for phrasing at the level of the p-phrase which he calls ‘focal phrases’. Since these phrases are not exclusively triggered by focus, but can also arise through syntactic movement and topicalization, we prefer to use a more neutral term ‘p-phrase’. This level of prosodic phrasing is indicated by round brackets in the following examples.

(17) Phonological rules applying at the level of the phrasing in p-phrases

- a. Penultimate Lengthening: The vowel in the penultimate syllable of a p-phrase is lengthened.
- b. Retraction: A H-tone is retracted from the final mora of a p-phrase to the penultimate syllable.
- c. Nonfinal Doubling: A H-tone is doubled (a mora is spreaded to the right), except if it is in the phrase final foot.

The word *ⁿjiⁿgá* ‘bicycle’ is realized unchanged in *ⁿjiⁿgÆyÆwiino* ‘good bicycle’ because there, it is not final. But it is pronounced as [*ⁿjiⁿga*] (with lengthening and retraction), when it is p-phrase final. Consider next the word *kugúlò* ‘buy.’ H-tone doubling applies in *kugaÆnyaama* ‘to buy meat’. When this word is p-phrase final it is realized as *kuguula*, with penultimate lengthening.

As far as prosodic phrasing is concerned, Kanerva claims that, in an all-new expression, a head is phrased together with a following complement, as well as with any other element within the same projection, as shown in (18) and (19).

(18) [V NP]
 (tinaba kaluulu)_p
 we-stole hare

(19) [X1 XP2 XP3]_{XP1}
 ()_p
 [V NP [P NP]_{PP}]_{VP}
 (anamenya nyumba ndi mwaala)_p
 he-hit house with rock

‘He hit the house with the rock.’

This means that, in a sentence without narrowly focused constituent, all constituents are pronounced in a single P-phrase. But focus restructures the phrasing of utterances. For instance, if the verb is focused, it forms its own phrase, and the subsequent phrases are phrased individually, as illustrated schematically in (20).

(20) [V_{FOC} NP PP]_{VP}
 ()_p ()_p ()_p

As one can see from (21), illustrated in Figure 1 (in the appendix),⁸ the same prosodic pattern as the one shown in (19) was reproduced in our recordings. As can be extracted from Figure 1, *njuchi izi* ‘these bees’ is separated from the rest of the sentence by a clear break. The first [i] of *izi* is lengthened, as predicted by rule (17a). Furthermore, the first p-phrase of a sentence is regularly terminated by a high tone, regardless of the underlying tone of the final syllable of this p-phrase. The second p-phrase in Figure 1 is uttered at a register which is altogether downstepped relatively to the first one. There is no break between the verb and the following direct object. On the contrary, the final *a* of the verb and the first *a* of *alenja* are fused together. The high tone of *opusa* is downstepped relatively to the high tone of the verb. The last characteristic of this phrase is the final low tone typical for declarative phrases.

- (21) (Njuchi izi) (zi-na-luma alenje awa opusa) Canonical SVO sentence
 10.bee 10.these 10-past-bite 2.hunter 2.these 2.foolish
 ‘These bees bit these foolish hunters.’

A right-dislocated object forms its own p-phrase separated from the verb, as illustrated in (22) and Figure 2. Once again, the same result is reported by Downing, et al (p.15, ex.(19b)). The difference between Figure 1 and Figure 2 is in the phrasing, which induces deaccenting of *awa opusa* (and of course in the presence of the OM *-wa-* on the verb).

- (22) (Njuchi izi) (zi-na-wa-luma) (alenje awa opusa) Object right dislocation
 10.bee 10.these 10-past-2-bite 2.hunter 2.these 2.foolish
 ‘These bees bit them, these foolish hunters.’

Similarly, a left-dislocated object forms its own p-phrase. Figure 3 is taken from Downing et al (p.15, ex.(19c)). In (23), the left-dislocated object *mbuzi izi* ‘these goats’ is a contrastive topic. The postverbal subject *mikango* ‘lion’, non-prominent information, is right-dislocated and forms a separate p-phrase from the verb.

- (23) (Mbúzi izi) (inázisaaka) (mikáango) Object left dislocation
 10.goat 10.these 4.past.10.hunt 4.lion
 ‘These goats, they (lions) hunted them (goats), the lions.’

Turning now to split constituents, example (24) in Figure 4, shows that the fronted part *alenje* ‘hunters’ forms its own p-phrase, like the left-dislocated whole object in (23). The postverbal remaining part of the split object NP is extraposed and appears in the right periphery.

- (24) (**Alenje**) (zinawaluma njuchi izi) (**awa opusa**) Splitting the head of the obj NP
 Hunters bite bees these these foolish
 ‘These bees bit these foolish hunters.’

Figure 5 shows a pitch track of example (14), reproduced in (25) with its phonological phrasing. In this example, the object is split and the subject is topicalized. Both fronted parts are phrased in separated p-phrases, and are not downstepped relatively to each other. In other words, *njuchi izi* ‘these bees’ is at the same pitch height as *alenje* ‘hunters,’ but the p-phrase containing the verb is again downstepped.

- (25) (**Alenje**) (njuchi izi) (zinawaluma) (**awa opusa**) =(4c)
 2.hunter 10.bee 10.these 10.past.2.bite 2.these 2foolish
 ‘These bees bit these foolish hunters’

⁸All the figures are attached in the appendix.

Finally, example (26) and Figure 6 illustrate that both the subject and the object may be split in a single sentence. The extraposed elements are phrased together pointing to the fact that there may be a restriction on the number of deaccented p-phrases.

- (26) (Izi) (awa opusa) (zinawaluma) (alenje njuchi)
 These these foolish bite hunters bees

The phonological phrasing of the split constructions in (23) to (26) clearly shows that, prosodically, no part of the split NP is inside the minimal clause nucleus VP.

Summary

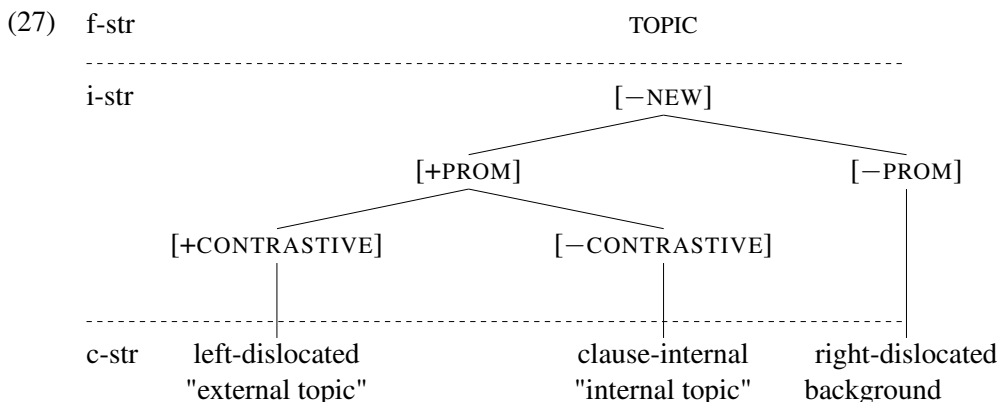
The available data suggest that DCs in Chicheŵa are best analyzed as involving both left-dislocation of the fronted element(s) and right-dislocation of the remaining element(s). Pragmatically the fronted part serves as a contrastive topic, as characteristic of external, left-dislocated topics in other languages. The external topic analysis of DCs in Chicheŵa is not in line with Reinholtz' (1996) analysis that DCs have focus and arise by way of *wh*-movement. We suggest here that languages that permit split NP constructions make use of them for discourse purposes, but exactly which function DCs have may depend on the information structuring of an individual language (see also Féry and Paslawska 2005 for a similar observation). While focus (or discourse-prominent elements in general) may be expressed clause-initially in Algonquian languages (cf. Aissen 1992), in Bantu languages clause-initial position is strictly reserved for topic, and focus is expressed postverbally (cf. Morimoto 2000). Thus, given the patterns of information structuring in Bantu, clause-initial DCs would naturally receive a topic interpretation.

4 Discourse Configurational Analysis

Taking the discourse functions and phonological phrasing as our basis, we now consider the syntactic structure of split NPs. The key analytical problems we wish to solve are the following: (i) functional identification of the DCs with the associated argument function, and (ii) configurational identification of the types of topic involved the split construction—namely the external, contrastive topic in the left-periphery, and the non-prominent, old topic in the right-periphery.

4.1 I-Structure

The parallel structures and their relations we assume in the present work are shown (27).

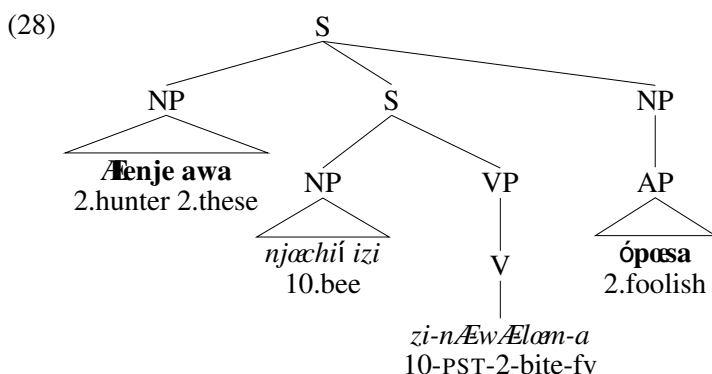


As shown, the f-structure TOPIC interfaces with i-structure [–NEW], and subsumes the distinct i-structural topics: contrastive and non-contrastive, and among the latter, prominent and non-prominent. Following Choi (1999, 2001), we use binary features to represent these types of i-structure topics. The c-structure positions of these nodes then determine the different types of i-structural topics: the (f-structural) TOPIC in the left-periphery is associated with the (i-structural) contrastive topic, and the one in the right-periphery with old non-prominent topic. In other words, it is the mapping between i-structure and c-structure that gives the f-structure notion of TOPIC particular discourse interpretations.

The idea that there is a grammaticized notion of TOPIC at the level of f-structure which subsumes different i-structure topics is supported by the fact that different i-structure topics are not distinctly marked by morphology (which is represented at the level of f-structure). For example, the Japanese and Korean topic markers *wa* and *nun* mark all types of i-structural topic: contrastive, continuing, and non-prominent old topic. These different i-structural topics are usually distinguished by structural position or prosody. We therefore assume a direct mapping between i-structure and c-structure to account for the discourse configurational nature of DCs.

4.2 Mapping between the Parallel Structures

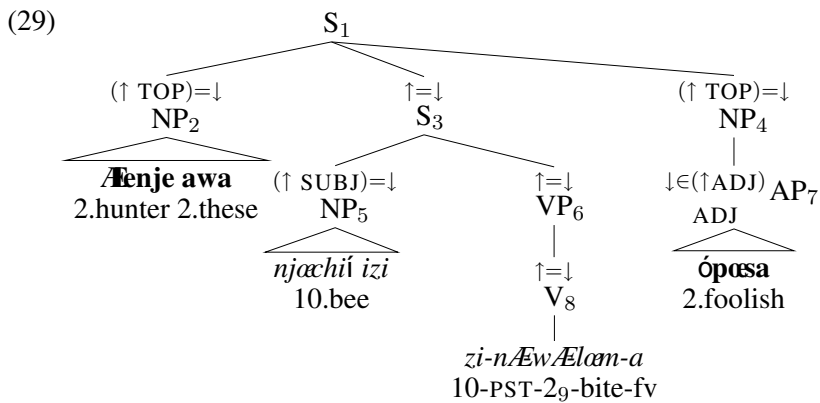
The proposed c-structure of the split construction is shown in (28). For the illustration, we use the example in (14) above.



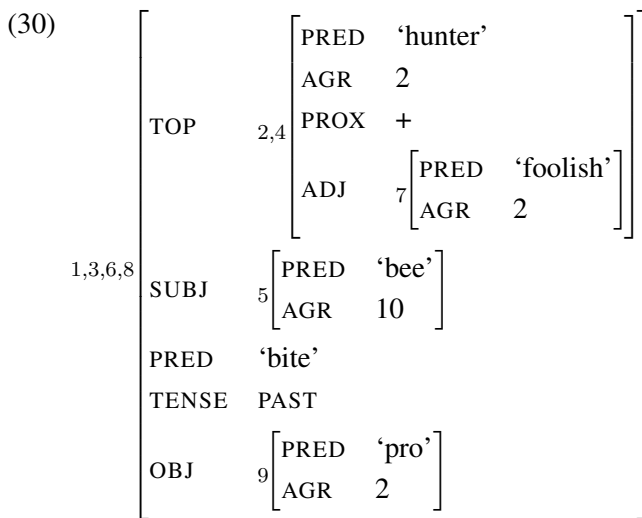
LFG posits two types of clausal organization in natural languages: the endocentric clausal organization with headed XPs, and the exocentric one with S. As in the structure in (28), we make use of the exocentric category S for languages that lack independent evidence for I. In Bantu languages, all verbs inflect uniformly like main verbs, and there is no particular class of inflectional verbs that behave otherwise. For this reason, it has been proposed that Bantu clauses consist of the exocentric category S rather than IP (e.g. Bresnan and Mchombo 1987, Morimoto 2000, 2001).

C- to f-structure mapping: The functional annotations on (28) are shown (29). The corresponding f-structure is shown in (30).⁹

⁹As the internal structure of the fronted NP in (29) is abbreviated, the corresponding f-structure in (30) is also simplified: the lexical information of *awa* has been reduced to a single feature PROX(IMATE).



The annotation on the fronted DC (NP₂) and the remnant part of the DC (NP₄) ‘(↑ TOP) = ↓’ states that the f-structure of the mother node (S₁) contains TOP, whose value is identified with the f-structure of the respective NP. The annotation on the AP builds an inner f-structure of the ADJ(unct) function inside the f-structure of TOPIC.



The functional identification of TOPIC with the argument function OBJ is ensured by the principles of COMPLETENESS and COHERENCE, or more precisely, EXTENDED COHERENCE. Completeness requires that every function designated by a predicate be present in the f-structure of that predicate (Bresnan 2001:63). Thus, completeness rules out examples like that in (31), where all the arguments selected by the predicate *give* are not present.

(31) *John gave a book.

Note that completeness is a requirement that applies at the level of f-structure, and does not require that all the arguments be present on c-structure. Null argument languages like Japanese and Korean, for examples, allow an utterance like that in (31), but at the level of f-structure, all the arguments selected by the predicate are represented and provide their morphosyntactic information and semantic content.

Now in examples like that in (29), part of the DC is the ADJUNCT function (AP) inside the object NP. Completeness is not sufficient to license such elements because it only requires that the selected arguments be properly represented in the f-structure. These adjuncts, not properly selected by the predicate, nonetheless must be properly integrated into the semantics of the predicate and its arguments.

COHERENCE, or the EXTENDED COHERENCE CONDITION, on the other hand, ensures just this type of well-formedness. Coherence requires that every argument function in an f-structure be designated by a PRED. The principle rules out ill-formed examples like that in (32) (Bresnan 2001:63).

(32) *We talked *the man* about that problem for days.

| | | |
|-------|------------------------|---|
| PRED | ‘talk <SUBJ, OBL>’ |] |
| SUBJ | [“we”] |] |
| *OBJ | [“the man”] | |
| OBL | [“about that problem”] | |
| ADJ | [“for days”] | |
| TENSE | PAST | |
| | | |

The intransitive verb *talk* takes an optional oblique argument, and PRED has the OBL designator in (32). It has no OBJ designator, however; having the extra argument violates the coherence condition and results in an ill-formed f-structure.

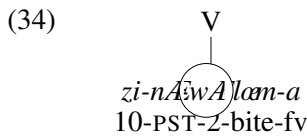
While the coherence condition applies only to argument functions (SUBJ, OBJ, OBJ_θ, OBL), the extended coherence condition applies to all syntactic functions, requiring them to be appropriately integrated into an f-structure (Fassi Fehri 1984, Zaenen 1985, Bresnan and Mchombo 1987). As stated above, argument functions are integrated when they are designated by the PRED. Adjuncts are integrated if their immediate f-structure contains a PRED. The grammaticized discourse functions TOP and FOC are integrated if they are functionally identified, or anaphorically linked to, an integrated function.

Semantic Coreference: Returning to our example in (29), the TOPIC function in the left- and right-periphery is properly integrated into the f-structure in (30) by the extended coherence condition, but completeness and extended coherence must be satisfied by one of the arguments identifying TOPIC as being associated with it. As we have seen, in a sentence with an object DC, the DC is cross-referenced by the obligatory presence of the object marker on the verb.

A standard way of identifying TOPIC with one of the arguments in LFG is the equation in (33), which simply states that the f-structure of the mother node contains the OBJ attribute, whose value is identified with the f-structure of TOP.

(33) (↑ OBJ) = (↑ TOP)

This equation would be a problem, however, for the present case of OM-TOPIC identification: the OM and the TOPIC NP have different PRED values, and PRED values cannot unify. What we want is to anaphorically link the TOP to an integrated function that shares the same agreement features with those of the TOP. To obtain this coreference, we assume that the OM carries the following information given in (34). For the illustration, we use the verb form in (29), repeated below as (34).

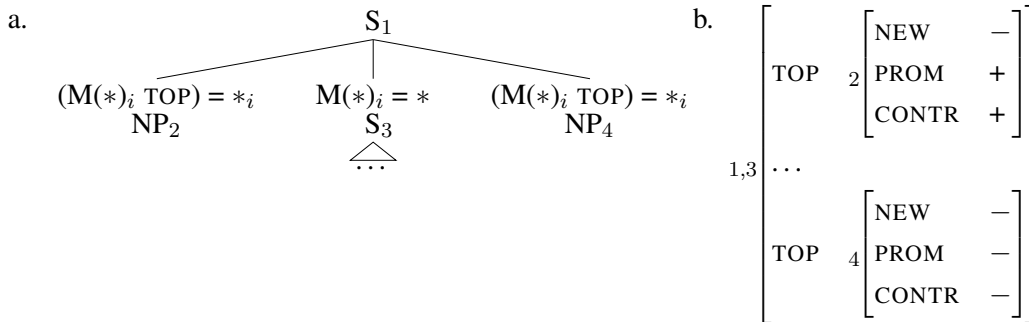


- (↑ OBJ)
 (↓ PRED) = 'pro'
 (↓ INDEX AGR) = 2
 (↓ INDEX) = (↑ TOP INDEX)

The functional annotations on the OM in (34) instantiates the f-structures within the f-structure of OBJ, contained in the f-structure of the mother node (V). The down arrow points to the f-structure of OBJ. The annotation '(↓ INDEX AGR) = 2' states that the f-structure of OBJ contains another f-structure whose attribute is INDEX. The value of INDEX is another f-structure, whose attribute is AGR and its value 2.

C- to i-structure mapping: To model the direct mapping between c- and i-structure, we simply replace the functional annotations ↓ and ↑ with the notations used in (35), which represent an abbreviated c-structure for the split NP construction.

(35) c- to i-structure mapping



The * refers to the current c-structure node, and the M(*) to the mother node (cf. Kaplan 1987). The subscripted *i* refers to the i-structure of that node. The annotation on NP₂, for example, states that the i-structure of the mother node contains a topic, whose value is identified with [−NEW, +PROMINENT, +CONTRASTIVE], due to the c-structure position of the annotated node (left-adjoined). The annotation on the right-peripheral NP, on the other hand, associates the node with the i-structure topic [−NEW, −PROMINENT, −CONTRASTIVE].

Of course, a further analytical problem we must consider for a more complete analysis is how to ensure that left-peripheral topic will be associated with the contrastive, prominent one and the right-peripheral topic with the non-contrastive, non-prominent one. An obvious solution might be to assume a set of mapping constraints like those in (36), which are to be interpreted as universal and violable as in OT. We then let these constraints interact with c-structure constraints to give us language particular c-structure realization of i-structure topics.

- (36) a. Realize –NEW at the left edge. (“old before new”)
 b. Realize +CONTRASTIVE at the left edge. (“iconicity”: prominent information = prominent position)
 c. Realize +PROMINENT at the left edge. (“iconicity”)
 d. Do not realize –PROMINENT at the left edge.

We will leave for future research the precise implementation of such a constraint system into our analysis of split constituents. Such an approach is already explored by Choi (1999, 2001) for various word order phenomena such as scrambling, detachment, topicalization, and focus preposing.

4.3 Further Consequences of the Right-Dislocation Analysis of the “Remnant”

We now return to the last restriction noted earlier in section 2 on the splitting of complex possessive NPs. The relevant examples from (13) are repeated here in (37). The observation was that of the various splitting possibilities of a complex possessive NP, the only grammatical instance is where the head noun remains and the rest is fronted, as in (37d).

- (37) a. Anyaní a-na-chi-mphwanya **chipanda chÆ alenje a mfumu.**
 2.baboons 2-PAST-7-smash 7.calabash 7.ASSOC 2.hunter 2.ASSOC 1.chief
 ‘The baboons smashed the calabash of the hunters of the chief.’
- b. ***Chipanda_i** anyaní a-na-chi-mphwanya ____i **chÆ alenje a mfumu.**
 7.calabash 2.baboons 2-PAST-7-smash 7.ASSOC 2.hunter 2.ASSOC 1.chief
 ‘The calabash, the baboons smashed of the hunters of the chief.’
- c. ***A mfumu_i** anyaní a-na-chi-mphwanya **chipanda chÆ alenje ____i.**
 2.ASSOC 1.chief 2baboons 2-PAST-7-smash 7.calabash 7.ASSOC 2.hunter
 ‘Of the chief, the baboons smashed the calabash of the hunters.’
- d. **ChÆ alenje a mfumu_i** anyaní a-na-chi-mphwanya **chipanda ____i.**
 7.ASSOC 2.hunter 2.ASSOC 1.chief 2.baboons 2-PAST-7-smash 7.calabash
 ‘Of the hunters of the chief, the baboons smashed the calabash.’

Our speculation on these data is that this is not due to some syntactic constraint, but it is constrained (at least partly) by phonological weight—namely that only one prosodic word is allowed in the right-dislocated position, where the constituent forms its own phonological phrase. A similar observation is made for non-discontinuous right-dislocation in other Bantu languages. For example in Kinyarwanda, Kimenyi (1980:203) observes that whereas multiple left-dislocated topics are possible, right-dislocated topics are restricted to only one constituent. The latter restriction is exemplified in (38).

- (38) *Umgabo y-a-ya-mu-haa-ye, **amafaraanga, umugóre.**
 man 1SM-PAST-it-give-PERF money woman
 ‘The man gave it to her, the money (to) the woman.’

Furthermore, we noted earlier in footnote 3 that (37b) would be grammatical if the fronted head noun *Chipanda* ‘calabash’ had an appositive interpretation. Crucially, in that case the verb cannot have

the OM. This suggests that the instance of what appears to be fronting with the appositive interpretation in fact involves neither fronting of any element nor right-dislocation of the “remnant” element(s), and that the clause-initial appositive element is simply added on to a canonical SVO sentence. Therefore, assuming that our right-dislocation analysis of the remnant is correct, we conjecture that this right-dislocated position imposes the constraint on phonological weight, and DCs involving ‘heavy’ remnants are dispreferred.¹⁰

5 Conclusion

In this paper, we have offered a discourse-configurational analysis of Chicheŵa split constituents, where the fronted element, the contrastive topic, occupies a left-dislocated topic position, and the remnant part of the split NP, the old, non-prominent topic, appears in a right-dislocated position. The analysis is consistent with the fact that every instance of object DCs requires the corresponding object marker on the verb, whose function is topic-anaphoric (Bresnan and Mchombo 1987). The structural analysis is also supported by the preliminary findings on phonological phrasing of DCs. Given the right-dislocated analysis of the remnant part of a split NP, we speculated that the constraint on splitting of complex possessive NPs has to do with phonological weight—that heavy elements are dispreferred in right-dislocated position.

Examining DCs beyond the Bantu family would naturally require looking at the various discourse functions that DCs serve in the languages in question and determining the structural correlates of such discourse elements. Nonetheless we hope that, in future research, our analysis of Chicheŵa split NPs will be a step in the right direction towards taking into account multiple levels of representations (discourse, syntax, phonology) in order to provide a comprehensive analysis of split constructions.

¹⁰We realize that this cannot be the whole story, as the right-dislocation of the entire complex possessive NPs in (37a), for example, would be possible. So there is something peculiar about the syntax of splitting complex possessive NP.

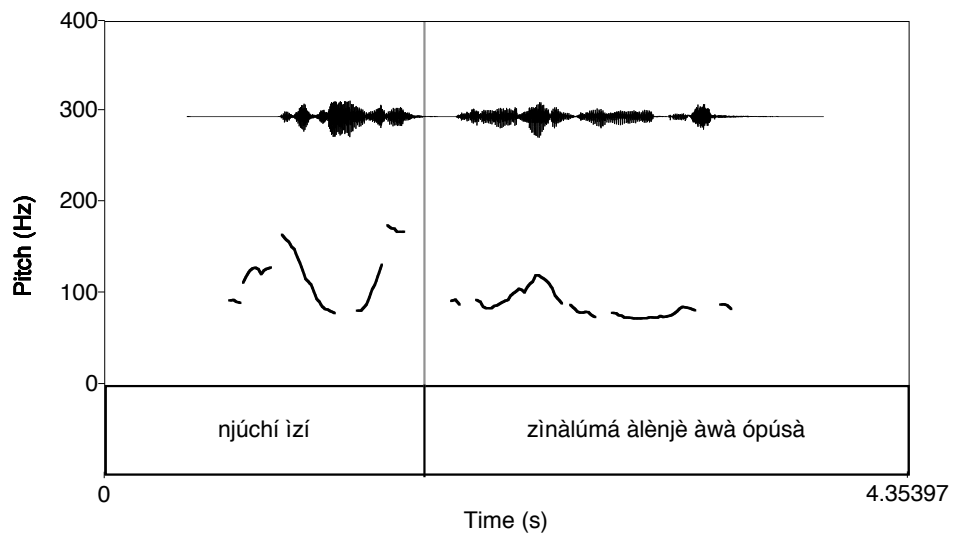


Fig 1

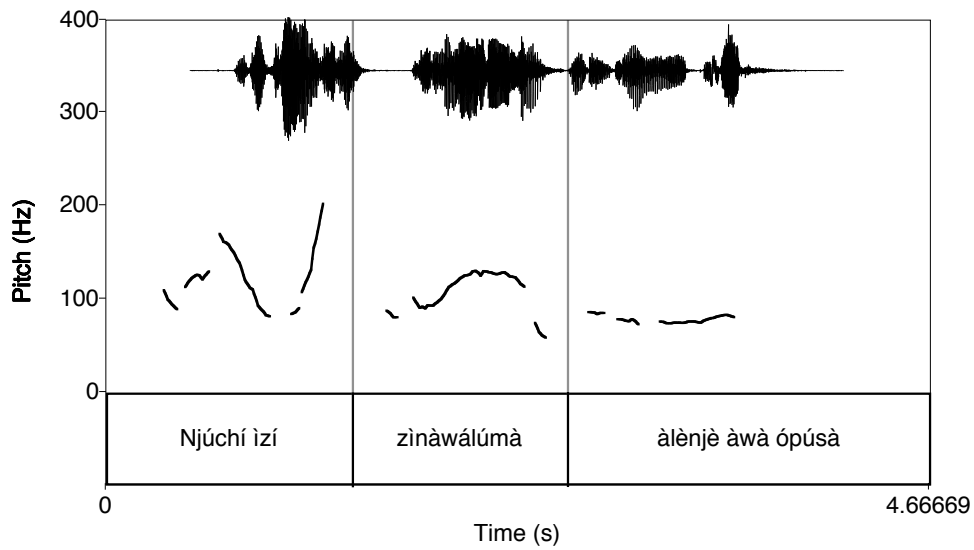


Fig.2

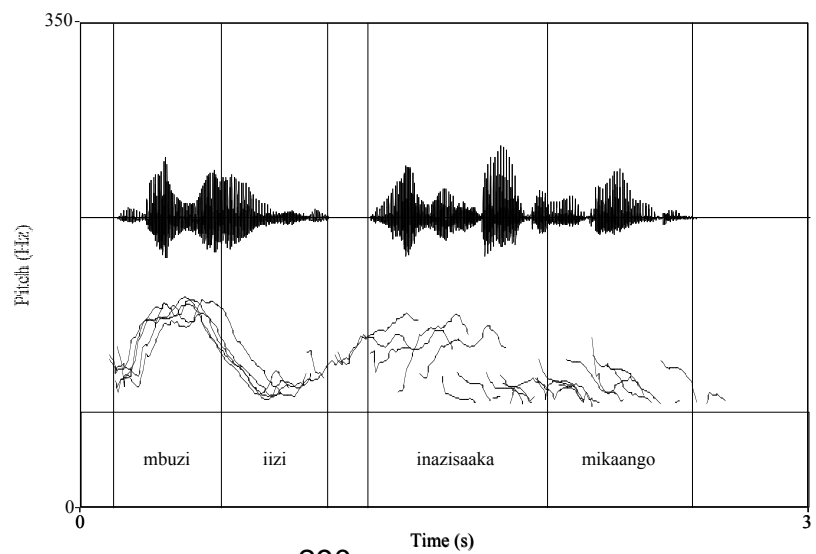


Fig.3

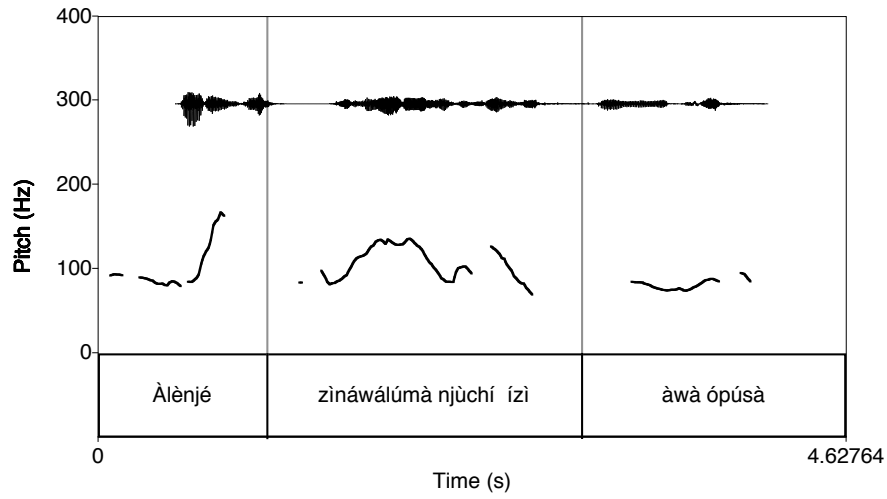


Fig 4

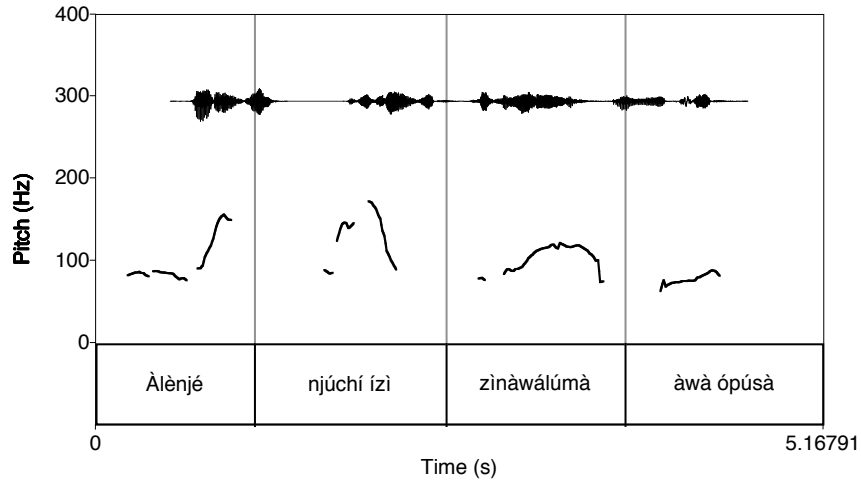


Fig. 5

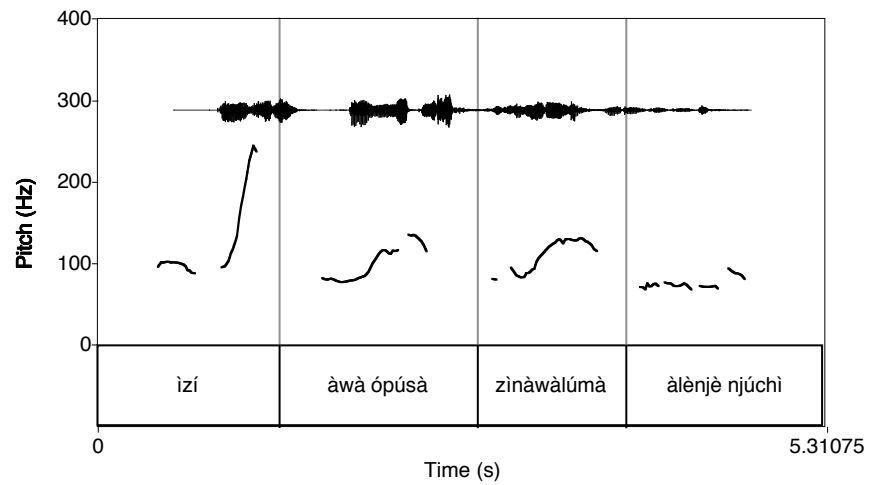


Fig. 6

References

- Alsina, Alex, and Sam A. Mchombo. 1993. Object asymmetries and the Chicheŵa applicative construction. In Sam A. Mchombo (ed.), *Theoretical Aspects of Bantu Grammar*, chapter 1, 17–45. Stanford, CA, CSLI Publications.
- Baker, Mark C. 1996. *The Polysynthesis Parameter*. New York/Oxford, Oxford University Press.
- Berman, Judith. 2000. *Topics in the Clausal Syntax of German*. Doctoral dissertation, Universität Stuttgart, Stuttgart.
- Birner, Betty J., and Gregory Ward. 1998. *Information Status and Noncanonical Word Order in English*. Amsterdam/Philadelphia, John Benjamins.
- Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford, Blackwell Publishers.
- Bresnan, Joan, and Jonni M. Kanerva. 1989. Locative inversion in Chicheŵa: A case study of factorization of grammar. *Linguistic Inquiry* 20(1), 1–50.
- Bresnan, Joan, and Sam A. Mchombo. 1987. Topic, pronoun, and agreement in Chicheŵa. *Language* 63(4), 741–782.
- Bresnan, Joan, and Lioba Moshi. 1990. Object asymmetries in comparative Bantu syntax. *Linguistic Inquiry* 21(2), 147–185.
- Choi, Hye-Won. 1999. *Optimizing Structure in Context: Scrambling and Information Structure*. Stanford, CA, CSLI Publications.
- Choi, Hye-won. 2001. Phrase structure, information structure, and resolution of mismatch. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality Theoretic Syntax*, 17–62. Stanford, CA, CSLI Publications.
- Cook, Philippa. 2001. *Coherence in German: An Information Structure Approach*. Doctoral dissertation, University of Manchester, Manchester, UK.
- Cooreman, Ann. 1992. The pragmatics of word order variation in Chamorro narrative text. In Payne (Payne 1992), 243–263.
- Downing, Laura, Al Mtenje, and Bernd Pompino-Marschall. 2005. Non-accentual prosodic cues to focus in a tone language: the case of Nteheu Chicheŵa. Paper presented at the Bantu tone and stress conference, 16–18 June, Leiden University.
- Downing, Pamela, and Machael Noonan (eds.). 1995. *Word Order in Discourse*. Vol. 30 of *Typological Studies in Language*. Amsterdam/Philadelphia, John Benjamins.
- É. Kiss, Katalin (ed.). 1995. *Discourse Configurational Language*. New York/Oxford, Oxford University Press.
- Fassi Fehri, Abdelkader. 1984. Agreement in Arabic, binding, and coherence. In Michael Barlow and Charles A. Ferguson (eds.), *Agreement in Natural Language*, 107–158. Stanford, CA, CSLI.
- Féry, Caroline, and Alla Paslawska. n.d. Discontinuous constructions in Ukrainian. MS. University of Potsdam and University of Lwiw.
- Grimshaw, Jane. 1991. Extended projection. MS. Department of Linguistics and Center for Cognitive Science, Rutgers University.
- Grimshaw, Jane. 1997. Projection, heads, and Optimality. *Linguistic Inquiry* 28(3), 373–422.
- Hale, Ken. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language & Linguistic Theory* 1(1), 5–47.
- Harold, Bruce B. 1995. Subject-verb word order and the function of early position. In Downing and Noonan (Downing and Noonan 1995), 137–161.

- Jelinek, Eloise. 1984. Empty categories, case, and configurationality. *Natural Language & Linguistic Theory* 2(1), 39–76.
- Kanerva, Jonni M. 1990. *Focus and Phrasing in Chichewa Phonology*. New York, Garland.
- 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*, 149–188. London, Academic Press.
- Kathol, Andreas, and Richard Rhodes. 2000. Constituency and linearization of Ojibwa nominals. MS. University of California Berkeley.
- Kimenyi, Alexandre. 1980. *A Relational Grammar of Kinyarwanda*. Berkeley, CA, University of California Press.
- King, Tracy Holloway. 1995. *Configuring Topic and Focus in Russian*. Stanford, CA, CSLI Publications.
- King, Tracy Holloway. 1997. Focus domains and information-structure. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG-97 Conference*, Stanford, CA. CSLI Publications. On-line at <http://www-csli.stanford.edu/publications>.
- King, Tracy Holloway, and Annie Zaenen. 2004. F-structures, information structure and discourse structure. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of LFG-04*. Stanford, CA, CSLI Publications. Extended abstract. <http://www-csli.stanford.edu/publications>.
- Lee, Chungmin. 1999a. Contrastive topic: A locus of the interface. In K. Turner (ed.), *The Semantics/Pragmatics Interface from Different Points of View*. Elsevier Science.
- Lee, Chungmin. 1999b. Topic, contrastive topic and focus: What's on our minds. plenary paper presented at the second international conference on cognitive science, Waseda, Tokyo.
- Morimoto, Yukiko. 2000. *Discourse Configurationality in Bantu Morphosyntax*. Doctoral dissertation, Stanford University, Stanford, CA.
- Morimoto, Yukiko. 2001. Verb raising and phrase structure variation in OT. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality-Theoretic Syntax*, 129–196. Stanford, CA, CSLI Publications.
- Ngonyani, Deogratias. 1998. Properties of applied objects in Kiswahili and Kindendeule. *Studies in African Linguistics* 27, 67–95.
- Payne, Doris L. 1990. *The Pragmatics of Word Order: Typological Dimensions of Verb Initial Languages*. Berlin/New York, Mouton de Gruyter.
- Payne, Doris L. (ed.). 1992. *Pragmatics of Word Order Flexibility*. Vol. 22 of *Typological Studies in Language*. Amsterdam/Philadelphia, John Benjamins.
- Sankoff, Gillian. 1993. Focus in Tok Pisin. In Francis Byrne and Donald Winford (eds.), *Focus and Grammatical Relations in Creole Languages*, Vol. 12 of *Creole Language Library*, 117–140. Amsterdam/Philadelphia, John Benjamins.
- Sells, Peter. 1998. Japanese postposing involves no movement. Paper presented at the meeting of the Linguistics Association of Great Britain, Lancaster, April 1998.
- Speas, Margaret. 1990. *Phrase Structure in Natural Languages*. Vol. 21 of *Studies in Natural Language and Linguistic Theory*. Dordrecht/Boston/London, Kluwer Academic Publishers.
- Takami, Ken-ichi (ed.). 1995. *Rightward Movement Constructions in English and Japanese*. Tokyo, Hituzi Syobo.
- Zaenen, Annie. 1985. *Extraction Rules in Icelandic*. New York, Garland.

NUMERALS, NOUNS AND NUMBER IN WELSH NPS

Ingo Mittendorf Louisa Sadler

University of Essex University of Essex

Proceedings of the LFG05 Conference

University of Bergen, Norway

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

<http://www-csli.stanford.edu/>

Abstract

Agreement mismatches in number, gender or case present an interesting challenge to any grammatical theory. We consider two styles of analysis for a number mismatch in Welsh, which arises when nouns are modified by cardinal numbers. In this construction, the nominal must be singular. One analysis pursues the idea that the nominal is an argument of the numeral, with some elements agreeing with the numeral and some with the noun. The other adopts the distinction between INDEX and CONCORD features, together with the proposal that in this construction it is the numeral which determines the plurality of the INDEX.¹

1 Introduction

Agreement mismatches in number, gender or case present an interesting challenge to any grammatical theory. We consider two styles of analysis for a number mismatch within NP arising when nouns are modified by cardinal numbers in Welsh. The paper is organised as follows. Section 2 lays out the data concerning noun phrase internal agreement including structures in which the head noun is modified by a numeral. Section 3 explores two possible analyses. The first of these, in section 3.1, treats the noun as an argument of the numeral (somewhat akin to a partitive construction). Section 3.2 then presents an alternative which makes use of the distinction between INDEX and CONCORD features. In section 4 we consider what further grounds there are for deciding between these alternative analyses for the Welsh data.

2 NP Internal Agreement

2.1 Basic Facts

Welsh distinguishes two grammatical genders (FEM, MASC) and two numbers (SG, PL). Number is inflectionally marked on the noun: plurals are formed by the addition of a suffix (of which there are several) and/or stem vowel modifications. Some illustrative examples are given in (1).²

¹The work reported on here was carried out in the project Verb Initial Grammars: a Multilingual, Parallel Approach: see <http://users.ox.ac.uk/~cpg10015/pargram/index.html>. We are grateful for the financial support of the ESRC (research grant RES-000-23-0505, to Dalrymple and Sadler) and also for comments and feedback from participants at LFG05, and in particular to Mary Dalrymple and Tracy Holloway King.

²Some Welsh nouns form the singular by suffixation from the plural: for example, PL *moch* 'pigs', SG *mochyn* 'pig'. Such nouns mostly denote animals or plants typically occurring in large groups. The plural in these cases often has a collective meaning, with the singular denoting a unit of the collective: PL, COLL *glaswellt* 'grass', SG, UNIT *glaswelltyn* 'blade of grass'. See (Thomas, 1996; King, 1993) for further details.

| | SG | PL | |
|-----|-----------------|---------------|----------|
| (1) | <i>afal</i> (M) | <i>afalau</i> | apple(s) |
| | <i>ceg</i> (F) | <i>cegau</i> | mouth(s) |
| | <i>ci</i> (M) | <i>cŵn</i> | dog(s) |
| | <i>cath</i> (F) | <i>cathod</i> | cat(s) |

Although some adjectives do still have plural forms in the modern language, as shown in (2), plural forms are most often found in fixed phrases (*mwyar duon* berry.PL black.PL ‘blackberries’). Beyond such phrases, the use of plural forms for adjectives, where they exist in the modern language, is not obligatory. The examples in (3) show the plural noun occurring with both plural and singular forms of the adjective *bychan* ‘small’.³

| | SG | PL | |
|-----|---------------|----------------|---------|
| (2) | <i>bychan</i> | <i>bychain</i> | ‘small’ |
| | <i>ifanc</i> | <i>ifainc</i> | ‘young’ |

- (3) a. *Gall busnesau bychain achosi llygredd am nifer o resymau.*
 can business.PL small.PL cause pollution for number from reasons
 Small businesses can cause pollution for a number of reasons.
- b. *Mae gwerthu digon i gynnal y busnes yn broblem fawr i fusnesau bychan*
 is sell enough to maintain the business PT problem big to
 business.PL small.SG
 Selling enough to maintain the business is a big problem for small businesses.

One exception is the adjective *arall* (pl. *eraill*) ‘other’ where the use of a plural form in agreement with a plural noun is obligatory.

(4) *merch arall*
 girl.F.SG other.SG
 another girl

(5) *merched eraill*
 girl.F.PL other.PL
 other girls

The vast majority of adjectives are not marked for gender, and only a very limited number of adjectives have distinct FEM forms. A representative selection of these is provided in (6).

| | MASC | FEM | | MASC | FEM | |
|-----|---------------|---------------|--------|-------------|-------------|--------|
| (6) | <i>gwyn</i> | <i>gwen</i> | white | <i>cryf</i> | <i>cref</i> | strong |
| | <i>melyn</i> | <i>melen</i> | yellow | <i>trwm</i> | <i>trom</i> | heavy |
| | <i>bychan</i> | <i>bechan</i> | small | <i>byr</i> | <i>ber</i> | short |

³Similarly, Google results for *fffermwyr ifainc/ifanc* (farmer.M.PL young.PL/SG) ‘young farmers’ shows that both occur quite regularly.

In addition, the use of these feminine forms is mostly optional and the “masculine” forms may appear with feminine nouns even when those particular adjectives have feminine forms.

(7) a. *cadair drom*
 chair.F.SG heavy.F.SG
 a heavy chair

b. *wythnos drwm*
 week.F.SG heavy.M.SG
 a heavy (busy) week

(8) a. *merch fer*
 girl.F.SG short.F.SG
 a short girl

b. *merch gryf*
 girl.F.SG strong.M.SG
 a strong girl

Demonstratives agree in GEN and NUM and follow any postnominal adjectives. As the examples below show, demonstratives require the presence of the definite article.

(9)

| | PROX | DIST |
|------|------------|---------------|
| M.SG | <i>hwn</i> | <i>hwennw</i> |
| F.SG | <i>hon</i> | <i>honno</i> |
| PL | <i>hyn</i> | <i>hynny</i> |

(10) *y ci hwn*
 the dog.M.SG this.M.SG
 this dog

(11) *y cathod hynny*
 the cat.F.PL that.PL
 those cats

Finally, most numerals are invariant, but the lower numerals (sometimes referred to as paucal numbers) have distinct FEM, MASC forms, as shown below:

(12)

| MASC | FEM | |
|---------------|---------------|---|
| <i>dau</i> | <i>dwy</i> | 2 |
| <i>tri</i> | <i>tair</i> | 3 |
| <i>pedwar</i> | <i>pedair</i> | 4 |

A particularly salient feature of Welsh and the other Celtic languages is the system of initial consonant mutations, or phonological alternations of the initial phoneme of a word. Welsh has three sets of mutations (see the table in (15)) in addition to the citation or radical form. Mutations are triggered by a variety of lexical and syntactic triggering environments. One such is that adjectives following a F.SG noun appear in soft mutated form. In (13), the adjective (which is not itself a feminine form) has undergone soft mutation (*mawr* → *fawr*) after the F.SG noun:

mawr in (14) is not mutated because the noun is plural rather than singular.⁴ The radical form is used after all other nouns.

- (13) *torth fawr (< mawr)*
 loaf.F.SG big.SG
 a big loaf
- (14) *torthau mawr*
 loaf.F.PL big.SG
 big loaves

| | Radical | Soft Mut. | Nasal Mut. | Aspirate Mut |
|------|---------|-----------|------------------|------------------|
| | c | g | ng | ch |
| | p | b | mh | ph |
| | t | d | nh | th |
| (15) | g | ∅ | ng | [= <i>Rad</i>] |
| | b | f | m | [= <i>Rad</i>] |
| | d | dd | n | [= <i>Rad</i>] |
| | m | f | [= <i>Rad</i>] | [= <i>Rad</i>] |
| | ll | l | [= <i>Rad</i>] | [= <i>Rad</i>] |
| | rh | r | [= <i>Rad</i>] | [= <i>Rad</i>] |

In a series of adjectives, each one undergoes soft mutation (SM). Note also that although both adjectives in (16) (*byr* ‘short’, *tywyll* ‘dark’) have feminine forms (*ber*, *tywell*), only the first of them occurs in this form, the second occurring in the (generalised) masculine form.⁵

- (16) *merch fer, dywyll (ber, tywyll)*
 girl.F.SG short.F.SG dark.M
 a short dark girl (Thorne: 134)
- (17) *cyfres fer flasus (ber, blasus)*
 series.F.SG short.F.SG interesting
 a short interesting series

Feminine singular nouns are also distinguished by the fact that they appear in soft-mutated form immediately following the definite article *y*⁶:

| | | | |
|------|-------------------|---------------------|------------|
| (18) | <i>bardd</i> (m) | <i>y bardd</i> (m) | bard, poet |
| | <i>baner</i> (f) | <i>y faner</i> (f) | flag |
| | <i>ci</i> (m) | <i>y ci</i> (m) | dog |
| | <i>cath</i> (f) | <i>y gath</i> (f) | cat |
| | <i>cŵn</i> (m) | <i>y cŵn</i> (m) | dogs |
| | <i>cathod</i> (f) | <i>y cathod</i> (f) | cats |

⁴*mawr*, *fawr* are glossed as SG because a plural form, *mawrion* does exist, though its use is infrequent.

⁵Again, the glossing reflects that fact that *byr/ber* has a plural form while *tywyll/tywell* does not.

⁶The definite article *y* has the (purely positional) variants *yr* and *'r*. It does not itself distinguish either number or gender.

In terms of basic agreement facts, then, the Welsh NP appears to be quite straightforward, with demonstratives, nouns and adjectives co-specifying constraints over the GEN and NUM features of the f-structure of the NP. For example, FEM.PL forms of adjectives (where they exist) require the f-structure of the NP to be specified as feminine plural:

- (19) *Adj_{fem}* ((ADJ ∈ ↑) GEND) =_c FEM
 ((ADJ ∈ ↑) NUM) =_c PL

2.2 Combining Numeral and Noun

The situation is more complex when nouns are modified by a cardinal numeral. Nouns following a numeral are obligatorily singular:

- (20) *pum ci*
 five dog.M.SG
 five dogs

- (21) *tair cath*
 three.F cat.F.SG
 three cats

Moreover adjectives agree with the noun in number, so that, for example, *arall* ‘other’, an adjective that obligatorily agrees with a nominal in number, must be singular in the presence of a numeral. Hence we see the following contrast:

- (22) *cŵn eraill*
 dog.M.PL other.PL
 other dogs

- (23) *pum ci arall*
 five dog.M.SG other.SG
 five other dogs

By contrast, demonstratives are always plural in the presence of a noun with a numeral premodifier, as illustrated by the examples below. Note that in these examples, though the head noun is singular in both, the demonstrative is singular in (24) and plural in (25).

- (24) *y gath hon*
 the cat.F.SG this.F.SG
 this cat

- (25) *y tair cath hyn*
 the three cat.F.SG this.PL
 these three cats

In (24) *cath* is soft-mutated (to *gath*) after the definite article because the article is followed by a F.SG NP. In contrast, *tair* in (25) does not show soft mutation (which would give the form *dair*). Recall that the definite article causes soft mutation of the following word only when the NP is feminine singular: the lack of soft mutation in (25) therefore indicates that an NP following the article qualifies as PL, from the perspective of the definite determiner! To recap so far: when a noun is premodified by a numeral, the noun and any adjectives are singular, but the demonstrative and determiner are plural.

Externally, the NP behaves as a plural in terms of other agreement processes which it controls, for example, pronominal anaphora.⁷

- (26) *Roedd y pum dyn yn gweld eu hunain yn y drych.*
 be.IMPERF.3S the five man.M.SG PT see 3PL self in the mirror
 The five men saw/were seeing themselves in the mirror.

- (27) *Cafodd y pum ci eu curo.*
 get.IMPERF.3S the five dog.M.SG 3PL beat
 The five dogs were beaten.

3 Analysis

3.1 Two-Tier F-structure

One possible approach to the agreement data illustrated in section 2.2 above involves taking the numeral as the head of the construction. On this view, the numeral sub-categorises for a singular nominal complement: adjectives are in construction with this complement and show singular agreement. Demonstratives and determiners, on the other hand, are in construction with the numeral phrase and agree with the (inherent plurality of) the numeral.

A sentence such as (28) would have the f-structure shown in (29): *y* and *hynny* contribute features to the SPEC function of the numeral while the AP provides an ADJ function to the f-structure of the nominal (for present purposes, we take the single argument of the numeral to be an OBJ without further discussion).

- (28) *y tair cath ddu hynny*
 the.PL three.F cat.F.SG black.SG that.PL
 those three black cats

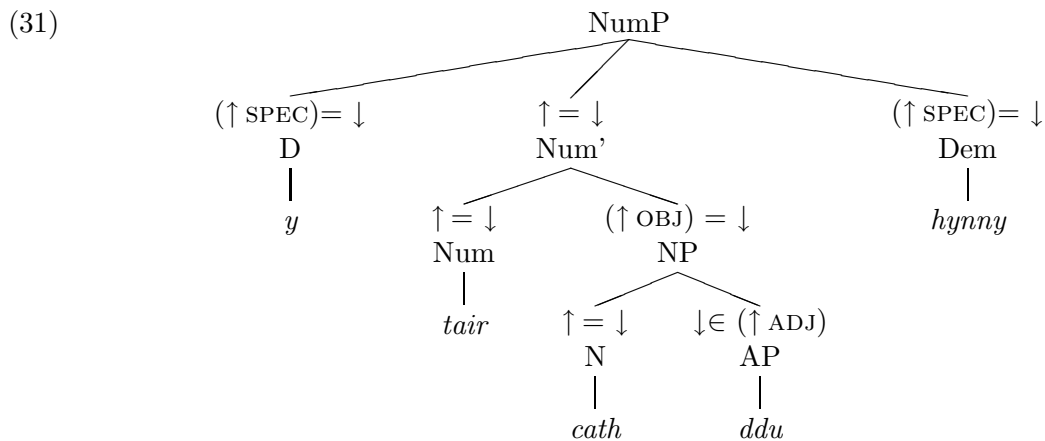
⁷In the personal passive construction illustrated in (27) a clitic pronoun occurs before the non-finite main verb (*curo*) coding the passive subject. This pronoun is plural, in agreement with *y pum ci* ‘the five dogs’.

$$(29) \left[\begin{array}{l} \text{PRED} \quad \text{'THREE<OBJ>'} \\ \text{NUM} \quad \text{PL} \\ \text{SPEC} \quad \left[\begin{array}{ll} \text{DEIX} & \text{DIST} \\ \text{DEF} & + \end{array} \right] \\ \text{OBJ} \quad \left[\begin{array}{ll} \text{PRED} & \text{CAT} \\ \text{NUM} & \text{SG} \\ \text{GEN} & \text{FEM} \\ \text{ADJ} & \{ [\text{PRED} \quad \text{'BLACK'}] \} \end{array} \right] \end{array} \right]$$

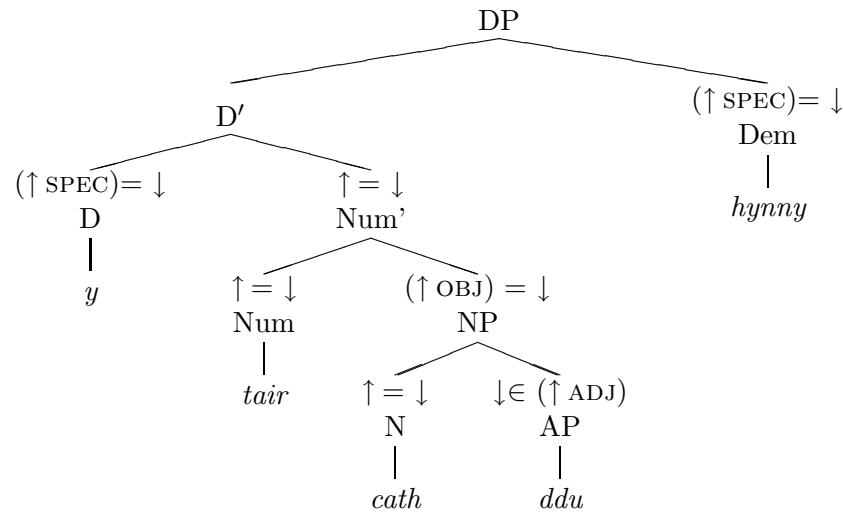
The numeral would define the number of its f-structure as plural, and require its argument, here taken to be an object, to be singular in number. As shown in (30), the lower numbers, that is, those which distinguish gender, also require their argument to be in the same gender as they themselves are.

$$(30) \begin{array}{l} \text{tair} \quad (\uparrow \text{ PRED}) = \text{'THREE<OBJ>'} \\ (\uparrow \text{ NUM}) = \text{PL} \\ (\uparrow \text{ OBJ GEN}) =_c \text{FEM} \\ (\uparrow \text{ OBJ NUM}) =_c \text{SG} \end{array}$$

Determining the precise c-structure is not our main concern here (see Sadler (2003); Willis (to appear) for discussion of the structure of Welsh NPs). The data discussed here might motivate a c-structure along the lines of (31) or (32), in which D is a projecting (functional) category, and the demonstrative is a structural specifier of the D head.



(32)



This analysis captures the agreement facts adequately. It essentially equates the construction in (20)-(25) with some sort of partitive. Welsh does in fact also have a partitive numeral construction, as shown in (33), but it is not clear whether the existence of this construction has any bearing on the plausibility of the two-tier analysis for the bare numeral-noun construction under discussion here. Note that in the partitive construction the nominal is plural. In many contexts, there seems to be little semantic difference between the numeral-noun and the partitive numeral construction.

- (33) *y tri o ddynion*
the three.M of man.M.PL
the three men

3.2 Index vs. Concord

3.2.1 Index and Concord Agreement

Various constraint-based approaches to syntactic agreement propose a distinction between two sets of agreement features within the NP. In HPSG Wechsler and Zlatić (2000) distinguish between a set of CONCORD features and a set of INDEX features, the former being more closely related to morphological (inflectional) classes and the latter to the semantics (and see also Kathol (1999) for an earlier and related proposal). For Wechsler and Zlatić, CONCORD features are typically relevant for NP-internal concord (between nouns, determiners and adjectives), while INDEX features are typically relevant to subject-verb agreement and for pronominal anaphora. In LFG, King and Dalrymple (2004) make a related but nonetheless distinct proposal to distinguish between CONCORD and INDEX features associated with nominal f-structures. The key difference for King and Dalrymple is that INDEX is a non-distributive feature while CONCORD is a distributive feature. Given an f-structure which is a set (for example, in the case of a coordinate structure), a distributive

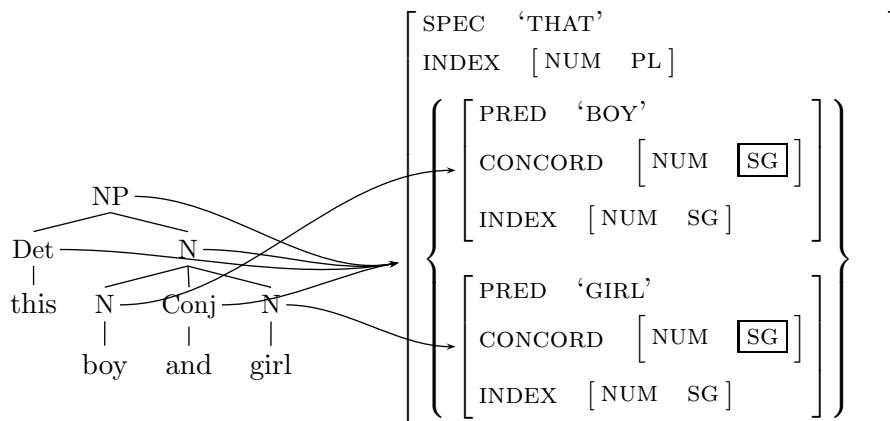
feature holding of this set will hold for every member of the set, whereas a non-distributive feature will hold of the set as a whole.

- (34) 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.
 Dalrymple and Kaplan (2000)

King and Dalrymple observe that languages differ as to whether they permit singular and/or plural determiners to combine with a coordination of singular nouns. The English demonstrative, for example, occurs in the singular with a coordination of singular nouns, indicating CONCORD agreement, as shown in (36), whereas subject-verb agreement in English accesses the INDEX feature (which in the case of a coordinate structure represents the resolution of the INDEX features of the coordinate daughters).

- (35) a. *that/*those boy and girl*
 b. *That boy and girl *is/are my friends.*

- (36) that: (\uparrow CONCORD NUM) = SG



In Russian, on the other hand, the plural determiner *éti* appears to require a plural INDEX and so can co-occur with a plural noun, a coordination of plural nouns, and, crucially, a coordination of singular nouns (37). It is this latter case which indicates that what is at issue here is INDEX agreement.

- (37) *éti mužčina i ženščina*
 these-PL man-M.SG and woman-F.SG
 this man and woman (Russian: King and Dalrymple 2004:95)

3.2.2 Numerals and Nouns: An Index/Concord Mismatch

The distinction between INDEX and CONCORD offers an alternative approach to the Welsh data, in which the f-structure of the entire phrase is headed not by the numeral but by the noun, and the numeral is treated as a modifier of the head noun. Under this analysis the f-structures of an adjectivally modified nominal phrase and a numerally modified nominal phrase will be (structurally) similar, differing only in their agreement features.

The fundamental idea is that numeral-noun combinations in Welsh project an f-structure (for the NP) with an INDEX-CONCORD mismatch: the numeral modifier contributes the INDEX NUM and the noun contributes the CONCORD NUM.

Consider first the f-structure of a simple NP such as (38).⁸

| | | | | | | | |
|------|-----------------|--------------|------------|---|-------|--------------------|---|
| (38) | <i>hen ddyn</i> | old man.M.SG | an old man | [| PRED | ‘MAN’ |] |
| | | | | | INDEX | [NUM SG] | |
| | | | | | CONC | [NUM SG] | |
| | | | | | ADJ | { [PRED ‘OLD’] } | |

Here the noun specifies a SG number value for INDEX and CONCORD and the (invariant) adjective *hen* places no GENDER or NUMBER restrictions. The contribution of the SG noun in this context is:

| | | | |
|------|------|-------------|----|
| (39) | N.SG | CONCORD NUM | SG |
| | | INDEX NUM | SG |

Similarly, a plural noun will specify a PL value for INDEX and CONCORD.

| | | | |
|------|------|-------------|----|
| (40) | N.PL | CONCORD NUM | PL |
| | | INDEX NUM | PL |

The f-structure of an NP containing a numeral modifier will be structurally similar, but we assume that it will differ in the values for INDEX and CONCORD NUM (see (41)). Recall that cardinal numerals occur with (obligatorily) singular nouns, but we know that the INDEX of the NP overall is PL. This is shown by the plural pronominal anaphors in (26) and (27) above.

| | | | | | | | |
|------|----------------|------------------|-----------|---|-------|----------------------|---|
| (41) | <i>tri dyn</i> | three.M man.M.SG | three men | [| PRED | ‘MAN’ |] |
| | | | | | INDEX | [NUM PL] | |
| | | | | | CONC | [NUM SG] | |
| | | | | | ADJ | { [PRED ‘THREE’] } | |

The respective contributions of the SG noun and the numeral in this context are as follows:

⁸The form *ddyn* in (38) is the soft mutated form of *dyn*. The mutation is triggered by the preceding adjective *hen* ‘old’, which belongs to a limited number of adjectives regularly occurring in prenominal position, most of which cause soft-mutation of the following noun.

(42)

| | | |
|------|----------|----|
| N.SG | CONC NUM | SG |
| Num | IND NUM | PL |

Thus, numerals assign a (plural) INDEX NUM value to the containing f-structure, and singular forms of nouns assign a (singular) CONCORD NUM value.⁹ At the same time, numerals also require this CONCORD NUM value to be SG (because numerals must be followed by singular nouns). For a singular noun form this means that while its CONCORD NUM value is fixed to SG, the INDEX NUM value is SG only by default. This is captured by the lexical entries along the following lines for cardinal numerals and (singular and plural) nouns:¹⁰

(43) *tri* (↑ PRED) = ‘THREE’
 ((ADJ ∈ ↑) IND NUM) = PL
 ((ADJ ∈ ↑) CONC NUM) =_c SG

(44) *dyn* (↑ PRED) = ‘MAN’
 (↑ CONC NUM) = SG
 {(↑ IND NUM) = SG | (↑ IND NUM) =_c PL}

(45) *dynion* (↑ PRED) = ‘MAN’
 (↑ CONC NUM) = PL
 (↑ IND NUM) = PL

3.2.3 NP Internal Agreement

Turning now to NP internal agreement, recall that determiners and demonstratives are plural (thus appearing to agree with the inherent plurality of the numeral) while adjectives are obligatorily singular in this construction, and hence appear to agree with the noun. This now has a straightforward explanation: determiners and demonstratives show INDEX agreement while adjectives show CONCORD agreement. This agreement type selection, one may be justified to assume, not only holds in the presence of a numeral, but quite generally. However, since ordinarily CONCORD NUM and INDEX NUM have the same values, the difference in agreement controllers is usually unnoticeable and becomes visible only where an NP contains a numeral. (46) will further illustrate the various constraints in operation.

(46) *y tair cath ddu arall hynny*
 the.PL three.F cat.F.SG black.SG other.SG that.PL
 those other three black cats

⁹The numeral *un* ‘one’ of course assigns both singular CONCORD NUM and INDEX NUM.

¹⁰Additionally, the numerals 2, 3 and 4, but not other numerals, specify GEN constraints.

The combination of the numeral and the noun will together ensure that the f-structure of the NP (headed by *cath*) has the following agreement features:

$$(47) \left[\begin{array}{l} \text{PRED} \quad \text{'CAT'} \\ \text{INDEX} \quad [\text{NUM} \quad \text{PL}] \\ \text{CONC} \quad [\text{NUM} \quad \text{SG}] \end{array} \right]$$

This NP f-structure, with mismatched NUM values, only ever arises in the presence of a numeral.

The demonstrative shows INDEX agreement, and thus only a plural form is grammatical when a numeral is present:¹¹

$$(48) \quad y \quad \textit{tair} \quad \textit{cath} \quad \textit{*honno/hynny}$$

the.PL three.F cat.F.SG that.F.SG/that.PL

those three cats

$$(49) \quad \textit{hynny} \quad (\uparrow \text{ DEIX}) = \text{DIST}$$

$$\quad \quad \quad ((\text{SPEC } \uparrow) \text{ INDEX NUM}) =_c \text{ PL}$$

The determiner *y* also selects INDEX NUM agreement. Recall that although the determiner *y* itself is invariant (apart from its positional variants *yr*, *'r*), it requires the next word to be soft mutated if the NP is F.SG, and is followed by the radical if the NP is M.SG or (M or F) PL.¹² The fact that *tair* is not mutated (to *dair*) in (46) indicates that from the perspective of the determiner, the NP is not F.SG but PL (since it cannot be M.SG, *cath* being feminine). The *y* which is followed by the radical has the following entry:

$$(50) \quad y \quad (\uparrow \text{ DEF}) = +$$

$$\quad \quad \quad \{ ((\text{SPEC } \uparrow) \text{ INDEX NUM}) =_c \text{ PL} \mid$$

$$\quad \quad \quad ((\text{SPEC } \uparrow) \text{ INDEX NUM}) =_c \text{ SG}$$

$$\quad \quad \quad ((\text{SPEC } \uparrow) \text{ INDEX GEN}) =_c \text{ M} \}$$

Turning now to adjectives, recall that only the adjective *arall* (plural *eraill*) ‘other’ obligatorily shows number agreement. Most other adjectives lack plural forms and even where such forms do exist, their use (in plural contexts) is not obligatory.¹³ The numeral and adjectives in (46) have the following lexical descriptions.¹⁴ The

¹¹For concreteness, we assume both Dem and Det nodes in the f-structure are annotated ($\uparrow \text{SPEC}$) = \downarrow , and thus demonstratives and determiners specify constraints over the containing f-structure (that of the NP) by means of inside-out equations.

¹²As may be clear from the data in this paper, there is virtually no gender distinction in the plural in Welsh.

¹³Plural forms are, of course, restricted to plural contexts. A similar distributional pattern is found for existing FEM gender forms.

¹⁴The adjective *ddu* ‘black’ in (46) is in soft mutated form ($d \rightarrow dd$) following a F.SG noun.

entry (53) defines the INDEX NUM value for the f-structure of the NP to be PL and constrains the CONCORD NUM value to be SG (and thus combines only with a (feminine) singular noun).

- (51) *du* (\uparrow PRED) = 'BLACK'
 (52) *arall* (\uparrow PRED) = 'OTHER'
 ((ADJ \in \uparrow) CONCORD NUM) =_c SG
 (53) *tair* (\uparrow PRED) = 'THREE'
 ((ADJ \in \uparrow) INDEX NUM) = PL
 ((ADJ \in \uparrow) CONCORD NUM) =_c SG
 ((ADJ \in \uparrow) CONCORD GEN) =_c FEM

These lexical entries, together with that for *cath* 'cat.F.SG', will define the following f-structure:

$$(54) \left[\begin{array}{l} \text{PRED} \quad \text{'CAT'} \\ \text{INDEX} \quad [\text{NUM} \quad \text{PL}] \\ \text{CONC} \quad \left[\begin{array}{l} \text{NUM} \quad \text{SG} \\ \text{GEN} \quad \text{FEM} \end{array} \right] \\ \text{SPEC} \quad \left[\begin{array}{l} \text{DEIX} \quad \text{DIST} \\ \text{DEF} \quad + \end{array} \right] \\ \text{ADJ} \quad \left\{ \begin{array}{l} [\text{PRED} \quad \text{'BLACK'}] \\ [\text{PRED} \quad \text{'THREE'}] \end{array} \right\} \end{array} \right]$$

It should be clear that the treatment rules out the occurrence of plural adjectives with the numeral-noun combination: such adjectives constrain the CONCORD NUM feature of the f-structure of the NP to be PL, a constraint that is not satisfied in a structure such as (47).

4 Discussion

We have presented two different approaches to the number mismatch arising in Welsh NPs containing a numeral. Both analyses seem independently viable for this set of data, but they are based on quite different intuitions. The question is whether there are any reasons for preferring one style of analysis over the other.

4.1 Other Mismatches

The INDEX vs CONCORD analysis has the advantage that the distinction between these two sets of features may also be motivated by other number mismatches unconnected with the numerals problem and may therefore be independently needed

in the grammar. For example, it has some potential for capturing the behaviour of the exceptional, idiosyncratic noun *pobl*, which in both its meanings ‘people, nation’ and ‘group of people’ behaves partly as a singular and partly as a plural.

In terms of its morphology, *pobl* is a singular form – the plural form is *pobloedd*, and it is also idiosyncratic. In terms of syntax, the form *pobl* behaves partly as a singular form, and partly as a plural form. In (55), its singular behaviour can be seen in the fact that it has undergone soft mutation (*pobl* > *bobl*) after the definite article (this is characteristic of FEM.SG; its plural behaviour is signalled by the fact that it occurs with a plural demonstrative. The 3PL verb form in (56) (*gwelsant*) also indicates that the subject is plural from the perspective of pronominal anaphora.¹⁵

(55) *y bobl hyn*
 the.F.SG people this.PL
 these people

(56) *Cododd y bobl a gwelsant....*
 rose.3SG the people and saw.3PL
 The people rose up and saw

4.2 A Special Noun

The CONCORD/INDEX analysis treats singular nouns being associated with a SG INDEX only by default and posits a disjunction to capture this fact:

(57) $\{(\uparrow \text{IND NUM}) = \text{SG} \mid (\uparrow \text{IND NUM}) =_c \text{PL}\}$

The intuition that this seeks to capture is that SG nouns are singular except when they are in construction with numerals. The intuition behind the two-tier analysis is rather different, namely that the numeral-noun construction is rather like a partitive, with the noun corresponding to an argument of the numeral. On this analysis, the NUM of the f-structure of the nominal is SG and that of the numeral is PL.

It is interesting to note that there is one noun in Modern Welsh, *blwyddyn* ‘year’, which has not two number forms, but three: *blwyddyn*, *blynnyddoedd*, *blynedd*. The third form, *blynedd*, is a special form used only in combination with numerals (except ‘one’), while the singular *blwyddyn* is barred from this environment:

(58) *Mae'r grantiau ar gael am un flwyddyn yn unig....*
 is-the grants on get for one year only
 The grants are available for only one year.

¹⁵Finite verbs in Welsh show full agreement only with pronominal subjects (which can be dropped). Full lexical NPs occur with 3SG finite verbs - hence the verb *cododd* in (56) itself tells us nothing about the number of the subject.

(59) *am y tair neu bedair blynedd nesa*
 for the three.F or four.F year.F next
 for the next three or four years

(60) *y blynyddoedd cynnar yn yr ysgol*
 the years early in the school
 the early years in school

If we follow the CONCORD/INDEX analysis, these three forms must have the lexical descriptions in (61) - (63). Note that the first two correspond to the one disjunctive singular entry for ordinary nouns. The selection of the correct form in the presence or absence of a numeral simply falls out naturally on this view.

(61) *blwyddyn* (↑ PRED) = ‘YEAR’
 (↑ CONC NUM) = SG
 (↑ IND NUM) = SG

(62) *blynedd* (↑ PRED) = ‘YEAR’
 (↑ CONC NUM) = SG
 (↑ IND NUM) =_c PL

(63) *blynyddoedd* (↑ PRED) = ‘YEAR’
 (↑ CONC NUM) = PL
 (↑ IND NUM) = PL

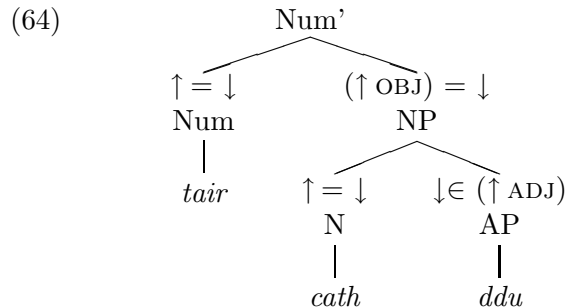
If we follow the two-tier analysis, both *blwyddyn* and *blynedd* would be singular forms. In order to prevent *blynedd* occurring without a numeral, and *blwyddyn* occurring with a numeral, it would be necessary to postulate an additional (book-keeping) feature in the f-structure of the numeral (for example, PRED-TYPE = NUMERAL), and add inside-out statements to the lexical descriptions of *blynedd* and *blwyddyn*, controlling for the presence (or absence) of that feature in the containing f-structure. Thus it seems that the INDEX/CONCORD analysis extends more gracefully to this additional data.

4.3 Further Nominal Structures

Although our main concern here is not with matters of c-structure, we note that some potential difficulties for the two-tier analysis might arise in more complex NPs.

The possible c-structures associated with the two-tier analysis in section 3.1 involved a constituent structure with the numeral as sister to the (possibly adjectivally modified) NP which corresponds to the OBJ of the numeral (see the relevant sub-tree, repeated in (64)). A crucial property of these structures was that the demonstrative was outside of the NP OBJ, consistent with the fact that it follows any adjectival

modifiers of the noun. This NP-external position crucially ensures that, when a numeral is present, the demonstrative contributes its agreement constraints to the f-structure of the numeral rather than to the f-structure of the noun:



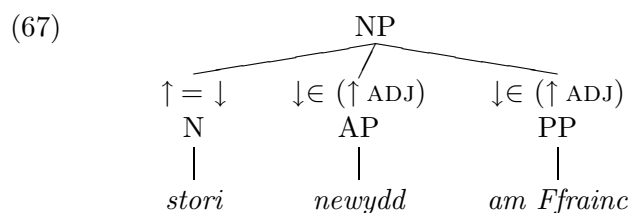
Postnominal adjectival modifiers in Welsh NPs intervene between the head N and any complements or adjuncts:

- (65) *stori newydd am Ffrainc*
 story new about France
 a new story about France

They also precede the NP possessor, which in turn precedes any complements or adjuncts:

- (66) *disgrifiad manwl y gyrrwr o'r ddamwain*
 description detailed the driver of-the accident
 the driver's description of the accident (Rouveret 1991: 193)

These can be accommodated by adopting either a hierarchical or a flat structure within NP:¹⁶



However, although the demonstrative always occurs after any postnominal adjectives, it is not always final in the NP. Willis (to appear) gives the order of elements within the NP as in (68).

- (68) Det - Num - N - Adj - Poss/Dem - Compl/Adjuncts

¹⁶For a treatment using a hierarchical approach to NP-internal structure, see Sadler (2003).

Crucially, the demonstrative may be followed by any complements or adjuncts of the head noun.

- (69) *y stori hon am Ffrainc*
the story.F.SG this.F.SG about France
this story about France
- (70) *y disgrifiad manwl hwn o'r ddamwain*
the description.M.SG careful this.M.SG of-the accident
this careful description of the accident
- (71) *yn y llythyr anghyflawn hwn o Ffrainc*
in the letter.M.SG incomplete this.M.SG from France
in this incomplete letter from France

Examples such as these appear quite problematic for the two-tier approach of section 3.1. Consider (71), where the demonstrative intervenes between AP and PP adjuncts. If this position is taken to indicate that a demonstrative attaches within NP, then the two-tier analysis, whereby the demonstrative agrees with the (plurality of the) numeral, founders, because the demonstrative would then (incorrectly) reflect the number of the noun (rather than the numeral). Alternatively, if this data is taken to indicate that complements and adjuncts of the noun are extraposed out of NP and under a higher projection (DP or NumP), then they will have to be associated with a disjunctive f-description to reflect the potential presence of the numeral. Neither of these routes is appealing.

4.4 Conclusion

While it seems clear that the two-tier analysis will certainly apply in a natural way to numeral-noun constructions in languages in which a numeral governs an object (or oblique) and assigns a case to it (genitive, for instance), our current view is that it is less appealing for the Welsh data which we have considered in this paper. In the PARGRAM Welsh grammar, we therefore currently implement the INDEX/CONCORD analysis of these constructions. In future work we hope to be able to provide a more substantial exposition of the role of the INDEX/CONCORD distinction in the grammar of Welsh.

References

- Dalrymple, Mary and Ronald M. Kaplan. 2000. Feature indeterminacy and feature resolution. *Language* 76(4):759–798.
- Kathol, Andreas. 1999. Agreement and the syntax-morphology interface in HPSG. In R. Levine and G. Green, eds., *Studies in Current Phrase Structure Grammar*, pages 223–274. Cambridge, UK: Cambridge University Press.

- King, Gareth. 1993. *Modern Welsh*. London: Routledge.
- King, Tracy Holloway and Mary Dalrymple. 2004. Determiner agreement and noun conjunction. *Journal of Linguistics* 40(1):69–104.
- Sadler, Louisa. 2003. Noun Phrase structure in Welsh. In M. Butt and T. H. King, eds., *Argument Realization*, pages 73–110. Stanford, CA: CSLI Publications.
- Thomas, Peter Wynn. 1996. *Gramadeg y Gymraeg*. Caerdydd: Gwasg Prifysgol Cymru.
- Thorne, David A. 1993. *A Comprehensive Welsh Grammar*. Oxford: Blackwells.
- Wechsler, Stephen and Larisa Zlatić. 2000. A theory of agreement and its application to Serbo-Croatian. *Language* 76(4):759–798.
- Willis, David. to appear. Against N-raising and NP-raising analyses of Welsh noun phrases. *Lingua* .

‘WH’-IN-SITU IN CONSTITUENT QUESTIONS

Louise Mycock

University of Manchester

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

All languages have constructions which enable speakers to ask constituent ('wh'-) questions. While cross-linguistically question formation strategies differ, these strategies may share certain features. One such feature is 'wh'-in-situ, the appearance of a question word in a position associated not with syntactic focusing but with a non-question word bearing the same grammatical function. This paper discusses 'wh'-in-situ as it is found in English and Japanese. By fully exploiting the architecture of Lexical-Functional Grammar with its distinct yet parallel projections 'wh'-in-situ is captured by a single non-derivational analysis, providing the basis for future analysis of constituent questions cross-linguistically.

1 Introduction¹

In terms of the traditional typology of constituent question formation strategies, English is a simple-fronting language and Japanese is a 'wh'-in-situ language.²

In a neutral Japanese constituent question, all question words appear in situ.³ This means that the word order of a neutral declarative sentence and a comparable constituent question will be identical. This is true regardless of whether the constituent question contains a single question word (1b) or multiple question words (2).⁴

- (1) a. Mari-ga depāto-de ojōsan-ni ranpu-o eranda.
Mari-NOM dept.store-LOC daughter-DAT lamp-ACC choose.PAST
SUBJ ADJ OBL_{BEN} OBJ VERB
'Mari chose a lamp for her daughter at the department store.'
- b. SINGLE CONSTITUENT QUESTION
Mari-ga depāto-de *dare*-ni ranpu-o eranda ka.
Mari-NOM dept.store-LOC who-DAT lamp-ACC choose.PAST Q
SUBJ ADJ OBL_{BEN} OBJ VERB
'Who did Mari chose a lamp for at the department store?'

- (2) MULTIPLE CONSTITUENT QUESTION
Dare-ga depāto-de *dare*-ni ranpu-o eranda ka.
who-NOM dept.store-LOC who-DAT lamp-ACC choose.PAST Q
SUBJ ADJ OBL_{BEN} OBJ VERB
'Who chose a lamp for who at the department store?'

In English when there is a single question word in a constituent question, the question word appears clause-initially regardless of its grammatical function. The result of this syntactic focusing is a different word order from that found in a comparable declarative sentence.

¹ I would like to thank Mimi Nakajima, Juri Saito and Yasue Toda for their insights and patience. This work is supported by funding from The Arts and Humanities Research Council.

² A distinction is often made between two sub-types of in-situ language usually exemplified by Mandarin Chinese and Japanese. It has been argued that Subjacency effects can be detected in one (Japanese), but not in the other (Mandarin Chinese.) See, for example, Watanabe (2001).

³ The issue of the syntactic operation of scrambling is set aside here as it occurs in addition to constituent question formation and does not appear to be in itself a question formation strategy.

⁴ Examples of 'wh'-in-situ appear in italics throughout.

- (3) a. Anna offered Lily oranges.
 SUBJ VERB OBL_{GOAL} OBJ
 b. SINGLE CONSTITUENT QUESTION
 What does Anna offer Lily?
 OBJ AUX SUBJ VERB OBL_{GOAL}

Only English multiple constituent questions contain examples of ‘wh’-in-situ.⁵

- (4) MULTIPLE CONSTITUENT QUESTION
 What does Anna offer *who*?
 OBJ AUX SUBJ VERB OBL_{GOAL}

The key issues which an analysis of ‘wh’-in-situ cross-linguistically must address are to an extent dependent on the theoretical framework adopted. For those working within a derivational framework, the question is how the notion of movement can satisfactorily explain constituent question formation which seemingly does not involve displacement. For researchers using the non-derivational framework of Lexical-Functional Grammar (LFG), the main points to consider are the apparent lack of both functional uncertainty and focusing of question words in cases of ‘wh’-in-situ.

In LFG, a statement of equivalence between the discourse function FOCUS and an argument or adjunct function is associated with the dislocated position occupied by a question word in a simple-fronting language such as English. The inside-out version of this functional uncertainty has been used to analyse the ‘wh’-in-situ language Mandarin Chinese (Huang, 1993). However, the defining characteristic of a ‘wh’-in-situ question word/phrase is the fact that it is quite clear which grammatical function it bears. I therefore contend that while functional uncertainty is appropriate for assignment of FOCUS when it has an identifiable c-structure correlate, that is when it is grammaticalized, a unified cross-linguistic analysis of ‘wh’-in-situ should not be based on this notion.

Viewed purely in syntactic terms, there is no evidence of focusing in cases of ‘wh’-in-situ. This is puzzling given that the constituent question formation strategy employed in simple-fronting languages involves (syntactic) focusing of one question word, while in multiple-fronting languages such as Russian all question words are (syntactically) focused. However, relevant Japanese and English data confirm that all question words are indeed focused. Though ‘wh’-in-situ question phrases are not syntactically focused, they are subject to prosodic focusing. When structural levels other than syntax are considered, it is therefore possible to define ‘wh’-in-situ more accurately.

One issue which is important to any analysis of constituent questions irrespective of the theoretical framework adopted though is a semantics for interrogatives.

2 Semantics for constituent questions

I adopt the semantics for interrogatives which Ginzburg & Sag (2000) propose in their HPSG analysis of English questions.

Ginzburg & Sag (2000) treat questions as ‘open propositions’, formally characterized as propositional abstracts.⁶ A propositional abstract, ‘constructed from’ a proposition, is a semantic object in its own right according to Ginzburg & Sag (2000). This is because the semantic universe assumed is one in

⁵ The issue of so-called ‘echo’ questions which include a single in-situ question word such as *Anna offered Lily what?* is beyond the scope of this work.

⁶ This is not a new idea. Ginzburg & Sag (2000) trace it back to Jespersen (1924) and Cohen (1929).

which abstracts have an ontological status comparable with that of ‘ordinary’ individuals. Abstraction is therefore a semantic operation along the lines of substitution, the ‘output’ of which is a member of the ontology as basic as any other semantic object.

Ginzburg & Sag (2000) claim that a question word makes a very specific, two-fold contribution to the meaning of a constituent question. First, it enables an abstraction to occur over the parameter which a question word associates with the argument role that it fills. Each parameter has an index value linking it to an argument in the body of the abstraction. In this sense, a question word functions as a place-holder.

- (5) ‘What does Anna offer Lily?’
 $\lambda\{x\}.\text{offer}(\text{anna}, x, \text{lily})$

Second, a question word introduces certain restrictions on the role-filler which can bear a particular argument role instead of a place-holder. For example, *who* requires the substitution of a human entity for a place-holder, while *what* requires the substitution of a non-human entity for a place-holder. The preliminary semantic representation in (5) must therefore be revised.

- (6) ‘What does Anna offer Lily?’
 $\lambda\{x\}.\text{offer}(\text{anna}, (x, \text{non-human}(x)), \text{lily})$

A parameter is a member of a set of restriction-bearing elements that links an abstracted argument to an argument position within the proposition from which a propositional abstract is ‘constructed’. The parameter set corresponds to the set of entities that gets abstracted away, that is the set of parameters in a parametric object. The scope of a parameter set is a proposition so, in the case of a question, the parametric object involved (the body of the abstraction) is a proposition containing place-holders linked to each parameter. A parameter set will have multiple members if the question is a multiple constituent question. This means that question words do not take scope over each other.

- (7) ‘Who offers what to who?’
 $\lambda\left\{\begin{matrix} x \\ y \\ z \end{matrix}\right\}.\text{offer}((x, \text{human}(x)), (y, \text{non-human}(y)), (z, \text{human}(z)))$

In summary, according to this analysis a question is a propositional abstract $\lambda\{\dots\}.P$. Each question word introduces a parameter which is a member of the parameter set $\{\dots\}$. P represents the proposition which is the body of the abstraction. A question word associates a parameter with the argument role it occupies in P .

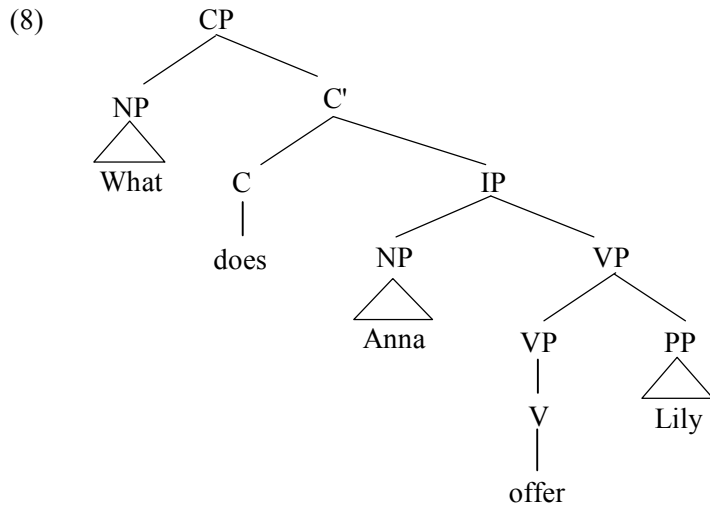
Ginzburg & Sag’s (2000) propositional abstract theory is the semantic basis for this LFG analysis of constituent question formation involving ‘wh’-in-situ in English and Japanese.

3 'Wh'-in-situ in English and Japanese

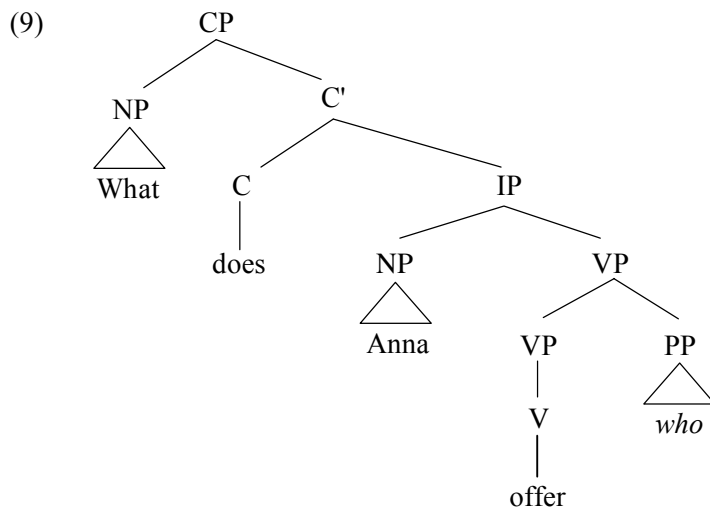
3.1 English constituent question formation

3.1.1 Syntax

Constituent questions are characterised by the presence of one and only one question word in clause-initial ('fronted') position. This question word is syntactically focused.



When there is more than one question word, only one is syntactically focused. All other question words appear in situ.



The syntactically focused question word is the only obligatory question word in a regular English constituent question in the sense that if there is only one question word, it will appear clause-initially.

In a sentence with more than one clause, the syntactically focused question word's position is key to indicating the scope of interrogativity involved. The interrogativity extends only as far as the end of the clause which begins with a question word.

- (10) a. [Charlie knows [WHAT Anna offered Lily]].
 b. [Charlie knows [WHO offered *what* to *who*]].
 c. [WHAT does Charlie know [Anna offered Lily]]?
 d. [WHO does Charlie know [offered *what* to *who*]]?

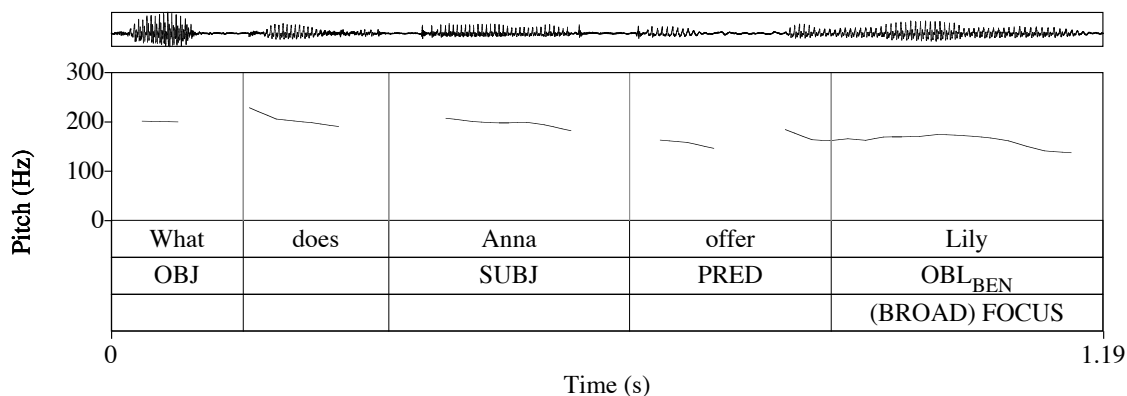
The linearly first question word's position therefore delimits the scope of interrogativity involved in a constituent question.

3.1.2 Prosody

English is a stress-accent, stress-timed language. Stress is a property of individual syllables within words. The location of these prominent syllables is marked in the lexicon. Post-lexically, accentuation is used to highlight a word in a discourse. Pitch accents are points of intonational prominence and so are prosodic properties of utterances. As a pitch accent is linked to one primary-stressed syllable, post-lexical pitch accent is derived from lexical stress in English. In quantitative terms, an accented syllable is the locus of a sharp change in pitch (a sudden fundamental frequency excursion), and generally appears to be longer and louder than a stressed syllable.

It is reasonable to assume that the focus of a single-clause constituent question is the clause-initial question word. However, Culicover & Rochemont (1983: 140) note that in constituent questions, "the location of primary stress ... does not coincide with the focused constituent". A fronted question word is therefore syntactically but not prosodically focused. For example, in (11) it is *Lily* and not the question word *what* that is accentually prominent.⁷ This is consistent with the default prominence pattern known as broad focus, in which the meaning of the whole clause is focused rather than specific words within it.

- (11) What does Anna offer **Lily**?



The prosody of a syntactically focused question word contrasts with that of in-situ English question words. As stated by Ladd (1996: 170-171), an in-situ question word will to some degree be accentually prominent.

- (12) a. Who orders **what when**?
 b. What does **who** order **when**?
 c. When does **who** order **what**?

⁷ Prosodically focused words appear in bold.

Therefore, when both syntactic and prosodic focusing are considered, it is true to say that all question words in English constituent questions are focused. In-situ question words differ only in terms of the level at which focusing is manifested – they are prosodically rather than syntactically focused.

3.1.3 Discourse Information

Following discussion of the prosody and syntax of English constituent question formation, Culicover and Rochemont (1983: 160) conclude, “it is impossible to correlate all instances of focus with stress”. When one approaches FOCUS as discourse information which may be realized at the level of syntax or prosody in a language, this stands to reason. It is clear that ‘fronting’ is focusing too, just at the level of syntax. Either level is capable of expressing FOCUS status.

Culicover and Rochemont (1983: 160) continue, “focus is represented as a unified phenomenon only at the level of F(ocus)-structure”. This is consistent with the notion of all discourse information (including FOCUS) being realized at one level in LFG: information (i-)structure. While at i-structure question words have the same (FOCUS) status, data show that in an English multiple constituent question FOCUS realization is not confined to one structural level.

3.1.4 Data Summary

In English constituent questions, all question words are focused at some level. One question word must be syntactically focused (‘fronted’), while any other question words which bear a grammatical function in the same clause must be prosodically focused (have accentual prominence). The position of the linearly first question word serves to delimit the scope of interrogativity involved.

3.2 Japanese constituent question formation

3.2.1 Syntax

Japanese is a non-configurational language with a rich morphology and relatively free word order. As (1) and (2) show, no syntactic focusing of question words occurs as part of the regular constituent question formation strategy. There is no difference therefore in the word order of neutral declarative sentences and constituent questions.

3.2.2 Prosody

Interrogativity in Japanese is indicated by the addition of a question particle such as *ka* and/or rising intonation at the end of a sentence. While a sentence-final question particle is optional in spoken Japanese (Hinds, 1986), the final rise is present even when the question particle is not. For example, Hirotsu (2003: 121), citing Maekawa's (1997) experimental data, notes that "His sentences did not have a [question particle] attached to the verb, but the rise of [fundamental frequency] was also observed at the end of the verb, a common characteristic of Japanese questions".⁸

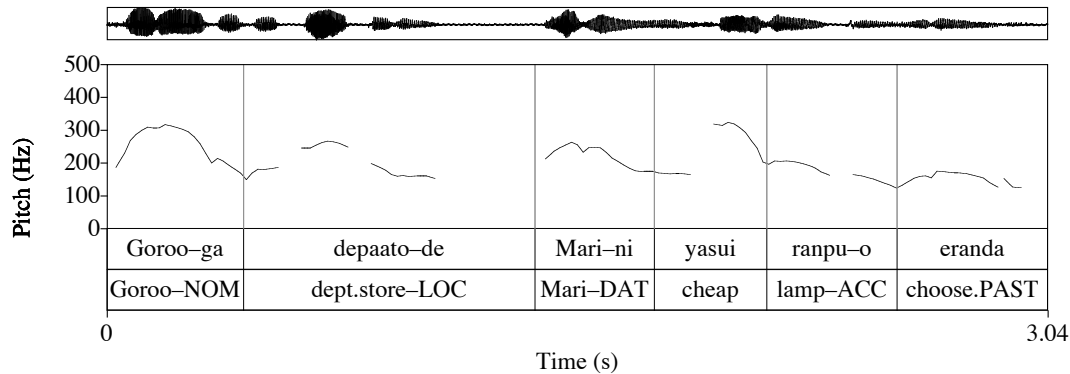
In spoken Japanese, there is prosodic rather than syntactic focusing of question words (Deguchi & Kitagawa, 2002; Ishihara, 2000, 2002, 2004; Maekawa, 1991). Prominence is marked prosodically by manipulation of pitch range, pitch range being the vertical fundamental frequency space within which a speaker realises individual tones.

Any question word receives accentual prominence, characterised by an expansion of pitch range. Subsequent to this pitch range expansion, there is a period of pitch range compression which affects the accents of all words following the question word. A sharp rise in intonation indicates the end of pitch range manipulation (PRM). The entire period of PRM coincides with the scope of any question word bearing a grammatical function in the same clause.

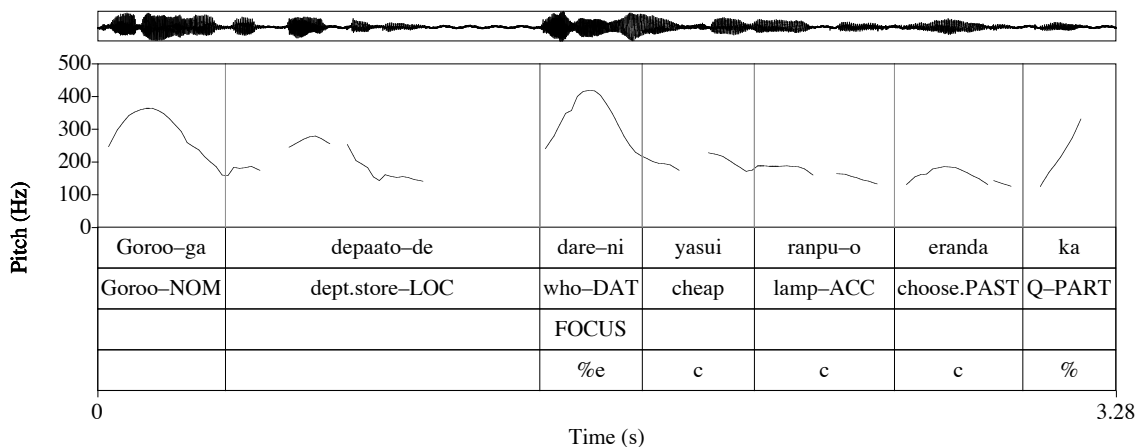
Compare the pitch range values of the declarative sentence (13a) with those of the single constituent question (13b). Focus information and details of PRM are presented in separate tiers for (13b). Pitch range expansion is labelled *e*, pitch range compression is labelled *c*, and the PRM period's boundaries are marked %.

⁸ Though long-distance 'wh'-dependencies in constituent questions are not the subject of this paper, it should be noted that subordinate-clause-final question particles are not optional in Japanese. An analysis of this particular use of question particles may therefore differ in some respects from that proposed for single-clause constituent questions, but it is anticipated that the fundamental LFG approach to 'wh'-in-situ outlined will not be undermined. Thanks to Peter Sells for raising the issue of question particles in Japanese subordinate clauses.

- (13) a. Gorō-ga depāto-de Mari-ni yasui ranpu-o eranda.
 Gorō-NOM dept.store-LOC Mari-DAT cheap lamp-ACC choose.PAST
 SUBJ ADJ OBL_{BEN} OBJ VERB
 ‘Gorō chose a cheap lamp for Mari at the department store.’



- (13) b. SINGLE CONSTITUENT QUESTION
 Gorō-ga depāto-de dare-ni yasui ranpu-o eranda ka.
 Gorō-NOM dept.store-LOC who-DAT cheap lamp-ACC choose.PAST Q
 SUBJ ADJ OBL_{BEN} OBJ VERB
 ‘Who did Gorō chose a cheap lamp for at the department store?’



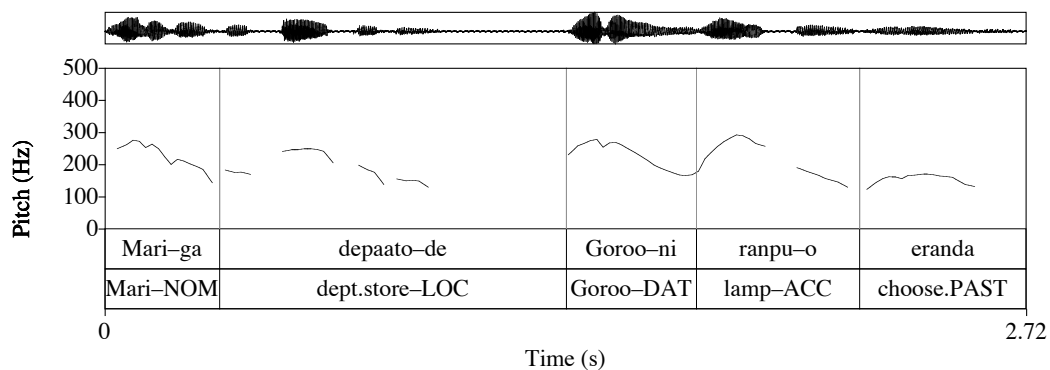
The question word *dare* in (13b) is subject to pitch range expansion: its pitch range (202Hz) is greater than that of the equivalent non-interrogative OBL_{BEN} *Mari* (93Hz) in (13a). The object phrase which follows the question word *dare* in (13b), *yasui ranpu-o*, is subject to pitch range compression. In (13a) the pitch range for this phrase is 199Hz, while in (13b) it is 85Hz.⁹ Pitch range manipulation in (13b) ends with the final sharp rise in intonation affecting the question particle *ka*. The period of PRM is therefore bookended by the pitch range expansion affecting the question word and the final sharp rise

⁹ The case for pitch range compression is less clear-cut with respect to the verbal predicate *eranda*. The relevant pitch range in (13a) is 52Hz, while in (13b) it is 56Hz. However, this is expected given that pitch range compression appears to affect all sentence-final predicates in Japanese declaratives (Venditti, 1997) while verbal predicates in questions are subject to a local rise (Poser, 1984; Pierrehumbert & Beckman, 1988). Both Poser (1984) and Pierrehumbert & Beckman (1988) claim the latter is due to the presence of a sharp rise in intonation at the end of a question in Japanese. Therefore, the pitch range for a verbal predicate in a question may not necessarily be smaller than the pitch range for a verbal predicate in a comparable declarative.

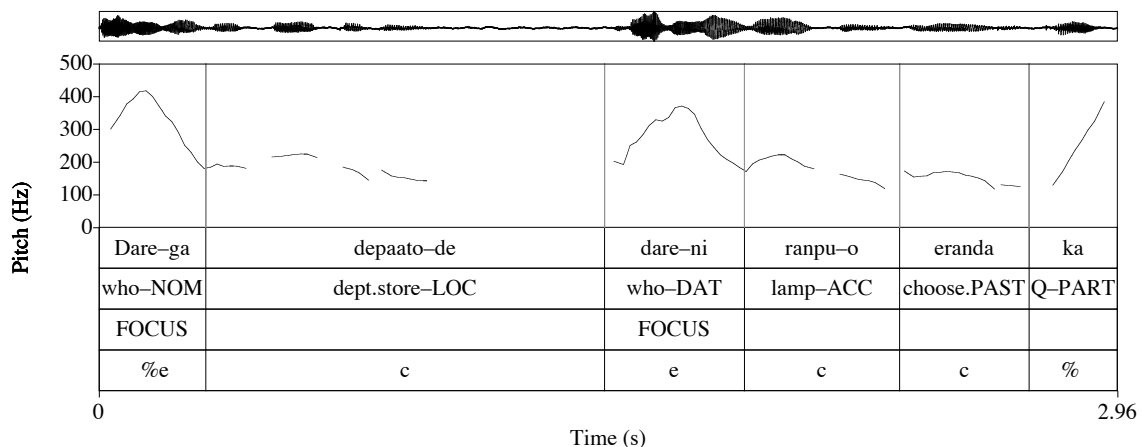
in intonation. This means that the question word represents one PRM boundary and the final sharp rise in intonation the other.

In a multiple constituent question, at least one prosodically focussed question word appears within the period of pitch range compression that follows the first question word to bear a grammatical function in the same clause. Compare the pitch range values for the declarative sentence (14a) with those for the multiple constituent question (14b).

- (14) a. Mari-ga depāto-de Gorō-ni ranpu-o eranda.
 Mari-NOM dept.store-LOC Gorō-DAT lamp-ACC choose.PAST
 SUBJ ADJ OBL_{BEN} OBJ VERB
 ‘Mari chose a lamp for Gorō at the department store.’



- (14) b. MULTIPLE CONSTITUENT QUESTION
 Dare-ga depāto-de dare-ni ranpu-o eranda ka.
 who-NOM dept.store-LOC who-DAT lamp-ACC choose.PAST Q
 SUBJ ADJ OBL_{BEN} OBJ VERB
 ‘Who chose a lamp for who at the department store?’



The SUBJ *dare* and the OBL_{BEN} *dare* in (14b) exhibit pitch range expansion – their respective pitch range values (240Hz and 197Hz) are greater than those of the SUBJ and the OBL_{BEN} in (14a) (121Hz and 113Hz respectively). Once more, the period of PRM ends with a final sharp rise in intonation.

Note that (14b) shows pitch range compression on either side of the OBL_{BEN} question word *dare*. In (14b) the relevant pitch range values are 81Hz (*depāto-de*) and 125Hz (*ranpu-o*) compared with 116Hz (*depāto-de*) and 159Hz (*ranpu-o*) in (14a). Rather than OBL_{BEN} *dare* starting a new period of PRM, it appears to be the case that any other question words are embedded within the period of pitch range compression that follows the linearly first question word. In both multiple constituent questions such as (14b) and single constituent questions there is a single PRM boundary at the left edge signified by the final sharp rise in intonation. Furthermore, the total period of PRM which begins with the linearly first question word coincides with the scope of all the question words that bear a grammatical function in the same clause. The parallels between the PRM in a single constituent question and the PRM in a multiple constituent question are clear. Therefore, I propose that only one question word represents a left-edge PRM boundary in a multiple constituent question (the linearly first question word) and only one period of PRM exists.

As in English, the linearly first question word evidently plays a part in delimiting the scope of interrogativity involved in a Japanese constituent question. However, in Japanese scope is not signalled at the level of syntax, but rather at the level of prosody by a period of PRM which begins with the linearly first question word.

3.2.3 Discourse Information

Data show that all question words are prosodically focused in Japanese. The only difference between Japanese and English constituent questions therefore is that one question word in English must be syntactically focused. When all question focus information is unified, neutral English and Japanese constituent questions are identical at the level of i-structure.

3.2.4 Data Summary

As in English all question words have been shown to be focused in Japanese constituent questions, meaning that the i-structures of these sentences will be identical in terms of question focus in both languages. The two languages differ though with respect to the level at which this focusing takes place. In Japanese, focusing of question words is purely prosodic.

Another feature which Japanese shares with English is that the linearly first question word is key to delimiting the scope of interrogativity. This is because it marks the beginning of the period of PRM which coincides with the scope of all question words that bear a grammatical function in the same clause.

4 LFG Analysis of 'wh'-in-situ

Given the semantics for interrogatives and data discussed previously, there are a number of key points that must be included in an LFG analysis of constituent question formation strategies which involve 'wh'-in-situ.

- All question words introduce parameters.
- All question words are focused, either prosodically or syntactically.
- Focus and scope may be indicated at c-structure and/or p-structure level.
- Key to interrogative scope cross-linguistically is the linearly first question word.

4.1 Characterizing Question Words

Question words are pronouns and as such have value ‘PRO’ for the attribute PRED. As they are parameter-introducing elements, all question words are positively specified for a semantic attribute PARAM. A parameter is co-indexed with the question word that introduces it. This links the parameter to the grammatical function which the question word bears in the proposition’s f-structure, consistent with the notion of place-holder. Therefore, each question word has an attribute INDEX with a different value.

As a question word also introduces further semantic restrictions beyond those imposed by the argument to which it is linked, each question word will also have an appropriate semantic attribute. For example, the semantic attribute HUMAN will have a positive value when the question word is *who* and a negative value when the question word is *what*. Other question words will have different semantic attributes depending on the specific restrictions they contribute.

Finally, all question words must be focused. I propose a separate FOCUS TYPE whose value is ‘question’ for these words, as it is not clear that question words meet the criteria of other types of focus such as contrastive or presentational focus.¹⁰

A typical question word entry is given in (15). Its f-structure in (16) follows straightforwardly.

$$(15) \quad \text{who} \quad N^0 \quad \begin{array}{l} (\uparrow\text{PRED}) = \text{'PRO'} \\ (\uparrow\text{PARAM}) = + \\ (\uparrow\text{INDEX}) = x \\ (\uparrow\text{HUMAN}) = + \\ ((\uparrow_i\text{FOCUS}) \text{TYPE}) = \text{question} \end{array} \quad (16) \quad \text{who} \quad \left[\begin{array}{ll} \text{PRED} & \text{'PRO'} \\ \text{PARAM} & + \\ \text{INDEX} & x \\ \text{HUMAN} & + \end{array} \right]$$

Focusing of question words is achieved by a mapping from c- and/or a ToBI-style p-structure (O’Connor, 2004) to i-structure.¹¹ With respect to the mapping from p- to i-structure, pitch range expansion appears to be a prosodic correlate of FOCUS in both English and Japanese.¹²

“Focus realization in English is fundamentally similar to that in Mandarin [and also Japanese], i.e., the pitch range of the focused item is expanded, the pitch range of the post-focus items, if any, is compressed and lowered, and the pitch range of the pre-focus items, if any, remains neutral.”

(Xu & Xu, 2005: 193)

This is not to say that pitch range expansion is the sole component of prosodic focusing in English or any other language, only that pitch range expansion has been identified as the prosodic correlate of FOCUS in a number of languages. This means that for both English and Japanese, a mapping from p- to i-structure can be provided which formalizes this characterization of prosodic focusing as crucially involving pitch range expansion (*e*) affecting an unspecified tone (*t*).

¹⁰ This proposal is discussed further in Mycock (in prep.).

¹¹ See Beckman & Ayers Elam (1997), Beckman & Hirschberg (1994) and Beckman *et al* (2005) for information on the ToBI system for transcribing the intonation and prosodic structure of English. For details of the variant developed for transcribing Japanese (J_ToBI), see Venditti (1997, 2005).

¹² This may also be true cross-linguistically. See Pierrehumbert & Beckman (1988: 99).

(17) PROSODIC FOCUSING

$$\begin{array}{c} t \\ \downarrow = e \\ \downarrow \in \{\text{FOCUS}\} \end{array}$$

A unified analysis of ‘wh’-in-situ in English and Japanese as prosodic focusing of a question word is therefore possible.

(18) ‘WH’-IN-SITU (prosodic focusing of a question word)

$$\begin{array}{c} t \\ \downarrow = e \\ \downarrow \in \{\text{FOCUS}\} \\ (\downarrow\text{PARAM}) = + \end{array}$$

There is also syntactic focusing in English, characterised at f-structure as functional uncertainty. This constitutes a mapping from c- to i-structure in English, given that FOCUS at f-structure is FOCUS at i-structure too. (19) is the familiar rule of focus fronting in English, which applies to constituent question formation.

(19)

$$\text{CP} \rightarrow \begin{array}{c} \text{XP} \\ (\uparrow \text{FOCUS}) = \downarrow \end{array} \text{ , } \begin{array}{c} \text{C}' \\ \uparrow = \downarrow \end{array}$$

To deal fully with Japanese Focus Prosody, the basic p- to i-structure mapping given in (17) must be expanded in order to allow for different possible configurations. Specifically, it is necessary to permit multiple instances of prosodic focusing as well as the intervention between prosodically focused items of non-prosodically focused material which will be subject to pitch range compression.

(20)

$$\dots \begin{array}{c} t \\ \downarrow = \%e \\ \downarrow \in \{\text{FOCUS}\} \end{array} \left(\left(\begin{array}{c} t \\ \downarrow = c \end{array} \right)^* \left(\begin{array}{c} t \\ \downarrow = e \\ \downarrow \in \{\text{FOCUS}\} \end{array} \right)^* \right)^* \begin{array}{c} t \\ \downarrow = c \end{array} \begin{array}{c} t \\ \downarrow = \% \end{array} \dots$$

Japanese Focus Prosody as expressed in (20) applies to spoken constituent questions as well as to declarative sentences containing focused items.

4.2 Question Word FOCUS-Q₁

One question word – the linearly first question word – has a special status cross-linguistically. Remember that in English scope is related to the c-structure position of the syntactically focused, linearly first question word. In Japanese, all question words with the same scope are embedded within the period of pitch range compression which follows the linearly first question word. While all question words have FOCUS TYPE value ‘question’ at i-structure, the linearly first question word appears to be slightly different. This particular question word will be referred to as FOCUS-Q₁.¹³

¹³ The ‘1’ in FOCUS-Q₁ is in no way intended to indicate ‘main’ or ‘primary’; it is only used to reflect the special status of this particular question word.

(21) FOCUS-Q₁ Rule

The linearly first parameter-introducing word alone has the FOCUS TYPE value Q₁ at i-structure. FOCUS-Q₁ is obligatory in all constituent questions.

This FOCUS-Q₁ Rule ensures that if there is only one question word in a constituent question, it will have FOCUS-Q₁ status. In a multiple constituent question, all other parameter-introducing words which bear a grammatical function in the same f-structure will have the FOCUS TYPE value ‘question’. This means that while FOCUS-Q₁ status is restricted to one and only one parameter, FOCUS TYPE ‘question’ parameters are members of a set at i-structure.

4.3 Interrogative Scope

Though FOCUS-Q₁ is key to indicating interrogative scope in a constituent question, it is only one component of the scope marking involved. It has been shown that scope is indicated at different levels in English and Japanese. In English scope is indicated at c-structure, while in spoken Japanese this occurs at p-structure for single-clause constituent questions. I propose that specific configurations at these levels, which crucially refer to FOCUS-Q₁, make a contribution to meaning. This is not a new proposal, at least with respect to c-structure. Rosén (1996) and Dalrymple (2001: 417) both identify constructions in which meaning contributions are associated with phrase structure configurations rather than lexical items. I propose that the specific configurations associated with constituent question formation contribute the meaning constructor **[interrog scope]**.

(22) **[interrog scope]** = λPλQ(P), where *P* is a proposition and *Q* is the set of parameters

According to Ginzburg & Sag’s (2000) semantics for interrogatives, parameters introduced by question words form a set whose scope is a proposition. **[interrog scope]** determines exactly which proposition a particular parameter set takes scope over.

In English, the configuration which contributes **[interrog scope]** is a c-structure configuration. It is now possible to produce a mapping which constitutes a rule of English constituent question formation (ECQF) by augmenting the basic rule of focus fronting in (19) with **[interrog scope]**. (23) involves a c- to s-structure mapping which represents the contribution of **[interrog scope]** by the c-structure configuration involved in ECQF, as well as a c- to i-structure mapping which represents syntactic focusing of a single question word (FOCUS-Q₁).

(23) ECQF RULE

$$\begin{array}{ccc}
 & \text{XP} & \text{C}' \\
 \text{CP} & \rightarrow & \begin{array}{l} (\uparrow \text{ FOCUS}) = \downarrow \\ (\downarrow \text{ PARAM}) = + \end{array} , \quad \begin{array}{l} \uparrow = \downarrow \\ \text{[interrog scope]} \end{array}
 \end{array}$$

All other question words in an English multiple constituent question will be prosodically focused according to the p- to i-structure mapping provided in (18). Simple fronting is therefore characterised as a constituent question formation strategy which requires that FOCUS-Q₁ be syntactically focused and all other question words be prosodically focused.

In spoken Japanese, the configuration which contributes the meaning constructor **[interrog scope]** is a p-structure configuration.¹⁴ The basic rule of Japanese Focus Prosody provided in (20) can thus be revised to provide a rule of Japanese Constituent Question Formation (JCQF).

(24) JCQF RULE

$$\begin{array}{ccccccc}
 \text{Iff} & & t & & \dots & t & \text{then} & & t \\
 & & \downarrow = \%e & & & \% & & & \% \\
 & & \downarrow \in \{\text{FOCUS}\} & & & & & & \uparrow = \downarrow \\
 & & (\downarrow\text{PARAM}) = + & & & & & & \mathbf{[\text{interrog scope}]}
 \end{array}$$

This rule constitutes a mapping not only from p- to i-structure, but also from p- to s-structure. **[interrog scope]** is contributed by the right-edge boundary of any period of PRM involving at least one question word. The pitch range expansion which FOCUS-Q₁ must be subject to represents the left-edge boundary of this period of PRM. In this way, FOCUS-Q₁ is essential to interrogative scope marking in Japanese. The ‘wh’-in-situ constituent question formation strategy can therefore be defined as requiring that all question words be focused at the level of p-structure.

Fundamentally, (23) and (24) are the same rule. They crucially rely on the notions of FOCUS-Q₁ and **[interrog scope]** to characterize constituent question formation in both languages. Where they differ is with respect to the level at which FOCUS-Q₁ and **[interrog scope]** can be identified. That is, English and Japanese constituent question formation strategies exploit different means involving distinct structural levels to achieve the same ends. A parallel architecture such as LFG’s, which treats different structural levels as equal rather than assuming the primacy of syntactic structure, is well suited to capturing this type of cross-linguistic generalization.

4.4 Example Analyses: Single-Clause Multiple Constituent Questions

In this section example analyses of two single-clause multiple constituent questions are provided, one English and one Japanese.

Each structural level is represented separately, the mappings between the levels being those described in the previous section. As PRM alone is of direct concern to this analysis, the only p-structure information presented concerns pitch range. As such, the p-structures provided should be regarded as partial representations of the envisaged ToBI-style p-structures. PRM is represented in a separate tier intended to ‘overlay’ the standard Tone tier found in a ToBI-style transcription. This is motivated by the fact that a Pitch Range tier will deal with the actual realization of tones which are characterized as phonological tone events (pitch contours) in the Tone tier.¹⁵

In the partial i-structures provided, only FOCUS TYPE ‘question’ information is presented. Beyond this discourse information, which alone has a direct bearing on neutral constituent question formation according to the proposed analysis, features of i-structure are not of immediate concern and so are set aside.¹⁶

¹⁴ This explains why question particles are obligatory in written Japanese. Without the prosodic form of a string, another level of structure is required to contribute **[interrog scope]** to s-structure instead.

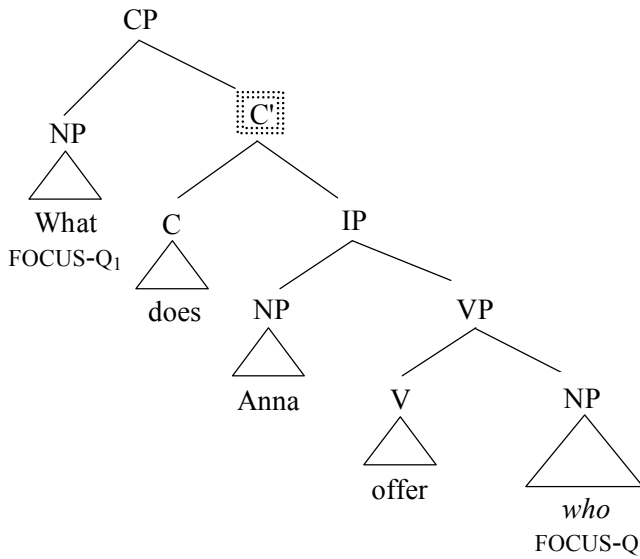
¹⁵ PRM is incorporated in the ToBI transcription system for Pan-Mandarin outlined in Peng *et al* (2005).

¹⁶ For more on i-structure in LFG see, for example, Butt & King (1996, 1997, 1998), Choi (1999) and O’Connor (2004).

4.4.1 English

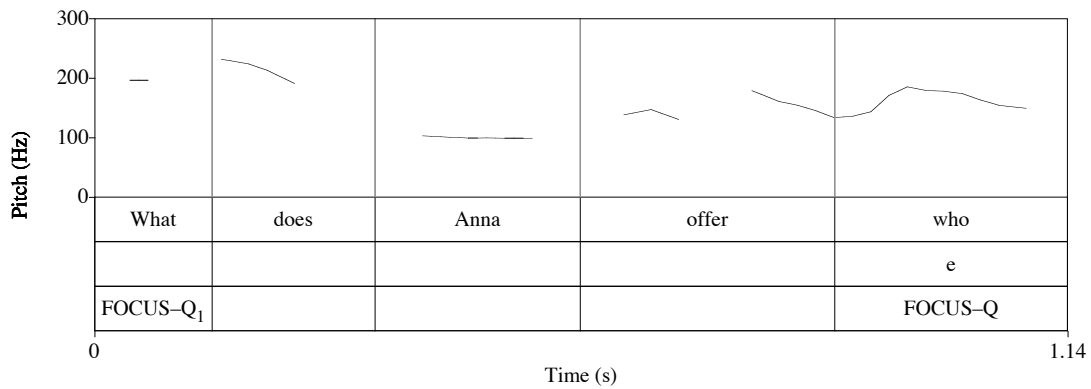
(25) ‘What does Anna offer *who*?’

c-structure

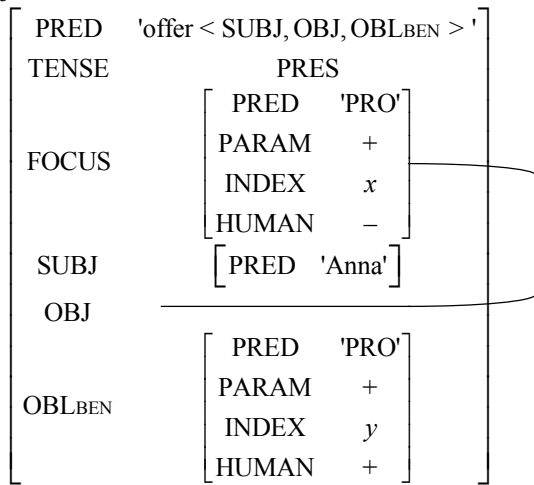


 marks the c-structure configuration which contributes the meaning constructor **[interrog scope]**

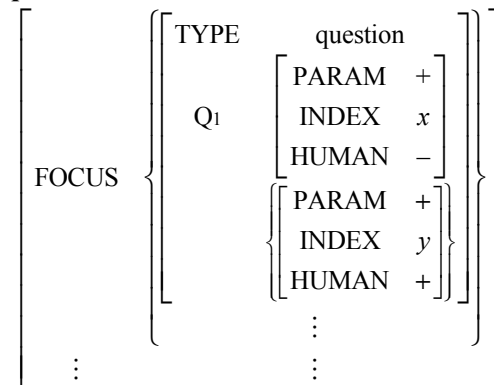
partial p-structure – pitch range manipulation



f-structure



partial i-structure



s-structure

An s-structure for the proposition part of this propositional abstract can be straightforwardly provided.

$$\text{offer}(\text{Anna}, (x, \text{non-human}(x)), (y, \text{human}(y)))$$

This representation does not include the parameters introduced by question words though and is therefore incomplete. The s-structure must be revised to include Q , where $Q = \begin{Bmatrix} x \\ y \end{Bmatrix}$ which is the set of parameters introduced by the two question words.

$$\text{offer}(\text{Anna}, (x, \text{non-human}(x)), (y, \text{human}(y))) \wedge Q$$

Beyond the semantic contributions of question words, constituent question formation involves the meaning constructor **[interrog scope]**. In the case of English, this meaning constructor is contributed by c-structure. **[interrog scope]** combines the meaning of a proposition with that of the parameter set Q , thus determining over which particular proposition the parameter set takes scope.

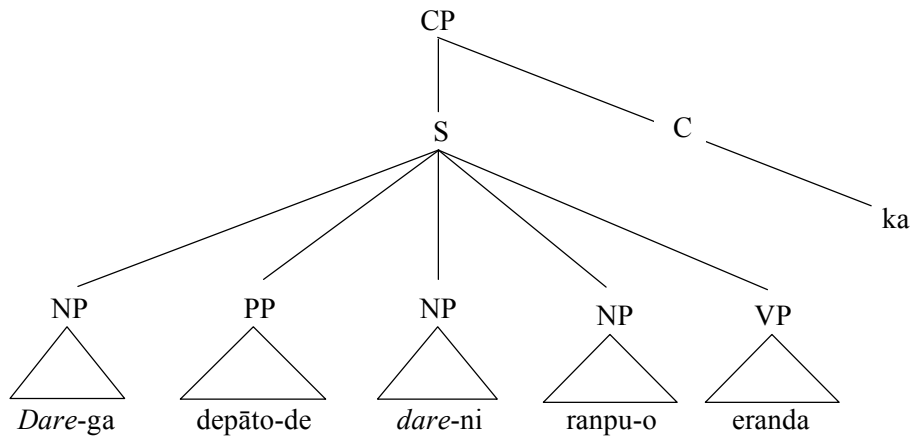
$$\text{offer}(\text{Anna}, (x, \text{non-human}(x)), (y, \text{human}(y))) \wedge Q \qquad \mathbf{[interrog\ scope]} \lambda P \lambda Q (P)$$

$$\lambda \begin{Bmatrix} x \\ y \end{Bmatrix} . \text{offer}(\text{Anna}, (x, \text{non-human}(x)), (y, \text{human}(y)))$$

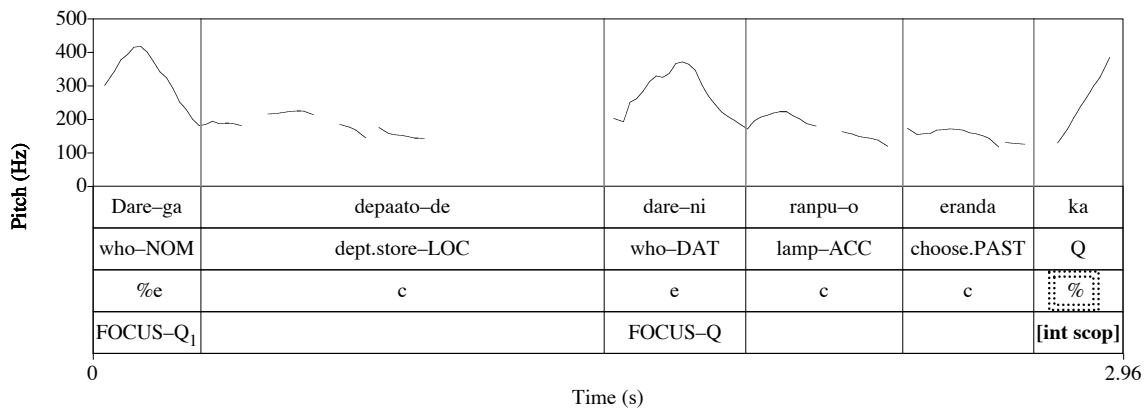
4.4.2 Japanese¹⁷


- (26) *Dare-ga* depāto-de *dare-ni* ranpu-o eranda ka.
 who-NOM dept.store-LOC who-DAT lamp-ACC choose.PAST Q
 ‘Who chose a lamp for who at the department store?’

c-structure



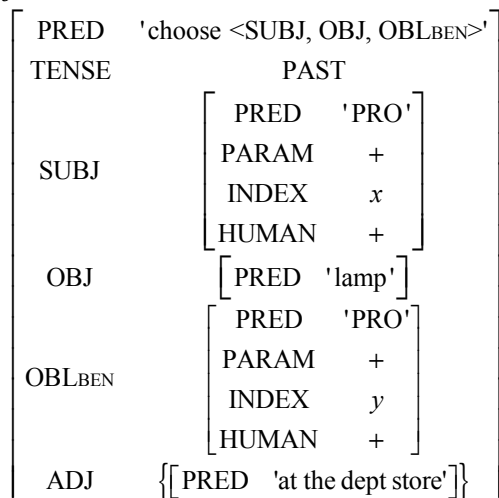
partial p-structure – pitch range manipulation



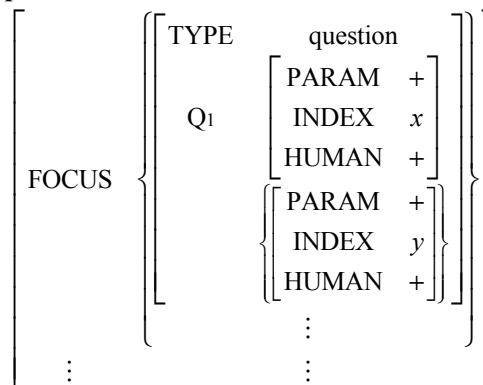
 marks the p-structure configuration which contributes the meaning constructor [interrog scope]

¹⁷ Japanese is one of the languages for which an LFG computational grammar is being developed as part of the Parallel Grammar (ParGram) Project. For information about proposed c- and f-structures, see <http://www.fujixerox.co.jp/research/eng/category/ii/document/01details.html#examples>

f-structure



partial i-structure



s-structure

Based purely on the semantic contribution of lexical items, (26) can be represented as:

$$(\text{at-the-dept-store } (\text{choose } ((x, \text{human}(x)), \text{lamp}, (y, \text{human}(y)))))) \wedge Q$$

where $Q = \left\{ \begin{matrix} x \\ y \end{matrix} \right\}$ which is the set of parameters introduced by the two question words.

However, in spoken Japanese it is not only lexical items which make a contribution to meaning. The meaning constructor **[interrog scope]** is contributed by p-structure in the case of a single-clause constituent question. **[interrog scope]** combines the meaning of a proposition with that of the parameter set Q , thus determining over which particular proposition the parameter set takes scope.

$$(\text{at-the-dept-store } (\text{choose } ((x, \text{human}(x)), \text{lamp}, (y, \text{human}(y)))))) \wedge Q \quad \mathbf{[interrog\ scope]} \lambda P \lambda Q (P)$$

$$\lambda \left\{ \begin{matrix} x \\ y \end{matrix} \right\} . (\text{at-the-dept-store } (\text{choose } ((x, \text{human}(x)), \text{lamp}, (y, \text{human}(y))))))$$

5 Conclusion

Data from Japanese, a ‘wh’-in-situ language, and English, a simple-fronting language, show that all question words are focused, though this focusing may be syntactic or prosodic. Prosodic focusing is the feature that characterises ‘wh’-in-situ in both languages, highlighting the important role that prosody may play.

Japanese and English data further show the linearly first question word in a clause to be key to scope. Scope is expressed in formal terms as the introduction of a meaning constructor **[interrog scope]** by a specific configuration including the linearly first question word (FOCUS-Q₁) at a particular structural

level. In English, **[interrog scope]** is contributed by a c-structure configuration, while in Japanese the relevant configuration is a p-structure one. According to Ginzburg & Sag's (2000) approach to the semantics of interrogatives, questions are propositional abstracts. **[interrog scope]** combines the meanings of a proposition and the parameter set to give the meaning of a question.

The parallel architecture of LFG enables generalizations about prosody and its contribution to other structural levels of language to be made. In this way, LFG provides important insights into the cross-linguistic features of and systematic differences between constituent question formation strategies. LFG's architecture with its distinct yet parallel projections enables 'wh'-in-situ, and indeed constituent question formation in general, to be captured by a single non-derivational analysis for two languages typically regarded as typologically distinct in terms of their constituent question formation strategies. This augurs well for the development of a unified LFG account of the full typological sorts of question formation.

References

- Beckman, Mary E & Gayle Ayers Elam (1997). *Guidelines for ToBI Labelling*. online ms, version 3, March 1997. http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/
- Beckman, Mary E & Julia Hirschberg (1994). *The ToBI Annotation Conventions*. online ms. http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html
- Beckman, Mary E, Julia Hirschberg & Stefanie Shattuck-Hufnagel (2005). The Original ToBI System and the Evolution of the ToBI Framework. In Sun-Ah Jun (ed.), *Prosodic Models and Transcription: Towards Prosodic Typology*. Oxford: Oxford University Press. 9-54.
- Butt, Miriam and Tracy Holloway King (1996). Structural Topic and Focus without Movement. In Miriam Butt & Tracy Holloway King (eds.) *Proceedings of the First LFG Conference*, Rank Xerox, Grenoble. <http://csli-publications.stanford.edu/LFG/1/toc-lfg1.html>
- Butt, Miriam & Tracy Holloway King (1997). Null Elements in Discourse Structure. In Karumuri Venkata Subbarao (ed.), *Papers from the NULLS seminar*. Delhi: Motilal Banarsidas.
- Butt, Miriam & Tracy Holloway King (1998). Interfacing Phonology with LFG. In *Proceedings of LFG98 Conference*. University of Queensland, Brisbane. Stanford, CA: CSLI Publications.
- Choi, Hye-Won (1999). *Optimizing Structure in Context: Scrambling and Information Structure*. Stanford, CA: CSLI Publications.
- Cohen, Felix S. (1929). What is a Question? In *The Monist* **39**. 350-364.
- Culicover, Peter W & Michael Rochemont (1983). Stress and Focus in English. In *Language* **59**. 123-165.
- Dalrymple, Mary (2001). *Lexical-Functional Grammar. Syntax and Semantics, vol. 34*. London: Academic Press.
- Deguchi, Masanori & Yoshihisa Kitagawa (2002). Prosody and Wh-questions. In Masakao Hirotani (ed.) *Proceedings of the 32nd Annual Meeting of the North-Eastern Linguistic Society*. University of Massachusetts at Amherst: GLSA. 73-92.
- Ginzburg, Johnathan & Ivan A Sag (2000). *Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives*. Stanford, CA: CSLI Publications.
- Hinds, John (1986). *Japanese*. London: Croom Helm.
- Hirotani, Masako (2003). Prosodic Effects on the Interpretation of Japanese Wh-questions. In Luis Alonso-Ovalle (ed.), *University of Massachusetts Occasional Papers in Linguistics, vol. 27: On Semantic Processing*. University of Massachusetts, Amherst, MA: GLSA. 117-137.
- Huang, Chu-Ren (1993). Reverse Long-distance Dependency and Functional Uncertainty: The Interpretation of Mandarin Questions. In Chungmin Lee and Boem-mo Kang (eds.), *Language, Information, and Computing*. Seoul: Thaeaksa. 111-120.

- Ishihara, Shinichiro (2000). Scrambling and its Interaction with Stress and Focus. In *MIT Working Papers in Linguistics* **38**. 95-110.
- Ishihara, Shinichiro (2002). Invisible but Audible Wh-scope Marking: Wh-constructions and Deaccenting in Japanese. In *Proceedings of the Twenty-first West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Press. 180-193.
- Ishihara, Shinichiro (2004). Prosody by Phase: Evidence from Focus Intonation – Wh-scope Correspondence in Japanese. In Shinichiro Ishihara, Michaela Schmitz & Anne Schwarz (eds.), *Interdisciplinary Studies on Information Structure* **1**. 77-119.
- Jespersen, Otto (1924). *A Modern English Grammar on Historical Principles, volumes 1-7*. London: Allen and Unwin.
- Ladd, D Robert (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.
- Maekawa, Kikuo (1991). Perception of Intonation Characteristics of WH and non-WH questions in Tokyo Japanese. In *Proceedings of the 12th International Congress of Phonetic Sciences* **4**. Université de Provence, Aix-en-Provence. 202-205.
- Maekawa, Kikuo (1997). The Intonation of Japanese Interrogatives [in Japanese]. In Spoken Language Working Group (eds.), *Speech and Grammar*. Tokyo: Kuroshio Publishers. 45-53.
- Mycock, Louise (in prep.). *The Typology of Constituent Questions: A Lexical-Functional Grammar Analysis of 'Wh'-Questions*. ms., University of Manchester.
- O'Connor, Rob (2004). *Information Structure in Lexical-Functional Grammar: The Discourse-Prosody Correspondence in English and Serbo-Croatian*. PhD thesis, University of Manchester.
- Peng, Shu-hui, Marjorie K M Chan, Chiu-yu Tseng, Tsan Huang, Ok Joo Lee & Mary E Beckman (2005). Towards a Pan-Mandarin System for Prosodic Transcription. In Sun-Ah Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press. 230-270.
- Pierrehumbert, Janet B & Mary E Beckman (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.
- Poser, William J (1984). *The Phonetics and Phonology of Tone and Intonation in Japanese*. ms., MIT.
- Rosén, Victoria (1996). The LFG Architecture and 'Verbless' Syntactic Constructions. In Miriam Butt & Tracy Holloway King (eds.) *Proceedings of the LFG '96 Conference*, Rank Xerox, Grenoble. CSLI Publications. <http://csli-publications.stanford.edu/LFG/1/rosen/rosen.html>
- Venditti, Jennifer J (1997). Japanese ToBI Labelling Guidelines. In Kim Ainsworth-Darnell & Mariopaolo D'Imperio (eds.), *Ohio State University Working Papers in Linguistics* **50**. 127-162.
- Venditti, Jennifer J (2005). The J_ToBI Model of Japanese Intonation. In Sun-Ah Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press. 172-200.
- Watanabe, Akira (2001). Wh-in-situ Languages. In Mark Baltin & Chris Collins (eds.), *The Handbook of Contemporary Syntactic Theory*. Oxford: Blackwell. 203-225.
- Xu, Yi & Ching X Xu (2005). Phonetic Realization of Focus in English Declarative Intonation. In *Journal of Phonetics* **33**: 159-197.

AUTOMATIC ACQUISITION OF SPANISH LFG RESOURCES FROM THE CAST3LB TREEBANK

Ruth O'Donovan, Aoife Cahill, Josef van Genabith and Andy Way

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

In this paper, we describe the automatic annotation of the Cast3LB Treebank with LFG f-structures for the subsequent extraction of Spanish probabilistic grammar and lexical resources. We adapt the approach and methodology of Cahill *et al.* (2004), O’Donovan *et al.* (2004) and elsewhere for English to Spanish and the Cast3LB treebank encoding. We report on the quality and coverage of the automatic f-structure annotation. Following the pipeline and integrated models of Cahill *et al.* (2004), we extract wide-coverage probabilistic LFG approximations and parse unseen Spanish text into f-structures. We also extend Bikel’s (2002) Multilingual Parse Engine to include a Spanish language module. Using the retrained Bikel parser in the pipeline model gives the best results against a manually constructed gold standard (73.20% preds-only f-score). We also extract Spanish lexical resources: 4090 semantic form types with 98 frame types. Subcategorised prepositions and particles are included in the frames.

1 Introduction

Manual construction of rich grammatical and lexical resources, particularly multilingual resources, is time-consuming, expensive and requires considerable linguistic and computational expertise. Previously in (Cahill *et al.*, 2004) and (O’Donovan *et al.*, 2004), we outlined an approach which exploits information encoded in treebank trees to automatically annotate each node in each tree with f-structure equations representing abstract predicate-argument structure relations. From the annotated treebank, we automatically extract large-scale unification grammar resources, namely probabilistic approximations of LFGs¹, and subcategorisation information, for parsing new text into f-structures. A growing number of treebanks for languages other than English (including Japanese, Chinese, German, French, Czech and Spanish) are becoming available. Cahill *et al.* (2003) and Burke *et al.* (2004) show how the lexical and grammatical extraction approaches described in (Cahill *et al.*, 2004) and (O’Donovan *et al.*, 2004) for English can be successfully migrated to typologically different languages (German and Chinese) and different treebank encodings (TIGER (Brants *et al.*, 2002) and Penn CTB (Xue, Chiou, and Palmer, 2002)). Here we describe the porting of the methodology to Spanish and the Cast3LB Treebank (Civit, 2003). We present an f-structure annotation algorithm for Cast3LB and describe how LFG grammars for Spanish can be induced from the f-structure-annotated treebank. We extract PCFG-based LFG approximations and report on a number of parsing experiments. We evaluate both the quality of the automatic f-structure annotation of the Cast3LB treebank, and the parser output. Finally, we describe how lexical resources can be extracted from the f-structure-annotated treebank and present sample lexical entries.

¹See (Cahill *et al.*, 2004) and (O’Donovan *et al.*, 2004) for details on how these resources differ from traditional LFGs.

2 From Cast3LB to a Spanish LFG

2.1 Cast3LB Treebank

The Cast3LB treebank (Civit, 2003) consists of 125,000 words (approximately 3,500 trees) taken from a wide variety of Spanish texts (journalistic, literary, scientific) from both Spain and South America. Despite the free word order of Spanish, constituency rather than dependency annotation is used in the Cast3LB treebank. Unlike the Penn-II Treebank which loosely complies with X-bar theory, the phrase-structure trees of the Spanish Treebank are essentially theory neutral. Only lexically realised constituents are annotated with the exception of elided subjects in pro-drop constructions. There are therefore no empty nodes and traces unlike in the Penn-II Treebank. Another policy of the Cast3LB creators was not to alter the surface word order of the constituents. Due to the free word order of Spanish, a verb phrase containing the verb and its arguments (other than subject) cannot always be established. As a result the main constituents of the sentence are daughters of the root node. The free word order of Spanish also means that phrase-structural position is not an indication of grammatical function, a feature of English which was heavily exploited in the automatic annotation of the Penn-II Treebank. Instead we take advantage of the rich Cast3LB functional annotation of verbal dependents and the fine-grained non-terminals to annotate the treebank with f-structure equations.

Figure 1 shows an example tree from the Cast3LB Treebank. The verbal elements of the sentence are realised by the *gv* (grupo verbal) subtree. The *sn* (sintagma nominal) subject of the sentence is marked as such using the functional tag *SUJ*. Any other verbal complements and adjuncts are marked in a similar way in the treebank. The full list of functional labels is provided in Table 1. Constituents which are not verbal complements do not receive functional annotations. The full list of phrasal category labels (i.e. excluding preterminals) is presented in Table 2. In addition to these, any of the clausal nodes may be annotated with an asterisk to indicate verbal ellipsis in coordinated structures. The tree in Figure 2 where the verb *es* is omitted from the second conjunct demonstrates this phenomenon. The preterminal tags in Cast3LB are fine-grained (see Figures 1 and 2) because they encode morphological as well as part of speech (POS) information. For example the tag *ncms000* indicates that *recurso* is a common noun which is masculine and singular. While there are some distinctions beyond POS encoded in the Penn-II tags, the limited inflectional morphology of English does not allow for or require the same level of detail as Spanish. In Penn-II there are just six verbal tags (excluding the modal tag) which suffice for English inflection. As a single Spanish verb morpheme carries information about person, number, tense, aspect and mood, the 147 verbal tags are by necessity considerably more complex.

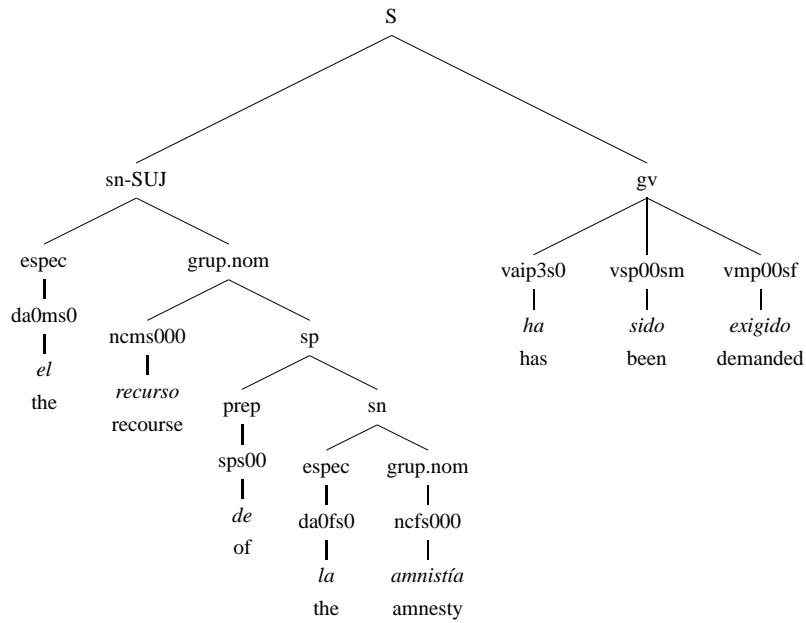


Figure 1: Example Tree from the Cast3LB Treebank

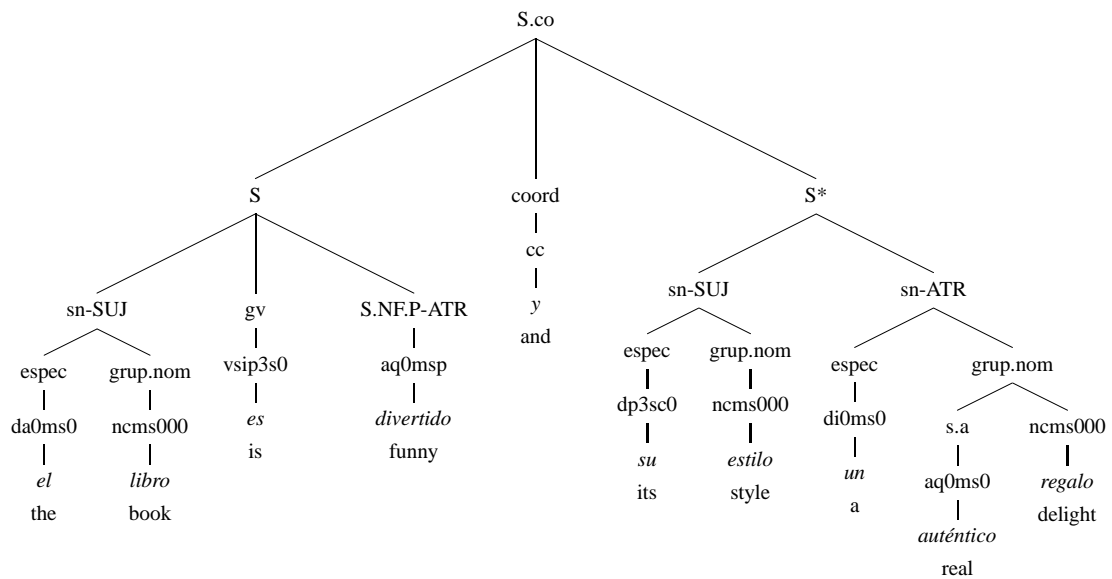


Figure 2: Cast3LB Annotation of Verbal Ellipsis in Coordinated Constructions

| | |
|--------|---------------------------------|
| SUJ | Subject |
| CD | Direct Complement |
| CI | Indirect Complement |
| ATR | Attributive |
| CPRED | Predicative Complement |
| CAG | Agentive Complement |
| CREG | Prepositional Phrase Complement |
| CC | Adjunct |
| ET | Textual Element |
| MOD | Modal Adverb |
| NEG | Negative |
| PASS | Passive |
| IMPERS | Impersonal |
| VOC | Vocative |

Table 1: Functional Annotations used in the Cast3LB Treebank

2.2 Automatic Annotation of Cast3LB Trees

The annotation algorithm for Spanish is constructed following the same methodology used for English, German and Chinese. We begin by automatically extracting all the rules and their associated frequencies from the treebank. We extract 7972 rules when we conflate preterminals containing morphological information to basic POS tags.² We then select the most frequent rule types for each left hand side (lhs) category which together give 85% coverage of all rule tokens expanding that category. This results in a reduced set of 3638 rules. The right hand sides (rhs) of these 3638 rules are then automatically assigned default annotations, e.g. any node with a *SUJ* functional annotation is assigned the functional equation $\uparrow\text{SUBJ}=\downarrow$. The rules are also head lexicalised following the head lexicalisation rules developed for Spanish. The reason for the relatively large number of CFG rules is the fine-grained tags for sentential nodes which are used in the treebank (Figure 2). Of the 3638 rule types, 3533 have a sentential node on the left hand side. As many of the daughters of sentential nodes are tagged with Cast3LB functional tags, the right hand sides of 2870 of the 3638 rules are unsurprisingly completely annotated after automatic head lexicalisation and default annotation. Out of a total of 15039 right hand side nodes, 14091 (93.70%) are assigned an annotation automatically. Next the remaining partially annotated rules (768 in total) are manually examined and used to construct annotation matrices which generalise to unseen rules. The annotation matrices encode information about the left and right context of a rule’s head. For example, an *espec* node to the left of the head of an *sn*’s head is a spec-

²For example the preterminals *ncms000* and *ncfs000* are conflated to the generic POS tag *n*.

| | |
|----------------|---|
| S.F.C | Subordinated Finite Complement |
| S.F.R | Subordinated Finite Adjectival |
| S.F.A | Subordinated Finite Adverbial |
| S.F.A.Cond | Subordinated Conditional Finite Adverbial |
| S.F.A.Conc | Subordinated Concessive Finite Adverbial |
| S.F.A.Cons | Subordinated Consecutive Finite Adverbial |
| S.F.A.Comp | Subordinated Comparative Finite Adverbial |
| S.NF.C | Subordinated Non-Finite Complement |
| S.NF.A | Subordinated Non-Finite Adverbial |
| S.NF.P | Subordinated Non-Finite Adjectival |
| S.NF.R | Subordinated Non-Finite Relative |
| INC | Parenthetical |
| sn(.e) | Noun Phrase (elided) |
| sa | Adjectival Phrase |
| sadv | Adverbial Phrase |
| sp | Prepositional Phrase |
| gv | Verbal Group |
| infinitiu | Infinitival |
| gerundi | Gerund |
| grup.nom | Nominal group |
| prep | Preposition |
| interjeccio | Interjection |
| neg | Negation (no) |
| relatiu | Relative Pronoun |
| numero | Number |
| morfema.verbal | Pronoun <i>se</i> in passive and impersonal constructions |
| morf.pron | Reflexive Pronoun |
| espec | Specifier |

Table 2: Phrasal categories from the Cast3LB Treebank

ifier while an `sp` node to the right of a `grup . nom`'s head is an adjunct. Lexical information is provided by macros which are written for the POS tags.

The f-structure algorithm is implemented in Java following a similar architecture to that used for English, German and Chinese. The automatic annotation of the entire treebank is essentially a four step process illustrated in Figure 3. First, the annotation algorithm attempts to assign an f-structure equation to each node in the tree based on the Cast3LB functional labels. We have compiled an f-structure equation look-up table which assigns default f-structure equations triggered by each Cast3LB functional label. For example, the default entry for the `SUJ` label is $\uparrow\text{SUBJ}=\downarrow$. Table 3 gives the complete set of default annotations. Next, the head of each local subtree of depth one is found following the head lexicalisation rules we have compiled. For example, the `prep` daughter of an `sp` node is its head and is assigned the f-structure equation $\uparrow=\downarrow$. In the third step, the annotation algorithm deals specifically with coordination as this phenomenon is not covered by the left-right generalisations for other constructions. Figure 4 provides an example of coordination in the Cast3LB Treebank. The `.co` suffix on the `grup . nom` node label indicates that the node is mother of two or more coordinated `grup . nom` nodes. The coordinating conjunction (`cc`) is annotated as the head of the coordinated noun phrase and the coordinated elements are annotated as elements of the noun phrase's conjunct set. In a final step, the annotation algorithm moves top-down left-to-right through each tree and any unannotated nodes in each local subtree of depth one are assigned f-structure equations using the left-right context principles constructed by examining the subset of most frequent treebank rules mentioned above. For example, an `sn` node to the right of the head of a prepositional phrase (`sp`) is annotated as the object of the prepositional phrase ($\uparrow\text{OBJ}=\downarrow$). The f-structure equations are then automatically collected and passed to a constraint solver which produces an f-structure. The annotated tree and resulting f-structure for the tree in Figure 1 is shown in Figure 5. The tense, number and gender information as well as root forms are derived from the lexical macros. At present we produce “proto” f-structures (with unresolved long distance dependencies) rather than “proper” f-structures as the Cast3LB does not contain trace information.

2.3 Evaluation of the Annotation Algorithm

We first evaluated the coverage of the annotation algorithm on the entire Cast3LB Treebank. The results are presented in Table 4. 96.04% of the sentences receive one covering and connected f-structure. Ideally, we wish to generate just one f-structure per sentence. A number of sentences (102) receive more than one f-structure fragment. This is due to cases where the algorithm cannot establish a relationship between all elements in the treebank sentence and leaves nodes unannotated. There are also a small number of sentences (36) which do not receive any f-structure. These are a result of feature clashes in the annotated trees, which are caused by inconsistent annotation.

We also evaluate the quality of the annotation against a manually constructed gold standard of 100 f-

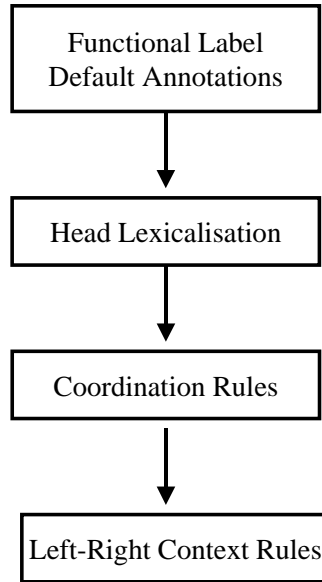


Figure 3: Architecture of Spanish Annotation Algorithm

| | |
|--------|---------------|
| SUJ | ↑SUBJ=↓ |
| CD | ↑OBJ=↓ |
| CI | ↑OBJ_THETA=↓ |
| ATR | ↑XCOMP=↓ |
| CPRED | ↑XCOMP=↓ |
| CAG | ↑OBLAG=↓ |
| CREG | ↑OBL=↓ |
| CC | ↓∈(↑ADJ) |
| ET | ↓∈(↑ADJ) |
| MOD | ↓∈(↑ADJ) |
| NEG | ↓∈(↑ADJ) |
| PASS | ↑PASSIVE=+ |
| IMPERS | ↑IMPERSONAL=+ |
| VOC | ↓∈(↑ADJ) |

Table 3: Functional tag triggered default annotations used in the Cast3LB Treebank

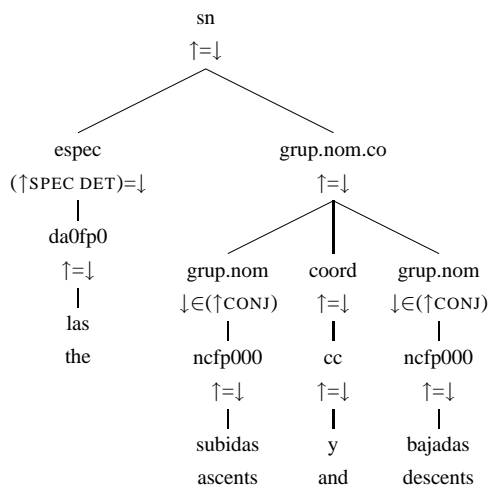


Figure 4: Coordination example from Cast3LB with automatically generated f-structure equations

| F-Structures | Trees | % Trees |
|--------------|-------|---------|
| 0 | 36 | 1.03 |
| 1 | 3347 | 96.04 |
| 2 | 96 | 2.75 |
| 3 | 5 | 0.14 |
| 4 | 1 | 0.03 |

Table 4: Coverage and Fragmentation results of Spanish f-structure annotation algorithm

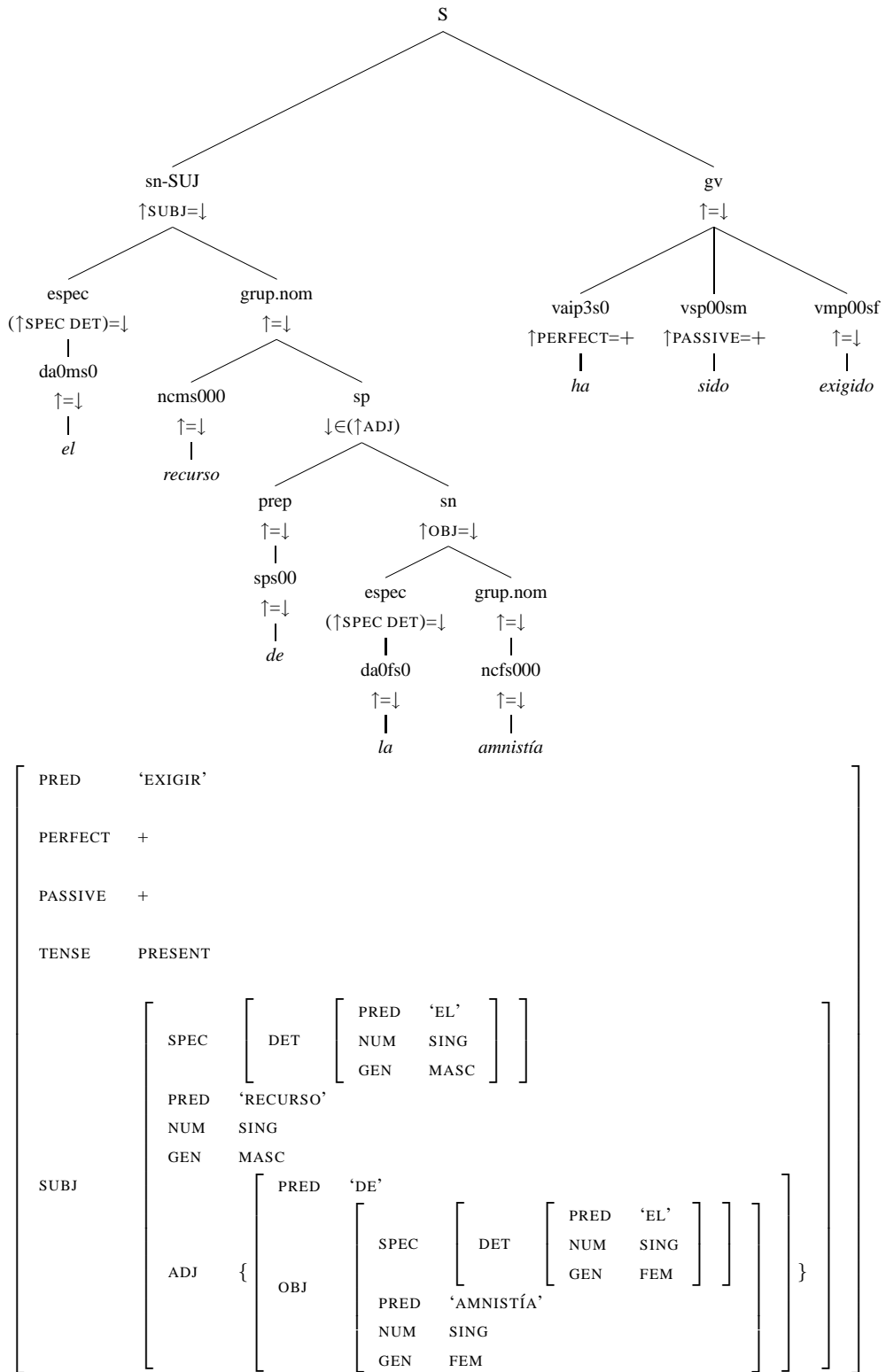


Figure 5: Automatically-annotated tree and f-structure for the example in Figure 1

| | Precision | Recall | F-Score |
|------------|-----------|--------|---------|
| All GFs | 98.40 | 93.56 | 95.92 |
| Preds Only | 97.90 | 92.31 | 95.02 |

Table 5: Evaluation of the automatically produced f-structures against the 100 gold-standard f-structures

structures. For our parsing experiments we set aside approximately 10% of the treebank (336 sentences) for testing purposes. This test set is selected randomly from the various text genres which make up the treebank. We extracted 100 sentences at random from the test set, to develop our gold standard. The f-structures from the original Cast3LB trees for these sentences generated by the automatic annotation algorithm were manually corrected and converted into dependency format. We use the triples encoding and evaluation software of Crouch *et al.* (2002). Table 5 shows that currently the automatic annotation algorithm achieves an f-score of 95.92% for all grammatical functions and 95.02% for preds only. In both cases, precision is about 5% higher than recall. Table 6 shows a more detailed analysis of how well the automatic f-structure annotation algorithm performs for each function in the all grammatical functions evaluation. The algorithm performs well on most features, e.g. the OBJ f-score is 94% and that for SUBJ is 92%. At present, we score worst on the OBLAG feature (the agent in a passive construction). There are only four occurrences of this feature in the gold standard. We expect this along with all the other figures to improve as the annotation algorithm is further refined.

3 Parsing Experiments

To parse raw text into f-structures, we use the **pipeline** and **integrated** parsing architectures of Cahill *et al.* (2004), illustrated in Figure 6. For the pipeline model, we first extract a PCFG from the Cast3LB treebank excluding the 336 test sentences. Cast3LB functional tags are retained in the grammar extraction. We use Helmut Schmid’s BitPar parser (Schmid, 2004) to parse new text with the grammar, using Viterbi pruning to obtain the most probable parse. The resulting parse trees are then automatically annotated using the annotation method described above. The f-structure equations are collected from the trees and passed to the constraint solver which produces an f-structure for each sentence. For the integrated model, we first automatically annotate the Cast3LB treebank with f-structure equations. We then read off a grammar from the annotated treebank, resulting in an *annotated* PCFG (A-PCFG) for Spanish. We again use BitPar to parse new text with this grammar producing annotated trees. Again the f-structure equations are collected from the parse trees and passed to the constraint solver to produce f-structures. We also transformed each grammar using a parent transformation (Johnson, 1999) to give us a P-PCFG and a PA-PCFG.

In addition, we extend Dan Bikel’s multilingual, parallel-processing statistical parsing engine (Bikel,

| DEPENDENCY | PRECISION | RECALL | F-SCORE |
|-------------|----------------|----------------|---------|
| ADJUNCT | 608/618 = 98 | 608/648 = 94 | 96 |
| AUX | 22/22 = 100 | 22/25 = 88 | 94 |
| CASE | 12/12 = 100 | 12/17 = 71 | 83 |
| COMP | 21/22 = 95 | 21/23 = 91 | 93 |
| CONJ | 185/190 = 97 | 185/196 = 94 | 96 |
| DET | 326/328 = 99 | 326/342 = 95 | 97 |
| FORM | 56/57 = 98 | 56/59 = 95 | 97 |
| GEN | 914/920 = 99 | 914/954 = 96 | 98 |
| IMPERSONAL | 3/3 = 100 | 3/3 = 100 | 100 |
| NUM | 1115/1130 = 99 | 1115/1174 = 95 | 97 |
| OBJ | 429/444 = 97 | 429/464 = 92 | 94 |
| OBJ_THETA | 17/17 = 100 | 17/19 = 89 | 94 |
| OBL | 13/14 = 93 | 13/15 = 87 | 90 |
| OBLAG | 2/3 = 67 | 2/4 = 50 | 57 |
| PART | 4/4 = 100 | 4/5 = 80 | 89 |
| PARTICIPLE | 27/27 = 100 | 27/30 = 90 | 95 |
| PASSIVE | 11/11 = 100 | 11/12 = 92 | 96 |
| PERS | 189/196 = 96 | 189/207 = 91 | 94 |
| REFLEX | 17/17 = 100 | 17/18 = 94 | 97 |
| RELMOD | 34/34 = 100 | 34/36 = 94 | 97 |
| SUBJ | 255/258 = 99 | 255/294 = 87 | 92 |
| SUBORD | 50/50 = 100 | 50/54 = 93 | 96 |
| SUBORD_FORM | 50/50 = 100 | 50/54 = 93 | 96 |
| TENSE | 183/187 = 98 | 183/196 = 93 | 96 |
| XCOMP | 62/66 = 94 | 62/73 = 85 | 89 |

Table 6: Breakdown of all grammatical functions annotation algorithm evaluation results by dependency

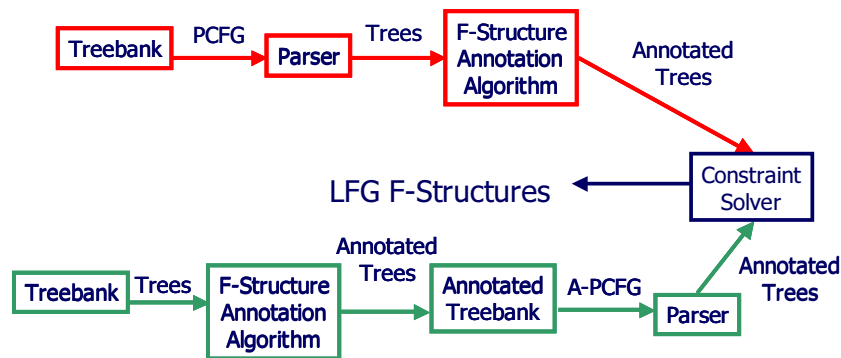


Figure 6: Pipeline (Red) and Integrated (Green) Parsing Architectures

2002) to include a language package for Spanish. Implemented in Java, the parsing engine is a history-based parser emulating Collins’ Model 2 (Collins, 1997). The language package is a collection of Java classes that are extensions of several of the abstract classes which provide the description of data and methods specific to a particular language and treebank annotation style. Aside from creating the Spanish classes, we added a data file specifying the head rules specific to the Spanish Cast3LB treebank to be read by the HeadFinder class. With this extension, we trained the parser on the training set of the treebank retaining Cast3LB functional tags and parsed the test set with the grammar. Following the pipeline model, we then automatically annotated the resulting parse trees, collected the f-structure equations and passed them to the constraint solver to produce f-structures.

As previously noted, the Cast3LB preterminals are very fine-grained, encoding extensive morphological detail in addition to POS information. For example, the tag `vaip3s0` denotes a verb (`v`) which is an auxiliary (`a`), used indicatively (`i`) in the present tense (`p`), and is third person (`3`) singular (`s`). In total there are 327 preterminal types in the treebank. This level of fine-grainedness together with our relatively small training set causes a data sparseness issue for parsing new text. With such a large number of POS tags, it is inevitable that certain tags appear in the test set which have not been seen in a similar context in training with adverse effects on coverage.³ To deal with this issue, initially we masked the morphological detail in the preterminals thereby conflating them to more generic POS tags.

3.1 Initial Results

We then parsed the 336 raw test sentences with the four grammars using BitPar and the retrained and extended Bikel parsing engine. The results are shown in Table 7. We evaluated the quality of the trees produced by the parsers using `evalb` and measured how many of the 336 sentences produce one covering

³If BitPar encounters a sentence in the test set containing a previously unseen tag, it will crash at that point.

| | PCFG | A-PCFG | P-PCFG | PA-PCFG | Bikel |
|---------------------------------------|-------|--------|--------|---------|-------|
| Parses (out of 336) | 334 | 330 | 305 | 264 | 328 |
| Labelled F-Score | 79.01 | 78.89 | 78.78 | 78.44 | 79.19 |
| Unlabelled F-Score | 82.64 | 82.45 | 82.61 | 81.86 | 82.28 |
| Fragmentation (336 F-Structures) | 96.11 | 93.64 | 85.90 | 71.21 | 88.41 |
| All GFs F-Score (100 F-Structures) | 59.70 | 57.99 | 55.75 | 46.93 | 60.13 |
| Preds-Only F-Score (100 F-Structures) | 69.38 | 68.01 | 66.02 | 55.88 | 72.11 |

Table 7: Initial Parsing Results

and connected f-structure. The PCFG performs best in terms of coverage and fragmentation with over 96% of sentences being assigned one covering and connected f-structure. Coverage drops for the A-PCFG with fragmentation of 93.64%. This trend continues when parent transformations are added (71.21% for PA-PCFG). This may be attributed to data sparseness problems. The PA-PCFG rules are very information-rich and it is possible that constructions encountered in testing will not have been seen during training. As before, we evaluated the automatically produced f-structures qualitatively against the manually constructed gold standard using the evaluation software of Crouch *et al.* (2002). The results of this evaluation reveal a problem with the use of preterminal conflation to avoid data sparseness problems in parsing. Usually an all-grammatical-functions evaluation is less rigid than a preds-only evaluation as the features with atomic values (such as person, number and gender) are typically associated with the correct local `pred` even if the `pred` is attached incorrectly in global f-structure. In the case of these experiments however, the grammars score very poorly (as low as 46.93% for the PA-PCFG) in the all-grammatical-functions evaluation. By conflating the preterminal tags we discard the morphological information required by the lexical macros in the f-structure annotation algorithm to project this information to the level of f-structure.

3.2 Final Results

In order to optimise both coverage and f-structure quality we refined our morphological masking process to include a subsequent unmasking step so as to correctly trigger the lexical macros. The masking-unmasking process works as follows. The trees in the treebank are transformed in two ways: the lemmas are removed leaving behind the surface forms of the words and the preterminal tags are conflated to more general POS tags. The masked information is not disposed of but stored in a tab delimited data file in the following format: full preterminal tag, surface form of word, lemma. For example: `vaip3s0 ha haber`. The grammars are extracted from the pre-processed morphologically masked trees and used to parse new text as before. The trees produced by the parser then go through a new post-processing unmasking stage. The lemma information is re-inserted and the conflated tags are expanded. Next the lexical macros are triggered

| | PCFG | A-PCFG | P-PCFG | PA-PCFG | Bikel |
|---------------------------------------|-------|--------|--------|---------|-------|
| Parses (out of 336) | 334 | 330 | 305 | 264 | 328 |
| Labelled F-Score | 79.01 | 78.89 | 78.78 | 78.44 | 79.19 |
| Unlabelled F-Score | 82.64 | 82.45 | 82.61 | 81.86 | 82.28 |
| Fragmentation (336 F-Structures) | 96.11 | 93.64 | 85.90 | 71.21 | 88.41 |
| All GFs F-Score (100 F-Structures) | 79.53 | 77.76 | 74.00 | 62.01 | 79.85 |
| Preds-Only F-Score (100 F-Structures) | 69.41 | 68.01 | 66.02 | 55.88 | 73.20 |

Table 8: Final Parsing Results

by the now fully unmasked POS tags and all f-structure equations are sent to the constraint solver as before. The f-structures produced now contain morphological information. The results are shown in Table 8. As expected, the `evalb` and fragmentation results are unchanged. When compared to initial f-structure results in Table 7, the improvement in the all-grammatical-functions due to this extra step is clear: between 15% and 20% for all of the grammars. There are also slight improvements for the preds-only scores of the PCFG and Bikel. The extended Bikel parsing engine performs best overall: all-grammatical-functions (79.85%) and preds only (73.20%). The PCFG, A-PCFG and P-PCFG produce f-structures of roughly similar quality. The results reported for the PA-PCFG are considerably lower. There is a general trend that the more fine-grained the grammar, the worse the coverage with PA-PCFG achieving only 71.21% fragmentation. This reflects data-sparseness problems due to the comparatively small data set. In contrast to English (Johnson, 1999), for Spanish the parent transformation has an adverse effect on parse quality.

4 Lexical Extraction

The method for automatically inducing semantic forms of O’Donovan *et al.* (2004) is highly suited to multilingual lexical extraction as it works on the level of the more language independent f-structure rather than the more language dependent c-structure. We can apply the extraction algorithm originally developed for English as is to the set of f-structures automatically generated from the Cast3LB in order to induce lexical resources for Spanish. We automatically extract 4090 semantic forms. As for English, we associate conditional probabilities with the extracted frames, differentiate between active and passive frames, parameterise frames with obliques for specific prepositions and optionally include details of syntactic category. Unlike English, the Spanish frames do not yet reflect long-distance dependencies. Of these extracted frames, 3136 are for 1401 verbal lemmas, i.e. 2.4 semantic forms per verb. The verbal semantic forms display all 98 of the frame types extracted. Table 9 provides an overview of the main extraction results broken down by category.

| | Semantic Form Types | Lemmas | Frame Types |
|------------|---------------------|--------|-------------|
| Total | 4090 | 2322 | 98 |
| Verbal | 3136 | 1401 | 98 |
| Nominal | 432 | 432 | 3 |
| Adverbial | 26 | 24 | 4 |
| Adjectival | 496 | 474 | 20 |

Table 9: Spanish semantic forms broken down by category

| Semantic Form | Frequency |
|--------------------------------|-----------|
| <i>ser</i> ([subj, xcomp]) | 1202 |
| <i>estar</i> ([subj, xcomp]) | 208 |
| <i>tener</i> ([subj, obj]) | 206 |
| <i>poder</i> ([subj, xcomp]) | 135 |
| <i>haber</i> ([obj]) | 109 |

Table 10: The most frequently occurring semantic forms extracted from Cast3LB

Table 10 shows the most frequently-occurring semantic forms extracted from the Cast3LB Treebank. The most frequent frame for the verb *haber* (auxiliary ‘have’) is *haber*[obj] due to the Spanish construction with an invariant form of this verb (*hay*) meaning ‘there is’ or ‘there are’ which never occurs with an overt subject. Table 11 shows the attested semantic forms for the verb *ver* (‘see’) with their associated conditional probabilities. Note that as for English, the passive frame is marked with p. The passive is realised in three ways in Spanish. The verb ‘to be’ (*ser*) is combined with a past participle in a manner similar to the English construction. Consider Figure 1 where the string *ha sido exigido* can be translated word for word to the English ‘has been demanded’. The annotation algorithm uses left-right context information to annotate *sido* with the f-structure equation $\uparrow\text{PASSIVE}=\text{+}$ which is exploited by the lexical extraction algorithm at f-structure level. A reflexive construction may also be used to express the passive. For example, ... *se registró un descenso*... (‘... a descent was registered...’) where *un descenso* is the surface subject of the normally transitive *registrar*. In Cast3LB the pronominal constituent (*se*) is tagged as a *morfema.verbal* and has an additional functional tag *-PASS* which is used by the annotation algorithm to assign the $\uparrow\text{PASSIVE}=\text{+}$ f-structure equation. Finally, the Spanish passive may be realised using the third person plural of the verb to be passivised with an empty subject. In this case the verb used passively will not be marked as such because it does not display the movement typically associated with the passive and is essentially an active construction with an empty subject.

| Semantic Form | Conditional Probability |
|-------------------------------|-------------------------|
| <code>ver([subj,obj])</code> | 0.468 |
| <code>ver([subj])</code> | 0.290 |
| <code>ver([subj,comp])</code> | 0.121 |
| <code>ver([subj],p)</code> | 0.072 |

Table 11: Automatically extracted lexical entries for *ver* (see) with associated conditional probabilities

5 Conclusions and Future Work

We have shown how the methodology for automatically annotating the Penn-II Treebank with LFG f-structure equations for the purpose of extracting grammatical and lexical resources can be adapted to Spanish. The methodology has also been successfully migrated to German and Chinese. Our methodology constitutes a novel approach to deep multilingual constraint-based grammar and lexical acquisition based on treebank resources and automatic f-structure annotation algorithms. As treebanks become available for a growing number of languages, we expect this method can deliver robust, wide-coverage multilingual resources with a substantial reduction in development cost. The multilingual work presented here is very much proof of concept. Just three months of development effort have been invested to induce the resources and further work is required to integrate long-distance dependency resolution and to refine the grammar and lexicon extraction.

We developed and applied an automatic f-structure annotation algorithm to the treebank and measured its coverage as well as the quality of the annotations. Over 96% of the trees in the treebank receive one covering and connected f-structure. When evaluated against a gold standard of 100 hand-crafted f-structures, the algorithm scores over 95% for preds-only and all-grammatical-functions. We extract four different PCFGs from the treebank and use them to parse 336 sentences set aside for testing. We also extend and retrain Bikel’s (2002) statistical parsing engine with a Spanish language package to parse the test set. The retrained Bikel parser integrated into the pipeline model performs best against the gold standard, achieving a preds-only f-score of 73.20% against the gold standard. We extract 4090 semantic forms from the annotated treebank using the same methodology applied to the Penn-II Treebank. Long-distance dependency resolution, refinement and extension of the annotation algorithm, grammar and lexicon extraction as well as the evaluation of the lexical resources remain as future work.

References

Bikel, Daniel M. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In *Pro-*

- ceedings of the Human Language Technology Conference*, pages 24–27, San Diego, CA.
- Brants, Thorsten, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In E Hinrichs and K Simov, editors, *Proceedings of the first Workshop on Treebanks and Linguistic Theories (TLT'02)*, pages 24–41, Sozopol, Bulgaria.
- Burke, Michael, Olivia Lam, Rowena Chan, Aoife Cahill, Ruth O'Donovan, Adams Bodomo, Josef van Genabith, and Andy Way. 2004. Treebank-Based Acquisition of a Chinese Lexical-Functional Grammar. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation*, pages 161–172, Tokyo, Japan.
- Cahill, Aoife, Michael Burke, Ruth O'Donovan, Josef van Genabith, and Andy Way. 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*, pages 320–327, Barcelona, Spain.
- Cahill, Aoife, M. Forst, M. McCarthy, R. O'Donovan, C. Rohrer, J. van Genabith, and A. Way. 2003. Treebank-Based Multilingual Unification-Grammar Development. In *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development, at the 15th European Summer School in Logic Language and Information*, pages 17–24, Vienna, Austria.
- Civit, Montserrat. 2003. *Criterios de etiquación y desambiguación morfosintáctica de corpus en español*. Ph.D. Thesis, Universitat Politècnica de Catalunya.
- Collins, Michael. 1997. Three Generative, Lexicalized Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 16–23.
- Crouch, Richard, Ron Kaplan, Tracy Holloway King, and Stefan Riezler. 2002. A comparison of evaluation metrics for a broad coverage parser. In *Proceedings of the LREC Workshop: Beyond PARSEVAL – Towards Improved Evaluation Measures for Parsing Systems*, pages 67–74, Las Palmas, Canary Islands, Spain.
- Johnson, Mark. 1999. PCFG models of linguistic tree representations. *Computational Linguistics*, **24**(4):613–632.
- O'Donovan, Ruth, Michael Burke, Aoife Cahill, Josef van Genabith, and Andy Way. 2004. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II Treebank. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 368–375, Barcelona, Spain.
- Schmid, Helmut. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2004)*, pages 162–168, Geneva, Switzerland.

Xue, Nianwen, Fu-Dong Chiou, and Martha Palmer. 2002. Building a Large-Scale Annotated Chinese Corpus. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

VERBAL CATEGORY AND NOMINAL FUNCTION:
EVIDENCE FROM HUNGARIAN SUBJECT CLAUSES

György Rákosi and Tibor Laczkó

University of Debrecen

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

The aim of this paper is to investigate the categorial and the functional status of the clausal arguments of modal and evaluative predicates in Hungarian. Such an argument can be realized either as a finite *that*-clause or as an optionally agreement-marked infinitival clause, and in both cases it is claimed to map onto SUBJ. Agreement-marked infinitives are shown to have no nominal properties, contra É. Kiss (1987, 2002) and in contrast with Portuguese agreement-marked infinitives. Clausal subjects are always verbal categorially in Hungarian, despite being mapped onto a canonically nominal function. *That*-clauses can have a pronominal associate, in which case this pronoun is the subject of the matrix predicate and the *that*-clause itself is an adjunct to it. Infinitival clauses cannot have pronominal associates because infinitives cannot be adjuncts of nominal categories in Hungarian.

1. Introduction

Modal and evaluative predicates in Hungarian allow for nominal (1a) as well as for sentential (1b-c) arguments in the same argument position:

- (1) a. *Nem sikerül-t* ***a*** ***start.***
not succeed-PAST.3SG the start.NOM
'The start was not successful.'
- b. *Nem sikerül-t* ***el-startol-n-unk*** / ***el-startol-ni.***
not succeed-PAST.3SG away-start-INF-1PL / away-start-INF
'(For us) to start was not successful.'
- c. *Nem sikerül-t,* ***hogy el-startol-j-unk.***
not succeed-PAST.3SG that away-start-SUJU-1PL¹
'It was not successful for us that we should start.'

On the basis of the apparent parallelism (but without any further empirical motivation), traditional descriptive grammars of Hungarian treat the arguments in bold as functionally identical: they take each of these constituents to be the subject of the modal/evaluative predicate. The parallelism is generally considered to be manifest at the categorial level, too – hence the frequent assumption that the subordinate clauses (especially the infinitival ones) have nominal properties.

Generative research has shown an increased interest in various aspects of the syntax of these constructions in the last two decades.² However, the functional and the categorial status of modal/evaluative predicates and their argument structure have received relatively little attention, and even the works which do discuss these issues fail to comment on how the three constructions in (1) relate to each other.

This paper investigates the clausal arguments of modal and evaluative predicates in Hungarian and presents an LFG-theoretic analysis of their functional and categorial properties. We claim that the three structures in (1) are functionally similar in the relevant respects, i.e., the clausal arguments (1b-c) are syntactic subjects, just like the nominal

¹ SUJU = subjunctive suffix.

² The most important works written in English are Dalmi (2002), Kenesei (2001), É. Kiss (1987, 2002), Komlósy (1994), and Tóth (2000, 2001, 2002).

argument in (1a). We also show that despite the functional similarity, the clausal arguments have no nominal properties, contrary to the position advocated not only in the descriptive literature, but also in the generative proposal of É. Kiss (1987, 2002).

We will also seek to provide an answer to why pronominal associates are compatible with *that*-clauses (2a), but not with infinitival clauses (2b) in Hungarian:

- (2) a. (Az) Nem sikerül-t (az), hogy el-startol-j-unk.
that.NOM not succeed-PAST.3SG that.NOM that away-start-SUJU-1PL
‘It was not successful for us that we should start.’
- b. (*Az) Nem sikerül-t (*az) el-startol-n-unk.
that.NOM not succeed-PAST.3SG that.NOM away-start-INF-1PL
‘For us to start was not successful.’

Our account is based on the claim that if the pronoun is present, then it acts as the argument of the matrix modal/evaluative predicate, the clause itself being an adjunct to it. We will show that Hungarian (pro)nominal expressions do not license infinitival adjunct modifiers, and that is why (2b) with the pronoun is ungrammatical.

The structure of the paper is as follows. First, we offer a brief descriptive overview of the syntax of Hungarian modal and evaluative predicates in Section 2. In Section 3, we present our analysis of the functional and categorial status of the infinitival construction type illustrated in (1b), drawing on the results of Rákosi (2004) and Rákosi & Laczkó (to appear). Next, we perform a similar investigation concerning the *that*-clause construction (1c) in Section 4. Special attention is given to the nature of the relation between the pronominal associate and the *that*-clause (2a). In Section 5, we address the problem of the incompatibility of this pronoun with infinitival clauses in Hungarian (2b). We close the paper with some concluding remarks in Section 6.

2. Modal and evaluative predicates in Hungarian: an overview

By *modal predicates* we mean predicates that express (different types of) necessity or possibility, and by *evaluatives* we mean predicates that express some kind of evaluation of an entity of a given semantic type. The two classes are not categorially uniform: we find verbal (3a), adjectival (3b), and nominal predicates (3c) in both.³

- (3) a. *kell*_v *tetszik*_v
‘must, have to, need (to)’ ‘appeal to, please’
- b. *lehetséges*_{adj} *jó*_{adj}
‘possible’ ‘good’
- c. *lehetetlenség*_n *ostobaság*_n
‘impossibility’ ‘silliness’

³ The citation form of Hungarian verbs is their third person singular, present tense, indicative form. This slot in the paradigm is unmarked morphologically. Adjectival and nominal predicates form a complex with the copula *van* ‘be’, which also has a zero form in present tense indicative if the subject is third person.

The individual anchor of the model in which the modal or the evaluative predicate is interpreted is the speaker by default. If it is not the speaker, then it is syntactically encoded as a dative-marked argument. This argument may appear in all the three constructions we have seen in (1). Consider the following examples with the modal predicate *kell* ‘must, need’:

- (4) a. *Egy új otthon kell-ett János-nak.*
 a new home.NOM must-PAST.3SG John-DAT
 ‘John needed a new home.’
- b. *János-nak otthon kell-ett len-ni-e / len-ni.*
 John-DAT home must-PAST.3SG be-INF-3SG / be-INF
 (i) ‘John must have been at home.’
 (ii) ‘John had to be at home.’
- c. *János-nak az kell-ett, hogy otthon legy-en.*
 John-DAT that.NOM must-PAST.3SG that home be.SUJU-3SG
 ‘What John needed was to be at home.’

In (4a) and (4c), the dative argument *Jánosnak* ‘for John’ is an argument of the modal predicate. (4b), however, is ambiguous between a monadic (epistemic) and a dyadic (deontic or circumstantial) reading. On the monadic reading (i), the dative expression is not a semantic argument of the modal. On the dyadic reading (ii), the dative is the semantic as well as the syntactic argument of the modal and it controls the subject slot of the infinitive. This paper focuses on the functional and the categorial properties of the non-dative argument, and the interested reader is referred to the literature listed in Footnote 1 for details concerning the behaviour of the dative argument.

The infinitive in these constructions can be marked for agreement, but this is optional. In actual fact, most modal and evaluative predicates are only seldom used with agreement-marked infinitives in current Hungarian, and the plain infinitive is generally preferred. If the infinitive is agreement-marked, then it shows the full agreement paradigm and it agrees with the dative argument. This agreement phenomenon is discussed in detail in É. Kiss (1987, 2002) and in Tóth (2000, 2001, 2002).

3. Infinitival arguments of modals and evaluatives

3.1. Categorial status

É. Kiss (1987, 2002) develops an account of agreement-marked infinitives in Hungarian which considers the surface similarity between possessive constructions and agreement-marked infinitives essential and treats them on a par. (5a) is a possessive and (5b) is an infinitival construction:

- (5) a. *János-nak sikerül-t a start-ja.*
 John-DAT succeed-PAST.3SG the start-POSS.3SG.NOM
 ‘John’s start was successful.’

- b. *János-nak sikerül-t* *el-startol-ni-a.*
 John-DAT succeed-PAST.3SG away-start-INF-3SG
 ‘John managed to start.’

É. Kiss assumes that the infinitival marker *-n(i)* is a nominalising suffix and she predicts that the attachment of this morpheme creates a nominal shell around the verbal core. Therefore, the syntax of agreement-marked infinitives should be identical in relevant respects to that of nouns.

One immediate problem with her approach is that the same infinitival marker appears on both agreement-marked and plain infinitives. If this marker is a nominalising suffix, then every infinitive is predicted to have nominal properties in Hungarian – but she explicitly restricts the nominal analysis to the agreement-marked domain. We present a number of arguments against the parallel analysis of possessive and agreement-marked infinitive constructions in Rákosi & Laczkó (to appear) and show that the latter differ both in their syntax and morphophonology from the former. Thus, even the allegedly nominal agreement-marked infinitives fail to pattern up with possessive noun phrases.

The evidence we provide here against the claim that agreement-marked infinitives are nominal concerns not the specific details of the parallel analysis but the general distributional asymmetries between nouns and infinitives. We nevertheless contrast agreement-marked infinitives with possessive constructions for expository purposes, but the latter are intended as representatives of noun phrases in general. Besides, agreement-marked infinitives do not differ in their distribution from plain infinitives in their licensing domain: as indicated above (1b & 4b), they are mostly interchangeable. Since É. Kiss restricts her nominal analysis to agreement-marked items, we focus on these, but it should be noted that plain infinitives show the same test results.

European Portuguese offers an interesting comparison as it licenses agreement marking on infinitives. Raposo (1987: 92-95) argues that the agreement marker on infinitives is “an overt pronominal realisation of the category N at the zero-bar level”, and as a consequence, the maximal projection of the infinitive, IP, is also “nondistinct” from NP.⁴

In Portuguese, however, there are good reasons to assume that the external syntax of agreement-marked infinitives is indeed nominal. First, they may take the definite article (Raposo 1987: 96):

- (6) *Nós lamentamos (o) eles terem recebido pouco dinheiro.*
 we.NOM regret the they.NOM have.INF.3PL received little money
 ‘We regret that they have received little money.’

In Hungarian, possessive phrases can co-occur with the definite article as expected, whereas infinitives never can:

⁴ This claim holds for at least Portuguese agreement-marked infinitival clauses of the following types: subject clauses, complements of factive predicates, and adjunct clauses introduced by a preposition.

- (7) a. **(A) start-om nehéz volt.*
 the start-1SG.NOM difficult was
 ‘My start was difficult.’
- b. **(A) startol-n-om nehéz volt.*
 the start-INF-1SG difficult was
 ‘It was difficult for me to start.’

Second, Portuguese agreement-marked infinitives can appear as complements of prepositions (Raposo 1987: 88):

- (8) *Eu entrei em casa [sem [os meninos verem]].*
 I entered the house without the children see.INFL.3PL
 ‘I entered the house without being noticed by the children.’

Hungarian has postpositions, which take nominal complements, but not infinitives:

- (9) a. *A start-om mellett a finis-em is nehéz volt.*
 the start-1SG.NOM besides the finish-1SG.NOM too difficult was
 ‘Besides my start, my finish was also difficult.’
- b. **Startol-n-om mellett be-fut-n-om is nehéz volt.*
 start-INF-1SG besides in-run-INF-1SG too difficult was
 intended reading: ‘Besides starting, it was also difficult for me to finish the race.’

Two coordinated non-plural noun phrases functioning as subjects optionally trigger plural agreement in the preverbal domain in Hungarian:

- (10) *A start-om és a finis-em jól sikerül-t(-ek).*
 the start-1SG.NOM and the finish-1SG.NOM well succeed-PAST.3SG(-3PL)
 ‘My start and my finish were very successful.’

An agreement-marked infinitive cannot be coordinated with a true noun phrase (11a), and two coordinated subject infinitives cannot trigger plural agreement (11b):

- (11) a. **Startol-n-om és a finis-em jól sikerül-t(-ek).*
 start-INF-1SG and the finish-1SG.NOM well succeed-PAST.3SG(-3PL)
 ‘*For me to start and my finish was/were very successful.’
- b. *Startol-n-om és be-fut-n-om jól sikerül-t(*-ek).*
 start-INF-1SG and in-run-INF-1SG well succeed-PAST.3SG(-3PL)
 ‘For me to start and (for me) to run in was/were very successful.’

Thus, in contrast with the Portuguese construction, the Hungarian agreement-marked infinitive does not show any nominal properties. Therefore, it should be treated as a verbal category.

What needs to be decided next for the purposes of this paper is whether Hungarian infinitival clauses are CPs or Ss.⁵ The former option is evidently available for the English infinitive:

- (12) a. *I didn't know [what to do].*
 b. *I didn't know [whether to go].*

Hungarian finite *that*-clauses, as expected, can take the complementizer *hogy* 'that' (see also Subsection 4.1). A WH-expression (generally in focus position) can immediately follow the complementizer (13a), unlike in English (13b).

- (13) a. *Nem tud-t-am [CP hogy hova men-t-él].*
 not know-PAST-1SG that where go-PAST-2SG
 'I didn't know where you had gone.'
 b. *I didn't know (*that/*whether) where you had gone.*

WH-expressions, which are thus within the S projection and not in [Spec, CP] in Hungarian, are licensed in the initial position of infinitival clauses. But infinitival clauses never take a complementizer:

- (14) a. *Nem tud-t-am (*hogy) [S hova men-ni].*
 not be.able-PAST-1SG that where go-INF
 lit. 'I couldn't go anywhere.'

We conclude on the basis of this evidence that Hungarian infinitival clauses are uniformly Ss, and not CPs.⁶

3.2. Functional status

There is no consensus in the literature on the functional properties of the infinitival clauses in question. (4b) is repeated as an illustrative example:

- (4) b. *János-nak otthon kell-ett len-ni-e / len-ni.*
 John-DAT home must-PAST be-INF-3SG / be-INF
 (i) 'John must have been at home.'
 (ii) 'John had to be at home.'

It has been suggested that the matrix predicate is subjectless, and the infinitival clause acts as its complement, cf. Komlósy (1994) and Kenesei (2001).⁷ Another possible analysis is

⁵ There is no evidence for the relevance of an IP projection in the c-structure of the Hungarian clause.

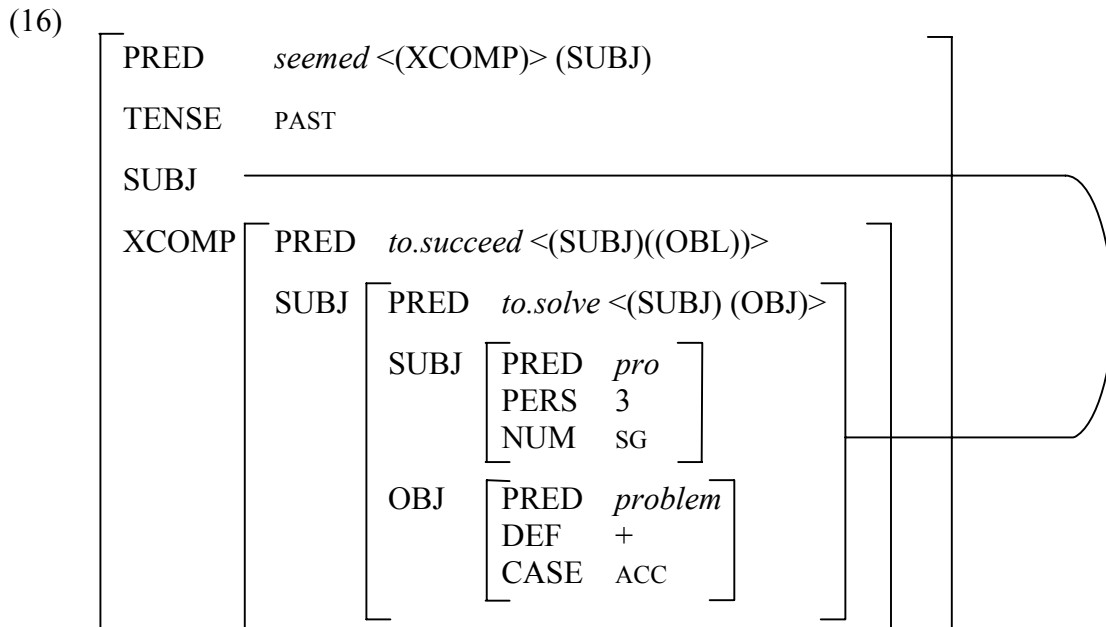
⁶ In future research, we plan to develop a detailed LFG analysis of the structure of finite and non-finite clauses in Hungarian.

⁷ Tóth (2000: 178) also comes to the conclusion that these infinitival clauses are arguments of the matrix modal/evaluative predicate. She refers to them as subject clauses in a footnote, nevertheless she generates them in a complement position.

to take the dative expression to be the matrix subject (Dalmi 2002), on analogy of Icelandic quirky subject constructions, cf., among other works, (Sigurðsson 2002). The infinitival clause is presumably an object then.

Rákosi (2004) argues against both these approaches and claims that it is the infinitival clause itself that is the syntactic subject of the matrix modal/evaluative predicate. The full argumentation can be found there, here we only present an example of a subject raising construction in which the infinitival clause *as a “raised” subject* functionally controls the subject slot of the evaluative predicate *sikerülni* ‘to succeed’. The raising predicate *látszik* ‘seems’ is stress-avoiding, which means it follows its complement to allow it to carry the main stress. The simplified f-structure representation of (15) is in (16).

- (15) *Sikerül-ni látsz-ott [megolda-ni-a a problémá-t].*
 succeed-INF seem-PAST.3SG solve-INF-3SG the problem-ACC
 ‘He seemed to succeed in solving the problem.’
 [lit. ‘For him to solve the problem seemed to succeed.’]



Rákosi (2004) shows that these infinitival arguments are targeted as subjects in a number of Hungarian raising constructions. Thus, in this respect they behave exactly like nominal subjects, cf. (15) and (17).

- (17) *Sikerül-ni látsz-ott a start.*
 succeed-INF seem-PAST.3SG the start.NOM
 ‘The start seemed to succeed.’

Notice that clausal subjects are not restricted to discourse functions in Hungarian, as opposed to, for instance, English (Koster 1978). Clausal and nominal subjects can occur in a preverbal topic position (18a), or they can both follow the verb, in which case they bear no discourse function (18b).

- (18) a. *Startol-n-om / A start-om tényleg nehéz volt.*
 start-INF-1SG the start-1SG.NOM indeed difficult was
 ‘As for starting/the start, it was indeed difficult for me.’
- b. *Tényleg nehéz volt startol-n-om / a start-om.*
 indeed difficult was start-INF-1SG the start-1SG.NOM
 ‘For me to start/the start was indeed difficult for me.’

4. Modals/evaluatives and *that*-clauses

4.1. Categorical status

Though descriptive grammars of Hungarian generally argue that infinitival clauses have nominal properties *because* they can bear nominal functions (SUBJ or OBJ), they do not carry the same reasoning over to the finite *that*-clause arguments of the same predicates, which are thus not considered to be nominal.⁸ We treat these clausal arguments as either CPs or Ss, depending on whether the complementizer *hogy* ‘that’ is present or not.

Whether or not the complementizer can be omitted depends first and foremost on the matrix predicate itself: some predicates license this omission, others disallow it.

- (19) a. *Lehet (hogy) János már megérkez-ett.*
 may.be that John.NOM already arrive-PAST.3SG
 ‘John may have already arrived.’
- b. *Tetsz-ik nek-em, *(hogy) János már megérkez-ett.*
 Please-3SG DAT-1SG that John.NOM already arrive-PAST.3SG
 ‘I like it that John has already arrived.’
 [lit. ‘That John has already arrived pleases me.’]

Besides inter-predicate variation of this kind, the omission of the complementizer is constrained by various other factors (see Kenesei (1994) for an overview). *Hogy* ‘that’ is obligatorily present if the matrix predicate is not verbal (20a), if it is in an adjunct clause (20b), if it is in a subordinate clause whose verbal head is in subjunctive mood (20c), or if its host clause is left-dislocated (20d):

- (20) a. *Lehetséges, *(hogy) János már megérkez-ett.*
 possible that John.NOM already arrive-PAST.3SG
 ‘It is possible that John has already arrived.’
- b. *János el-men-t, *(hogy) hoz-z-on valami-t.*
 John.NOM away-go-PAST.3SG that bring-SUJU-3SG something-ACC
 ‘John left to bring something.’

⁸ The Minimalist analysis of Lipták (1998) relies, among other things, on the assumption that the C head of an argument clause carries a +D/N categorial feature. This feature is utilized in her account of long focus raising. We present a different account of this phenomenon in Subsection 4.2. Besides, the general LFG principles do not demand the nominal treatment of a clausal category even if it bears a SUBJ function.

- c. (*Nek-ünk*) *sikerül-t*, *(*hogy*) *jól* *el-startol-j-unk*.
 DAT-1PL succeed-PAST.3SG that well away-START-SUJU-1PL
 ‘It was successful for us that we should start well.’
- d. *(*Hogy*) *János* *már* *megérkez-ett*, *az* *nem* *lehet*.
 that John.NOM already arrive-PAST.3SG that.NOM not may.be
 ‘That John has arrived is not possible.’

We assume that in the presence of the complementizer, finite clauses are CPs in Hungarian, and they are Ss in the lack of it. As is evident, the choice between the two categorial types is subject to lexical as well as configurational factors.⁹

4.2. That-clauses and pronominal associates at f-structure

As already pointed out in the introduction (see example (2b)), *that*-clauses can have pronominal associates in Hungarian. Here is another example:

- (21) *Tényleg szükséges* (*az*), *hogy János* *itt* *legy-en*.
 indeed necessary that.NOM that John.NOM here be.SUJU-3SG
 ‘It is indeed necessary that John be here.’

There are two approaches available to the categorial status of *az* ‘that’ in the literature. On the one hand, it can be regarded as an expletive (Kenesei (1994); Lipták (1998)). On the other hand, one can take it to be a bona fide pronoun (É. Kiss (1987, 2002); Tóth (2000)). We follow this latter approach and take *az* ‘that’ to be not an expletive in these constructions, but a fully-fledged pronoun with a PRED feature of its own.¹⁰ This claim is based on the following considerations.

First, *az* can indeed occur on its own as an ordinary demonstrative pronoun:

- (22) *Csak AZ* *szükséges*.
 only that.NOM necessary.
 ‘Only THAT is necessary.’

Second, as a *that*-clause associate, it typically occurs in discourse functions. For instance, it is the FOCUS of the matrix clause in (23a) and as such, it carries sentential stress. The *that*-clause itself cannot be focussed for prosodic reasons (23b):

- (23) a. *Csak AZ* *szükséges*, *hogy János* *itt* *legy-en*.
 only that.NOM necessary that John.NOM here be.SUJU-3SG
 ‘What is only necessary is that John be here.’
- b. **Csak* [*hogy János* *itt* *legy-en*] *szükséges*.
 only that John.NOM here be.SUJU-3SG necessary
 intended reading: ‘What is only necessary is that John be here.’

⁹ We continue to use the term *that*-clause to refer to the clause type in question, irrespective of the presence or absence of the complementizer.

¹⁰ Similar claims can be found with respect to other languages in, among other works, Hoekstra (1983) and Bennis (1986) for the Dutch *het*, and in Berman (2001) for the German *es*.

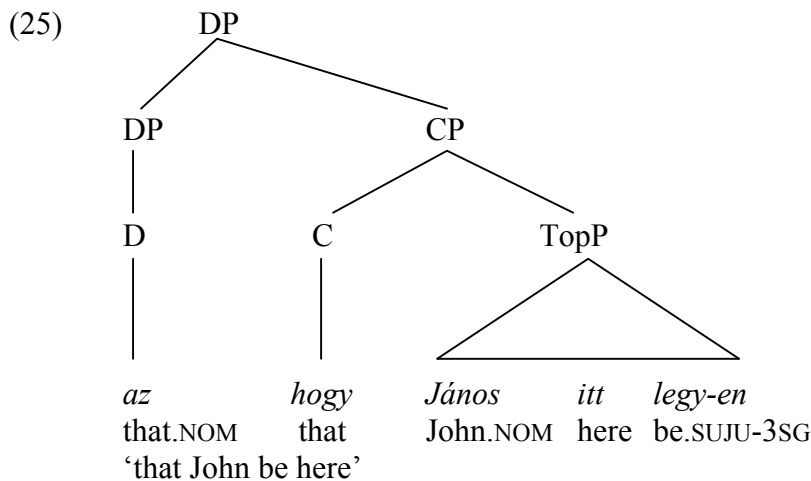
As Tóth (2000) also points it out, true expletives cannot be stressed, whereas pronouns obviously can.

Third, in appropriate discourse settings, *az* can be replaced by its proximal counterpart, *ez* ‘this’:

- (24) *Tényleg EZ szükséges, hogy János itt legy-en?*
 really this.NOM necessary that John.NOM here be.SUJU-3SG
 lit.: ‘Is this really necessary that John be here?’

These data all point towards the conclusion that *az* is a pronoun and not an expletive in these constructions.

It is reasonable to think that this pronoun is the subject argument and the *that*-clause is an adjunct to it. This claim is made by Tóth (2000), who assumes that if no overt pronoun is present, then the subject of the modal/evaluative predicate is a *pro*. Thus the *that*-clause is always an adjunct, whether there is an overt pronominal *az* subject or not. In É. Kiss’s (1987, 2002) analysis, the pronoun and its associate *that*-clause form a complex noun phrase, the latter being an argument clause that bears an appositive relation to the former. The structure she would assign to the complex in (21) conceived of in this way is as follows (based on the structure she provides in É. Kiss (2002: 235)):



She assumes that the clause can be extraposed from this complex noun phrase – an operation that proves to be the norm rather than an exception. The DP-layer is also projected if *az* is not present, in which case a phonologically unrealized pronoun occupies the D head (as in (19), for instance). Consequently, the *that*-clause is always in apposition inside a DP-shell.

Contra both these approaches, but in line with Berman’s (2001) analysis of related German constructions, we propose that in the presence of the pronoun, the *that*-clause is indeed an adjunct, but in the absence of it, the *that*-clause is the SUBJ of the matrix

predicate. This gives a straightforward explanation for long focus raising facts in Hungarian, which are briefly summarized below.¹¹

Long focus raising is the descriptive term denoting the operation in which material from an embedded clause is focussed in the matrix clause. It is possible from the finite *that*-clause arguments of bridge-verbs in Hungarian, but it is always incompatible with the presence of *az*:¹²

- (26) *Csak JÁNOS-t sikerül-t (*az), (*#) hogy lerajzol-j-am.*
 only John-ACC succeed-PAST.3SG that.NOM that draw-SUJU-1SG
 ‘It is only John who I succeeded in taking a picture of.’

In É. Kiss’s (1987, 2002) analysis, the presence of the pronoun blocks extraction (ie., long focus raising), since it violates the complex noun phrase constraint, cf. (25). She needs to stipulate, however, that focus raising is grammatical in the absence of the pronoun as “a projection containing no phonologically realized material is transparent for subjacency” (É. Kiss: 2002, 253).

For a different perspective, consider the following. As indicated in (26), there cannot be an intonational break between the two clauses if focus raising takes place (Gervain 2002: 48-49). Such a prosodic boundary is grammatical, however if the pronoun associate is present (27), and the same is true of adjunct clauses (28):

- (27) *Csak AZ sikerül-t, (#) hogy lerajzol-j-am János-t.*
 only that.NOM succeed-PAST.3SG that draw-SUJU-1SG John-ACC
 ‘What succeeded only was for me to draw a picture of John.’
- (28) *Jö-tt-em, (#) hogy lerajzol-j-am János-t.*
 come-PAST-1SG that draw-SUJU-1SG John-ACC
 ‘I have come to take a picture of John.’

Furthermore, as is well-known, adjunct clauses do not license long focus raising:

- (29) **Csak JÁNOS-t jö-tt-em, hogy lerajzol-j-am.*
 only John-ACC come-PAST-1SG that draw-SUJU-1SG
 intended reading: ‘I have come to take a picture only of John.’

On our account, the pronoun associate is predicted to be ungrammatical in a focus raising construction (26) because in its presence the *that*-clause is an adjunct and not an argument of the matrix predicate. If the pronoun is absent, the clause is the subject argument of the matrix predicate, and the possibility of focus raising follows. The prosodic similarity between (27) and (28) also derives from the fact that both *that*-clauses are adjuncts.

¹¹ Some of the works that contain more detailed descriptions of long focus raising phenomena in Hungarian are É. Kiss (1987, 2002), Gervain (2002), Kenesei (1994) and Lipták (1998).

¹² Unlike in German, where at least psych-verbs require the presence of *es* in long focus raising constructions, cf. Berman (2001).

It has to be added though that long focus raising from subject clauses is much less acceptable if the clause is in indicative mood than if it is in subjunctive mood, cf. (26) and (30):

- (30) *??Csak JÁNOS-t tetsz-ik, hogy lerajzol-od.*
 only John-ACC please-3SG that draw-2SG
 ‘It is only your drawing a picture of John that pleases me.’

(30) also contrasts with indicative object clauses that license focus raising:

- (31) *Csak JÁNOS-t mond-t-am, hogy lerajzol-om.*
 only John-ACC say-PAST-1SG that draw-2SG
 ‘It is only John that I said I would draw a picture of.’

Nevertheless, subject clauses can license focus raising in Hungarian. This can be explained if we assume with Davis & Dubinsky (2001) that subjects are not islands in languages in which they are not required to have nominal properties. As we have demonstrated that Hungarian clausal subjects are not required to have nominal properties (Subsection 3.1.), it follows that they are not necessarily islands. Still, the degraded acceptability of focus raising from indicative subject clauses needs to be explained. It seems to be a feasible generalization that indicative mood, as opposed to subjunctive mood, is not a sufficient trigger for clause union effects if the subordinate clause is a SUBJ. An appropriate account of this variation, however, lies beyond the confines of the present paper. What is important to notice is that the grammar of Hungarian allows for long focus raising from subject clauses.¹³

Finally, let us notice that the behaviour of *az* and its associate *that*-clause parallels the behaviour of pronouns and their postmodifying adjunct relative clauses in Hungarian:

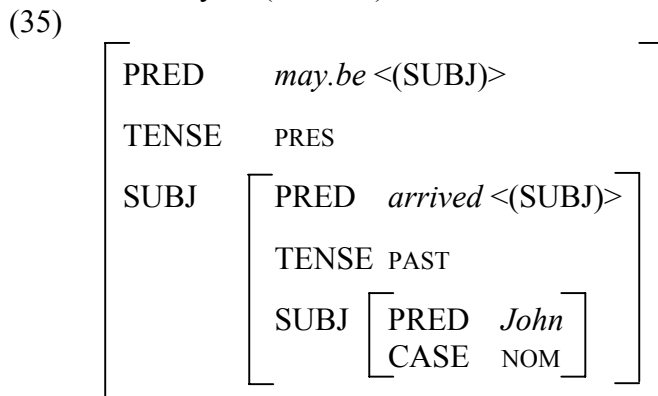
- (32) a. *Az, ami-t én csinál-t-am, nem sikerül-t.*
 that.NOM which-ACC I.NOM do-PAST-1SG not succeed-PAST.3SG
 ‘What I did was not successful.’
 b. *Az, hogy jól startol-j-unk, nem sikerül-t.*
 that.NOM that well start-SUJU-1PL not succeed-PAST.3SG
 ‘It was not successful that we should start well.’
- (33) a. *Csak AZ nem sikerül-t, ami-t én csinál-t-am.*
 only that.NOM not succeed-PAST.3SG which-ACC I.NOM do-PAST-1SG
 ‘What wasn’t successful was only what I did.’
 b. *Csak AZ nem sikerül-t, hogy jól startol-j-unk.*
 only that.NOM not succeed-PAST.3SG that well start-SUJU-1PL
 ‘What wasn’t successful was only that we should start well.’

¹³ Lipták (1998:96) also observes that native speakers do not uniformly accept long focus raising from subject clauses. She notes that it is possible that the matrix predicates in these constructions are on the way to becoming bridge-predicates. We think it is more probable that it is the subjecthood of these clauses that makes focus raising somewhat more marked than in the case of object clauses. This problem, however, needs further investigation.

Both clause types can occur string adjacent to their associate pronouns (32), or they can be separated at c-structure (33). This, we believe, provides further support for our analysis, which treats the *that*-clause uniformly as an adjunct in the presence of an associate pronoun.

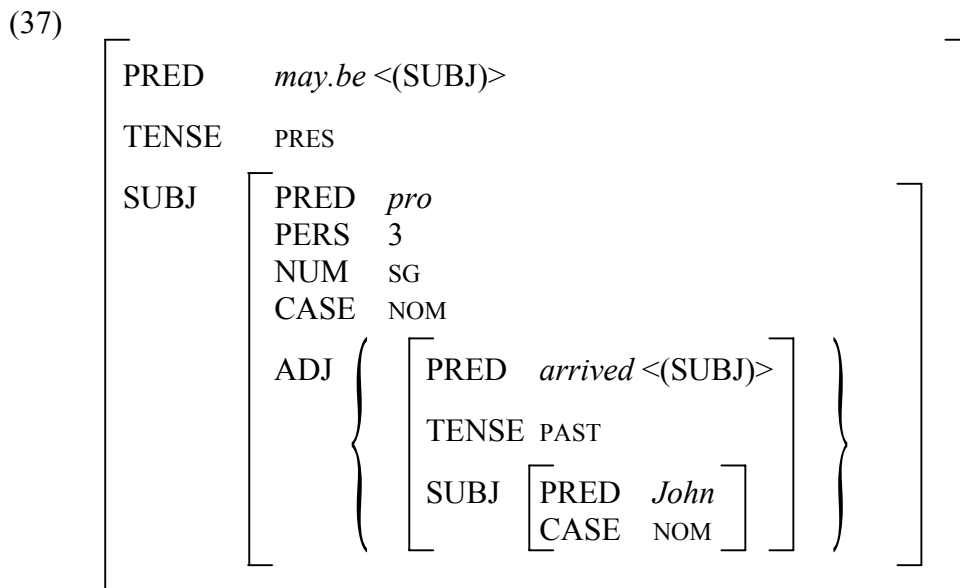
For the sake of an interim summary and the demonstration of the analysis so far, let us consider the following two modal constructions. In (34), the subordinate clause is the subject of the matrix modal predicate *lehet* ‘may be’. The simplified functional structure of this sentence is in (35).

- (34) *Lehet, hogy János megérkez-ett.*
 may.be that John.NOM arrive-PAST.3SG
 ‘It may be (the case) that John has arrived.’



In (36), the pronoun *az* ‘that’ is the subject argument of the matrix predicate, and the *that*-clause functions as an adjunct to it. (37) is the f-structure we assign to this construction.

- (36) *Az lehet, hogy János megérkez-ett.*
 that.NOM may.be that John.NOM arrive-PAST.3SG
 ‘It may be (the case) that John has arrived.’



We can draw a parallel between (37) and É. Kiss's structure in (25). There are, however, two important differences between our approach and hers. First, for us, the absence of the pronoun *az* does matter and the construction with the pronoun is assigned a completely different functional analysis than the one with it, cf. (35). Second, the fact that the pronoun and its associate *that*-clause form a functional unit does not require the two to form a constituent since the relation between f-structures and their corresponding c-structure is not a function. As we have seen, the pronoun and the *that*-clause do not usually occur string-adjacent at c-structure, which É. Kiss can only account for by assuming that the latter regularly undergoes extraction from the complex noun phrase in which it is claimed to be generated.

5. Modals/evaluatives and *that*-clauses

It has been noted in the introduction that infinitives cannot have pronominal associates in Hungarian. (2) is repeated here to illustrate this point.

- (2) a. (*Az*) *Nem sikerül-t* (*az*), *hogy el-startol-j-unk.*
 that.NOM not succeed-PAST.3SG that.NOM that away-start-SUJU-1PL
 ‘It was not successful for us that we should start.’
- b. (**Az*) *Nem sikerül-t* (**az*) *el-startol-n-unk.*
 that.NOM not succeed-PAST.3SG that.NOM away-start-INF-1PL
 ‘For us to start was not successful.’

Now we are in the position to reconsider this problem from the perspective of the analysis presented in the previous section and formulate it as the question of why it is not possible in Hungarian for infinitival clauses to be adjuncts (or adpositions) to pronouns.

In English, infinitival clauses can have an associate pronominal-type expletive (38). Notice that it is also grammatical for nouns (39) and even for pronouns (40) to be modified by an infinitival adjunct:

- (38) *It is good to read books.*
- (39) a. *The book to read by tomorrow is on the table.*
 b. *They obeyed the command to evacuate.*
- (40) a. *The ones to watch are the ones you never hear about.*
 b. *The workers are the first ones to suffer.*

The Hungarian equivalents of these constructions are all ungrammatical. (41) contrasts with (38) in the way already explicated, and (42)-(43) contrast with (39). The two sentences in

(39) do not have a structurally equivalent Hungarian counterpart, whether the infinitive is placed behind (*a*-examples) or in front of (*b*-examples) the noun head.¹⁴

- (41) **Az jó könyv-ek-et olvas-ni.*
 it.NOM good book-PL-ACC read-INF
 intended reading: ‘It is good to read books.’
- (42) a. **Teljesít-ett-ék a parancs-ot evakuál-ni.*
 obey-PAST-3PL the order-ACC evacuate-INF
 b. **Teljesít-ett-ék az evakuál-ni parancs-ot.*
 obey-PAST-3PL the evacuate-INF order-ACC
 intended reading of both: ‘They obeyed the command to evacuate.’
- (43) a. **A könyv holnap-ra el-olvas-ni az asztal-on van.*
 the book.NOM tomorrow-SUBL PV-read-INF the table-SUP is
 b. **A holnap-ra el-olvas-ni könyv az asztal-on van.*
 the tomorrow-SUBL PV-read-INF book.NOM the table-SUP is
 intended reading of both: ‘The book to read by tomorrow is on the table.’

On the other hand, participial clauses are allowed to premodify noun phrases (44), and *that*-clauses, as expected, can also form a constituent with a preceding nominal head (45):

- (44) a. *a könyv-et olvas-ó fiú*
 the book-ACC read-ÓPART boy
 ‘the boy reading a book’
 b. *a János által olvas-ott könyv*
 the John by read-TPART book
 ‘the book read by John’
- (45) *Teljesít-ett-ék a parancs-ot, hogy evakuál-j-anak.*
 obey-PAST-3PL the order-ACC that evacuate-SUJU-3PL
 lit. ‘They obeyed the order that they should evacuate.’

The appropriate descriptive generalisation is that nominal categories can have clause-level adjuncts in Hungarian (in a pre- or post-head position, depending on the categorial properties of the head of the clause) as long as this clause is not headed by an infinitive. This incompatibility is best encoded in the lexical form of infinitives as a categorial constraint on the *f*-structure which includes that of the infinitival clause. This constraint can be expressed with the CAT predicate of Kaplan and Maxwell (1996):

¹⁴ The English pronoun *one(s)* has no Hungarian equivalent. In Hungarian a special elliptical construction is used instead. The closest Hungarian counterpart of (40b), for instance, is (i), but even that is ungrammatical.

(i) **A munkás-ok az első-k szenved-ni.*
 The worker-PL the first-PL suffer-INF
 ‘The workers are the first ones to suffer.’

(46) $V_{INF}: \{D, N\} \notin \text{CAT}((\text{GF}\uparrow))$

(46) constrains infinitives not to have a pronominal associate in Hungarian, as infinitival clauses cannot be adjuncts to nominal categories. Thus, the ungrammaticality of (2b) and (41) can be reduced to more general regularities in the grammar of Hungarian.

6. Conclusions

We have argued in this paper that modal and evaluative predicates in Hungarian subcategorize for a subject argument which can be realized categorially as a noun phrase, as a finite *that*-clause, or as an infinitival clause. That there are no special categorial restrictions on the realisation of an argument is expected in LFG, as predicates subcategorise for arguments of a particular functional, and not of a particular categorial type. This functional uniformity behind the categorial diversity is generally not acknowledged by most generative approaches to this Hungarian construction.

In fact, it has been suggested that these subject clauses have a nominal shell and, therefore, show the external syntax of noun phrases (É. Kiss 1987, 2002). It has been shown here that the nominal analysis of either agreement-marked or plain infinitives in Hungarian fails to give the right predictions. Whereas the nominal analysis of infinitives has strong empirical support in, for instance, Portuguese, the Hungarian infinitive has to be regarded as a construction of solely verbal properties. Subject *that*-clauses may have a pronoun associate, in which case the clause is an adjunct and forms a functional unit with the pronoun. In the absence of the pronoun, the clause itself is the subject argument and it is not considered to have any nominal properties. This analysis gives the right predictions for focus raising phenomena, and, together with the observation that infinitival clauses cannot be adjuncts to noun phrases in Hungarian, it helps to explain why a pronominal associate is not licensed with argument infinitival clauses in Hungarian.

We assume that Hungarian is a mixed language in the sense of Dalrymple & Lødrup (2000), ie., both CPs and $[\pm\text{fin}]$ Ss can have either nominal (SUBJ, OBJ), or propositional (COMP, XCOMP, ADJ, XADJ) functions. We believe in the usefulness of the COMP function and intend to demonstrate in future research that OBJ and COMP clauses need to be distinguished in Hungarian. We have not shown how propositional arguments can be treated in LMT for Hungarian, and, in particular, how clausal arguments map onto SUBJ, but work on the mapping proposal is in progress.

References

- Bennis, Hans 1986. *Gaps and dummies*. Dordrecht: Foris.
Berman, Judith 2001. "On the Cooccurrence of Es with Finite Clause in German: an LFG Analysis". In Meurers, Walt Detmar & Kiss, Tibor eds. *Constraint-Based Approaches to Germanic Syntax*. Stanford: CSLI. 7-30.
Bresnan, Joan 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.
Dalmi, Gréte 2002. *The Role of AgrP in Non-Finite Predication*. PhD dissertation, Budapest, ELTE.

- Dalrymple, Mary & Lødrup, Helge 2000. "The Grammatical Functions of Complement Clauses". In Butt, Miriam & King, Tracy Halloway eds. *Proceedings of the LFG00 Conference, University of California, Berkeley*. Stanford: CSLI Publications. <http://csli-publications.stanford.edu/LFG/5/lfg00dalrympl-lodrup.pdf>
- Davis, William D. & Stanley Dubinsky 2001. "Functional Architecture and the Distribution of Subject Properties". In Davis, William D. & Dubinsky, Stanley eds. *Objects and Other Subjects – Grammatical Functions, Functional Categories and Configurationality*. Dordrecht: Kluwer. 247–279.
- Hoekstra, Teun 1983. "The Distribution of Sentential Complements". In Bennis, Hans & Klooke, W. U. S. van Lessen eds. *Linguistics in the Netherlands*. Dordrecht: Foris, 93–103.
- Kaplan, Ronald M. & Maxwell, John T. 1996. *LFG Grammar Writer's Workbench*. Technical Report, Xerox Palo Alto Research Center. <ftp://ftp.parc.xerox.com/pub/lfg/lfgmanual.ps>
- Kenesei, István 2001. "Criteria for Auxiliaries in Hungarian". In *Argument structure in Hungarian*. Kenesei, István eds. Budapest: Akadémiai Kiadó. 79–111.
- É. Kiss, Katalin 1987. *Configurationality in Hungarian*. Dordrecht: Reidel.
- É. Kiss, Katalin 2002. *The Syntax of Hungarian*. Cambridge: CUP.
- Komlósy, András 1994. "Complements and Adjuncts". In Kiefer, Ferenc & É. Kiss, Katalin eds. *The Syntactic Structure of Hungarian (= Syntax and Semantics 27)*. San Diego: Academic Press. 91–178.
- Koster, Jan. 1978. "Why Subject Sentences don't Exist". In Keyser, Jay S. ed. *Recent Transformational Studies in European Languages*. The MIT Press: Cambridge, Massachusetts. 53–64.
- Rákosi, György 2004. "Infinitival Clauses as Syntactic Subjects in Hungarian". In Blaho, Sylvia; Vicente, Luis & de Vos, Mark ed. *Proceedings of Console XII*. http://www.sole.leidenuniv.nl/content_docs/ConsoleXII2003pdfs/rakosi-2003.pdf
- Rákosi, György & Laczkó, Tibor. To appear. "The Categorical Status of Agreement-Marked Infinitives in Hungarian". In Piñón, Christopher; Siptár, Péter & Szentgyörgyi, Szilárd eds. *Approaches to Hungarian 10*.
- Raposo, Eduardo 1987. "Case Theory and Infl-to-Comp: The Inflected Infinitive in European Portuguese". *Linguistic Inquiry* 18(1). 85–109.
- Sigurðsson, Haldór Ármann. 2002. "To Be an Oblique Subject: Russian vs. Icelandic". *Natural Language and Linguistic Theory* 20. 691–724.
- Tóth, Ildikó 2000. *Inflected Infinitives in Hungarian*. PhD dissertation, Tilburg University.
- Tóth, Ildikó 2001. "Impersonal Constructions and Null Expletives". In Kenesei, István ed. *Argument structure in Hungarian*. Budapest: Akadémiai Kiadó. 51–78.
- Tóth, Ildikó 2002. "Can the Hungarian Infinitive Be Possessed?" In Kenesei, István & Siptár, Péter eds. *Approaches to Hungarian 8*. Budapest: Akadémiai Kiadó. 133–160.

CONSTRUCTING A PARSED CORPUS WITH A LARGE LFG GRAMMAR

Victoria Rosén, Paul Meurer and Koenraad de Smedt
University of Bergen and AKSIS

Proceedings of the LFG'05 Conference
University of Bergen
Miriam Butt and Tracy Holloway King (Editors)

2005
CSLI Publications
<http://csli-publications.stanford.edu>

Abstract

The TREPIL project (Norwegian treebank pilot project 2004-2008) is aimed at developing and testing methods for the construction of a Norwegian parsed corpus. Annotation of c-structures, f-structures and mrs-structures is based on automatic parsing with human validation and disambiguation. Parsing is done with a large LFG grammar and the XLE parser. We propose a method for efficient disambiguation based on discriminants and we have implemented a set of computational tools for this purpose.

1 Treebanks and parsed corpora

We use the term *treebank* to mean a corpus annotated with sentence structures beyond the part of speech level. Even though the term refers to syntactic tree structures, it is in current usage extended to corpora with all kinds of structural annotation at syntactic and even semantic levels, such as constituent structures, grammatical functions, or predicate-argument relations (Nivre, De Smedt, and Volk, 2005). Our work in the context of the TREPIL project (Norwegian treebank pilot project 2004-2008) is aimed at developing and testing methods for the construction of a Norwegian treebank based on deep parsing. Before going into details about this project and its results so far, we first provide some background on previous related work.

Currently, linguists and language engineers have easy access to large text and speech corpora, many of which are annotated at the word level, mostly by parts of speech. Although searching in large corpora for certain words and sequences of words with given categories may yield valuable information, two problems can be discerned (Abeillé, 2003). Firstly, part of speech tagging is of limited use to syntacticians, as it fails to distinguish boundaries of clauses, of phrases and even of compound words that are written separately. Secondly, as automatic part of speech tagging is normally based on a shallow analysis or statistical processor, the quality of the annotated corpus is likely to be unsatisfactory.

To overcome the limitations of corpora with word-level annotation only, efforts have been made towards more sophisticated linguistic annotation of corpora. Whereas the first syntactically annotated corpora were developed mostly with manual methods, the development of more sophisticated linguistic models prompted the application of such models to treebank construction (Abeillé, 2003). This has led to the term *parsed corpus* which is usually reserved for a treebank that is grounded in a computational grammar model. Treebank construction on the basis of automatic parsing with a computational grammar is desirable for both practical and theoretical reasons. Indeed, manual annotation has the disadvantage of being costly and prone to human error, and it is difficult to achieve satisfactory consistency both within and between human annotators (van der Beek et al., 2002a). Moreover, an annotation scheme which is only verbally defined and is not grounded in a computational grammar model risks isolating the corpus from the very applications for which it could be useful.

Fully automatic annotation, on the other hand, only works to the extent that the analyses chosen by the parser are correct. Since perfect coverage is not attainable in practice, many current approaches to treebank construction are semi-automatic, in the sense that parser output is validated by a human annotator. Furthermore, automatic parsing usually produces more than one possible analysis, since many sentences can be analyzed in a variety of ways that may be infelicitous and can neither be excluded on purely syntactic grounds nor completely avoided by statistical learning techniques. Therefore, manual disambiguation is a necessity. In the Alpino treebank, for instance, the corpus is automatically parsed with a dependency grammar, assisted by interactive tools for manual checking, including disambiguation and extension of the lexicon (van der Beek et al., 2002b). The TREPIL project has devoted significant efforts to disambiguation methods, as discussed below.

Several treebank projects have prominently used the LFG and HPSG formalisms. The PARC 700 Dependency Bank (King et al., 2003) was constructed by a two-step approach. First, a corpus was parsed with an LFG grammar and the best parse for each sentence was chosen manually and stored. Then, from each

stored functional structure, a corresponding dependency structure was automatically derived, modified as needed, and validated by a human annotator. The PARC 700 has subsequently been used as an external standard in the evaluation of other f-structure annotations (Burke et al., 2004b). One of the points illustrated by the PARC 700 is that a treebank constructed by parsing with a certain language model (in this case, based on the LFG formalism) nevertheless can be convertible into different linguistic models. A different example of treebanking by conversion is the derivation of Estonian phrase structures on the basis of Constraint Grammar function tags (Bick, Uibo, and Müürisep, 2004). It may perhaps be concluded that aspirations for neutrality with respect to grammatical theory are as unnecessary as they are illusory.

Treebanks are currently receiving a lot of attention because they provide highly valuable empirical data for many research questions in linguistics and language technology (Nivre, De Smedt, and Volk, 2005). Provided they are composed and annotated as reference corpora rather than special purpose collections, treebanks allow for multiple uses in the various sciences of language as well as in language technology. Linguists may want to search for examples or counterexamples of syntactic constructions under investigation, whereas psycholinguists may be interested in relative frequencies of various possible attachments of prepositional phrases or relative clauses (Abeillé, 2003). Formal and computational linguists can evaluate the correctness and coverage of grammars and lexicons against the analyses stored in a treebank, and at a more general level, the adequacy of linguistic theories and formalisms can be assessed (Bouma, 2004).

From the grammatical information stored in treebanks, other resources such as grammars and lexicons can be induced. Stochastic grammars can be trained using frequency information about the parse choices. Other research has focused on the induction of LFG grammars from existing manually annotated treebanks, with the goal of deriving robust, wide-coverage grammars from treebanks rather than having to hand code them (Burke et al., 2004a). Claims that the performance of automatically induced LFG grammars may surpass that of hand-coded LFG grammars (Cahill, 2004) have to be weighed against questions concerning the generality and explanatory power of the linguistic theories embodied by the induced grammars.

More important than the particular formalism used is the level of detail of the analyses in treebanks. While some of the earliest syntactically annotated corpora contain syntactic boundaries, others contain for instance constituent structures (Abeillé, Clément, and Toussnel, 2003), functional dependency structures (Hajič, 1998) or, in addition to syntactic structures, also predicate-argument structures (Marcus et al., 1994). TREPIL is cooperating with the LOGON project on Norwegian-English machine translation (Oepen et al., 2004), which has produced a small treebank containing semantic structures. In the context of translation and contrastive linguistics, we also want to mention the potential of parallel treebanks of translated texts, where detailed and deep analyses offer an interesting domain of study.

2 Treebanking goals in the TREPIL project

The TREPIL project is a research project on treebanking methods, aimed at building a Norwegian parsed corpus. The current project is a preparatory project; it will not produce a full-scale treebank, but a methodology, a set of computational tools, and a demonstration corpus. Our hope is that the resulting methods and tools will be put to use in a subsequent project for building a large scale Norwegian treebank which will form part of a future Norsk Språkbank (Norwegian Language Bank).

The TREPIL project uses the LFG formalism and explores a tight relation between a grammar and a corpus, but our focus is different from earlier LFG-banking projects. Our method for treebank construction is based on the testing and further improvement of an existing hand-coded grammar and parser, and its extension with additional treebanking tools, primarily for disambiguation. For this purpose, we use the NorGram LFG grammar for Norwegian, together with the Xerox Linguistic Environment (XLE). Our motivation for using NorGram as a starting point is twofold. Firstly, NorGram is currently the only deep grammar for Norwegian with large coverage. Secondly, the grammar is developed in the international ParGram project

(Butt et al., 2002) which attains a certain level of generality across languages through agreements on similar feature structures and the existence of a transfer formalism for f-structure based translation. However, we do not want to overemphasize the choice of formalism for reasons outlined above.

An innovative characteristic of TREPIL is that, in contrast to the single stratum approaches of most other treebanks, the Norwegian grammar generates three separate but interrelated structures for each sentence: a constituent structure, a functional structure, and a semantic structure. The semantic projection is based on Minimal Recursion Semantics (MRS) (Copestake et al., in preparation), which allows a deeper level of semantic description than the predicate-argument coding in the Penn treebank (Marcus et al., 1994). MRS represents the semantics of a sentence as a bag of elementary predications, underspecified for scope. The mrs-structures are derived by co-description and may contain information that cannot be derived from the c- or f-structures, such that the mrs-projection represents an autonomous level of structure.

The triple stratum annotation generated by our grammar represents a rich, layered description of the syntax and semantics of each sentence, which allows for multiple uses. However, this sophistication comes at a price, because disambiguation and validation are nontrivial and manual annotation would be quite difficult and inefficient. Thus, our treebanking method is strongly dependent on computational systems, including an efficient parser and a grammar of which the coverage and precision is being continually improved.

One sometimes comes across scepticism with respect to the possibility of deep (full) parsing, often by adherents of shallow parsing as an approximation. However, it has also been pointed out that much of the scepticism is unwarranted for the XLE parser (Zaenen, 2004). Although full parsers may be slower, the XLE parser is still fast enough for off-line parsing of a corpus. Also, it has been pointed out that some full parsers are too brittle to deal with anomalous input, but the XLE parser allows fragment parses, so any input can receive some form of analysis. It has also been claimed that full parsers may yield so many parses that applications have difficulty coping with them. While this problem is being addressed by the inclusion of probability based disambiguation components, such automatic disambiguation is not feasible in our situation, since it would need to be bootstrapped from a treebank which is not yet built. Instead, we are focusing our attention on an altogether different method aimed at highly efficient manual disambiguation. This method, which will be the primary focus of the next section of this paper, is supported by a computational tool which we have implemented in a working first version.

Also, we are working towards a system which allows the automatic reanalysis of the corpus as the grammar develops. In as far as TREPIL involves the synchronous evolution of a treebank and a grammar, our approach is similar to that of LinGO Redwoods (Oepen et al., 2003), which is based on HPSG and the LKB parser environment. LinGO has developed a set of advanced tools that allow the automated update of the treebank after reparsing with a new version of the grammar, but without having to fully disambiguate the corpus over again. This is achieved by reapplying earlier recorded choices by the annotator in the selection of the preferred parse, based on techniques proposed by Carter (1997). A crucial point is that not only the preferred analysis of the sentence is recorded, but all decisions made as part of the annotation in the database.

We believe there are important methodological advantages to our approach. Instead of building a treebank incrementally and improving the grammar independently, we develop an efficient way to successively reannotate the corpus with each version of the grammar, thus obtaining a parsed corpus that is fully consistent with the grammar. The end result is therefore not only a treebank, but also a grammar that can be deployed in other applications, for example machine translation, especially since it produces semantic analyses. In view of such applications, we believe it is advantageous to retain manual control over the grammar in order to obtain the kind of abstraction and readability required by a linguist, rather than inducing an entirely new grammar from the treebank.

3 Disambiguation with XLE

One of the main challenges in using parser output for treebanking is selecting the desired parse among a potentially large number of parses. It is worth remembering that the number of parses is exponential to the number of ambiguities, such that up to 2^n analyses may be produced for n binary choices. Six unresolved, independent parsing choices can for instance give rise to 64 analyses. Consequently, a disambiguation strategy that concentrates on local ambiguities might be more efficient than one that only looks at the whole set of resulting analyses.

XLE has a built-in facility for disambiguation in the form of packed representations. When a sentence is parsed, XLE displays the analyses one at a time in the c-structure and f-structure windows. This allows the user to browse through all the analyses and inspect each c-structure and its corresponding f-structures in turn. In addition, *f-structure chart* windows show *packed* representations of all analyses. There are two different formats in which this compact information is shown, but we will concentrate on the f-structure chart window that indexes the analyses by constraints, providing a view of choices listed as alternatives. When a sentence contains a single ambiguity, this type of representation makes it easy to spot the source of the ambiguity, as shown in figure 1 for example 1.

- (1) *Hun er barn.*
 she is child/children
 “She is a child.” / “She is children.”

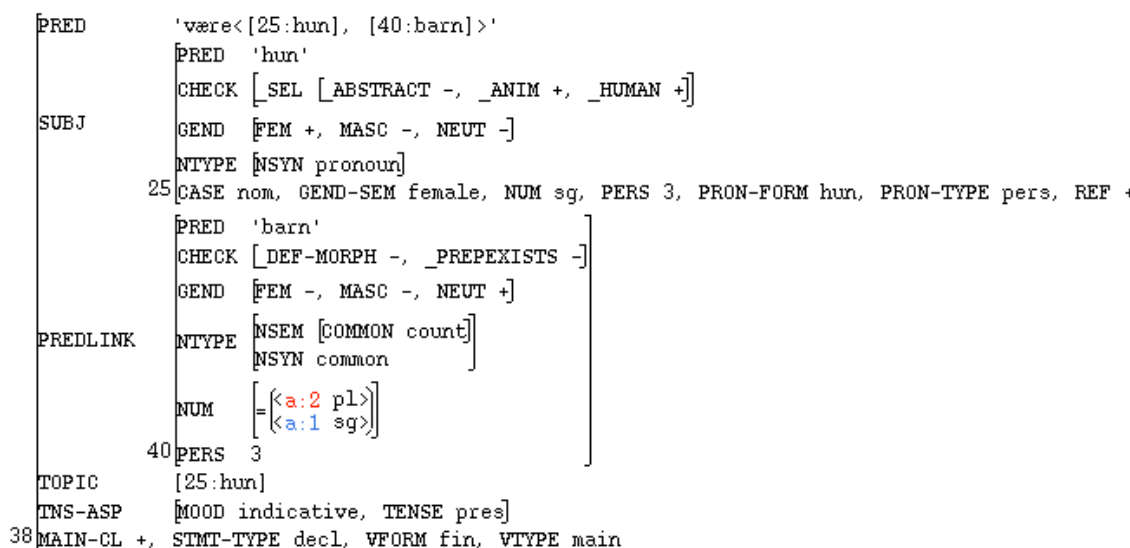


Figure 1: F-structure chart for (1) *Hun er barn.*

This sentence has two analyses, identical except for the value of the feature number, which may be singular or plural. In the f-structure chart, the two values are displayed as alternatives, labeled with indices *a:1* and *a:2*. The choices in these windows are active, so that the user can click on a choice and have a solution corresponding to it displayed in the c-structure and f-structure windows. This facility is easy to use for disambiguation when there are only a few choices.

The sentence in example 2 has two local ambiguities, resulting in four analyses. The packed f-structure chart is shown in figure 2.

- (2) *Hun kjøper klær i den dyre butikken.*
 she buys clothes in the expensive store
 “She buys clothes in the expensive store.”

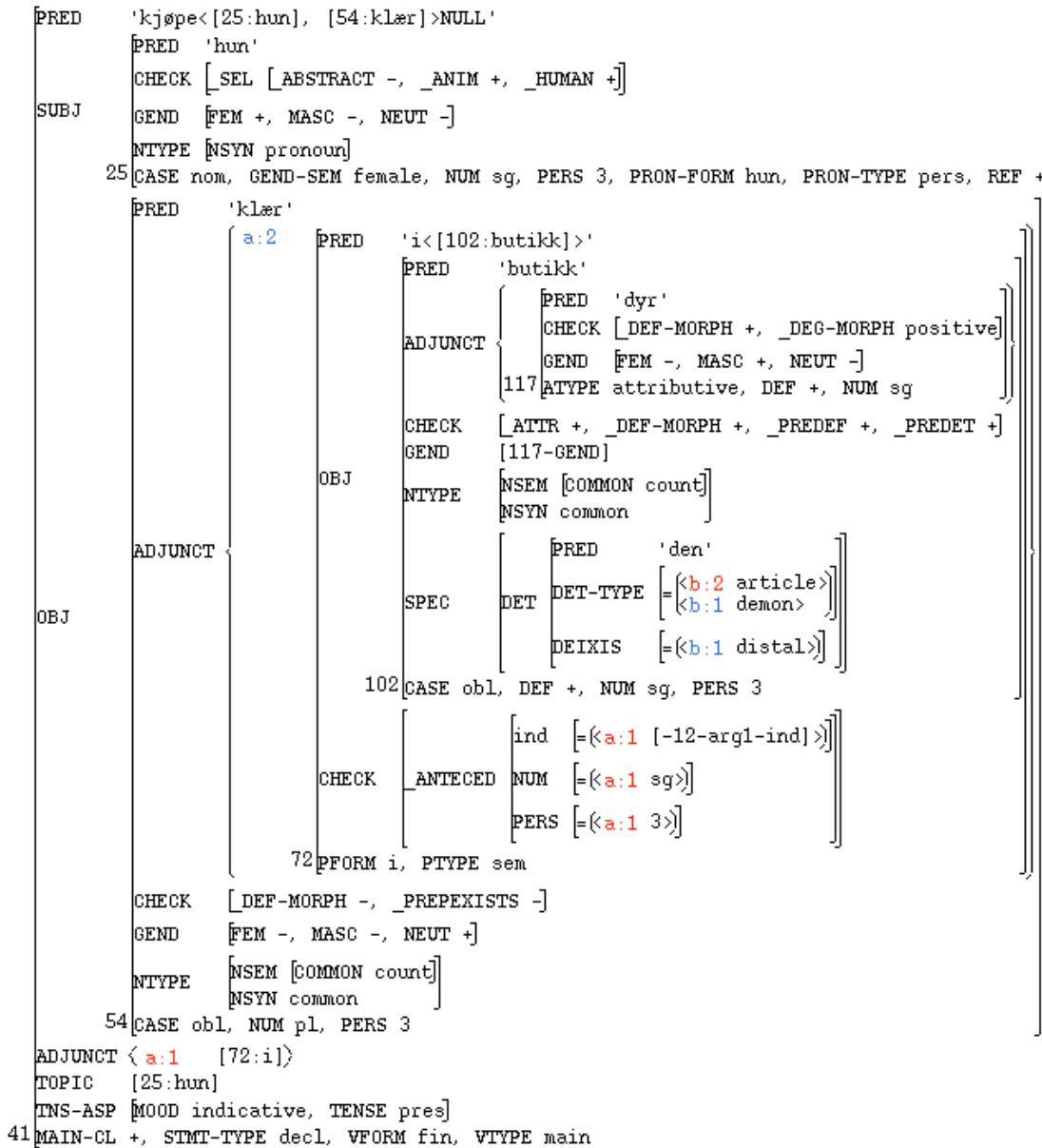


Figure 2: F-structure chart for (2) *Hun kjøper klær i den dyre butikken.*

The word *den* may either be a simple definite determiner or a demonstrative. There is also a syntactic ambiguity: the prepositional phrase may be attached to the NP or to the VP. When the PP is part of the NP, it is an ADJUNCT in the OBJ; when it is part of the VP, it is an ADJUNCT on the outer level of the f-structure. In the f-structure chart shown in figure 2, this is represented by having the attribute ADJUNCT in two places, each with an alphabetic index, *a:1* for the sentential ADJUNCT and *a:2* for the ADJUNCT in the OBJ.

An only slightly more complicated example is given in example 3.

- (3) *En jente kjøper klær i den dyre butikken.*
 a girl buys clothes in the expensive store
 “A girl buys clothes in the expensive store.”



Figure 3: Partial f-structure chart for (3) *En jente kjøper klær i den dyre butikken.*

In addition to the lexical ambiguity and the PP-attachment ambiguity we observed in example 2, there are two more ambiguities here. The word *en* is either the indefinite article or the numeral “one”. And since Norwegian is a V2 language, and there is no person or number inflection on verbs, the phrases before and after the finite verb may each either be SUBJ or OBJ. These four local ambiguities result in 16 possible analyses. For this sentence, the f-structure chart requires more space than will fit on even a large computer screen, and we show only a small part of it in figure 3.

A somewhat more compact display may be obtained from an alternative f-structure chart showing a tree of choices, but the main problem for a human disambiguator is that regardless of which method of display is chosen, the packed representations always contain all the information in all analyses. The disambiguator can choose to mark certain choices as dispreferred, turning their labels gray. But the information still remains in the display, forcing the disambiguator to keep track of many kinds of information about the analyses at the same time. These few examples should make it clear that this method of disambiguation requires expert competence in using XLE and detailed knowledge of the grammar. Even for an annotator with that kind of competence, disambiguating sentences with hundreds of analyses in this way can be a formidable task.

4 The XLE Web Interface

The XLE Web Interface (XLE-Web) is a web-based tool for parsing with XLE and viewing c-structures, f-structures and mrs-structures. Initially developed in the LOGON project (Oepen et al., 2004), it has been

the starting point for the work in TREPIL discussed in the following section. XLE-Web allows the user to choose a grammar and type in a sentence to be analyzed. The sentence is then processed by the XLE parser, and the resulting c-structures, f-structures and mrs-structures are displayed, either one solution at a time, or all solutions together in the form of packed c- and f-structure representations (there is no packed mrs-representation at present).

Packed f-structures were first implemented in XLE in order to provide a compact internal representation of the set of solutions of a sentence. The XLE display system uses this packing to simultaneously display all f-structures in one graph, and the packed f-structures in XLE-Web have been tightly modeled after XLE's packed f-structure display. An innovation in XLE-Web is the display of *packed c-structures* as directed acyclic graphs, namely, a set of c-structure trees where nodes that are equal across solutions are identified and where additional nodes indicate in which contexts their subnodes are valid.

The XLE-Web server software runs on Linux and MacOS; it is implemented in Common Lisp and uses a shared library version of the XLE core parsing engine which is dynamically linked into the server program. Internally, the interface web pages are generated as XML files, which are converted on the server side to HTML by means of XSLT. The interactive features of the displayed structures are implemented in Javascript. For instance, when mousing over a QEQ relation in an mrs-structure, the corresponding variables in the elementary predicates are highlighted. In the same fashion, structure sharing in f-structures is made visible, and mousing over a c-structure node highlights both the f-structure projection of that node and all other nodes having the same projection. The c-structure trees (and graphs in the case of packed representations) are drawn using the XML-compliant standard SVG (Scalable Vector Graphics).

Screenshots from XLE-Web for the analysis of the ambiguous sentence in example 1 are provided in figure 4. The analyses are by default shown successively on separate web pages. By clicking on *Previous* and *Next* buttons, the user browses through the various c- and f-structures, as well as the mrs-structures. The c- and f-structures are displayed side by side. The mrs-structure is displayed by clicking on the *Show MRS* button.

There are a number of options which may be chosen by ticking off the boxes on the main page. Most LFG grammars that have been implemented in XLE make use of optimality marks to prefer some analyses over others. By ticking the *Disable Optimality marks* box, the user chooses to have all analyses displayed rather than just the 'optimal' analyses. For the purpose of treebanking it may be desirable to run the grammar without the optimality marks, to make sure that all possible analyses are presented for manual disambiguation.

Another way of examining ambiguous analyses in XLE-Web is chosen by ticking off the *Packed representation* box. This displays all analyses of a sentence on one page. For the two-way ambiguous sentence in example 1, we get the packed representation in figure 5. There is no c-structure distinction between the two analyses, so that only one c-structure is displayed. The difference is, as noted above, the value of the attribute NUM, which may be either *sg* or *pl*. This is displayed in the packed f-structure in a way very similar to the standard XLE packed representation shown above in figure 1. In addition, the display in 5 shows a list of discriminants, which are the topic of the following section.

5 Disambiguation with discriminants

The technique of disambiguating with discriminants was first described by Carter (1997), who points out that there may be "dozens of analyses that are variations on a small number of largely independent themes: choices of word sense, modifier attachment, conjunction scope and so on." Carter argues that disambiguation may be achieved quickly and without expert competence if it is based on elementary linguistic properties which the disambiguator may accept or reject independently of other properties. The disambiguation is then done on the individual properties rather than on the analyses themselves.

XLE Web Interface

Grammar: bokmal-mrs

Write a sentence, ending it with punctuation (. ? or !):

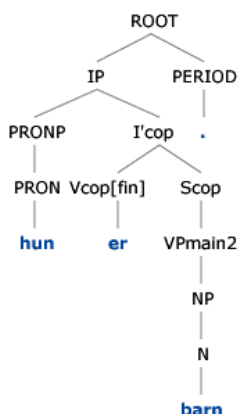
Hun er barn.

Packed representation CG-prepare
 Disable Optimality marks CG-prepare on fragment analysis Include non-top F-structures

Solution 1 of 2:

C-structure

For viewing, you'll need SVG support (i.e. Firefox 1.5 or Adobe's [SVG Viewer plugin](#).)



F-structure

| | |
|-----------------|--|
| PRED | 'være<[1:hun], [2:barn]>' |
| TNS-ASP | 3 TENSE pres, MOOD indicative |
| TOPIC | PRED 'hun' NTYPE 7 NSYN pronoun GEND 6 NEUT -, MASC -, FEM + REF +, PRON-TYPE pers, PRON-FORM hun, PERS 3, NUM sg, GEND-SEM female, CASE nom |
| PREDLINK | PRED 'barn' NTYPE 30 COMMON count NSYN common GEND 27 NEUT +, MASC -, FEM - PERS 3, NUM sg |
| SUBJ | [1] |
| 0 | VTYP e main, VFORM fin, STMT-TYPE decl, MAIN-CL + |

MRS

| | | | | | | |
|---|--|---|--|--|---|--|
| TOP | h1 | | | | | |
| INDEX | e10 | | | | | |
| RELS | <table border="0" style="border-collapse: collapse;"> <tr> <td style="border-right: 1px solid black; padding-right: 5px;"> stative_asp_rel LBL h9 ARG1 e10 </td> <td style="border-right: 1px solid black; padding-right: 5px;"> bare_sg_q_rel LBL h12 ARGO x3 BODY h13 RSTR h11 LNK 20 </td> <td style="border-right: 1px solid black; padding-right: 5px;"> prpstn_m_rel LBL h1 ARGO e10 MARG h14 </td> <td style="border-right: 1px solid black; padding-right: 5px;"> cop_id_rel LBL h9 ARGO e10 ARG1 x7 ARG2 x3 LNK 13 </td> <td style="padding-right: 5px;"> pronoun_q_rel LBL h5 ARGO x7 BODY h6 RSTR h4 LNK 0 </td> </tr> </table> | stative_asp_rel LBL h9 ARG1 e10 | bare_sg_q_rel LBL h12 ARGO x3 BODY h13 RSTR h11 LNK 20 | prpstn_m_rel LBL h1 ARGO e10 MARG h14 | cop_id_rel LBL h9 ARGO e10 ARG1 x7 ARG2 x3 LNK 13 | pronoun_q_rel LBL h5 ARGO x7 BODY h6 RSTR h4 LNK 0 |
| stative_asp_rel LBL h9 ARG1 e10 | bare_sg_q_rel LBL h12 ARGO x3 BODY h13 RSTR h11 LNK 20 | prpstn_m_rel LBL h1 ARGO e10 MARG h14 | cop_id_rel LBL h9 ARGO e10 ARG1 x7 ARG2 x3 LNK 13 | pronoun_q_rel LBL h5 ARGO x7 BODY h6 RSTR h4 LNK 0 | | |
| HCONS | { h4 QEQ h8, h11 QEQ h2, h14 QEQ h9 } | | | | | |

Figure 4: Screenshots from XLE-Web

2 solutions:

Discriminants

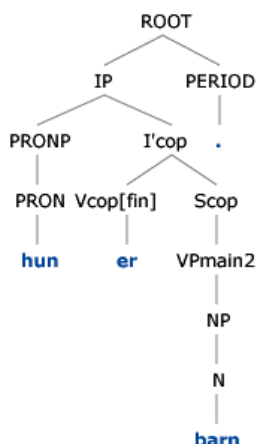
Selected solutions: 2

F-structure discriminants

| | | |
|---------------|-------|---|
| 'barn' NUM sg | compl | 1 |
| 'barn' NUM pl | compl | 1 |

C-structure

For viewing, you'll need SVG support (i.e. Firefox 1.5 or Adobe's [SVG Viewer plugin](#) .)



F-structure

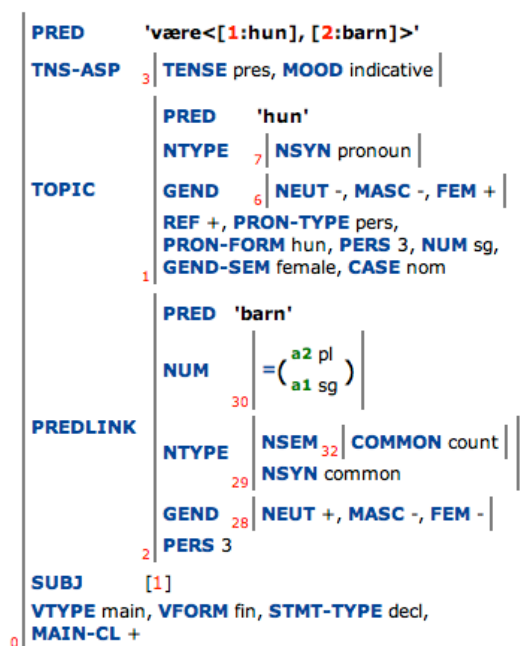


Figure 5: Packed representation of (1) *Hun er barn*.

In LFG terms a discriminant is, in general, any local property of a c-structure or f-structure that not all analyses share. We have implemented three types of discriminants in TREPIL: c-structure discriminants, f-structure discriminants and morphology discriminants. A c-structure discriminant is the segmentation of a surface constituent string induced by a minimal subtree (a node with its immediate subnodes); in addition, the rule that gives rise to this subtree is a discriminant. An f-structure discriminant is a direct path in an f-structure from a PRED value to an embedded PRED value or from a PRED value to an atomic value. A morphology discriminant is a word with the tags it receives from morphological preprocessing. Examples of all three types of discriminants will be given below.

As a first step towards developing a treebanking tool for disambiguation, we have implemented discriminants in XLE-Web. To the left in figure 5 is a list of discriminants. This sentence has only f-structure discriminants. In this example, the discriminants are paths from a PRED value to two alternative atomic values. Next to each discriminant in the display there are two columns, one where it says *compl* (for ‘complement’) and one with a number. The disambiguator may choose a discriminant by clicking on it, or reject a discriminant by clicking on *compl*. The number in the third column gives the number of analyses that will remain if that discriminant is chosen. When a discriminant choice has been made, the chosen discriminant is boldfaced, and only the discriminants still compatible with that choice are redisplayed. Since there is only one local ambiguity in this sentence, it will be fully disambiguated after one discriminant choice has been made.

The sentence in example 2 has, as mentioned above, two local ambiguities and four analyses. Figure 6 shows the seven discriminants which distinguish these analyses from each other and, in addition to a packed

Discriminants

Selected solutions: 4

F-structure discriminants

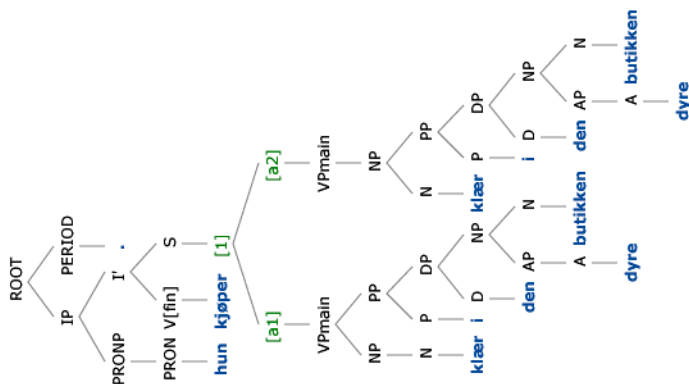
| | | |
|---------------------------------------|-------|---|
| 'klær' ADJUNCT > 'i<[]>' | compl | 2 |
| 'kjøpe<[]>[[]>NULL' ADJUNCT > 'i<[]>' | compl | 2 |
| 'den' DET-TYPE demon | compl | 2 |
| 'den' DET-TYPE article | compl | 2 |
| 'den' DEIXIS distal | compl | 2 |

C-structure discriminants

| | |
|-----------------------------|---------|
| klær i den dyre butikken | |
| NP -> N PP | compl 2 |
| VPmain -> NP PP | compl 2 |

C-structure

For viewing, you'll need SVG support (i.e. Firefox 1.5 or Adobe's SVG Viewer plugin.)



F-structure

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--|---|-------------|------------------|-----------------|---------------------------------------|----------------------------|--|--|--|-----------------|------------------|----------------------------|--------------|---------------------------|--------------------------------------|---------------------------------|------------------|-----------------------------|--|-------------|-------|------------|---------------------------------------|------------|--|-------------|---------|--------------|--------------------------------------|------------|----------------------------|-----------------------------|--|-------------|------------------|------------|-----------|---------------------------|-----------|---------------------------------|-----------|-------------|-----|----------|--|
| PRED | 'kjøpe<[1:hun], [2:klær]>NULL' | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| TNS-ASP | 4 TENSE pres, MOOD indicative | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| TOPIC | <table border="1"> <tr> <td>PRED</td> <td>'hun'</td> </tr> <tr> <td>NTYPE</td> <td>8 NSYN pronoun</td> </tr> <tr> <td>GEN</td> <td>7 NEUT -, MASC -, FEM +</td> </tr> <tr> <td>REF +, PRON-TYPE pers, PRON-FORM hun, PERS 3, NUM sg, GEND-SEM female, CASE nom</td> <td></td> </tr> </table> | PRED | 'hun' | NTYPE | 8 NSYN pronoun | GEN | 7 NEUT -, MASC -, FEM + | REF +, PRON-TYPE pers, PRON-FORM hun, PERS 3, NUM sg, GEND-SEM female, CASE nom | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PRED | 'hun' | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| NTYPE | 8 NSYN pronoun | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| GEN | 7 NEUT -, MASC -, FEM + | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| REF +, PRON-TYPE pers, PRON-FORM hun, PERS 3, NUM sg, GEND-SEM female, CASE nom | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| ADJUNCT_{a1} | <table border="1"> <tr> <td>PRED</td> <td>'i<[35:butikk]>'</td> </tr> <tr> <td>PRED</td> <td>'butikk'</td> </tr> <tr> <td>SPEC</td> <td> <table border="1"> <tr> <td>DET</td> <td>'den'</td> </tr> <tr> <td>DET-TYPE</td> <td>71 demon</td> </tr> <tr> <td>DEIXIS_{b1}</td> <td>71 article</td> </tr> </table> </td> </tr> <tr> <td>NTYPE</td> <td>42 NSEM₆₉ COMMON count</td> </tr> <tr> <td>GEN</td> <td>41 NSYN common</td> </tr> <tr> <td>ADJUNCT_{a2}</td> <td> <table border="1"> <tr> <td>PRED</td> <td>'dyr'</td> </tr> <tr> <td>GEN</td> <td>40 NUM sg, DEF +, ATYPE attributive</td> </tr> </table> </td> </tr> <tr> <td>OBJ</td> <td> <table border="1"> <tr> <td>PRED</td> <td>'kjøpe'</td> </tr> <tr> <td>NTYPE</td> <td>32 NSEM₇₅ COMMON count</td> </tr> <tr> <td>GEN</td> <td>31 NEUT +, MASC -, FEM -</td> </tr> <tr> <td>ADJUNCT_{a2}</td> <td> <table border="1"> <tr> <td>PRED</td> <td>'i<[35:butikk]>'</td> </tr> <tr> <td>OBJ</td> <td>35 [35]</td> </tr> <tr> <td>PTYPE sem, PFORM i</td> <td>34 [34]</td> </tr> </table> </td> </tr> <tr> <td>PERS 3, NUM pl, CASE obl</td> <td>29 [29]</td> </tr> </table> </td> </tr> <tr> <td>SUBJ</td> <td>[1]</td> </tr> <tr> <td>0</td> <td>VTYPE main, VFORM fin, STMT-TYPE decl, MAIN-CL +</td> </tr> </table> | PRED | 'i<[35:butikk]>' | PRED | 'butikk' | SPEC | <table border="1"> <tr> <td>DET</td> <td>'den'</td> </tr> <tr> <td>DET-TYPE</td> <td>71 demon</td> </tr> <tr> <td>DEIXIS_{b1}</td> <td>71 article</td> </tr> </table> | DET | 'den' | DET-TYPE | 71 demon | DEIXIS_{b1} | 71 article | NTYPE | 42 NSEM ₆₉ COMMON count | GEN | 41 NSYN common | ADJUNCT_{a2} | <table border="1"> <tr> <td>PRED</td> <td>'dyr'</td> </tr> <tr> <td>GEN</td> <td>40 NUM sg, DEF +, ATYPE attributive</td> </tr> </table> | PRED | 'dyr' | GEN | 40 NUM sg, DEF +, ATYPE attributive | OBJ | <table border="1"> <tr> <td>PRED</td> <td>'kjøpe'</td> </tr> <tr> <td>NTYPE</td> <td>32 NSEM₇₅ COMMON count</td> </tr> <tr> <td>GEN</td> <td>31 NEUT +, MASC -, FEM -</td> </tr> <tr> <td>ADJUNCT_{a2}</td> <td> <table border="1"> <tr> <td>PRED</td> <td>'i<[35:butikk]>'</td> </tr> <tr> <td>OBJ</td> <td>35 [35]</td> </tr> <tr> <td>PTYPE sem, PFORM i</td> <td>34 [34]</td> </tr> </table> </td> </tr> <tr> <td>PERS 3, NUM pl, CASE obl</td> <td>29 [29]</td> </tr> </table> | PRED | 'kjøpe' | NTYPE | 32 NSEM ₇₅ COMMON count | GEN | 31 NEUT +, MASC -, FEM - | ADJUNCT_{a2} | <table border="1"> <tr> <td>PRED</td> <td>'i<[35:butikk]>'</td> </tr> <tr> <td>OBJ</td> <td>35 [35]</td> </tr> <tr> <td>PTYPE sem, PFORM i</td> <td>34 [34]</td> </tr> </table> | PRED | 'i<[35:butikk]>' | OBJ | 35 [35] | PTYPE sem, PFORM i | 34 [34] | PERS 3, NUM pl, CASE obl | 29 [29] | SUBJ | [1] | 0 | VTYPE main, VFORM fin, STMT-TYPE decl, MAIN-CL + |
| PRED | 'i<[35:butikk]>' | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PRED | 'butikk' | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| SPEC | <table border="1"> <tr> <td>DET</td> <td>'den'</td> </tr> <tr> <td>DET-TYPE</td> <td>71 demon</td> </tr> <tr> <td>DEIXIS_{b1}</td> <td>71 article</td> </tr> </table> | DET | 'den' | DET-TYPE | 71 demon | DEIXIS_{b1} | 71 article | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| DET | 'den' | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| DET-TYPE | 71 demon | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| DEIXIS_{b1} | 71 article | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| NTYPE | 42 NSEM ₆₉ COMMON count | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| GEN | 41 NSYN common | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| ADJUNCT_{a2} | <table border="1"> <tr> <td>PRED</td> <td>'dyr'</td> </tr> <tr> <td>GEN</td> <td>40 NUM sg, DEF +, ATYPE attributive</td> </tr> </table> | PRED | 'dyr' | GEN | 40 NUM sg, DEF +, ATYPE attributive | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PRED | 'dyr' | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| GEN | 40 NUM sg, DEF +, ATYPE attributive | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| OBJ | <table border="1"> <tr> <td>PRED</td> <td>'kjøpe'</td> </tr> <tr> <td>NTYPE</td> <td>32 NSEM₇₅ COMMON count</td> </tr> <tr> <td>GEN</td> <td>31 NEUT +, MASC -, FEM -</td> </tr> <tr> <td>ADJUNCT_{a2}</td> <td> <table border="1"> <tr> <td>PRED</td> <td>'i<[35:butikk]>'</td> </tr> <tr> <td>OBJ</td> <td>35 [35]</td> </tr> <tr> <td>PTYPE sem, PFORM i</td> <td>34 [34]</td> </tr> </table> </td> </tr> <tr> <td>PERS 3, NUM pl, CASE obl</td> <td>29 [29]</td> </tr> </table> | PRED | 'kjøpe' | NTYPE | 32 NSEM ₇₅ COMMON count | GEN | 31 NEUT +, MASC -, FEM - | ADJUNCT_{a2} | <table border="1"> <tr> <td>PRED</td> <td>'i<[35:butikk]>'</td> </tr> <tr> <td>OBJ</td> <td>35 [35]</td> </tr> <tr> <td>PTYPE sem, PFORM i</td> <td>34 [34]</td> </tr> </table> | PRED | 'i<[35:butikk]>' | OBJ | 35 [35] | PTYPE sem, PFORM i | 34 [34] | PERS 3, NUM pl, CASE obl | 29 [29] | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PRED | 'kjøpe' | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| NTYPE | 32 NSEM ₇₅ COMMON count | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| GEN | 31 NEUT +, MASC -, FEM - | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| ADJUNCT_{a2} | <table border="1"> <tr> <td>PRED</td> <td>'i<[35:butikk]>'</td> </tr> <tr> <td>OBJ</td> <td>35 [35]</td> </tr> <tr> <td>PTYPE sem, PFORM i</td> <td>34 [34]</td> </tr> </table> | PRED | 'i<[35:butikk]>' | OBJ | 35 [35] | PTYPE sem, PFORM i | 34 [34] | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PRED | 'i<[35:butikk]>' | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| OBJ | 35 [35] | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PTYPE sem, PFORM i | 34 [34] | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PERS 3, NUM pl, CASE obl | 29 [29] | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| SUBJ | [1] | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0 | VTYPE main, VFORM fin, STMT-TYPE decl, MAIN-CL + | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Figure 6: Packed representation for (2) *Hun kjøper klær i den dyre butikken*.

f-structure, also a packed c-structure. For this example, the PP attachment ambiguity is reflected in both c-structure and f-structure discriminants. At the f-structure level, the disambiguator can choose the first discriminant, which may be read: the PRED ‘klær’ (“clothes”) has an ADJUNCT whose PRED is ‘i’ (“in”), or the second discriminant, which may be read: the PRED ‘kjøpe<[],[]>NULL’ (“buy” with two arguments) has an ADJUNCT whose PRED is ‘i’ (“in”).

Alternatively, the disambiguator may choose one of the c-structure discriminants. The bracketing (represented by ||) of the string in the c-structure discriminant table in figure 6 may be attributed either to the NP rule or the VPmain rule listed underneath this string. If the annotator decides to disambiguate the PP attachment by choosing the *complement* of the first c-structure discriminant (NP → N PP), a new set of discriminants and structures will be displayed as in figure 7. There are now only f-structure discriminants left for distinguishing between the two readings of *den* as either demonstrative or article. If the article reading is chosen, the interface redisplayes the fully disambiguated structures as in figure 8.

In addition to c-structure and f-structure discriminants, we have also implemented morphology discriminants. As mentioned above, a morphology discriminant is a word with the tags it receives from morphological preprocessing. Consequently, only words that have morphological features receive morphology discriminants. Consider example 4, which has many possible readings due to multiple lexical ambiguities. The *noun* readings of the ambiguous words receive the morphological discriminants shown in 5. By choosing the complement of each of these discriminants, we can eliminate all noun readings, thereby reducing the number of analyses from 45 to 6.

- (4) *To av disse ga henne tre.*
 two/stuff of these/swing gave her three/wood

| | | | |
|-----|------------------------|-------|----|
| | to+SP+Noun+Neut+Indef | compl | 18 |
| (5) | disse+Sg+Noun+MF+Indef | compl | 18 |
| | tre+SP+Noun+Neut+Indef | compl | 30 |

We have demonstrated that even in sentences with a small number of analyses, discriminant disambiguation is easier and more efficient for a human disambiguator than examining full analyses. The true power of discriminant analysis becomes apparent when one considers sentences with a large number of analyses. The previous example showed that 39 analyses could be eliminated through three simple discriminant decisions. Consider also sentence 6, which has many local ambiguities.

- (6) *Sjefen har drevet og sendt invitasjoner til alle han kjenner.*
 boss-the has driven and sent invitations to everyone he knows
 “The boss has been sending invitations to everyone he knows.”

This sentence gets 86 solutions. The number of discriminants is also large: there are 5 c-structure discriminants, 14 morphology discriminants, and 77 f-structure discriminants, which is too large a number to be shown here. However, it may be fully disambiguated to the intended analysis by making choices concerning only two discriminants, for instance those shown in examples 7 and 8. The discriminant in 7 chooses the pseudo-coordination analysis of the progressive, while the one in 8 specifies that the noun *sjef* “boss” is the subject of the verb *sende* “send”.

- (7) ‘sende<[],[],[]>’ TNS-ASP ASP progressive

- (8) ‘sende<[],[],[]>NULL’ SUBJ ‘sjef’

Discriminants

Selected solutions: 2

F-structure discriminants

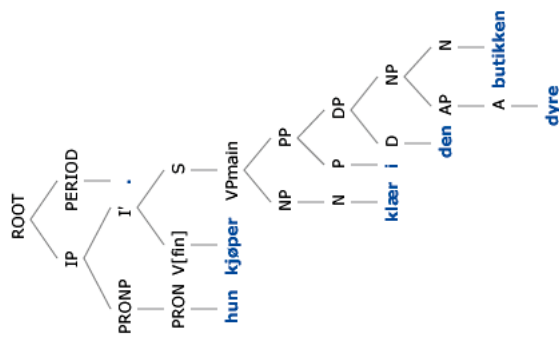
| | | |
|------------------------|-------|---|
| 'den' DET-TYPE demon | compl | 1 |
| 'den' DET-TYPE article | compl | 1 |
| 'den' DEIXIS distal | compl | 1 |

C-structure discriminants

| | |
|-----------------------------|--------------|
| klær i den dyre butikken | |
| NP -> N PP | compl |

C-structure

For viewing, you'll need SVG support (i.e. Firefox 1.5 or Adobe's [SVG Viewer plugin](#).)



F-structure

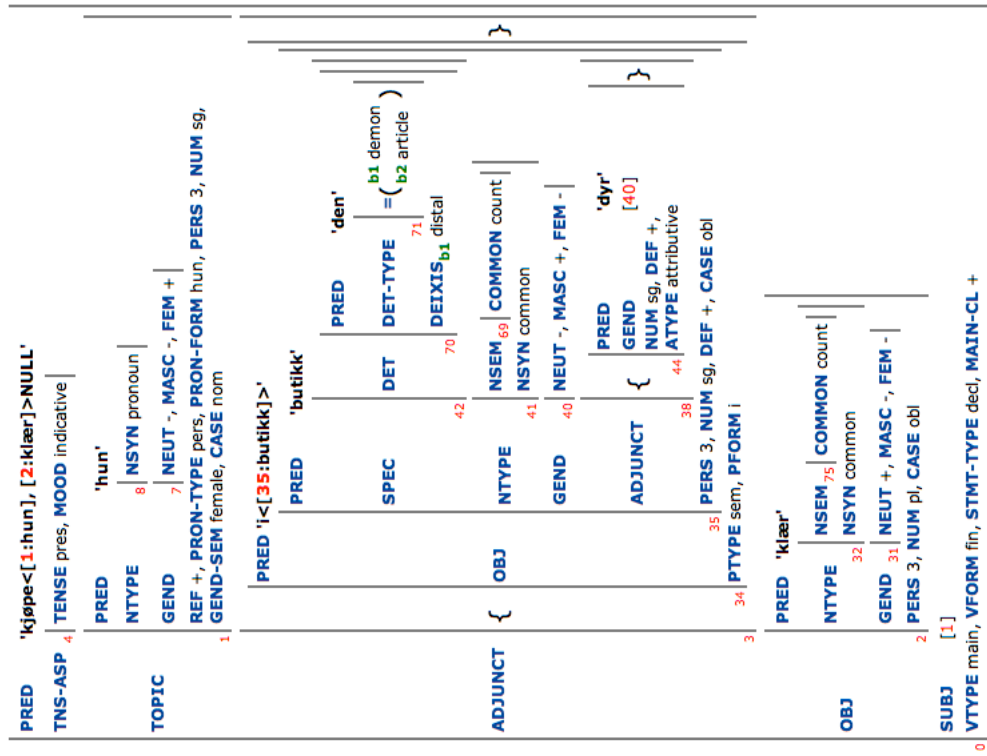


Figure 7: Partially disambiguated structures for (2) *Hun kjøper klær i den dyre butikken.*

Discriminants

Selected solutions: 1

F-structure discriminants

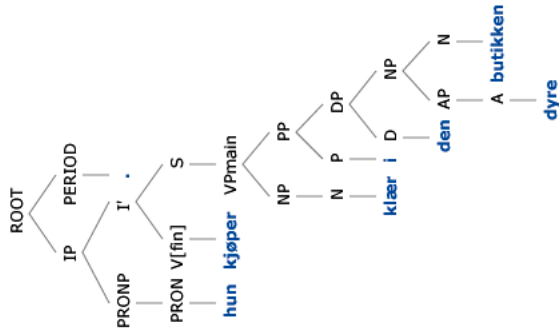
| |
|------------------------|
| 'den' DET-TYPE article |
|------------------------|

C-structure discriminants

| |
|-----------------------------|
| klær i den dyre butikken |
| NP -> N PP |
| compl |

C-structure

For viewing, you'll need SVG support (i.e., Firefox 1.5 or Adobe's [SVG Viewer plugin](#).)



F-structure

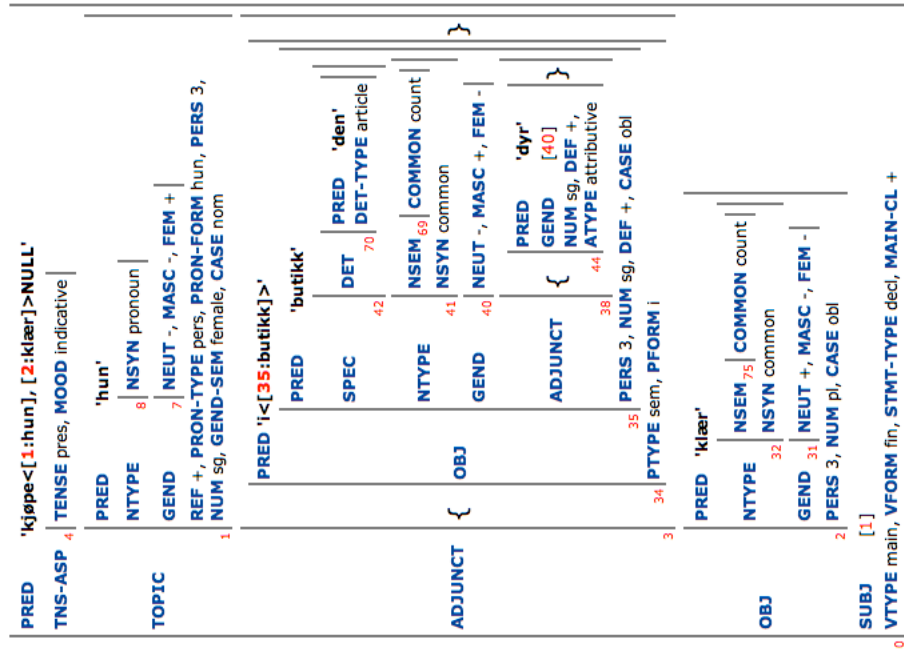


Figure 8: Fully disambiguated structures for (2) *Hun kjøper klær i den dyre butikken*.

6 Conclusion and further work

The TREPIL project is aimed at a methodology for the incremental development of a treebank in synchrony with a wide-coverage LFG grammar. This methodology is dependent on the XLE parser in conjunction with a disambiguation tool for recording all structural choices when the annotator selects one parse over other parses. To our knowledge, the TREPIL project is the first to develop a discriminant-based tool for LFG.

The XLE Web Interface provides either a browsing display or a packed representation, and has been extended with discriminants in TREPIL. When the packed representation is used for disambiguation, it is gradually unpacked as discriminants are chosen until disambiguation is complete and only one analysis is left. This tool is grammar and language independent, so that other LFG grammars developed on the XLE platform will be able to use it to create their own parsed corpora.

We have shown that a small number of discriminant choices may be sufficient to disambiguate a large number of analyses. It is therefore unnecessary to display all discriminants at one time. We will experiment with different ways of presenting selected discriminants to the annotator. Another important kind of functionality will be that the annotator should be able to record the degree of confidence with which decisions have been made.

One of the most interesting aspects of disambiguation by local discriminants is that such annotator decisions may be saved in a database and reused for automatic disambiguation after the grammar has been revised (Carter, 1997; Oepen et al., 2003). The decisions on local properties have been shown to be remarkably stable over revisions of the grammar. This means that the treebank may be produced in new versions as the grammar develops. This solves one serious problem with many treebanks, namely that they become obsolete as linguistic theories evolve. A treebank that can be updated semiautomatically as the grammar (and the theory behind the grammar) evolves is therefore dynamic. This approach has been followed in the Redwoods initiative (Oepen et al., 2003), and it will also be the aim in TREPIL.

7 Acknowledgments

This work was supported in part by a grant from the Research Council of Norway. We would like to thank John Maxwell at PARC, who has always been willing to discuss implementation issues and has provided extensions to the XLE software as we needed them.

References

- Abeillé, Anne. 2003. Introduction. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers, Dordrecht, pages xiii–xxvi.
- Abeillé, Anne, Lionel Clément, and François Toussenet. 2003. Building a treebank for French. In *Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers.
- Bick, Eckhard, Heli Uiibo, and Kaili Müürisep. 2004. Arborest – a VISL-style treebank derived from an Estonian Constraint Grammar corpus. In *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories*. Seminar für Sprachwissenschaft, Universität Tübingen.
- Bouma, Gosse. 2004. Treebank evidence for the analysis of PP-fronting. In *Third Workshop on Treebanks and Linguistic Theories, Seminar für Sprachwissenschaft, Tübingen, 2004*, pages 15–26.
- Burke, Michael, Aoife Cahill, Ruth O’ Donovan, Josef Van Genabith, and Andy Way. 2004a. Treebank-based acquisition of wide-coverage, probabilistic LFG resources: Project overview, results and eval-

- uation. In *The First International Joint Conference on Natural Language Processing (IJCNLP-04), Workshop “Beyond shallow analyses – Formalisms and statistical modeling for deep analyses”*, March 22-24, 2004 Sanya City, Hainan Island, China.
- Burke, Michael, Aoife Cahill, Ruth O’Donovan, Josef Van Genabith, and Andy Way. 2004b. Evaluation of an automatic f-structure annotation algorithm against the parc 700 dependency bank. In *Proceedings of the LFG04 Conference, Christchurch, New Zealand*.
- 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. 2004. *Parsing with Automatically Acquired, Wide-Coverage, Robust, Probabilistic LFG Approximations*. Ph.D. thesis, School of Computing, Dublin City University.
- Carter, David. 1997. The TreeBanker: A tool for supervised training of parsed corpora. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 598–603, Providence, Rhode Island.
- Copestake, Ann, Dan Flickinger, Ivan A. Sag, and Carl Pollard. in preparation. Minimal Recursion Semantics: An introduction. Manuscript.
- Hajič, Jan. 1998. Building a syntactically annotated corpus: The Prague dependency treebank. In *Issues of Valency and Meaning*. Karolinum, Praha, pages 106–132.
- King, Tracy Holloway, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ronald M. Kaplan. 2003. The PARC 700 dependency bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora, held at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL’03), Budapest*.
- Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*.
- Nivre, Joakim, Koenraad De Smedt, and Martin Volk. 2005. Treebanking in northern europe: A white paper. In Henrik Holmboe, editor, *Nordisk Sprogteknologi 2004. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. Museum Tusulanums Forlag, Copenhagen, pages 97–112.
- Oepen, Stephan, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, and Victoria Rosén. 2004. Som å kapp-ete med trollet? Towards MRS-based Norwegian–English Machine Translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October.
- Oepen, Stephan, Dan Flickinger, Kristina Toutanova, and Christopher D. Manning. 2003. LinGO Redwoods, a rich and dynamic treebank for HPSG. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories*, pages 117–128. Växjö University Press.
- van der Beek, Leonoor, Gosse Bouma, Jan Daciuk, Tanja Gaustad, Robert Malouf, Gertjan van Noord, Robbert Prins, and Begoña Villada. 2002a. Algorithms for linguistic processing: NWO PIONIER progress report, August 2002. Technical report, NWO.

van der Beek, Leonoor, Gosse Bouma, Robert Malouf, and Geertjan van Noord. 2002b. The Alpino dependency treebank. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University.

Zaenen, Annie. 2004. . . . But full parsing is impossible. *Elsnews*, 13(1):9–10.

A BROAD-COVERAGE, REPRESENTATIONALLY MINIMAL
LFG PARSER: CHUNKS AND F-STRUCTURES ARE
SUFFICIENT

Gerold Schneider

Institute of Computational Linguistics, University of Zurich
English Department of the University of Zurich
Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

A major reason why LFG employs c-structure is because it is context-free. According to Tree-Adjoining Grammar (TAG), the only context-sensitive operation that is needed to express natural language is Adjoining, from which LFG functional uncertainty has been shown to follow. Functional uncertainty, which is expressed on the level of f-structure, would then be the only extension needed to an otherwise context-free processing of natural language. We suggest that if f-structures can be derived context-freely, full-fledged c-structures are not strictly needed in LFG, and that chunks and dependencies may be sufficient for a formal grammar theory. In order to substantiate this claim, we combine a projection of f-structures from chunks model with statistical techniques and present a parser that outputs LFG f-structure like representations. The parser is representationally minimal, deep-linguistic, robust, and fast, and has been evaluated and applied. The parser addresses context-sensitive constructions by treating the vast majority of long-distance dependencies by approximation with finite-state patterns, by post-processing, and by LFG functional uncertainty.

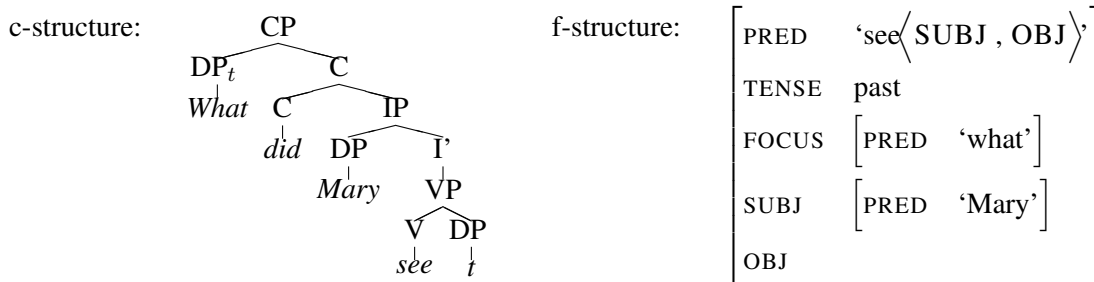
1 Introduction

In this paper we argue that full-fledged c-structures can be obviated for the syntactic analysis of natural language. We present and evaluate a broad-coverage statistical parser, Pro3Gres, that substantiates this claim. By reducing grammar complexity (Frank, 2002; Frank, 2004), by reducing parsing complexity to mostly context-free parsing and finite-state based chunking (Cahill et al., 2004; Schneider, 2003; Schneider, 2004), by bridging the gap between language engineering and formal grammar (Kaplan et al., 2004) by aiming for a representationally minimal theory (Jurafsky, 1996) we argue that chunks and dependencies (Abney, 1995; Frank, 2003) may be sufficient for a gormal grammar theory.

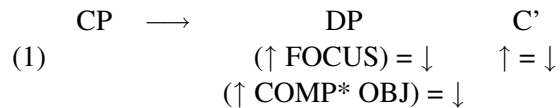
Two major factors that make broad-coverage parsing hard are (1) long-distance dependencies, as they break c-structure context-freeness, and (2) natural language ambiguity, which leads to immense search spaces during the parsing operation. We discuss long-distance dependencies in section 2 and ambiguity resolution in section 3.

1.1 Long-distance Dependencies

Long-distance dependencies as f-structure level mechanism The original LFG treatment of long-distance dependencies (Kaplan and Bresnan, 1982) used empty c-structure constituents, traces. For example, the relation between the DP node dominating *what* and the DP node dominating its trace *t* ensures that the wh-phrase *what* is both the FOCUS and OBJ of the sentence:



Subsequently, Kaplan and Zaenen (1989) proposed that long-distance dependencies are best expressed in functional and not phrasal terms. *Functional uncertainty* expresses long-distance dependency on the level of f-structure and obviates the need for trace-like devices in the theory of grammar, which has been described as descriptively more adequate and theoretically less redundant (Dalrymple, Kaplan, and King, 2001). A rule like the one in example 1 establishes two roles for the NP daughter of CP: it is the FOCUS, and it plays the grammatical role defined by the functional uncertainty path COMP* OBJ:



Constituency and Dependency Considerations of theoretical redundancy and linguistic accuracy can also give rise to questions concerning the necessity for c-structure. The grammar theory of Dependency Grammar (DG) is based on functional, grammar role dependencies in the spirit of LFG f-structure. Bröker, Hahn, and Schacht (1994) refers to DG as an LFG that only knows f-structure. Tesnière (1959)'s original DG concept aims at being a proto-semantic, monostratal, language-independent theory rather than merely a syntactic theory. In LFG terms, he challenged the need for c-structure. His view is to parse surface text (*ordre linéaire*) directly to f-structure (*ordre structurale*) in which word order plays no primary role, but may of course help disambiguate as a secondary role, for example by preferring projectivity. A theory that does not constrain dependency directions and allows non-projectivity (which is equivalent to using structure-sharing or movement) can express the same structures as constituency (Covington, 1994; Miller, 1999).

Discussions on headedness (Zwicky, 1985; Hudson, 1987), the prevalence of Chomskyan configurationalism and the desire to distinguish between different levels of analysis led to multistratal versions of DG (Mel'čuk, 1988) on the one hand, and influenced important DG based formal grammars, notably LFG and HPSG, on the other hand. LFG is an answer to the question of whether constituency or dependency should be exclusive – by respecting both: on the one hand the constituency-based context-free c-structure, on the other hand a non-configurational f-structure which expresses functional dependencies between constituents.

Parsing Complexity Dependency Grammar in its original conception allows non-projectivity which makes it computationally hard to process. Parsing algorithms able to treat completely unrestricted long-distance dependencies are NP-complete (Neuhaus and Bröker, 1997). In order to make broad-coverage DG parsing tractable, context-sensitivity needs to be maximally restricted. We discuss in section 2 how this can be done by using finite-state long-distance dependency approximations and functional uncertainty. Completely context-free traceless parsing only requires parsing algorithms with $O(n^3)$ complexity (Eisner, 1997), for example CYK (Younger, 1967). From a language-engineering perspective, context-freeness is a major appeal of c-structure. LFG constrains context-sensitivity by using a context-free c-structure backbone and then mapping to non-configurational f-structure. We follow arguments from Tree-Adjoining Grammar (TAG) (Joshi, 1985) to show that functional uncertainty is the only context-sensitive device needed to achieve the expressiveness exhibited by natural language. LFG functional uncertainty has been shown to follow as a corollary from TAG Adjoining (Joshi and Vijay-Shanker, 1989).

Context-free parsing was already recognised as potential candidate for broad-coverage application. When coupled with a probabilistic disambiguation, it turned out to be very successful (Collins, 1999; Charniak, 2000). But these parsers typically produce context-free data as output, trees that do not express long-distance dependencies. Although grammatical function and empty node annotation expressing long-distance dependencies are provided in Treebanks such as the Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993), these probabilistic Treebank trained parsers fully or largely ignore them (Collins (1999) Model 2 uses some of the functional labels, and Model 3 some long-distance dependencies). This entails two problems: first, the training cannot profit from valuable annotation data. Second, the extraction of long-distance dependencies (LDD) and the mapping to shallow semantic representations is not always possible from the output of these parsers. This limitation is aggravated by a lack of co-indexation information and parsing errors across an LDD.

Typical formal grammar parser complexity is much higher than the $O(n^3)$ for context-free grammar. The complexity of some formal grammars is still unknown. For Tree-Adjoining Grammars (TAG) it is $O(n^7)$ or $O(n^8)$ depending on the implementation (Eisner, 2000). Sarkar, Xia, and Joshi (2000) state that the theoretical bound of worst time complexity for Head-Driven Phrase Structure Grammar (HPSG) parsing is exponential. From a language engineering perspective, deep-linguistic formal grammars as a whole proved computationally too costly until recently; research thus successfully focused on finite-state based approaches such as chunking or

cascaded shallow parsing. Abney (1995) suggests a chunks & dependency model, but his chunks and cascaded parsing model (Abney, 1996) proved more successful.

We discuss in section 2 that most LDDs can be expressed in a context-free way (Schneider, 2003), and the remaining ones, if we follow TAG argumentation, by functional uncertainty. The vast majority of traces in the Penn Treebank can be treated as local dependencies by (1) using and modeling dedicated finite-state patterns across several levels of constituency subtrees partly leading to dedicated but fully local dependency syntactic relations and by (2) lexicalized post-processing rules. We also discuss that (3) some non-local dependencies are artifacts of configurational grammatical representations. The remaining long-distance dependencies can (4) be modelled with mild context-sensitivity by LFG functional uncertainty.

1.2 Ambiguity resolution

A Probabilistic Beam Search Approach Many approaches including ours profit from statistical data to prune unlikely partial analyses at parse-time, for example with a beam search. Parser performance decreases only marginally while time behaviour improves by at least an order of magnitude if reasonable pruning is used (Brants and Crocker, 2000) and allows us to explain psycholinguistic phenomena (Jurafsky, 1996; Crocker and Brants, 2000). A beam search approach also closes the gap between deterministic parsing (Nivre, 2004) and full parsing. Section 3 introduces our probability model and compares it to (Collins, 1999).

Shallow Chunking and F-Structure Parsing Some approaches (Kaplan et al., 2004; Schneider, 2004) include POS tagging preprocessing to reduce parsing ambiguity. Some systems include chunking preprocessing (Schneider, 2004) as is often used in probabilistic context-free parsing (Collins, 1999). The parser stays as shallow as is possible for each task, using finite-state based techniques for base phrase recognition. Parsing only takes place between the chunks of heads. Such chunks & dependency models can be attributed to Abney (1995). A chunk largely corresponds to a *nucleus* (Tesnière, 1959).

1.3 Related approaches

Recently, thanks to advances in exploiting and integrating statistics, the first deep-linguistic formal grammar based parsers have achieved the coverage and robustness needed to parse large corpora: Riezler et al. (2002) show how a hand-crafted LFG grammar can scale to the Penn Treebank with Maximum Entropy probability models. Hockenmaier and Steedman (2002) acquire a wide-coverage CCG grammar from the Penn Treebank automatically, Burke et al. (2004) an LFG grammar. Miyao, Ninomiya, and Tsujii (2004) semi-automatically acquire a broad-coverage HPSG grammar from the Penn Treebank and describe its efficiency (Tsuruoka and Tsujii, 2004).

Kaplan et al. (2004) compare speed and accuracy of a successful probabilistic context-free parser (Collins, 1999) to a robust LFG system based on (Riezler et al., 2002). They show that the gap between probabilistic context-free parsing and deep-linguistic full LFG parsing can be closed. On a random test set of 560 sentences from the Penn Treebank (4/5th of the PARC700 corpus¹) their full LFG grammar gives an overall improvement in F-score of 5% over (Collins, 1999) at a parsing time cost factor of 5. They also show that a limited LFG grammar (so called core system) still achieves a considerably higher f-score at a parsing time cost factor of only 1.5: about 200 seconds for Collins (1999) and about 300 seconds for the LFG core system. A conclusion that can be drawn from their and our results is that research in simplifying, restricting and limiting formal grammar expressiveness is bridging the gap between probabilistic parsing and formal grammar-based parsing.

Another important reason why deep-linguistic formal grammar parsing has become feasible and relatively fast is because long-distance dependencies are being approximated by deterministic or context-free approaches. Johnson (2002) shows that simple pattern-based approaches to obtaining LDDs from context-free probabilistic

¹www2.parc.com/istl/groups/nlitt/fsbank/

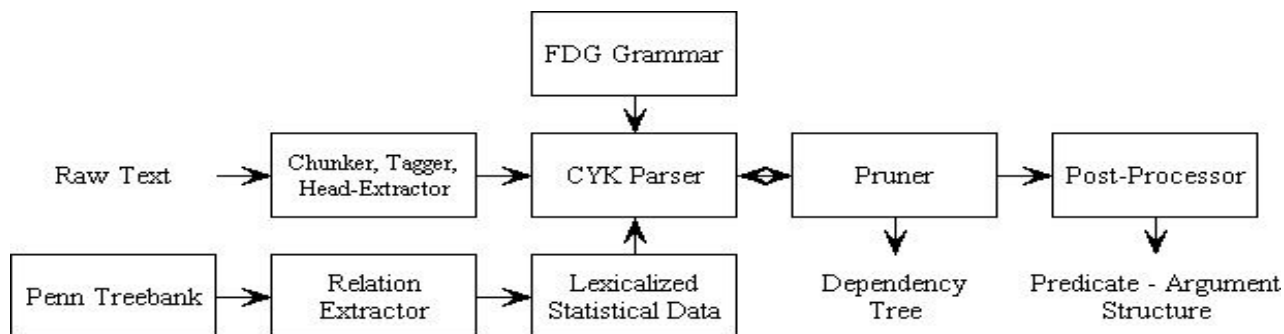


Figure 1: Pro3Gres flowchart

| Relation | Label | Example | Relation | Label | Example |
|---------------------|---------|--------------------------|---------------------|---------|-----------------------|
| verb-subject | subj | <i>he sleeps</i> | verb-prep. phrase | pobj | <i>slept in bed</i> |
| verb-first object | obj | <i>sees it</i> | noun-prep. phrase | modpp | <i>draft of paper</i> |
| verb-second object | obj2 | <i>gave (her) kisses</i> | noun-participle | modpart | <i>report written</i> |
| verb-adjunct | adj | <i>ate yesterday</i> | verb-complementizer | compl | <i>to eat apples</i> |
| verb-subord. clause | sentobj | <i>saw (they) came</i> | noun-preposition | prep | <i>to the house</i> |

Table 1: The most important dependency types used by the parser

parsers such as Collins (1999) are not successful. Jijkoun (2003) has used similar patterns, but containing LDD information, on the Penn Treebank in order to convert it to a Dependency format. We use a similar approach, assigning dedicated dependency labels to dependencies involving LDDs and statistical post-processing so that deep-linguistic parsing can mostly stay context-free (Schneider, 2003). Burke et al. (2004; Cahill et al. (2004) use a similar approach in LFG.

Frank (2003) suggests a (albeit non-probabilistic) chunks & dependencies model for LFG. Chunks can be freely combined subject to adjacency and projectivity (contiguity) constraints, which leads to a context-free parsing algorithm. Except for the added book-keeping functional annotations, her parsing algorithm is akin to CYK, which we use.

1.4 Our Parser

We present Pro3Gres, a parser that has been implemented to substantiate our claims. It has a highly modular architecture, shown in figure 1. It has been designed to keep search spaces and parsing complexity low while only taking minimal linguistic compromises (Schneider, 2004) and to be robust for broad-coverage parsing (Schneider, Dowdall, and Rinaldi, 2004). In order to keep parsing complexity as low as possible, aggressive use of shallow techniques and of context-free parsing is made. For low-level syntactic tasks, we use the shallow techniques of tagging and chunking, thus combining shallow and full parsing. We reduce the majority of context-sensitive tasks to context-free tasks by the use of patterns that are deep-linguistic because they are non-local, but shallow because they are fixed. For the few remaining context-sensitive tasks, mild context-sensitivity is sufficient.

We report evaluations of Pro3Gres on the 500 sentence Carroll corpus (Carroll, Minnen, and Briscoe, 1999). Special attention is given to long-distance dependencies and a linguistic analysis of errors. Comparisons to other parsers show that its performance is competitive.

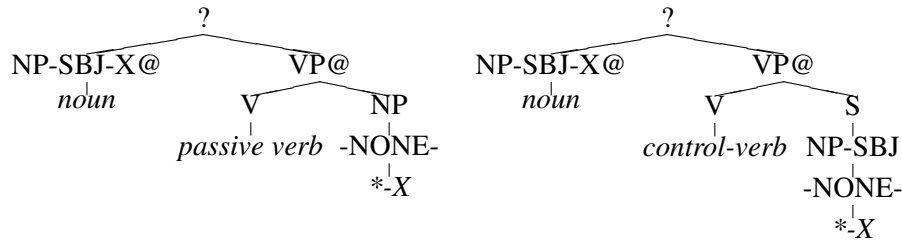


Figure 2: The extraction patterns for passive subjects (left) and subject control (right)

2 Long-distance dependencies

Treating long-distance dependencies is very costly (Neuhaus and Bröker, 1997), as they are context-sensitive. Most statistical Treebank trained parsers thus fully or largely ignore them. Johnson (2002) presents a pattern-matching algorithm for post-processing the Treebank output of such parsers to add empty nodes expressing long-distance dependencies to their parse trees. Encouraging results are reported for perfect parses, but performance drops considerably when using parser output trees.

We have applied structural patterns to the Treebank, where like in perfect parses precision and recall are high, and where in addition functional labels and empty nodes are available, so that patterns similar to Johnson's but relying on functional labels and empty nodes reach precision close to 100%. Unlike in Johnson, patterns for local dependencies are also used; non-local patterns simply stretch across more subtree-levels. We use the extracted lexical counts as lexical frequency training material. Every dependency relation has a group of structural extraction patterns associated with it. This amounts to a partial mapping of the Penn Treebank to Functional DG (Hajič, 1998), similar to the mapping described in Jijkoun (2003). Table 1 gives an overview of the most important dependencies.

The *subj* relation, for example, has the head of an arbitrarily nested NP with the functional tag *SBJ* as dependent, and the head of an arbitrarily nested VP as head for all active verbs. In passive verbs, however, a movement involving an empty constituent is assumed, which corresponds to the extraction pattern in figure 2 (left), where *VP@* is an arbitrarily nested VP, and *NP-SBJ-X@* the arbitrarily nested surface subject and *X* the co-indexed, moved element. Representing passive as movement, however, does not suggest long-distance movement. A close investigation confirms that passive movement is fixed, always local to a verbal domain, inside one clause. It can thus be represented by a single, local dependency.

Similar local restrictions can be formulated for other relations involving empty nodes in the Treebank, for example control structures, which have the extraction pattern shown in figure 2 (right), which are across two (possibly cascaded) clauses.

Grammatical role labels, empty node labels and tree configurations spanning several local subtrees are used as an integral part of some of the patterns. This leads to flatter trees, as typical for DG, which has the advantages that it helps to alleviate sparse data by mapping several nested structures that express the same dependency relation onto one dependency, that fewer decisions are needed at parse-time, which may reduce complexity and the risk of errors (Johnson, 2002), and that the costly overhead for dealing with unbounded dependencies can be largely avoided.

Let us consider the quantitative coverage of these patterns in detail. The ten most frequent types of empty nodes cover more than 60,000 of the approximately 64,000 empty nodes of sections 2-21 of the Penn Treebank. Table 2, reproduced from Johnson (2002) [line numbers and counts from the whole Treebank added], gives an overview.

Empty units, empty complementizers and empty relative pronouns [lines 4,5,9,10] pose no problem for DG as they are optional, non-head material. For example, a complementizer is an optional dependent of the subordinated verb.

Moved clauses [line 6] are mostly PPs or clausal complements of verbs of utterance. Only verbs of utterance

| | Antecedent | POS | Label | Count | Description | Example |
|------|------------|--------|-------|--------|-------------------------|--|
| 1 | NP | NP | * | 22,734 | NP trace | <i>Sam</i> was seen * |
| 2 | | NP | * | 12,172 | NP PRO | * to sleep is nice |
| 3 | WHNP | NP | *T* | 10,659 | WH trace | the woman <i>who</i> you saw *T* |
| (4) | | | *U* | 9,202 | Empty units | \$ 25 *U* |
| (5) | | | 0 | 7,057 | Empty complementizers | Sam said 0 Sasha snores |
| (6) | S | S | *T* | 5,035 | Moved clauses | <i>Sam had to go</i> , Sasha said *T* |
| 7 | WHADVP | ADVP | *T* | 3,181 | WH-trace | Sam explained <i>how</i> to leave *T* |
| (8) | | SBAR | | 2,513 | Empty clauses | <i>Sam had to go</i> , said Sasha (SBAR) |
| (9) | | WHNP | 0 | 2,139 | Empty relative pronouns | the woman 0 we saw |
| (10) | | WHADVP | 0 | 726 | Empty relative pronouns | the reason 0 to leave |

Table 2: The distribution of the 10 most frequent types of empty node and their antecedents in the Penn Treebank (adapted from Johnson2002). Bracketted lines designate long-distance dependencies that are local in DG

| Type | Count | prob-modeled | Treatment |
|--------------------------|--------|--------------|-------------------------|
| passive subject | 6,803 | YES | local relation |
| indexed gerund | 4,430 | NO | Tesnière translation |
| control, raise, semi-aux | 6,020 | YES | post-parsing processing |
| others / not covered | 5,481 | | |
| TOTAL | 22,734 | | |

Table 3: Coverage of the patterns for the most frequent NP traces [row 1]

allow subject-verb inversion in affirmative clauses [line 8]. In a dependency framework, none of them involve non-local dependencies or empty nodes, [line 6] and [line 8] are covered by rules that allow an inversion of the dependency direction under well-defined conditions.

NP Traces A closer look at NP traces ([line 1] of table 2) reveals that the majority of them are recognized by the grammar, and except for the indexed gerunds, they participate in the probability model. In control, raising and semi-auxiliary constructions, the non-surface semantic arguments, i.e. the subject-verb relation in the subordinate clause, are created based on lexical probabilities at the post-parsing stage, where minimal predicate-argument structures are output. In LFG terms, the probabilistic information on how likely a subordinate verb is to subcategorize for a control subject or object if they are unrealized is furnished by the matrix verb.

Unlike in control, raising and semi-auxiliary constructions, the antecedent of an indexed gerund cannot be established easily. The parser does not try to decide whether the target gerund is an indexed or non-indexed gerund nor does it try to find the identity of the lacking participant in the latter case. This is an important reason why recall values for the subject and object relations are lower than the precision values.

NP PRO As for the 12,172 NP PRO [line 2] in the Treebank, 5,656 are recognized by the *modpart* pattern (which covers reduced relative clauses), which means they are covered in the probability model. The dedicated *modpart* relation typically expresses the object function for past participles and the subject function for present participles.² A further 3,095 are recognized as non-indexed gerunds. Infinitives and gerunds may act as subjects, which are covered by translations (Tesnière, 1959), although these rules do not participate in the probability model. Many of the structures that are not covered by the extraction patterns and the probability model are still parsed correctly, for example adverbial clauses as unspecified subordinate clauses. Non-indexed adverbial phrases of the verb account for 1,598 NP PRO, non-indexed adverbial phrases of the noun for 268. As the NP is non-indexed, the identity of the lacking argument in the adverbial is unknown anyway, thus no semantic information is lost.

²The possible functional ambiguity is not annotated in the Treebank, hence the reduced relative clause is an unindexed empty NP

WH Traces Only 113 of the 10,659 WHNP antecedents in the Penn Treebank [line 3] are actually question pronouns. The vast majority, over 9,000, are relative pronouns. For them, an inversion of the direction of the relation they have to the verb is allowed if the relative pronoun precedes the subject.

Only non-subject WH-question pronouns and support verbs need to be treated as “real” non-local dependencies. In question sentences, before the main parsing is started, the support verb is attached to any lonely participle chunk in the sentence, and the WH-pronoun pre-parses with any verb, as we discuss in the following section.

2.1 Localising Long-Distance Dependencies

LDDs are traditionally grouped into two classes (see e.g. (Pollard and Sag, 1994, p. 157)). In the first class, there is an overt constituent in a nonargument position that can be thought of as strongly associated with (or filling) the gap or trace. An argument is fronted to a non-argument position. In this class we find topicalisations, WH-questions and relative clauses. In the second class there is no overt filler in a nonargument position, instead there is a constituent in an argument position that is interpreted as coreferential with the trace. Functionally speaking, a constituent that is realized once appears more than once as a semantic argument of a predicate. In the second class we find control and raising and *it*-cleft constructions.

For the second class, context-free parsing is sufficient, because the coreference of the argument positions is resolved at the post-processing stage by means of a statistical method. For control and raising, if a subordinate clause is subjectless and is in the infinitive, a decision based on the lexical probability of the superordinate verb or adjective to introduce subject or object control constructs a coreference. Parsing can stay context-free because there is no dependence between syntactic ambiguity and control or relative clause antecedent resolution.

We have discussed that most LDDs of the first class, with the notable exception of non-subject WH questions, can be treated locally in Dependency Grammar. We now discuss the mild context-sensitive approach that Tree-Adjoining Grammar (Joshi, 1985) uses for such WH questions. It has been suggested that mild context-sensitivity is expressive enough for natural language processing (Frank, 2002).

2.1.1 TAG Adjoining and mild context-sensitivity

The TAG formalism (Joshi, 1985; Joshi and Kroch, 1985) has developed a mathematically restrictive formulation of phrase structure grammar. In contrast to the string-rewriting systems of the Chomsky hierarchy, TAG is a system of tree-rewriting. Structural representations are built up from pieces of phrase structure, so-called *elementary trees*, which are taken as atomic. These trees can be combined by using one of two operations: *Substitution* and *Adjoining*.

Substitution Substitution involves the rewriting of a non-terminal node at the frontier of one elementary tree as another elementary tree with the requirement that the rewritten node must have the same label as the root of the elementary tree that rewrites it. Substitution can be understood as a traditional rewriting operation. Substitution accomplishes effects similar to those of the Merge operation form (Chomsky, 1995): it inserts XPs into the argument positions of syntactic predicates. Crucially, it is a context-free operation: context-free elementary trees combined by substitution only yield context-free structures. An example of Substitution is in fig. 3

Elementary trees are context-free by definition. “Every syntactic dependency is expressed locally within a single elementary tree” (Frank, 2002, p. 22)

Adjoining The Adjoining operation rewrites a non-terminal node anywhere within an elementary tree as another elementary tree. Unlike substitution, which rewrites or expands trees only along the frontier, Adjoining

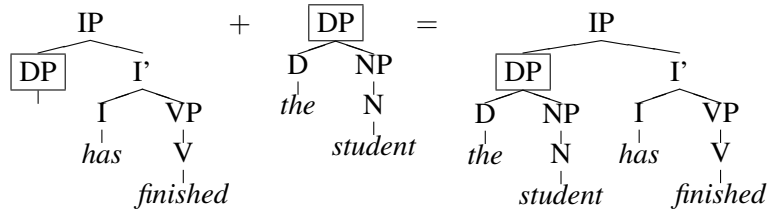


Figure 3: An example of the Substitution operation. The rewritten node is boxed

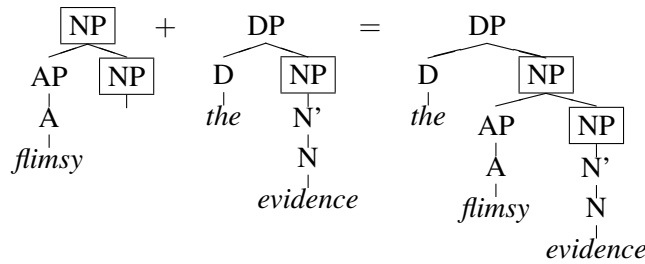


Figure 4: An example of the Adjoining operation. The foot node is boxed

uses a special class of recursive trees, so-called *auxiliary trees*. The root of an auxiliary tree is labeled identically to some node along its frontier, the *foot node*.

Given an auxiliary tree A with foot node X , Adjoining rewrites as A a node N that is labeled as X in an elementary tree T , and attaches the node that was under N in T at the foot node of the auxiliary tree. Adjoining thus works by rewriting some node of an elementary tree as a recursive piece of structure (the auxiliary tree). An example is in figure 4. Trees that have undergone Adjoining can be subject to subsequent Adjoining operations.

Adjoining on the one hand makes Chomsky adjunction possible. In this case, the recursion of the foot node in the auxiliary tree is across one level only, the recursive nodes are immediate mothers/daughters of each other, as in 4. On the other hand, TAG also allows the use of auxiliary trees in which the recursion stretches across several nodes. In this fashion, auxiliary trees that contain terminal nodes between the two recursive nodes can be inserted into elementary trees and thus stretch out local dependencies. An example is in figure 5.

TAG treats this sentence as follows: First, the dependency between the WH-element and its base position is established locally, within a single elementary tree, according to TAG principles. The effect of dislocating the WH-element into a higher clause is accomplished by means of Adjoining in fig. 5. Further embedding of instances can be derived analogously by further Adjoining operations.

Such stretching by Adjoining with recursive auxiliary trees is the one and only way in which context-sensitive constructions can be generated in TAG. This fact is known as the nonlocal dependency corollary: “Nonlocal dependencies always reduce to local ones once recursive structure is factored out.” (Frank, 2002, p. 27). Current research in TAG reveals that the severely restricted type of context-sensitivity generated by Adjoining, so-called *mild context-sensitivity*, accurately characterizes the non-locality present in natural language (Frank, 2002).

2.1.2 The Nature of Elementary and Auxiliary Trees

While the basic operations over elementary and auxiliary trees have been outlined now, nothing has been said about the nature of these trees. We will follow Frank (2004) and “assume that elementary trees are built around a single lexical element, that is, a semantically contentful word like a noun, verb or adjective” (Frank, 2004, p. 11).

This means that elementary trees are similar to DG nuclei or chunks (if we allowed attributive adjectives to be part of elementary trees). Elementary trees are assumed to provide argument slots and are closely related to predicate-argument structure:

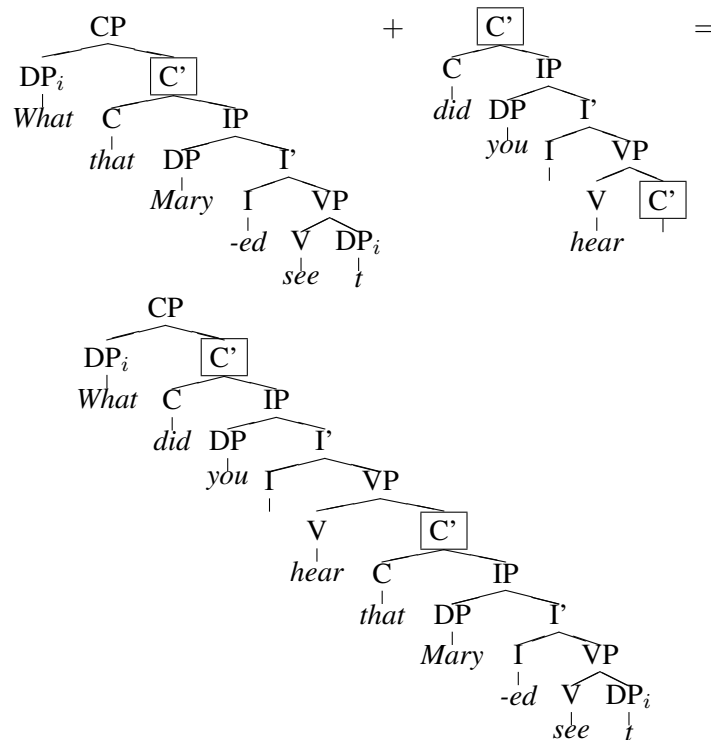


Figure 5: Adjoining for WH-questions. The deep recursion of the auxiliary trees introduces mild context-sensitivity. The foot node is boxed.

A great deal of work in syntactic theory has assigned a privileged status to the syntactic analogue of predicate argument structure. Such a domain, which we call a *thematic domain*, consists of a single lexical predicate along with the structural context in which it takes its arguments. This notion takes a variety of forms and names, but the same idea seems to underlie kernel sentences in Harris (1957) and Chomsky (1955; Chomsky (1957), cyclic domains in Chomsky (1965), strata in Relational Grammar (Perlmutter, 1983), f-structure nuclei in LFG (Bresnan, 1982) and governing categories in Government-Binding Theory (Chomsky, 1981). (Frank, 2002, p. 38)

DG parses directly for a predicate argument structure and DG structures have been described as the f-structure part of LFG (Bröker, Hahn, and Schacht, 1994). DG and TAG thus take a very similar stance on the inherent aims and structures of syntactic theory. Following work by Grimshaw (1991), elementary trees are assumed to include extended projections. “Grimshaw (1991) characterizes the linkage between between lexical and functional projections via a notion she labels *extended projection*. In essence, the extended projection of a lexical head includes the projections of all those functional heads that embed it (up through but not including the next lexical head).” (Frank, 2002, p. 43). Auxiliary trees are defined as elementary trees that show the recursive characteristics described.

TAG uses transformations to generate elementary trees. Grimshaw (1991) and Frank (2002) discuss that in head-movement the base position and the ultimate landing site lie within a single extended projection. This entails that head-movement generally is not unbounded. We have discussed in 2.1 for English how finite-state patterns can be used to cover them. Elementary trees, which include extended projections, are much larger than the production rules that are used in phrase-structure (PSG) frameworks. Therefore, many dependencies (for example head-movement) that stretch across more than a mother-daughter node relation and are thus non-local for PSG remain local in TAG, as they only involve a single elementary tree. The extended projections of a TAG elementary tree (Grimshaw, 1991) are also called *extended domain of locality* (EDOL) (Carroll et al., 1999). Much of the reduction in TAG grammar complexity is owed to EDOL. Features do not need to be percolated,

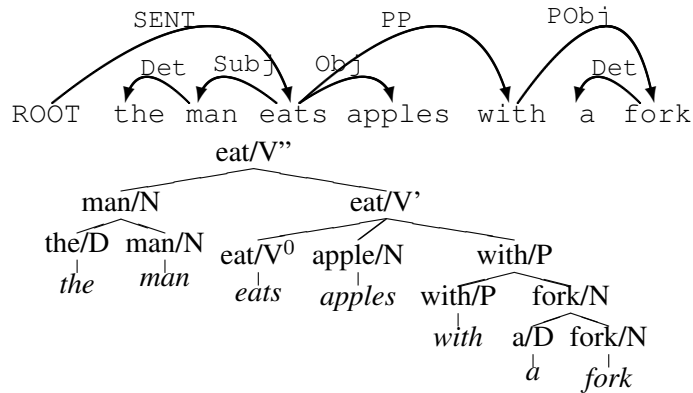


Figure 6: A DG representation and a principled conversion of DG to X-bar representation

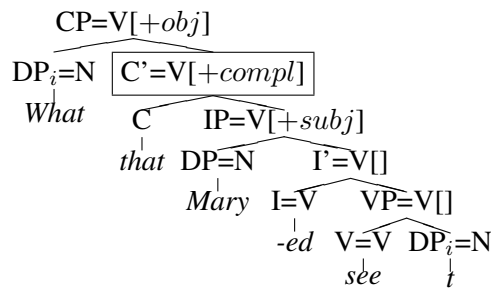


Figure 7: A DG to TAG tree mapped representation. DG relation labels are in square brackets.

and parsing algorithms of lower complexity can be applied. EDOL has a practical benefit for broad-coverage parsing, reducing search spaces and the number of unifications needed in a unification-based grammar.

2.1.3 TAG Adjoining in DG

DG shares EDOL with TAG, because it only knows content word projections (*nuclei*). At the same time, because DG grammar rules are binary, grammar size, which is a parameter in parsing complexity, stays low.

In LFG f-structure, HPSG and Functional DG, where functional projections appear below the content-word head as what HPSG has termed markers, the elementary tree of a word W that falls into a content word class and the maximal projection of W coincide. All bar-levels are isomorphic to the head word W in DG (Miller, 1999). The important difference between W s at different bar-levels is that they have attached more or less dependents. Different projections of W can be seen as different stages of derivation. A possible conversion from DG to X-bar for example distinguishes between a projection or derivation state of V with all dependents except subject attached (V' , internal arguments), and a projection or derivation state of V with all dependents attached (V'' , including the external argument). Such a conversion, and the equivalence of DG and X-bar is described in (Covington, 1994) and illustrated in fig. 6. A DG to TAG tree mapped representation following from that is shown in fig. 7. Unlike in TAG, the equivalent of elementary trees are also constructed without transformations in DG. The verb has local access to the fronted object in the elementary tree, i.e. in a non-embedded WH-question, just like in LFG f-structure, where all arguments appear flat under the verb predicate.

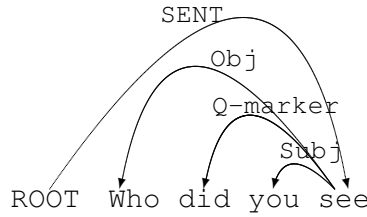
Because functional words are attached as markers, all DG equivalents of functional projections (the combination of a head word and a function word) are also all projections of the head word. The only possible foot node N in DG is therefore a projection W of the head word W . Adjoining inserts a recursive structure at some projection N which is called the foot node. The head of the inserted structure is N , and the part of the elementary tree that appeared below N occurs below the recursive N . Since the foot node N of the inserted

auxiliary tree appears above the N of the original elementary tree, Adjunction inserts new governors into an existing structure and thus breaks the context-freeness. In a nutshell, the DG difference between Substitution and Adjoining is: Substitution inserts dependents, Adjoining inserts governors.

In DG, Adjoining inserts an auxiliary tree into some projection or derivation stage of W . Adjoining to maximal projections (in which all dependents are attached) is pointless, because then Adjoining A to B is equivalent to Substituting B to A . The point is that the auxiliary tree is inserted at a derivation stage in which not all dependents have been attached, at a partial projection.

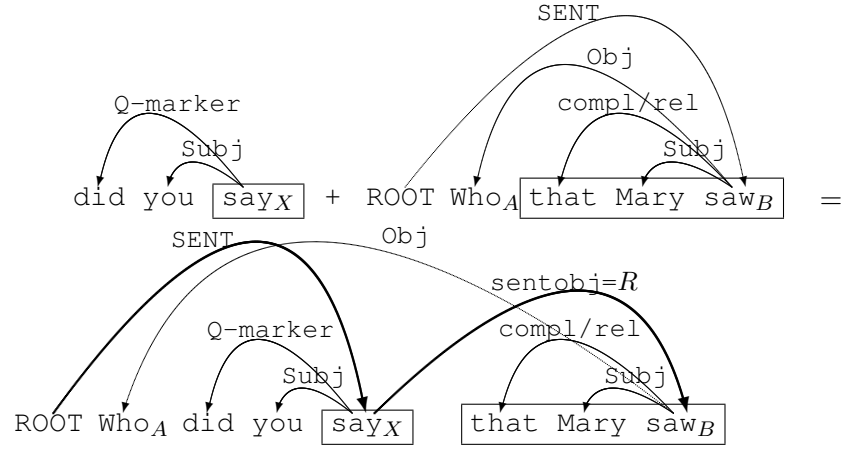
While in the example of 6 derivation order coincides with the internal/external argument ordering, that may not necessarily be so. If a standard CYK algorithm is employed for

(2) *Who did you see ?*



the subject is attached before the object, as can be seen in fig. 7. At the stage where all dependents except for the object are attached, Adjoining can occur.

(3) *Who did you see ?*



Adjoining can be described as follows in Functional DG: Given a local relation (of a type falling inside a TAG elementary tree, hence non-clausal) from B to A , if there is a maximal projection equivalent to a TAG auxiliary tree X , and if there can be a relation (or a relation chain) from X to both A and B such that

- the relation type R from X to B is across elementary trees, hence clausal,
- the possible relation from X to A would be of the same type as the one from B to A
- the governor of B is also licenced to be governor of X , and has the same relation type

then X can adjoin into the structure formed by A and B . Adjoining inserts X between A and B , constructs a relation R from X to B , and the governor of B becomes the governor of X – this is the mildly context-sensitive relation. As a result, the possible surface relation from X to A remains unrealized, delegated to the head of a lower clause (Nivre and Nilsson, 2005).

If we apply the principled conversion suggested in Covington (1994) for the conversion between a labelled DG relation and a constituent tree we can conclude that if every projective DG relation corresponds to a TAG elementary tree and every trigger for a non-projective DG relation corresponds to a TAG auxiliary tree, then DG and TAG are equivalent.

2.1.4 TAG Adjoining in LFG

LFG uses functional uncertainty for mild context-sensitivity (Kaplan and Zaenen, 1989; Dalrymple, Kaplan, and King, 2001). Functional uncertainty allows LDDs to extend across an unlimited, recursive path. Subordinate clauses appear as a COMP or XCOMP (the latter for control) dependent in f-structure, hence the recursion, expressed by the Kleene star, is COMP* or XCOMP*, but this is equivalent to TAG recursion on C-bar or DG recursion on V.

Modelling the recursion on the functional level, as in LFG or the suggested DG approach leads to a representationally minimal theory (Jurafsky, 1996).

2.1.5 Implementation

An implementation for the treatment of such embedded WH-dependencies exists in Pro3Gres. TAG Adjoining recursively inserts local trees into the middle of other trees. Due to this characteristic, only LDDs from the beginning of one elementary tree to the end of the originally same (elementary) tree can be generated.

In non-subject WH-questions, the WH-pronoun appears at the front of the sentence rather than in its usual post-verbal position. The implemented approach is based on pre-parsing: in WH-pronoun question sentences, before the main parsing is started, the WH-pronoun pre-parses with each verb, which may constitute the end of the originally same (elementary) tree.

We have thus implemented a simple version of TAG Adjoining or equivalently LFG functional uncertainty by using mild context-sensitivity in order to fulfill the goal of reducing grammatical complexity and expressiveness.

3 Probability Model

Pro3Gres is a probabilistic parser that parses between heads of chunks and thus profits from a combination of finite-state techniques and parsing. The chunks & dependencies model has been suggested by (Abney, 1995). It is described as psycholinguistically adequate (Crocker and Corley, 2002), especially when combined with a statistical model by (Jurafsky, 1996). (Frank, 2003) presents a (albeit non-probabilistic) chunks & dependencies model for LFG. Chunks can be freely combined subject to adjacency and projectivity (contiguity) constraints, which leads to a context-free parsing algorithm. Except for the added book-keeping functional annotations, her parsing algorithm is akin to CYK, which we use. Unlike (Frank, 2003), Pro3Gres is probabilistic. This is an important asset for a robust, broad-coverage and practically applicable parser. The statistical model that we suggest cannot be said to be probabilistic in the sense that it captures the probability of *generating* a sentence (Charniak, 1996; Collins, 1999), but rests on the psycholinguistically adequate assumption that parsing is a decision process. The probabilities of possible decisions at an ambiguous point in the derivation are assumed to add up to 1 (Crocker and Brants, 2000). In this sense, its probability estimation is closer to *discriminative* models (Johnson, 2001).

We will explain Pro3Gres' main probability model by way of comparing it to (Collins, 1996). Both Collins (1996) and Pro3Gres are mainly dependency-based statistical parsers parsing over heads of chunks, a close relation can therefore be expected. The Collins (1996) MLE and the main Pro3Gres MLE can be juxtaposed as follows:

$$(4) \text{ Collins (1996) MLE estimation: } P(R|\langle a, atag \rangle, \langle b, btag \rangle, dist) \cong \frac{\#(R, \langle a, atag \rangle, \langle b, btag \rangle, dist)}{\#(\langle a, atag \rangle, \langle b, btag \rangle, dist)}$$

$$(5) \text{ Main Pro3Gres MLE estimation: } P(R, dist|a, b) \cong p(R|a, b) \cdot p(dist|R) \cong \frac{\#(R, a, b)}{\#(a, b)} \cdot \frac{\#(R, dist)}{\#R}$$

The following differences are observed:

- Pro3Gres does not use tag information. The first reason for this is because the licensing, hand-written grammar is based on Penn tags.

- The second reason for not using tag information is because Pro3Gres backs off to semantic WordNet classes (Fellbaum, 1998) which has the advantage that it is more fine-grained³.
- Pro3Gres uses real distances, measured in chunks, instead of a vector of features. While the type of relation R is lexicalized, i.e. conditioned on the lexical items, the distance is assumed to be dependent only on R . This is based on the observation that some relations typically have very short distances (e.g. verb-object), others can be quite long (e.g. Verb-PP attachment). This observation greatly reduces the sparse data problem. (Chung and Rim, 2003) have made similar observations for Korean.
- The co-occurrence count in the MLE denominator is not the sentence-context, but the sum of competing relations. For example, the *object* and the *adjunct* relation are in competition, as they are licensed by the same tag sequence ($VB^* NN^*$). Pro3Gres models attachment probabilities as decision probabilities, which is in accordance with the view that parsing is a decision process.
- Relations (R) have a Functional Dependency Grammar definition, including long-distance dependencies.

4 Evaluation

In traditional constituency approaches, parser evaluation is done in terms of the correspondence of the bracketing between the gold standard and the parser output. Lin (1995) suggested evaluating on the linguistically more meaningful level of syntactic relations. For the current evaluation, a hand-compiled gold standard following this suggestion is used (Carroll, Minnen, and Briscoe, 1999). It contains the grammatical relation data of 500 sentences from the Susanne corpus⁴.

| | Percentage Values for | | | |
|-----------|--|--------|---------------|---------|
| | Subject | Object | noun-PP | verb-PP |
| Precision | 91 | 89 | 73 | 74 |
| Recall | 81 | 83 | 67 | 83 |
| | Comparison to Lin (on the whole Susanne corpus) | | | |
| | Subject | Object | PP-attachment | |
| Precision | 89 | 88 | 78 | |
| Recall | 78 | 72 | 72 | |
| | Comparison to Buchholz (Buchholz, 2002), according to Preiss | | | |
| | Subject | Object | | |
| Precision | 86 | 88 | | |
| Recall | 73 | 77 | | |
| | Comparison to Charniak (Charniak, 2000), according to Preiss | | | |
| | Subject | Object | | |
| Precision | 82 | 84 | | |
| Recall | 70 | 76 | | |

Table 4: Results of evaluating the parser output on subject, object and PP-attachment relations and a partial comparison

³For the semantic backoff of verbs, a version in which verbs use a Levin class (Levin, 1993) backoff has been tested. But Wordnet backoff performs better, possibly due to the fact that Levin coverage is lower

⁴The 500 sentences are a random sample of all those sentences from the Susanne corpus which their system was able to parse

| Relation | RASP | | Pro3Gres | | | |
|----------|-----------|--------|-----------|--------------|--------|--------------|
| | Precision | Recall | Precision | | Recall | |
| | % | % | % | # | % | # |
| nmod | 78 | 73 | 75.0 | 1590 of 2119 | 70.6 | 1690 of 2391 |
| arg_mod | 84 | 41 | 76.1 | 16 of 21 | 51.2 | 21 of 41 |
| ncsubj | 85 | 88 | 92.6 | 825 of 891 | 81.1 | 775 of 956 |
| dobj | 86 | 84 | 88.7 | 425 of 479 | 84.5 | 317 of 375 |
| obj2 | 39 | 84 | 90.0 | 9 of 10 | 56.3 | 9 of 16 |
| iobj | 42 | 65 | 74.8 | 80 of 107 | 56.1 | 88 of 157 |

Table 5: Comparison of evaluation results to RASP

| | LDD relations results for | |
|---|---------------------------|------|
| WH-Subject Precision | 57/62 | 92% |
| WH-Subject Recall | 45/50 | 90% |
| WH-Object Precision | 6/10 | 60% |
| WH-Object Recall | 6/7 | 86% |
| Anaphora of the rel. clause subject Precision | 41/46 | 89% |
| Anaphora of the rel. clause subject Recall | 40/63 | 63% |
| Passive subject Recall | 132/160 | 83% |
| Precision for subject-control subjects | 40/50 | 80% |
| Precision for object-control subjects | 5/5 | 100% |
| Precision of <i>modpart</i> relation | 34/46 | 74% |
| Precision for topicalized verb-attached PPs | 25/35 | 71% |

Table 6: Available results for relations traditionally considered to involve LDDs

Comparing these results to Lin (1998) and Preiss (2003) as far as is possible shows that the performance of the parser is state-of-the-art (see table 4). Carroll, Minnen, and Briscoe (2003) have evaluated their own parser (RASP) using this evaluation scheme. Their reported performance is compared to the Pro3Gres in table 5. We have used a simple post-processor to recover chunk-internal relations and do an argument/adjunct distinction for PPs. It appears that Pro3Gres performs better on chunk-external, RASP better on chunk-internal relations.

The new local relations corresponding to LDDs in the Penn Treebank have been selectively evaluated as far as the annotations permit, shown in table 6. For NP traces and NP PRO, the annotation does not directly provide all the necessary data. Passivity is not currently expressed in the predicate-argument parser output, thus only recall values can be delivered. Since Carroll, Minnen, and Briscoe (2003)'s annotation does not directly express control, reduced relative clauses or the dependency direction, only reliable precision values are available in those cases. As for gerunds, neither Carroll nor the parser output retains tagging information, which makes a selective evaluation of them impossible. The fact that performance for the new local relations corresponding to LDDs is not generally lower than in the dependencies corresponding to local constituency, although they correspond to a sequence of decisions in a traditional statistical parser, indicates that our LDD approach improves parsing performance. Absolute values are given due to the low counts of these relatively rare relations.

Table 7 shows that about half of the PP-attachment errors are real attachment errors. The second most frequent error is deficient tagging or chunking – the price to pay for shallowness.

| Error Classification of PP-Attachment Errors of the first 100 evaluation corpus sentences | | | | | | |
|---|-----------------------|---------------------------|------------------|-------------------------------------|--------------------|--|
| Attachment Error | Head Extraction Error | Chunking or Tagging Error | compl/prep Error | Grammar Mistake or incomplete Parse | Grammar Assumption | |
| Noun-PP Attachment Precision | | | | | | |
| 22 | 1 | 8 | 0 | 3 | 3 | |
| Verb-PP Attachment Precision | | | | | | |
| 12 | 1 | 5 | 1 | 1 | 2 | |
| Noun-PP Attachment Recall | | | | | | |
| 25 | 1 | 14 | 0 | 12 | 5 | |
| Verb-PP Attachment Recall (on PP arguments only) | | | | | | |
| 2 | 0 | 1 | 0 | 0 | 0 | |
| Percentages | | | | | | |
| 51% | 3% | 24% | 1% | 13% | 12% | |

Table 7: Analysis of PP-Attachment Errors

5 Conclusions

We have presented a fast, lexicalized broad-coverage parser delivering simple f-structures as output. An evaluation at the grammatical relation level shows that its performance is state-of-the-art.

We have shown that the parser stays as shallow as is possible for each task, combining shallow and deep-linguistic methods by integrating chunking and by expressing long-distance dependencies in a mostly context-free way, thus offering on the one hand a parsing complexity as low as for a probabilistic parser, but on the other hand a deep-linguistic analysis as with a type of formal grammars.

We have discussed that the vast majority of long-distance dependencies can be modelled locally in a functional representation. We have discussed the nature of the remaining truly context-sensitive cases, namely mild context-sensitivity as recursion over syntactic structures in TAG or equivalently, but representationally minimal, recursion over f-structures in LFG or DG. Unlike in TAG elementary trees, movement is obviated.

Following these theoretical considerations, the LFG suggestion by Frank (2003), as well as our broad-coverage evidence (Schneider, Dowdall, and Rinaldi, 2004; Rinaldi et al., 2004a; Rinaldi et al., 2004b; Weeds et al., 2005), we suggest that c-structures or other configurational “surface” representations may be obviated for the syntactic analysis of natural language. By reducing grammar complexity (Frank, 2002; Frank, 2004), by reducing parsing complexity to mostly context-free parsing and finite-state based chunking (Schneider, 2003; Schneider, 2004), by bridging the gap between language engineering and formal grammar (Kaplan et al., 2004) and by aiming for a representationally minimal theory (Jurafsky, 1996) we conclude that chunks and dependencies (Abney, 1995; Frank, 2003) may be sufficient for a formal grammar theory.

References

- Abney, Steven. 1995. Chunks and dependencies: Bringing processing evidence to bear on syntax. In Jennifer Cole, Georgia Green, and Jerry Morgan, editors, *Computational Linguistics and the Foundations of Linguistic Theory*, pages 145–164. CSLI.
- Abney, Steven. 1996. Partial parsing via finite-state cascades. In John Carroll, editor, *Proc. of the Workshop on Robust Parsing at the 8th Summer School on Logic, Language and Information*, number 435 in CSRP, pages 8–15. University of Sussex, Brighton.
- Brants, Thorsten and Matthew Crocker. 2000. Probabilistic parsing and psychological plausibility.

- In *Proceedings of 18th International Conference on Computational Linguistics COLING-2000*, Saarbrücken/Luxembourg/Nancy.
- Bresnan, Joan, editor. 1982. *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, Massachusetts.
- Bröker, Norbert, Udo Hahn, and Susanne Schacht. 1994. Concurrent lexicalized dependency parsing: The ParseTalk model. In *Coling 94*, pages 379–385.
- Buchholz, Sabine. 2002. *Memory-Based Grammatical Relation Finding*. Ph.D. thesis, University of Tilburg, Tilburg, Netherlands.
- Burke, M., A. Cahill, R. O'Donovan, J. van Genabith, and A. Way. 2004. Treebank-based acquisition of wide-coverage, probabilistic LFG resources: Project overview, results and evaluation. In *The First International Joint Conference on Natural Language Processing (IJCNLP-04), Workshop "Beyond shallow analyses - Formalisms and statistical modeling for deep analyses"*, Sanya City, China.
- Cahill, Aoife, Michael Burke, Ruth O'Donovan, Josef van Genabith, and Andy Way. 2004. Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations. In *Proceedings of ACL-2004*, Barcelona, Spain.
- Carroll, John, Guido Minnen, and Edward Briscoe. 2003. Parser evaluation: using a grammatical relation annotation scheme. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*. Kluwer, Dordrecht, pages 299–316.
- Carroll, John, Guido Minnen, and Ted Briscoe. 1999. Corpus annotation for parser evaluation. In *Proceedings of the EACL-99 Post-Conference Workshop on Linguistically Interpreted Corpora*, Bergen, Norway.
- Carroll, John, Nicolas Nicolov, Olga Shaumyan, Martine Smets, and David Weir. 1999. Parsing with an extended domain of locality. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Charniak, Eugene. 1996. Tree-bank grammar. Technical Report Technical Report CS-96-02, Department of Computer Science, Brown University.
- Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the North American Chapter of the ACL*, pages 132–139.
- Chomsky, Noam. 1955. *The logical structure of linguistic theory*. Distributed by University of Indiana Linguistics Club.
- Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.
- Chomsky, Noam. 1981. *Lectures on Government and Binding*. Foris Publications, Foris Publications.
- Chomsky, Noam. 1995. *The Minimalist Program*. The MIT Press, Cambridge, Massachusetts.
- Chung, Hoojung and Hae-Chang Rim. 2003. A new probabilistic dependency parsing model for head-final, free word order languages. *IEICE Transaction on Information & System*, E86-D, No. 11:2490–2493.
- Collins, Michael. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 184–191, Philadelphia.

- Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.
- Covington, Michael A. 1994. An empirically motivated reinterpretation of Dependency Grammar. Technical Report AI1994-01, University of Georgia, Athens, Georgia.
- Crocker, Matthew and Thorsten Brants. 2000. Wide coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, 29(6):647–669.
- Crocker, Matthew and Steffan Corley. 2002. Modular architectures and statistical mechanisms: The case from lexical category disambiguation. In Paola Merlo and Suzanne Stevenson, editors, *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*. John Benjamins, Amsterdam.
- Dalrymple, Mary, Ronald Kaplan, and Tracy Holloway King. 2001. Weak crossover and the absence of traces. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG01 Conference*, Hong Kong. CSLI.
- Eisner, Jason. 1997. Bilexical grammars and a cubic-time probabilistic parser. In *Proceedings of the 5th International Workshop on Parsing Technologies*, pages 54–65, MIT, Cambridge, MA, September.
- Eisner, Jason. 2000. Bilexical grammars and their cubic-time parsing algorithms. In Harry Bunt and Anton Nijholt, editors, *Advances in Probabilistic and Other Parsing Technologies*. Kluwer.
- Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Frank, Anette. 2003. Projecting F-structures from chunks. In Miriam Butt and Traci Holloway King, editors, *Proceedings of the LFG03 Conference*, Albany, NY. CSLI.
- Frank, Robert. 2002. *Phrase Structure Composition and Syntactic Dependencies*. MIT Press, Cambridge, MA.
- Frank, Robert. 2004. Restricting grammatical complexity. *Cognitive Science*, 28(5).
- Grimshaw, Jane. 1991. Extended projection. manuscript.
- Hajič, Jan. 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*. Karolinum, Charles University Press, Prague, pages 106–132.
- Harris, Zellig. 1957. Co-occurrence and transformation in linguistic structure. *Language*, pages 283–340.
- Hockenmaier, Julia and Mark Steedman. 2002. Generative models for statistical parsing with combinatorial categorical grammar. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia.
- Hudson, Richard. 1987. *Journal of Linguistics*, 23:109 – 132.
- Jijkoun, Valentin. 2003. Finding non-local dependencies: beyond pattern matching. In *Proceedings of the ACL 03 Student Workshop*, Budapest.
- Johnson, Mark. 2001. Joint and conditional estimation of tagging and parsing models. In *Proceedings of the 39th Meeting of the ACL*, pages 314–321, Toulouse, France.
- Johnson, Mark. 2002. A simple pattern-matching algorithm for recovering empty nodes and their antecedents. In *Proceedings of the 40th Meeting of the ACL*, University of Pennsylvania, Philadelphia.

- Joshi, Aravind. 1985. How much context-sensitivity is required to provide reasonable syntactic descriptions: Tree Adjoining Grammars. In David Dowty, Lauri Karttunen, and Arnold Zwicky, editors, *Natural Language Parsing: Psychological, computational, and theoretical perspectives*. CUP, Cambridge, pages 206–250.
- Joshi, Aravind and Anthony Kroch. 1985. The linguistic relevance of Tree Adjoining Grammar. Technical Report MS-CS-85-16, Department of Computer and Information Sciences, University of Pennsylvania.
- Joshi, Aravind K. and K. Vijay-Shanker. 1989. Treatment of long-distance dependencies in LFG and TAG: Functional uncertainty in LFG is a corollary in TAG. In *Proceedings of ACL '89*.
- Jurafsky, Daniel. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*.
- Kaplan, Ron, Stefan Riezler, Tracy H. King, John T. Maxwell III, Alex Vasserman, and Richard Crouch. 2004. Speed and accuracy in shallow and deep stochastic parsing. In *Proceedings of HLT/NAACL 2004*, Boston, MA.
- Kaplan, Ronald and Annie Zaenen. 1989. Long-distance dependencies, constituent structure, and functional uncertainty. In Mark Baltin and Anthony Kroch, editors, *Alternative Concepts of Phrase Structure*. Chicago University Press, pages 17 – 42.
- Kaplan, Ronald M. and Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA, pages 173–281.
- Levin, Beth C. 1993. *English Verb Classes and Alternations: a Preliminary Investigation*. University of Chicago Press, Chicago, IL.
- Lin, Dekang. 1995. A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of IJCAI-95*, Montreal.
- Lin, Dekang. 1998. Dependency-based evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*, Granada, Spain.
- Marcus, Mitch, Beatrice Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.
- Mel'čuk, Igor. 1988. *Dependency Syntax: theory and practice*. State University of New York Press, New York.
- Miller, Philip H. 1999. *Strong Generative Capacity*. CSLI, Stanford, CA.
- Miyao, Yusuke, Takashi Ninomiya, and Jun'ichi Tsujii. 2004. Corpus-oriented grammar development for acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. In *Proceedings of IJCNLP-04*.
- Neuhaus, Peter and Norbert Bröker. 1997. The complexity of recognition of linguistically adequate dependency grammars. In *Proceedings of the 35th ACL and 8th EACL*, pages 337–343, Madrid, Spain.
- Nivre, Joakim. 2004. Inductive dependency parsing. In *Proceedings of Promote IT*, Karlstad University.
- Nivre, Joakim and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 99–106, Ann Arbor, Michigan, June. Association for Computational Linguistics.

- Perlmutter, David. 1983. *Studies in Relational Grammar I*. Chicago University Press, Chicago.
- Pollard, Carl and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago University Press, Chicago, Illinois.
- Preiss, Judita. 2003. Using grammatical relations to compare parsers. In *Proc. of EACL 03*, Budapest, Hungary.
- Riezler, Stefan, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA.
- Rinaldi, Fabio, James Dowdall, Gerold Schneider, and Andreas Persidis. 2004a. Answering Questions in the genomics domain. In *ACL-2004 workshop on Question Answering in restricted domains*, Barcelona, Spain.
- Rinaldi, Fabio, Gerold Schneider, James Dowdall, Christos Andronis, Andreas Persidis, and Ourania Konstanti. 2004b. Mining relations in the GENIA corpus. In *Proceedings of the Second workshop on Data Mining and Text Mining for Bioinformatics*, Pisa, Italy.
- Sarkar, Anoop, Fei Xia, and Aravind Joshi. 2000. Some experiments on indicators of parsing complexity for lexicalized grammars. In *Proceedings of COLING '00*.
- Schneider, Gerold. 2003. Extracting and using trace-free Functional Dependencies from the Penn Treebank to reduce parsing complexity. In *Proceedings of Treebanks and Linguistic Theories (TLT) 2003*, Växjö, Sweden.
- Schneider, Gerold. 2004. Combining shallow and deep processing for a robust, fast, deep-linguistic dependency parser. In *ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP (ComShaDeP 2004)*, Nancy, France, August 2004.
- Schneider, Gerold, James Dowdall, and Fabio Rinaldi. 2004. A robust and deep-linguistic theory applied to large-scale parsing. In *Coling 2004 Workshop on Robust Methods in the Analysis of Natural Language Data (ROMAND 2004)*, Geneva, Switzerland, August 2004.
- Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Librairie Klincksieck, Paris.
- Tsuruoka, Yoshimasa, Yusuke Miyao and Jun'ichi Tsujii. 2004. Towards efficient probabilistic HPSG parsing: integrating semantic and syntactic preference to guide the parsing. In *Proceedings of IJCNLP-04 Workshop: Beyond shallow analyses - Formalisms and statistical modeling for deep analyses*.
- Weeds, Julie, James Dowdall, Gerold Schneider, Bill Keller, and David Weir. 2005. Using distributional similarity to organise BioMedical terminology. *Terminology*.
- Younger, D. H. 1967. Recognition and parsing of context-free languages in time n^3 . *Information and Control*, 10:189-208.
- Zwicky, Arnold. 1985. Heads. *Journal of Linguistics*, 21:1-30.

THE PERIPHERALITY OF THE ICELANDIC EXPLETIVE

Peter Sells

Stanford University

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

I argue here that the special distribution of the Icelandic expletive *það*, which restricts it to an intuitive ‘first position’, is due to the interaction of general constraints on Icelandic clause structure and the pragmatic function of a clause containing an expletive. The expletive is not restricted to V2 finite clauses, but can appear in principle in all matrix and embedded finite clauses, as well as certain non-finite clauses. I present an LFG analysis of the full range of facts which adopts a much more linear, relational view of Icelandic syntax than has been widely assumed in generative syntax studies.

1. Introduction

The data in (1) illustrate the distribution of the expletive *there* in English:

- (1) a. *(There) was dancing in the living room.
- b. Was *(there) dancing in the living room?
- c. When was *(there) dancing in the living room?

Assuming that basic clauses are rooted in IP, we conclude from this data that the expletive is in SpecIP, a position which must be filled.

The expletive *það* in Icelandic shows a different distribution, for which it has received much attention in the literature (e.g., Zaenen (1985), Rögnvaldsson and Thráinsson (1990), Sigurðsson (1990), Hornstein (1991), Jónsson (1996), among others). While the expletive is grammatical in (2)a, it is ungrammatical in (2)b-c, which is surprising if the expletive is needed to fill a specific position in the clausal structure.

- (2) a. Það var dansað í stofunni.
expl was danced in the.living.room
- b. Var (*það) dansað í stofunni?
was (**expl*) danced in the.living.room?
- c. Þess vagna var (*það) dansað í stofunni.
therefore was (**expl*) danced in the.living.room

These examples illustrate what I refer to as the ‘peripherality’ of *það*; Sigurðsson (2004) considers the expletive to obey ‘First Position Constraint’. We can compare the ungrammaticality of (2)b with the expletive to a corresponding Swedish example (3), from Platzack (1983):

- (3) Satt det en fågel på taket? Swe.
sat *expl* a bird on the.roof
‘Was there a bird sitting on the roof?’

Swedish is like English in terms of the distribution of the expletive.

Assuming an approach in which all V2 clauses are rooted in CP, with an initial XP in SpecCP followed by a finite verb in C, one account of the data in (2) is that *það* appears only in SpecCP, and hence must be maximally peripheral in the clause (Hornstein (1991), Holmberg and Platzack (1995), Wurmbrand (2004), Richards and Biberauer (2005), among others; Berman (2003, 65) suggests that German *es* is only in

Special big thanks go to Jonas Kuhn for providing the raw material for section 5.2, and helping me get the details right. I also received good advice on the presentation of this material from Eve Clark, Bruno Estigarribia, and Laura Staum.

SpecCP). Such an approach might also implicate the presence of *það* with V2 – perhaps, one might suppose that *það* is only necessary to provide the first element in a V2 structure, if nothing else is available.

I will argue against this positional approach; in fact, *það* is never in SpecCP. As I show in section 2, it is sometimes in SpecIP, and sometimes in a non-specifier clause-medial position. As one might expect in LFG, *það* has no c-structure positional restriction per se. My alternative proposal builds on a view of Icelandic clause structure which is not fixated on hierarchical positions, but rather a very simple view in terms of linear positions relative to the (finite) verb. Such an account is independently motivated for the famous Icelandic phenomenon of Stylistic Fronting (Sells (2002)). Section 3 presents the various parts of the linear analysis of Icelandic syntax.

My analysis follows in section 4, based on the intuition that the reason that *það* never follows the first verb of its clause, is that it would have no (pragmatic) function if it did. Some key parts of the specific LFG analysis are that *það* bears the SUBJ function, and therefore can be in SpecIP; and *það* may not bear a DF, and therefore cannot be in SpecCP. As it bears the GF SUBJ, *það* is not merely a c-structure place-holder.

2. The Distribution of *það*

2.1. Finite Clauses

As noted by some authors, there is considerable evidence that *það* can actually surface in SpecIP – (2)a is consistent with this. In embedded clauses, *það* can appear even where it could not be in SpecCP, and where it has nothing to do with V2. Rögnvaldsson and Thráinsson (1990) document a variety of embedded clause types where *það* may appear, and where the surface position of the expletive is clearly SpecIP – following the simple declarative complementizer *að* is one such context. (4) is a relevant similar example, and (5) shows *það* with the main clause complementizer *ætli*, which forms a matrix question without verb movement to C.

- (4) Ég spurði hvort _[IP] það hefðu margir komið í veisluna].
 I asked whether _[IP] *expl* had many.people come to the.party]
 ‘I asked whether many people had come to the party.’

- (5) Ætli _[IP] það verði talað við Jón a morgun]? (Jónsson (1996))
 wonder _[IP] *expl* will.be talked to John tomorrow]
 ‘Will John be interviewed tomorrow?’

Ottóson (1989) proposed that *það* is in SpecIP, and this idea has been adopted by Jónsson (1996) and Sigurðsson (2004), among others. Jónsson (1996) proposes the following account of the data in (2). First, he adopts an IP/CP analysis of V2, in which subject-initial V2 clauses like (2)a are rooted in IP while non-subject-initial V2 clauses like (2)c are rooted in CP. He then proposes that there is a competition between a null expletive (*pro-expl*) and the overt expletive, and that the Avoid Pronoun Principle favors the null expletive. Finally, he argues that *pro-expl* is only licensed under (canonical) government from I, and this is only possible when I has moved to C. Hence, in (2)a, *pro-expl* cannot be licensed, and so the overt expletive is used, in SpecIP. However, in (2)b-c, the finite verb has moved via I to C, so *pro-expl* is licensed and *það* is ungrammatical. Indeed, omitting *það* from those examples gives a fully grammatical sentence, and the account automatically extends to (4)–(5), which have no I-to-C movement.

Sigurðsson (2004) enforces the peripherality of *það* by proposing that main clauses have a null complementizer which attracts the expletive to immediately follow it. The position of this complementizer would be lexicalized in examples like (5) by *ætli*. He notes that any account which puts *það* (necessarily) in SpecCP would have to treat (4)–(5) as examples of CP recursion. This would predict a correlation between clauses

allowing ‘embedded topicalization’ and those allow *það* in the initial position. However, there are several embedded clause types which do not allow embedded topicalization, but which do allow *það*:

- (6) a. Ég verð hissa ef [_{IP} það hefur verið talað um þetta].
 I will.be surprised if [_{IP} *expl* has been talked about this].
 ‘I will be surprised if this has been talked about.’
- b. Ég verð glaður þegar [_{IP} það hefur verið talað um þetta].
 I will.be glad when [_{IP} *expl* has been talked about this].
 ‘I will be glad when this has been talked about.’

From the perspective of LFG, one might wonder whether a positional restriction to a specific c-structure position within CP or IP is a very natural condition. I argue that *það* bears the SUBJ function, but as there are three potential c-structure positions for the SUBJ in Icelandic (see section 3), this does not constrain the linear position of *það*. I will account for the apparently peripheral distribution of *það* by considering interacting functional constraints – in particular, the signalling effects that *það* has in clause-initial position.

2.2. *það* in Raising Structures

Important evidence about the constraints on *það* come from certain examples involving subject-to-object raising (SOR) structures such as (7)a, as any hypothesized function of *það* in main clauses does not carry over to such a context. The expletive is possible as the object of an SOR verb, as originally noted by Thráinsson (1979, 482); see also Platzack (1983, 87) and Bures (1992, 26).

- (7) a. Jón telur (það) vera mys í baðkerinu.
 John believes (*expl*) be mice in the.bathtub
- b. *Jón telur (það) hafa einhver étið hákarlinn.
 John believes (*expl*) have someone eaten the.shark

The expletive is optional in (7)a, as in all embedded contexts (see e.g., (33)).¹ If the lower predicate is transitive, as in (7)b, and if no (thematic) argument is raised, the example is ungrammatical regardless of the presence of *það*.

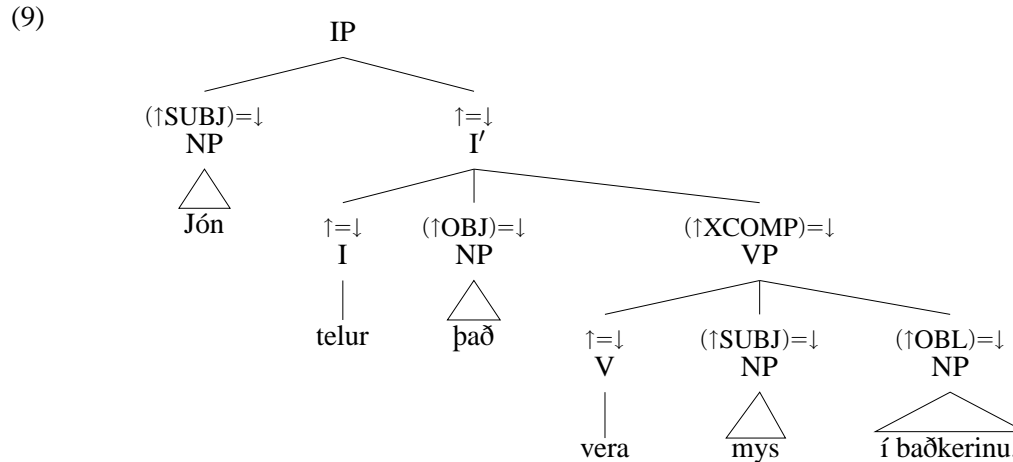
Examples similar to (7)a are given in Andrews (1990, 173):

- (8) a. Ég tel (það) hafa verið dansað á skipunu.
 I believe (*expl*) have been danced on the.ship
 ‘I believe there to have been dancing on the ship.’
- b. Ég tel (það) kveða að honum.
 I believe (*expl*) important to him
 ‘I believe him to be important.’
- c. Ég tel (það) hafa verið beðið eftir honum.
 I believe (*expl*) have been waited after him
 ‘I believe him to have been waited for.’

¹Additionally, only some speakers accept the raising examples (noted by Maling (1988)).

In these examples *það* is actually internal to *I'*, in a non-thematic object position, and is not even a constituent with the following VP. Hence it is unlikely that any analysis which restricts the expletive to a specific c-structure position can cover the full range of data.

As a raising verb, ‘believe’ takes complements that are NP and VP (functionally, OBJ and XCOMP). Crucially, there is no IP structure above the infinitival VP (see Thráinsson (1984, 1993)), which means that there is no ‘medial position’, nor a specifier position, in front of the infinitival verb. Hence the structure of (7)a is the (unsurprising) (9):



As ‘believe’ takes NP and VP complements, it is immediately clear why no version of (7)b can be acceptable: the infinitival verb heads a structure no bigger than a VP, and Icelandic does not allow both direct arguments of the verb to be VP-internal.² Hence the only possible structure involves *einhver* in the raising position:

- (10) Jón telur [einhver] [hafa étið hákarlinn]. (cf. (7)b)
John believes [someone] [have eaten the.shark]

2.3. Summary

We have now seen the following possibilities for the position of *það*, and these exhaust the cases:

- (11) a. *það*: first element in a V2 clause
b. *það*: first element in IP in an embedded non-V2 clause
c. *það*: a ‘raised’ subject under a subject-to-object raising (SOR) verb

The question is now, what unifies exactly these three positions? The descriptive generalization for *það* is simple: it must precede the (every) verb of which it is the SUBJ. This covers the initial examples (2), (4), (5), and the raising examples (7)a and (8). While the expletive follows *telur* in (7)a, it precedes *vera*, the verb of which it is the SUBJ. Note that the generalization cannot be that *það* precedes all coheads in its clause, because it follows the complementizer *að*, for instance, which would be a co-head in C:

- (12) Ég veit [CP að [IP það hefur enginn lesið bókina]].
I know [CP that [IP *expl* has no one read the.book]]
‘I know that no one has read the book.’

²More specifically, Icelandic does not allow Agents and Experiencers to be VP-internal (Maling (1988)).

The first verb in every clause in Icelandic marks whether that clause is finite or not, and there are conditions on clausal structure which make direct linear reference to the first verb, as described below. *það* must precede the exponent of finiteness in its nucleus:

- (13) *það* precedes the exponent of finiteness.

I discuss this condition more thoroughly in section 5.1.

3. Icelandic Clause Structure

In this section I sketch an LFG analysis of Icelandic clausal c-structures, showing that the structures are less hierarchically organized than has been assumed in most of the recent literature, and that major constraints on clausal constituent order derive from linear conditions, not hierarchical ones.

3.1. Hierarchical Positions

Icelandic clause structure has figured prominently in the development of the Minimalist Program (e.g., Chomsky (1995)) due to the various positions that subject and object may take in main clauses, especially in the Transitive Expletive Construction (TEC), which shows two subject positions, either side of the finite verb ((14) is from Bobaljik and Jonas (1996)):

- (14) *Það hafa margir jólasveinar borðað búðinginn.*
 there have many Christmas-trolls eaten the.pudding
 ‘Many Christmas trolls have eaten the pudding.’

Following the finite verb, there is certainly evidence in Icelandic for what we might call a ‘Mittelfeld’: an area of the V2 clause following the initial phrase and the finite verb, where various arguments and adjuncts may appear, to the left of the edge of the canonical VP. It is relatively uncontroversial for the Scandinavian languages that that left edge is marked by the position of negation. Hence in (15), from Jonas and Bobaljik (1993, 90), the constituents *sennilega margir stúdentar þessar bækur aldrei*, including the subject and object, all follow the finite verb in INFL and precede VP:

- (15) *Á bókasafninu settu sennilega margir stúdentar þessar bækur aldrei [VP á borðið].*
 in.the.library put probably many students these books never on.the.table
 ‘In the library, probably many students never put these books on the table.’

The relative order of arguments and adverbials in the medial area is somewhat free, but there is at most only one occurrence of subject and object.

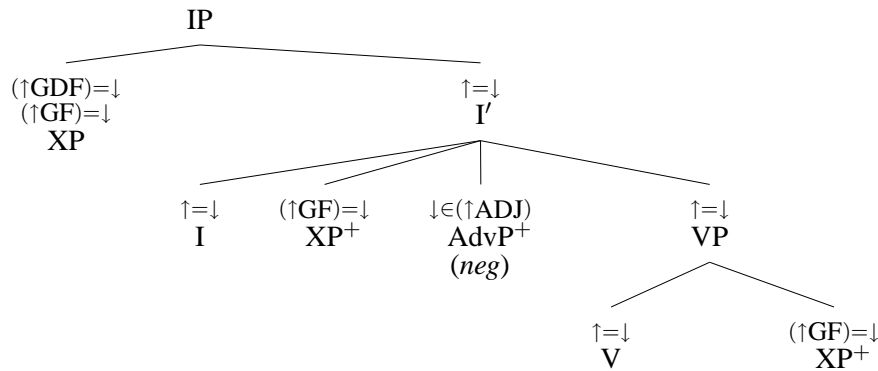
Sells (2001) proposed an analysis of Scandinavian clause structure built around the standard CP-IP-VP spine, which allowed intermediate elements between INFL and the left edge of VP (e.g., negation and other medial adverbs). I argued that, due to the lack of any positive evidence for further hierarchical structure, the medial elements should be analyzed as immediate descendants of I', following a sister INFL and preceding a sister VP. Icelandic allows any kind of definite or quantificational objects, as well as subjects, and adjuncts of many kinds, in the medial domain.³ In fact, from this perspective, we can say that what Bobaljik and Jonas (1996) showed was that Icelandic has a medial domain following the finite verb where all kinds of subjects, objects and adjuncts may appear. Work in the Minimalist Program following on from their proposals has assumed that there are several specifier positions within the clause (e.g., SpecAgrSP, SpecTP, SpecAgrOP,

³In this regard, Icelandic may be more liberal than Swedish, although Börjars et al. (2003) effectively argue for clausal structures like (16) in Swedish, suggesting that the account of Swedish in Sells (2001) was too structurally conservative.

SpecVP – see (46) below), but many of the predicted positions cannot be supported empirically. I briefly discuss the problems with the proposal of Bobaljik and Jonas (1996) in the Appendix at the end of this paper.

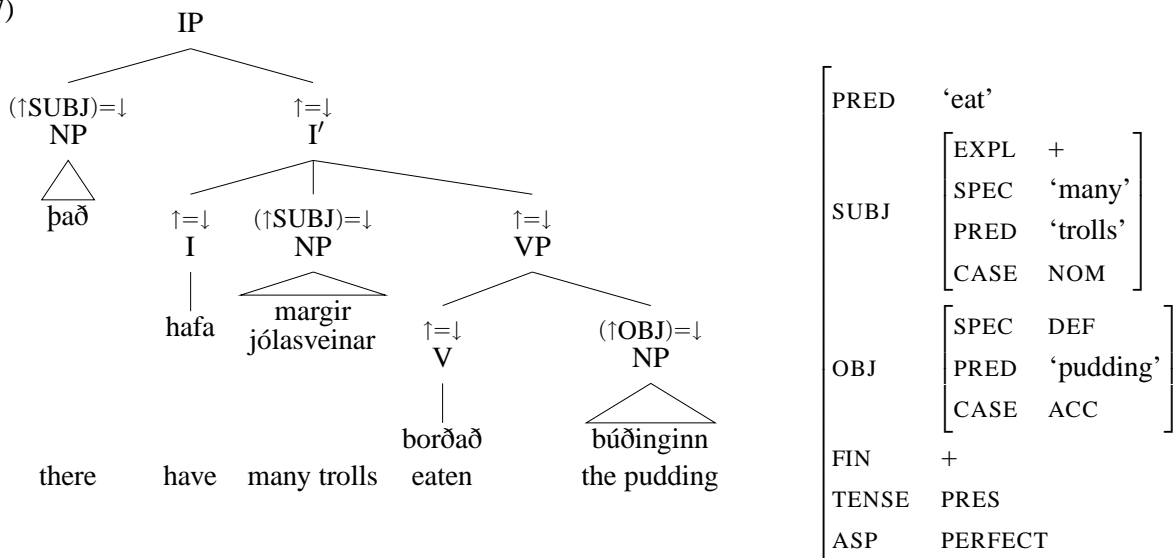
From the LFG perspective, the overall structural possibilities for Icelandic are given in (16) (from Sells (2001, 191)), a relatively flat structure:⁴

(16) Icelandic Clause Structure (Sells (2001, 191):



In the TEC (see (14) above), the expletive *það* is in SpecIP, associated with a thematic subject elsewhere in the clause. By appearing in SpecIP, the expletive prevents any other constituent from being ‘topicalized’, and as it is an expletive, it cannot bear a DF itself. Hence the annotation on SpecIP when the expletive appears there is (↑SUBJ)=↓, and in the TEC the expletive unifies in f-structure with the thematic SUBJ which appears lower in the c-structure. Hence (17) is the structure of (14); the expletive and its ‘associate’ both map to the SUBJ function, though only the latter provides contentful information; the f-structure is in (18).⁵

(17)



For my purposes here, the key point is that there are just 3 linear positions for subjects in Icelandic, the GF positions in (16). This structure illustrates all of the hierarchical properties that are necessary for Icelandic (with CP on top of IP). The main constraints on clause structure are linear, as I now demonstrate.

⁴I assume that the GDFs include SUBJ, and the true DFs TOP and FOC.

⁵Sells (2005) argues that finiteness is an f-structure attribute independent of tense. Finiteness is an essential part of the generalization about the distribution of *það*.

3.2. The V2 Constraint

The approach is one which factors out different and interacting parts of syntactic constructions, based on the general pattern in (16). For example, I will claim that V2 is satisfied in regular finite clauses by a sequence of overt elements in SpecIP and INFL, while V2 can also be satisfied by a sequence of two heads, INFL and V, in Stylistic Fronting clauses. Therefore, V2 cannot be associated with a single hierarchical structural configuration (at least in Icelandic). Rather, it is a constraint which must unify with some sequence of positions in (16); it is given in (18), which looks for two constituents, the first of which is at the left edge of the relevant domain, and the second of which is a finite verb.⁶ For ease of reference below, I refer to the two positions in the V2 structure as V2-1 and V2-2, respectively.

- (18) The V2 constraint: $[\quad \alpha \quad - \quad \underset{\text{V2-1}}{\text{V}} \quad - \quad \dots \quad]$
[+fin]
V2-1 V2-2

A V2 clause will then be characterized by the unification of (18) with some structure conforming to (16). When V2-1 is SpecIP and V2-2 is INFL, the canonical structure, there will be certain pragmatic information associated with the structure (see (23) and (34) below).

3.3. The I⁰ Constraint

There is one more constraint that is part of the definition of Icelandic clauses. INFL is in fact overtly present in all finite clauses which are IPs. Even in embedded clauses, the finite verb always precedes a medial adverb such as the negative *ekki*, as seen in (19), (examples from Holmberg (1986)):

- (19) a. Það var gott að [hann keypti ekki bókina].
 it was good that [he bought not the.book]
- b. Ég veit ekki hvers vegna [Sigga setur aldrei hlutina á réttan stað].
 I know not why [Sigga puts never the.things in the right place]
 ‘I do not know why Sigga never puts the things in the right place.’

These embedded clauses are simple subject-initial non-V2 clauses, in which there is not even an option for the finite verb to follow negation, meaning that the finite verb cannot be internal to VP. This motivates the constraint in (20). The fact that this constraint holds of all finite clauses is what makes embedded clauses in Icelandic look like they are V2 clauses even though, logically, (20) and (18) are separate constraints.

- (20) The I⁰ Constraint: INFL is present in every IP.

IP is present in all finite clauses, and even in some non-finite ones (Thráinsson (1984, 1993)) – in particular, control complements – in which case the first verb is in INFL, as (21) illustrates:

- (21) Risarnir lofa að [_{IP} éta ríkisstjórnir ekki].
 the.giants promise COMP [_{IP} eat.INF governments not]
 ‘The giants promise not to eat the government.’

In the bracketed embedded clause, the verb appears in the INFL position, allowing the object to shift out of the VP (‘Object Shift’), leaving the final word *ekki* marking the left edge of the would-be VP. If the

⁶Cf. Maling and Zaenen (1990), who propose that “the simplest statement of V2 is as a single positive template”.

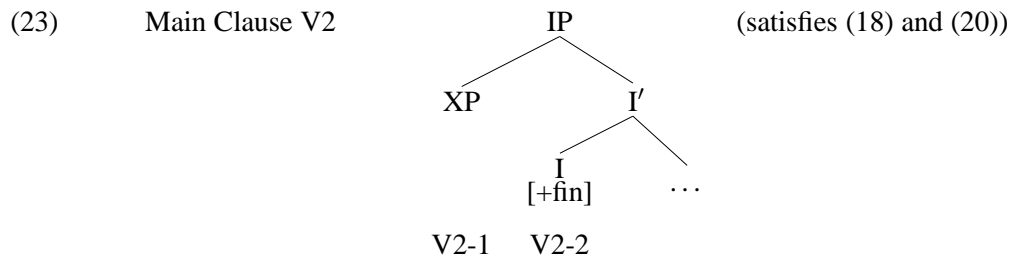
complement to a Control verb such as ‘promise’ is of category IP, as Thráinsson argues, then (20) will require that the INFL head is also present, hosting the non-finite verb. This is exactly what we find in (21).

Finally, in impersonal embedded finite clauses, no subject need precede the finite verb (in SpecIP), yet the verb still must be in INFL. The examples from Sigurðsson (1990, 53) show this clearly:

- (22) a. Ég veit ekki hvers vegna [_{IP} kemur ekki að þessu]. (Sigurðsson (1990, 53))
 I know not why [_{IP} comes not to this]
 ‘I do not know why it does not come to this.’
- b. Við förum ef [_{IP} rignir ekki mikið].
 we will.go if [_{IP} rains not much]

These examples demonstrate the independence of the I⁰ Constraint and the V2 Constraint, as clearly V2 is irrelevant, but the verb must nevertheless be in the INFL position.

Returning to V2 main clauses, (18) and (20) are respected in virtue of the fact that INFL itself hosts the finite verb; the INFL position is the V2-2 part of V2, as shown in (23).



3.4. Stylistic Fronting

In Sells (2002), I argued that Icelandic has another way of simultaneously satisfying (18) and (20), which is manifest in Stylistic Fronting clauses. The Stylistic Fronting construction was brought to the attention of generative syntacticians by Maling (1990) (first published in 1980). Stylistic Fronting is restricted to clauses in which the subject is missing from the canonical initial position, and involves the inversion of a word which would normally follow the finite verb to a position just in front of that verb:

- (24) Stylistic Fronting (Maling (1990) [1980])
- the subject of the clause must be a ‘gap’ (or at least not in the canonical subject position)
 - the clause must be finite
 - the fronted element is a word, not a phrase

A representative set of examples which illustrate Stylistic Fronting involve relative clauses where the subject is relativized, and therefore absent. In (25), the a/c examples are canonical, and the b/d examples involve Stylistic Fronting. I use underlining to indicate the element that is a (potential) target for the fronting, and ‘_’ marks the usual position of the fronted word:

- (25) a. Þetta er tilboð sem [er ekki hægt að hafna].
 this is an.offer that [is not possible to reject]
 ‘This is an offer which it is not possible to reject.’
- b. Þetta er tilboð sem [ekki er _ hægt að hafna]. ← Stylistic Fronting
 this is an.offer that [not is possible to reject]

- c. Þetta er maður sem [hefur leikið níutíu leiki].
 this is a.man that [has played ninety games]
 ‘This is a man who has played ninety games.’
- d. Þetta er maður sem [leikið hefur — níutíu leiki].
 this is a.man that [played has ninety games]

The Stylistic Fronting clauses have a structure that satisfies the verb-second (V2) constraint (see Maling (1990, 73); also Anderson (1997, 20ff.)). However, they do not easily assimilate to canonical SpecIP – INFL structures (as in (23)), as the first element is a word, not a phrase.

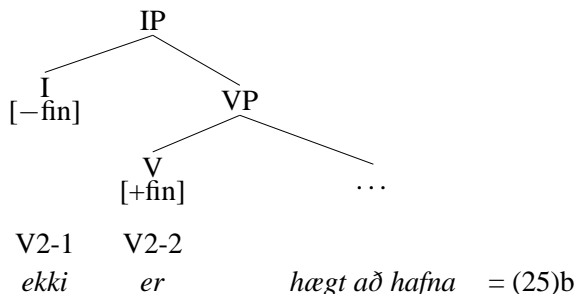
Stylistic Fronting is also possible in main clauses (examples in (26) from Jónsson (1991, 24)), where the affinity with V2 is clear:

- (26) a. Keypt hefur verið — tölva fyrir starfsfólkið.
 bought has been a.computer for the.staff
 ‘A computer has been bought for the staff.’
- b. Fallið hafa — margir hermenn í þessu stríði.
 died have many soldiers in this war
 ‘Many soldiers have died in this war.’

Such clauses are like those introduced by *það* in lacking any topical argument. Rögnvaldsson and Thráinsson (1990) discuss the similarities and differences between main clause ‘topicalization’ (V2 clauses) and Stylistic Fronting. Considering the mechanisms that derive the two kinds of structure, they “conclude that they are *syntactically* a unified process, even though they are certainly different *functionally*” (p. 28). In Sells (2002) I presented an LFG account of Stylistic Fronting, which also adopts the idea that regular V2 clauses and Stylistic Fronting clauses share a syntactic similarity, but in a different way from Rögnvaldsson and Thráinsson (1990): while they analyze the common syntactic process as being movement of some element to SpecIP in both V2 clauses and Stylistic Fronting clauses, my approach is that the two types of clause both instantiate the abstract V2 pattern (18).

For Stylistic Fronting clauses, suppose that INFL is present but hosts a non-finite element, as a marked property. As long as a finite verb is in some head position within the c-structure, the clause will be characterized as finite at f-structure, and of course the possibility of ‘head mobility’ in head positions such as C, INFL, or V is part of the design of the theory (see e.g., Bresnan (2001)). So if a non-finite element is in INFL, this will be the V2-1 part of V2, and then it must be that the next element is a finite verb. As INFL is already filled, the finite verb must appear as the first element in VP, *which is the next available head position*. This same insight is also sketched in Anderson (2000, 328–9). In other words, the structure is as in (27).

- (27) Stylistic Fronting: As a marked option, a non-finite element is generated in INFL. The element in INFL occupies the first position of the V2 constraint.



This satisfies the V2 constraint (18), the INFL constraint (20) and conforms to the structural possibilities in (16) just as well as the canonical SpecIP – INFL– rest-of-clause structure, but as it does this in a different way, we can assume a different functional or stylistic value. This account explains the fact that what fronts is an X^0 , the subject gap restriction, other constraints on Stylistic Fronting, and the restriction to finite clauses.

This account can only be stated if linear and hierarchical conditions are separated, in an analysis which guarantees the structural generalizations in (28) (such as the LFG analysis presented here):

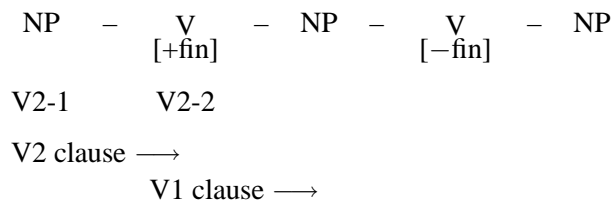
- (28) a. Except for I^0 , **no hierarchical position** is privileged in the clause.
 b. Even the finite V in V2 clauses **is not fixed** in its hierarchical position.

4. Functions in the Clause

4.1. Linear Positions in the Clause

Following on with the reasoning developed in the previous section, I will show here that the linear properties of Icelandic clause structure have certain semantic and pragmatic values, determined by the structural possibilities in (16). As in other V2 languages, the basic contrast is simply between whether a single constituent precedes the finite verb, or whether the finite verb is clause-initial (a ‘V1 clause’):

(29) Linear Positions (cf. (16)):



V2 and V1 clauses have the linear properties shown shown in (29). If V2-1 is absent, we have a verb-initial clause which is interpreted as a polar interrogative (if a main clause), or as a ‘V1 Declarative’ (see (31)b):

- (30) V1 Declaratives (from Sigurðsson (1990))
- a. Það voru oft langar umræður á fundunum.
expl were often long discussions at the.meetings
 ‘There were often long discussions at the meetings.’
- b. Voru oft langar umræður á fundunum.
 were often long discussions at the.meetings

For any argument GF, there are 3 relevant positions, summarized in (31), where V_f and V_l refer to the two V positions in (29) (‘first’ and ‘last’). NPs have different semantic and pragmatic properties in each of these 3 positions:

- (31) ‘Functional’ positions in Icelandic, for some GF
- a. Before the first/finite verb, V_f (‘first position’).
 b. Somewhere after the first/finite verb, V_f , but before the last verb V_l (‘medial position’).
 c. After the last verb, V_l (complement of V position).

Returning to the main topic of this paper, the position of *það*, suppose that it follows V_f . This would force an associate NP to be in the ComplV position. However, the (necessary) presence of V_l indicates the associate NP is in the ComplV position, regardless of the presence of *það*. As detailed by Vangnes (2002), the medial and final positions are only associated with different quantificational properties:

(32)

| Expletive | Intermediate Position (SpecTP) | Postverbal position |
|------------|--------------------------------|------------------------|
| <i>það</i> | *unembedded definite | *unembedded definite |
| <i>það</i> | *generic | *generic |
| <i>það</i> | * \forall /partitive | * \forall /partitive |
| <i>það</i> | indefinite | indefinite |
| <i>það</i> | *non-Q bare indefinite | non-Q bare indefinite |

(Vangnes (2002), Table 1, (his terminology))

Vangnes shows that NPs have their semantics restricted as shown in (32) when in medial and final positions, regardless of whether the initial position is filled by *það* or something else. In other words, while the first two positions of (29) indicate something about clause-type, the last 3 positions serve to indicate quantificational properties of NPs (and presumably, other subtle informational-structural properties).

Recall that there is no phrasal position in Icelandic, such as SpecIP, which needs to be filled (cf. (28)a). This is fundamentally why *það* has a restricted distribution. As nothing about the pragmatics of the clause is signalled by the medial or final NP positions, *það* would have no function if it appeared there.

When *það* does appear, it does carry some pragmatic information about the (sub-)clause in which it appears. For example, *það* in an SOR structure like (7)a indicates that the speaker has chosen not to raise the thematic subject of the infinitival complement. In finite embedded clauses, *það* is never structurally required, but its presence or absence in the initial position has semantic and pragmatic effects, and may be related to whether the clause is asserted or presupposed (see e.g., Rögnvaldsson (1984, 17ff.)):

- (33) a. Ég vissi að það/ \emptyset væri ekið vinstra megin í Ástralíu.
 I knew that *expl*/ \emptyset were driven left side in Australia
 ‘I knew that (there) were driven on the left side in Australia.’
- b. Ég veit að það/* \emptyset er ekið vinstra megin í Ástralíu.
 I know that *expl*/ \emptyset is driven left side in Australia
 ‘I know that *(there) is driven on the left side in Australia.’

The embedded verb in (33)a is past subjunctive, while the verb in (33) is present indicative, and in that case *það* is (pragmatically) obligatory. (Rögnvaldsson suggests that the more strongly a clause is asserted, the less felicitous is the expletive-less version.)

4.2. More on *það*

I have suggested above that *það* may have some pragmatic or signalling function when it precedes the V_f of its clause; in other positions, it has no function, and therefore is dispreferred on general grounds of structural economy. In this subsection, I explain this latter claim a little more. Returning to V2 clauses, we can identify 6 sub-types in Icelandic, depending on the nature of the element in V2-1:

(34) Pragmatic functions in main (V2) clauses:

| Element in V2-1 | Pragmatic Value | Clause Type |
|--------------------------------|--|-----------------|
| subject NP | subject is more topical than any other XP | (declarative) |
| non-subject XP | non-subject is more topical than any other XP | (declarative) |
| <i>það</i> | no XP is topical | (declarative) |
| non-referential X ⁰ | no XP is topical (Stylistic Fronting; e.g., (27)b) | (declarative) |
| subject NP[+wh] | constituent question | (interrogative) |
| non-subject XP[+wh] | constituent question | (interrogative) |

það has the function of indicating a V2 clause in which nothing is topical.

Now I consider in more detail the properties of clauses containing *það* in different positions. (35) shows the schematic distribution in clauses with *það* and a definite subject. In fact, *það* is incompatible with a [+def] subject:

(35) *það* and a [+def] subject

| | | | | |
|---|--------------|----------------|--------------|-----------------------------|
| ✓ | NP [+def] | V _f | | V _l |
| * | <i>það</i> | V _f | NP [+def] | V _l |
| * | <i>það</i> | V _f | | V _l NP [+def] |

This looks like a classic case of the ‘Definiteness Effect’ on existential constructions, though as noted above, Vangsnes (2002) shows that this pattern is not due to *það*, for the same distributional facts hold when the initial position is occupied by a referential non-subject such as the adverb ‘today’. (35) is in fact the kind of case analyzed by Mikkelsen (2002) in an Optimality Theory (OT) approach: a definite NP must be topical, so the first structure in (35) ‘wins’ over the others (this account is effectively anticipated for Icelandic in Sigurðsson (1989, 296ff.)). Mikkelsen proposed an analysis which I have summarized in (36), based on this idea of a priority for initial position:

(36) Priority for initial position: Definite > { Expletive, Indefinite } (Mikkelsen (2002))

In the context of an OT system, the effect of (36) is the following: if a definite is present, it will be in the initial position; if an indefinite is present it may alternate with an expletive for the initial position. A bare indefinite can be in the initial position ((37) from Vangsnes (2002)):

(37) Bjór hefur hellst á golfið.
beer has been.poured on the.floor

Roughly speaking, indefinites can appear in all 3 NP positions, though with some semantic differences between the two non-initial positions (see (32)). What are the options for *það* with an indefinite subject?

- (38) *það* and a [-def] subject
- | | | | | |
|---|--------------|-------|--------------|---|
| ✓ | NP [-def] | V_f | V_l | (V2 clause) |
| ✓ | | V_f | NP [-def] | V_l (V1 clause) |
| ✓ | | V_f | V_l | NP [-def] (V1 clause) |
| ✓ | <i>það</i> | V_f | NP [-def] | V_l (V2 clause) |
| ✓ | <i>það</i> | V_f | V_l | NP [-def] (V2 clause) |
| | | V_f | <i>það</i> | V_l NP [-def] ← structure is blocked, by Economy |

All of these are well-formed in structural terms, and potentially semantically interpretable. However, the last structure here loses to the third one, on grounds of Economy – there is no information for the hearer contributed by *það*– it is simply a V1 clause with an indefinite subject.

4.3. Summary

Crucially, all of the structural inferences just considered are interpreted relative to the finite verb V_f and the last verb V_l , and there is only one subject position between the two. It is a mistake to think that there are two or more medial positions, as is the case in an analysis in which the finite V can be in C, followed by SpecAgrSP and SpecTP (see (47) below). The distribution of *það* follows from an analysis with the properties summarized in (39):

- (39) a. No c-structure position in Icelandic needs to be present except for I^0 in IP.
 b. The position before the first verb V_f may signal a pragmatic property of the clause (nucleus) headed by that verb, across clause-types; no other position signals such a property.
 c. *það* has no function unless it precedes V_f .

5. Formalizing the Analysis

5.1. The Linear Constraint on *það*

We might wish to formalize the generalization in (13), as shown:

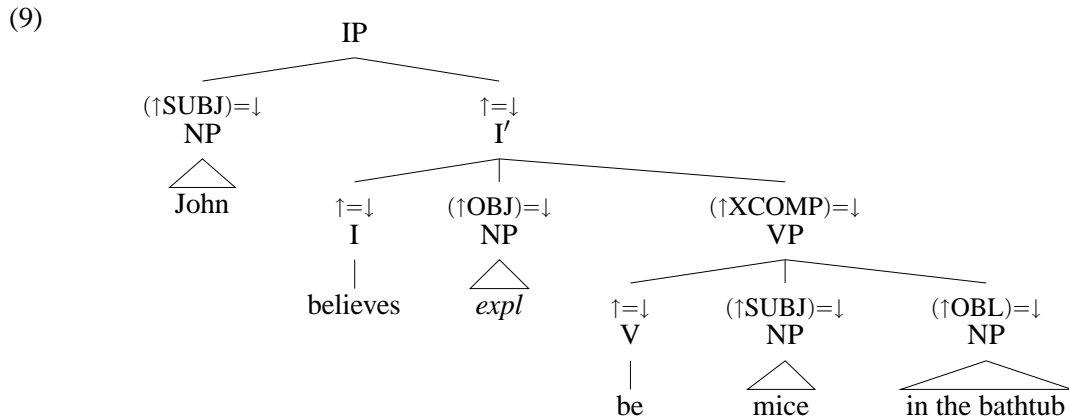
- (13) *það* precedes the exponent of finiteness:
 ¬ FIN f-precedes SUBJ[EXPL]

If we consider this generalization to be a formal property of the grammar, we should state it as a constraint introduced by the expletive (the (rest of the) lexical entry is below in (45)).

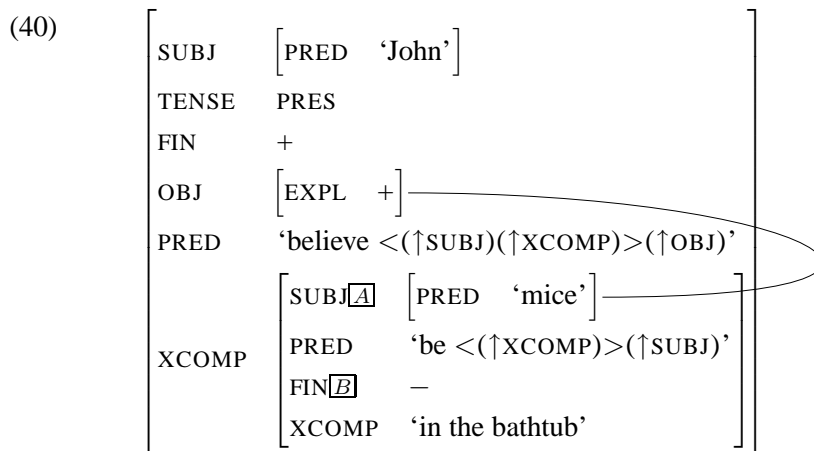
It is always the first verb in the clause which indicates the finiteness [\pm] of its nucleus. This verb may be located in C, I, or V, but it is always the first verb – hence the notation V_f used above.⁷ Let us assume

⁷In finite clauses, the verb is usually in I; it could also be in C, depending on the analysis of V2. Some non-finite clauses are IPs, with the verb in I (e.g., (21)), while some are VPs, with the verb in V (e.g., (9)).

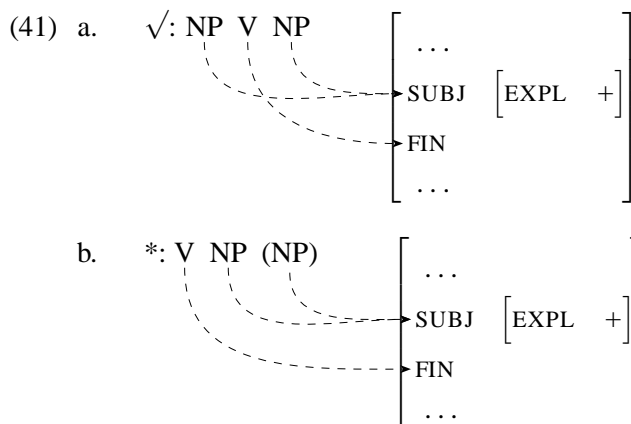
an attribute [EXPL +] introduced by *pað*, which will distinguish a clausal nucleus which corresponds to a c-structure with an expletive in it from one that lacks an expletive. Then the relevant linear condition is that the node instantiating FIN cannot precede the node instantiating SUBJ with an EXPL attribute. (41) is the f-structure for (9):



In (40), the exponent of **B** cannot precede the exponent(s) of **A**, according to the condition above:



More generally, (13) allows (41)a but not (41)b:



This analysis motivates the use of the attributes FIN and EXPL in f-structure, for the statement of the f-precedence condition.

5.2. Linear Constraints on C-Structure

The constraint specific to the expletive in (13) is stated in terms of f-precedence. The other major constraints on Icelandic syntax require more detailed access to c-structure properties, but can be encoded in the Monadic Second-Order Logic system of Kuhn (2003). The *CoProj'* predicate used below is defined in Kuhn (2003), the symbol \triangleleft means ‘immediately dominates, and $\exists!$ means ‘there is exactly one’.

The I^0 constraint is stated as follows:

(42) The I^0 Constraint

$$(\forall x)[IP(x) \rightarrow (\exists y)[I^0(y) \wedge CoProj'(x, y)]]$$

The formula says that every IP node has a I^0 node with which it is a coprojection – both map to the same f-structure, and a contiguous c-structure path connects the two nodes (see Kuhn (2003)).

The V2 constraint is an existential constraint, one which needs to find a finite verb with exactly one element preceding it in the clause:

(43) V2 Constraint:

$$(\exists x)[Fin(x) \wedge X^0(x) \wedge (\exists z)[(\exists! y)[CoProj'(x, z) \wedge z \triangleleft y \wedge y \prec x] \\ \wedge \neg(\exists w)[CoProj'(w, z) \wedge w \triangleleft z]]]$$

where x is the Finite element in second position and y is any element in first position, dominated by z , which coproject's with x . The formula assumes that the precedence relation \prec can be defined between adjacent constituents, even if they are not sisters. (43) says, “There is a node x which is the exponent of FIN and which is zero-level, and there is a node z such that there is exactly one node y such that x and z coproject and z immediately dominates y , and y precedes x , and there is no node w which coproject's with z such that w dominates z .” This has the consequence that node z is the top of the coprojection path, immediately dominating y , which is the one element which precedes x , which is the finite verb. (Compare with (23) and (27).)

6. The Syntax of the Expletive

My approach here is that the expletive *það* lacks an independent PRED, yet bears the SUBJ function. Hence, if the clause has a thematic SUBJ, this will be the associate of the expletive (e.g., (14)).⁸ *það* may also appear in impersonal clauses, in which it would be the only expression of SUBJ. If *það* bears a GF, rather than simply being a c-structure place-holder, the data are straightforwardly accounted for.

As I have mentioned above, the function of *það* in main clauses is essentially to present a V2 clause in which nothing is given special prominence. Rögnvaldsson and Thráinsson (1990, 29) observe “what the dummy actually does is to allow for the sentence type in which nothing is topicalized, not even the subject that in general acts as a discourse topic by default”; see also Zaenen (1983, 496). However, Sigurðsson (1990, 54) offers a slightly different diagnosis of the facts, and considers various embedded clauses, suggesting that the right condition on the expletive is that it itself cannot be associated with a DF (see also Sigurðsson (2004)). He shows that examples in which the subject associate of *það* is itself associated with a DF (in a question, a relative clause, etc.) are robustly ungrammatical (see (44)a), but that a DF associated with some non-SUBJ GF is not so bad, and impersonal clauses like (44)b are relatively acceptable (see also Rögnvaldsson and Thráinsson (1990, 30–31)). In the examples in (44), $_$ indicates the ‘gap’.

⁸I assume that *það* lacks a PRED value and is optionally specified for 3rd singular agreement features; in the absence of any associate to provide a PRED value for the clausal SUBJ attribute, the 3rd singular agreement features of the expletive (or the finite verb in its default form) will suffice for the formal condition of Completeness. This follows the analysis of German developed in Berman (2003) (see especially pp.56ff.).

- (44) a. maður sem (*það) — elskar margar konur
 a.man who (**expl*) — loves many women
 (það = SUBJ, gap = TOP = SUBJ)
- b. ?maðurinn sem (það) var talað við —
 the.man who (*expl*) was talked to —
 (það = SUBJ, gap = TOP = OBL OBJ)

If *það* bears a GF, in particular SUBJ, then by association in (44)a, *það* is also an expression of the TOP function, and the example is ungrammatical. (44)b lacks this association of *það* with a DF, and is somewhat acceptable. Consequently, I propose that *það* must appear as the value of SUBJ, and cannot also be the value of a DF-structure (see (45)). These functional specifications guarantee that *það* is only generated in positions in which it can be associated with a SUBJ, and the prohibition against a DF means that it cannot be associated with a DF in f-structure, or generated in SpecCP, which always associated with a DF (see e.g., Bresnan (2001)).

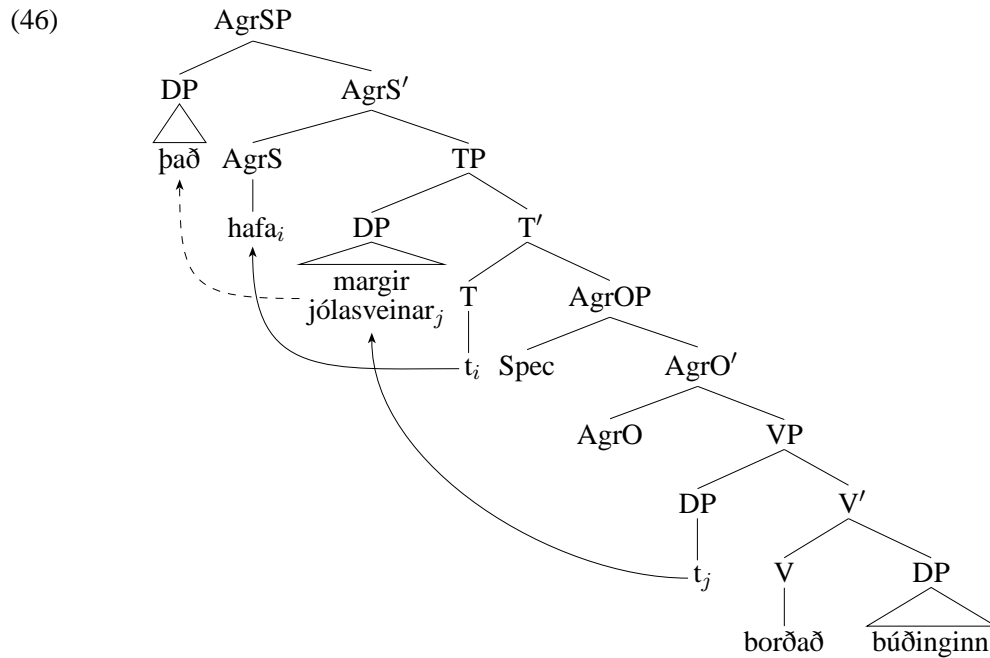
- (45) *það*: (SUBJ↑)
 ¬((SUBJ↑) DF) = ↑ (rules out (44)a, allows (44)b)
 (↑ EXPL) = + (see (41))

7. Conclusion

The restrictions on the distribution of *það* are not due to positional restrictions per se, but rather, are due to the pragmatic signalling functions of *það*: if it follows the exponent of finiteness in its clause, it has no possible function. The specific analysis that I developed necessarily involves treating Icelandic clause structure as less hierarchical than has been widely assumed, as well as stating the main properties of clause structure in terms of relative linear position. These linear properties cannot be ‘reduced’ to hierarchical properties (see the following Appendix), showing that such linear properties are indeed constitutive of syntax.

Appendix: ‘Minimalist’ Structures

The structure proposed in (17) contrasts with the structure proposed by Bobaljik and Jonas (1996):



While the technical details have changed, the basic configuration assumed for Icelandic has persisted in Minimalist syntax, with at least a CP on top of the structure shown in (46).

It is precisely this concentration on hierarchical structure, neglecting relative linear properties, which has led many researchers astray in considering the problems posed by *það*. For example, if we consider the schema in (47) as representing the positions occupied in a regular declarative existential in Icelandic, the lower line ‘int’ shows the pattern that we would expect for a polar interrogative, with the finite verb fronted to C. The ! notations in (47) show positions assumed by the theory but which are never overtly filled:

(47) Subject Positions in the Minimalist Clause

| | SpecCP | C | SpecAgrSP | AgrS | SpecTP | T | SpecVP | V | ComplV |
|---------|--------|-------------|------------|-------------|--------|---|--------|---|--------|
| (decl.) | | | <i>það</i> | V [+fin] | Subj | ! | ! | V | |
| (int.) | | V [+fin] | <i>það</i> | ! | Subj | ! | ! | V | |

The lower interrogative structure makes it look like there are two ‘subject positions’ following the finite verb in C and preceding a non-finite verb in V (technically, three subject positions if SpecVP is counted), and one of these positions, SpecAgrSP, is the grammatical position of *það* in a declarative. Hence, there seems to be no reason to suspect that *það* would be impossible in the interrogative.

However, Icelandic is not organized this way. Depending on the clause type, there may be one XP in front of the finite verb, and then in the ‘Mittelfeld’ area, following the finite verb and before any non-finite verb, there is just one potential subject position (as well as a potential object position, and many other adjunct positions). The hierarchy-only approach to syntax instantiated by (46) is not suited to expressing these clear generalizations.

References

- Anderson, Stephen R. 1997. Towards an optimal account of second position phenomena. In Lizanne Kaiser (ed.), *Yale A-Morphous Linguistics Essays*. New Haven, Department of Linguistics, Yale University, 1–27.
- Anderson, Stephen R. 2000. Towards an optimal account of Second-Position phenomena. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, Syntax and Acquisition*. Oxford, Oxford University Press, 302–333.
- Andrews, Avery D. 1990. The VP-Complement analysis in modern Icelandic. In J. Maling and A. Zaenen (eds.), *Syntax and Semantics 24: Modern Icelandic Syntax*. New York, Academic Press, 165–185.
- Berman, Judith. 2003. *Topics in the Clausal Syntax of German*. Stanford, CSLI Publications.
- Bobaljik, Jonathan, and Dianne Jonas. 1996. Subject positions and the role of TP. *Linguistic Inquiry* 27, 195–236.
- Börjars, Kersti, Elisabet Engdahl, and Maia Andréasson. 2003. Subject and object positions in Swedish. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG03 Conference*. Stanford, CSLI Publications, 43–58. (At <http://csli-publications.stanford.edu/LFG/8/lfg03.html>).
- Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford, Blackwell Publishing.
- Bures, Anton. 1992. There is an argument for an LF cycle here. In C. Canakis, G. Chan, and J. Denton (eds.), *The Cycle in Linguistic Theory. Chicago Linguistics Society 28, Parasession*. Chicago, Department of Linguistics, 14–35.
- Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MIT Press.
- Holmberg, Anders. 1986. *Word Order and Syntactic Features in the Scandinavian Languages and English*. Stockholm, University of Stockholm, Department of Linguistics.
- Holmberg, Anders, and Christer Platzack. 1995. *The Role of Inflection in Scandinavian Syntax*. New York, Oxford University Press.
- Hornstein, Norbert. 1991. Expletives: A comparative study of English and Icelandic. *Working Papers in Scandinavian Syntax* 47, 1–88.
- Jonas, Dianne, and Jonathan Bobaljik. 1993. Specs for subjects: The role of TP in Icelandic. In Jonathan Bobaljik and Colin Phillips (eds.), *Papers on Case and Agreement I*. (MIT Working Papers in Linguistics Vol. 18), Dept. of Linguistics, MIT, 59–98.
- Jónsson, Jóhannes Gísli. 1991. Stylistic fronting in Icelandic. *Working Papers in Scandinavian Syntax* 48, 1–44.
- Jónsson, Jóhannes Gísli. 1996. *Clausal Architecture and Case in Icelandic*. Doctoral dissertation, University of Massachusetts, Amherst.
- Kuhn, Jonas. 2003. Generalized tree descriptions for LFG. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG03 Conference*. Stanford, CSLI Publications, 269–289. (At <http://csli-publications.stanford.edu/LFG/8/lfg03.html>).

- Maling, Joan. 1988. Variations on a theme: Existential sentences in Swedish and Icelandic. In D. Fekete and Z. Laubitz (eds.), *McGill Working Papers in Linguistics: Special Issue on Comparative Germanic Syntax*. Montréal, Department of Linguistics, McGill University, 168–191.
- Maling, Joan. 1990. Inversion in embedded clauses in modern Icelandic. In J. Maling and A. Zaenen (eds.), *Syntax and Semantics 24: Modern Icelandic Syntax*. New York, Academic Press, 71–91.
- Maling, Joan, and Annie Zaenen. 1990. The nonuniversality of a surface filter. In J. Maling and A. Zaenen (eds.), *Syntax and Semantics 24: Modern Icelandic Syntax*. New York, Academic Press, 383–408. [Reprinted from *Linguistic Inquiry* 9, 475–497, 1978].
- Mikkelsen, Line. 2002. Reanalyzing the definiteness effect: Evidence from Danish. *Working Papers in Scandinavian Syntax* 69, 1–75.
- Ottóson, Kjartan. 1989. VP-specifier subjects and the CP/IP distinction in Icelandic and Mainland Scandinavian. *Working Papers in Scandinavian Syntax* 44, 89–100.
- Platzack, Christer. 1983. Existential sentences in English, German, Icelandic and Swedish. In Fred Karlsson (ed.), *Papers from the 7th Scandinavian Conference of Linguistics*. Univ. of Helsinki, Dept. of General Linguistics, 80–100.
- Richards, Marc, and Theresa Biberauer. 2005. Explaining EXPL. In Marcel den Dikken and Christina Tortora (eds.), *The Function of Function Words and Functional Categories*. Amsterdam/New York, John Benjamins, to appear.
- Rögvaldsson, Eiríkur. 1984. Icelandic word order and *það*-insertion. *Working Papers in Scandinavian Syntax* 8, 1–21.
- Rögvaldsson, Eiríkur, and Höskuldur Thráinsson. 1990. On Icelandic word order once more. In J. Maling and A. Zaenen (eds.), *Syntax and Semantics 24: Modern Icelandic Syntax*. New York, Academic Press, 3–40.
- Sells, Peter. 2001. *Structure, Alignment and Optimality in Swedish*. Stanford, CSLI Publications.
- Sells, Peter. 2002. Stylistic fronting in Icelandic: A base-generated construction. *Gengo Kenkyuu (Journal of the Linguistic Society of Japan)* 123, 257–297.
- Sells, Peter. 2005. Morphological and constructional expression and recoverability of verbal features. In C. Orhan Orgun and Peter Sells (eds.), *Morphology and the Web of Grammar: Essays in Memory of Steven G. Lapointe*. Stanford, CSLI Publications, 197–224.
- Sigurðsson, Halldór Ármann. 1989. *Verbal Syntax and Case in Icelandic: In a Comparative GB Approach*. Doctoral dissertation, University of Lund.
- Sigurðsson, Halldór Ármann. 1990. V1 declaratives and verb raising in Icelandic. In J. Maling and A. Zaenen (eds.), *Syntax and Semantics 24: Modern Icelandic Syntax*. New York, Academic Press, 41–69.
- Sigurðsson, Halldór Ármann. 2004. Argument features, clausal structure and the computation. In T. Bhattacharya, E. Reuland, and K. V. Subbarao (eds.), *Argument Structure*. Amsterdam, John Benjamins, to appear (citation according to web ms.).
- Thráinsson, Höskuldur. 1979. *On Complementation in Icelandic*. Doctoral dissertation, Harvard University; published by Garland Publishing Inc., New York.

- Thráinsson, Höskuldur. 1984. Different types of infinitival complements in Icelandic. In W. de Geest and Y. Putseys (eds.), *Sentential Complementation*. Dordrecht, Foris Publications, 247–255.
- Thráinsson, Höskuldur. 1993. On the structure of infinitival complements. In H. Thráinsson et al. (ed.), *Harvard Working Papers in Linguistics 3*. Cambridge, Department of Linguistics, Harvard University, 181–213.
- Vangsnes, Øystein. 2002. Icelandic expletive constructions and the distribution of subject types. In Peter Svenonius (ed.), *Subjects, Expletives, and the EPP*. New York, Oxford University Press, 43–70.
- Wurmbrand, Susi. 2004. Licensing case. Ms. University of Connecticut.
- Zaenen, Annie. 1983. On syntactic binding. *Linguistic Inquiry* 14, 469–504.
- Zaenen, Annie. 1985. *Extraction Rules in Icelandic*. Doctoral dissertation, Harvard University [1980]; published by Garland Publishing Inc., New York.

CASE IN HINDI

Andrew Spencer

University of Essex

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

I argue that Hindi clitic postpositions are not markers/realizations of case. Hindi has a genuine case system represented by the direct, oblique and vocative inflected forms of nouns. So-called case markers such as *ne* ‘Ergative’ or *ko* ‘Accusative/Dative’ are better thought of as postpositions which are non-projecting words (Toivonen 2003), selecting the oblique case form of their noun complements. Since the postpositions fail to project a phrase the case property of the head noun will be inherited by their NP/DP argument, so that any NP/DP marked with a postposition will itself be in the oblique case. Predicate agreement can now be stated very simply as ‘agree with the direct case marked NP’.

1. Introduction¹

The question of what counts as a case is one which has not been at the forefront of recent morphosyntactic research, yet it remains one of the more puzzling questions in theoretical linguistics. The prototypical case system is the type illustrated by Indo-European languages such as Latin, Greek, Sanskrit and most of the contemporary Slavic languages. In such a system nouns bear inflections which subserve various grammatical functions, such as the marking of subjects and objects. Sometimes a case will have an essentially semantic use, say, as a locative, or for the vocative case a special discourse function. Often particular prepositions or postpositions govern specific case forms. Finally, attributive modifiers (and more rarely predicates) will often agree in case with the head noun they modify. In Indo-European languages the form of a case marked nominal will often depend on the grammatical number and on inflectional class. However, the inflectional endings are entirely different from each other (and different again from those of various other inflectional classes). Therefore, it is necessary to set up an abstract CASE attribute which can permit us to generalize over these forms. However, I will make the assumption that the situation with the English translation equivalents of the Latin genitive is rather different. The preposition *of* is neither a case itself nor a marker of case. In order to state the fact that, say, possessive constructions are expressed by *of* we need simply make direct reference to *of* as a lexical item, just as we make direct reference to the preposition *with* without invoking a comitative, instrumental or whatever case.

Considerations of this sort have lead Beard (1995) to question whether an attribute of case is needed even in languages in which nouns appear to inflect for case, but in which there are no inflectional class differences. In languages such as Turkish the same case suffixes with the same allomorphy are used for all nominals. According to Beard, this means that a CASE attribute is redundant in the grammar of such languages. We can state the distribution of case-inflected nominals by referring directly to the form of the nominal. Thus, rather than speaking of, say, the genitive case form of Turkish *ev* ‘house’ we can speak of the *-In* form, *ev-in* (or in the plural *evler-in*). Internal to the grammar of Turkish nothing is lost in doing this (see Beard 1995: 259f). In Spencer and Otoguro (2005) this is referred to as ‘Beard’s Criterion’. Even if we balk at the idea that an agglutinative affixal paradigm fails to define a case system in Turkish, it is difficult to

¹ I am grateful to Ryo Otoguro, Tara Mohanan and Miriam Butt for useful comments and to an anonymous LFG05 abstracts reviewer for cajoling me into providing explicit discussion of the inflected pronouns.

fault Beard’s logic where adpositional systems are concerned, whether those adpositions are expressed as syntactic heads or as phrasal affixes.

2. Morphosyntactic preliminaries

Following Zograf (1960, see also Masica 1991) we can distinguish three ‘layers’ of functional category marking on Indo-Aryan nominals. The first layer is inflection proper. In Hindi nouns may inflect for singular/plural number and for three forms, which I shall call the direct form, the oblique form and the vocative form. Later I shall refer to these three forms as ‘cases’. However, for the present I shall call them ‘forms’ so as not to introduce terminological confusion. Inflection is illustrated for a representative sample of nouns in (1)²:

(1) Inflected noun forms (Zograf Layer I)

| | | | | | |
|----------|-------------------|----------|--|---------------------|---------|
| | Singular | Plural | | Singular | Plural |
| Direct | laRkaa | laRke | | makaan | makaan |
| Oblique | laRke | laRkō | | makaan | makaanō |
| Vocative | laRke | laRko | | | |
| | ‘boy’ (Masculine) | | | ‘house’ (Masculine) | |
| | Singular | Plural | | Singular | Plural |
| Direct | laRkii | laRkiyāã | | mez | mezē |
| Oblique | laRkii | laRkiyō | | mez | mezō |
| Vocative | laRkii | laRkiyo | | | |
| | ‘girl’ (Feminine) | | | ‘table’ (Feminine) | |

Adjectives may take similar inflections, except that the vocative form is always identical to the oblique form. Forms for *acchaa* ‘good’ and the demonstrative *yah* ‘this’ are given in (2) (the demonstrative does not inflect for gender):

(2) Hindi adjective inflection

| | | | | |
|----|-----|--------|--------|--------|
| | | Masc | Fem | |
| Sg | Dir | acchaa | acchii | yah |
| | Obl | acche | acchii | is |
| Pl | Dir | acche | acchii | ye |
| | Obl | acche | acchii | in |
| | | ‘good’ | | ‘this’ |

Inflecting modifiers agree with the noun head in number, gender and direct/oblique form, as seen in (3) (based on Dymšits 1986a: 78, 79):

² In the Hindi transcriptions, ‘R’ represents a retroflex rhotic and doubled vowels are long.

(3) Examples of Hindi adjective agreement

| | | | | | |
|---------|----|-------|---------------|--------|----------------|
| direct | Sg | accha | laRkaa | acchii | laRkii |
| | Pl | acche | laRke | acchii | laRkiyãã |
| oblique | Sg | acche | laRke | acchii | laRkii |
| | Pl | acche | laRkõ | acchii | laRkiyõ |
| | | | ‘good boy(s)’ | | ‘good girl(s)’ |

The same pattern of agreement is found when a declinable modifier is in construction with an indeclinable noun such as *ghar* ‘house’, so that ‘good house’ in the oblique singular form is *acche ghar*, while the direct singular form is *accha ghar*. This shows that the inflectional system forms a paradigm in which some forms for some lexical classes are syncretic.

These desinences show all the typical behaviour of inflectional affixes, as outlined in (4) (see Payne 1995: 284):

- (4) Properties of Zograf Layer I desinences
- They must be repeated on each noun of a conjoined phrase (though see below for asyndetic compounds)
 - They trigger agreement on attributive modifiers and must be repeated on all (inflecting) modifiers within the NP

Although the oblique stem form is frequently found in construction with a Layer II simple postposition, this form can exist as an inflected word in its own right, generally with a locational destination meaning, as seen in (5, 6) (Mohan 1994a: 88, 89):

(5)

| | | |
|------------------------|--------------|---------|
| raam | kalkatte | gayaa |
| Ram.NOM | Calcutta.OBL | go.PERF |
| ‘Ram went to Calcutta’ | | |

(6)

| | | | |
|------------------------|----------------|-------------------|-----------|
| raam | mere | ghar | aayaa |
| Ram.NOM | my.OBL.MASC.SG | house.OBL.MASC.SG | come.PERF |
| ‘Ram came to my house’ | | | |

Notice that in (10) the head noun *ghar* does not overtly inflect for the oblique form, but its obliqueness is unambiguously signalled by the form of the modifier ‘my’.

The second of Zograf’s layers is found with a small number of postpositional clitics (phrasal affixes). I shall follow traditional descriptive practice and refer to these as ‘simple postpositions’. The simple postpositions are used to realize grammatical functions such as (transitive) subject (*ne*), direct object (*ko*), indirect object (also *ko*) as well as a variety of adverbial functions. In the recent literature this has been taken to reflect a fully fledged case system, as illustrated in (7) (taken from Mohanan 1994a: 66):

| | | |
|-----|--------------|--------|
| (7) | nominative | (zero) |
| | ergative | ne |
| | accusative | ko |
| | dative | ko |
| | instrumental | se |
| | genitive | kaa |
| | locative1 | mě |
| | locative2 | par |

Each of these postpositions is invariable except for *kaa*, which agrees with the possessed noun. That behaviour is extremely unusual for a case marker though it parallels the morphosyntax of the Albanian ‘genitive clitic’ and the Bantu ‘A-of-association’, neither of which are cases. The *ko* postposition is systematically ambiguous, in the sense that it marks either a direct object or an indirect object (though speakers generally reject clauses containing two adjacent *ko*-marked phrases. For detailed discussion of ‘double *ko*’ clauses see Mohanan 1994b).

There is general agreement that the postpositions are clitics (not affixes) (Butt and King 2003, Mohanan 1994a, Payne 1995). There is, however, one interesting twist in the behaviour of one of the postpositions, *ko*. With personal pronouns there is an alternative realization of the sequence ‘pronoun + *ko*’, as illustrated in (8) (Dymšits 1986a: 99):

| | | | | | | | |
|-----|-----------------------|-------|-------|-----|------|-------|------|
| (8) | | 1sg | 2sg | 3sg | 1pl | 2pl | 3pl |
| | Direct | mãĩ | tuu | yah | ham | tum | ye |
| | Oblique | mujh | tujh | us | ham | tum | in |
| | Dative/ Accusative | mujhe | tujhe | use | hamě | tumhě | inhě |

The synthetic forms are synonymous with the more regular forms constructed from the oblique form and *ko*: *mujh=ko*, *tujh=ko* and so on. This is discussed below.

3. Why postpositions are not cases

In order for a formative within a noun phrase to be considered a case within the grammatical description of a language that formative must minimally serve as a marker of a grammatical relationship of some kind between some other head and that noun phrase as its dependent. However, this is only a necessary condition, since it would admit English-type prepositions as case markers (whether cliticized or not). There is no point in setting up a [CASE] attribute, either in syntax or morphology, unless that attribute generalizes over sets of distinct forms in some way. The most obvious need occurs with inflectional classes such as those of Latin, where we need to generalize across distinct morphological forms (e.g. the genitive singular and plural, not to mention distinct declension classes). A more subtle requirement is found in syntax: if we find that attributes must agree with their modified heads in a noun phrase then, *prima facie*, we would miss important generalizations unless we appealed to a [CASE] attribute in the syntax, so that we could state the recurrent case marking as a general phenomenon, and

so that we could distinguish the case formatives from other, non-agreeing, formatives such as postpositions. These latter instances of noun marking constitute sufficient criteria for casehood.

The Layer II postpositions show no properties which can be taken as sufficient criteria for casehood. In particular they fail to trigger the kind of agreements on modifiers that the Layer I inflections trigger. The only reason for labelling Layer II elements as cases is that they serve to mark grammatical functions, including the function of SUBJECT, but this is only a necessary criterion, not a sufficient one. To be sure, by giving the postpositions case names we can state typological generalizations. For instance, we can say that certain classes of verbs take ‘dative’ subjects. However, this is not a very good reason for setting up a (second) CASE attribute in the grammar of Hindi. Similar reasoning would force us to claim that the preposition *of* is a genitive case marker in English, for instance. Actually, matters are worse than this. It is not just that nothing is gained by ascribing a CASE attribute to Hindi postpositions. The postpositions-as-cases thesis actually prevents us from making generalizations. To see this we must consider predicate agreement.

Verbs agree with (highest ranking) nominative-marked argument (i.e. unmarked, no postposition, direct case). This may be SUBJ or OBJ.

- (9) a. *acchaa* *laRkaa* *gaRii* *calaataa* *hai*
 good.M.NOM.SG boy.NOM.SG car.NOM drive.IMPV.M.SG AUX
 ‘The good boy drives a car’
- b. *acche* *laRke=ne* *gaRii* *calaayii* *hai*
 good.M.OBL.SG boy.OBL.SG=NE car.NOM drive.PERF.F.SG AUX
 ‘The good boy has driven a car’

There are several points to bear in mind. First, the agreement process cannot be defined solely in terms of the pure forms of nouns. For instance, a form such as *ghar* ‘house’ can be either NOM SG or OBL SG. When followed by a postposition such as *ko* it will not trigger agreement, but when it appears in its clause as a bare noun subject or object it may trigger agreement. Thus, agreement must make reference to some kind of CASE feature.

Second, note that predicate agreement on lexical verbs is defined in terms of the attributes CASE NOM, GENDER and NUMBER attributes. Now, adjectival modifiers may agree with their head noun for these attributes, too. This poses no problems in the case of GENDER and NUMBER, since these attributes are clearly the same whether they trigger agreement on modifiers or on predicates. In other words, we can use the same features to express the agreement of the predicate with the MASC SG noun ‘boy’ in (9a), that we use to express the GENDER/NUMBER agreement between the adjective *acchaa* ‘good’ and *laRkaa* ‘boy’ in that clause.

However, when we come to examine CASE agreement we encounter a problem. The reason that the forms *acchaa* and *acche* agree in case with *laRkaa* and *laRke* respectively in (9a, b) is because the forms {*laRkaa*, *laRke*} are in a paradigmatic (inflectional) opposition to each other. However, given the postpositions-as-case analysis the verb form

in (9) shows agreement with the subject or object of the clause by virtue of the fact that the subject or object NP is not in construction with a ‘case’ postposition. The direct/oblique distinction plays absolutely no role in predicate agreement on that approach. Put differently, the value NOM is being used with a systematic, but unacknowledged, ambiguity. In predicate agreement it is part of the paradigm {NOM, ACC, DAT, GEN, ...}. In modifier agreement it is part of the paradigm {NOM, OBL, VOC}. But this means that we are dealing with two distinct case features and two distinct sets of case values, CASE1 {NOM1, ACC, DAT, GEN, ...} and CASE2 {NOM2, OBL, VOC}.

Yet it seems more than perverse to treat CASE2 NOM as being a distinct attribute from CASE1 NOM with respect to agreement. In all other respects, agreement is defined over the same features sets (reflecting the adjectival, participial origin of the agreeing verb forms). We seem to be losing a generalization if we concede that we are operating with two distinct notions of case. In addition, recall that a bare oblique-marked noun such as *kalkatte* ‘Calcutta.OBL’ can be used with a locative-directional function. But what is the relationship between that bare oblique form and a NP furnished with a locative postposition such as *mē* ‘in’? Specifically, what set of oppositions is being presupposed here. Are we going to be obliged to say that CASE2 OBL forms are also in a paradigmatic opposition to CASE1 {NOM1, ACC, ...} forms? In that case it would seem that we have just a single CASE attribute after all. But then how do we account for the fact that a postposition such as the ‘ergative’ *ne* or the ‘accusative/dative’ *ko* selects the oblique (‘locative?’) case form and not the ‘nominative’ case form? Surely that would be little different from saying that the German preposition *von* ‘of’ is a genitive case marker which selects the dative case of its NP.

In short, the different behaviour of predicate agreement and modifier agreement with respect to the case attribute leads to complete conceptual confusion if we adopt the simplest version of postpositions-as-cases approach. In the next section, I outline explicit discussion of this problem in the LFG literature.

4. Previous treatments

Some of the issues raised here have been discussed in the work of Butt and King (especially 2004). They treat the Layer II postpositions as members of a (projecting) category K, distinct from P. They do not address the question of modifier agreement for direct/oblique/vocative case. The only discussion in the LFG literature I know of which takes seriously the questions I have raised here is that of Mohanan (1993). She notes that a direct object in Hindi may be marked by *ko* if it is regarded as animate or if it is definite, otherwise the bare form of the noun is used. She contrasts two ways of looking at this situation. On the ‘morpheme alternation’ analysis we would say that the bare NP object and the *ko* marked object were in the accusative case and that this case marking is realized differently in different contexts. On the ‘feature alternation’ analysis some objects are accusative (those with *ko*) and some are nominative (bare NPs). Mohanan argues persuasively in favour of the ‘feature alternation’ view over the ‘morpheme alternation’ view.

Crucial for our purposes is the way that Mohanan treats modifier agreement vis-à-vis predicate agreement. She assumes a CASE attribute with the standard values {nom, acc, dat, ...}. She then sets up a property NON-NOM which essentially means ‘any value of CASE except NOM’. The NON-NOM property thus corresponds to the inflectional oblique case form of nouns and the corresponding agreement form in modifiers. Modifier agreement appeals to both the NOM and NON-NOM properties, while predicate agreement is sensitive solely to the NOM property.

It is not entirely clear from Mohanan’s exposition how the {NOM, NON-NOM} distinction is to be interpreted formally. For Mohanan a NP is ‘in a case’ by virtue of the clitic postposition to its right edge. But that means that the NON-NOM form of a noun (or adjective for that matter) is not ‘in’ any case until the NP is furnished with a postposition (see Mohanan’s example (25b)). But this means that it is rather misleading to speak of NON-NOM as a case value, rather it picks out a set of forms which receive a case value from a postposition. But that makes it difficult to see how modifier agreement can be stated. On the one hand NON-NOM is a property of a head noun which is not inherited by the NP as a whole. There are no NP-external syntactic processes in Hindi which appeal to the NON-NOM property. On the other hand, NOM is a property both of head nouns and of complete NPs and it is this property that governs predicate agreement.

There are two interpretations of the NON-NOM property which would elucidate this analysis. Under one interpretation we would say that NON-NOM stands for a special feature which is defined as the negation of NOM. On the other interpretation, we complicate the feature geometry for CASE and regard NON-NOM as a set-valued attribute which takes the other cases {ACC, DAT, GEN, ...} as values. Neither of these interpretations seems to be a desirable extension of standard practice. The first interpretation means that modifier agreement essentially says ‘use the NOM form of the modifier unless the NP has a non-zero, non-locative postposition, in which case use the NON-NOM form’. Note that on this interpretation it is necessary to assume two distinct zero case postpositions, one for NOM, the other for the bare noun locational. The second interpretation essentially says ‘use the NOM form of the modifier if the NP is marked ‘NOM’ and use the NON-NOM form of the modifier if the NP appears in any other case’. This principle, too, has to be supplemented by reference to a NOM zero case postposition as opposed to a locational zero postposition. A bare NON-NOM noun is not ‘in a case’ purely by virtue of being in that form and its case has to be provided by a zero marker in order for the modifier to recognize that the NP as a whole is in a non-nominative case.

The complications are only needed because of the desire to conflate two distinct sets of properties, namely, the Layer I inflectional system of nominative and oblique case and the Layer II system of clitic postpositions. But by Beard’s Criterion it is only the Layer I system which has any of the important properties of a case system.

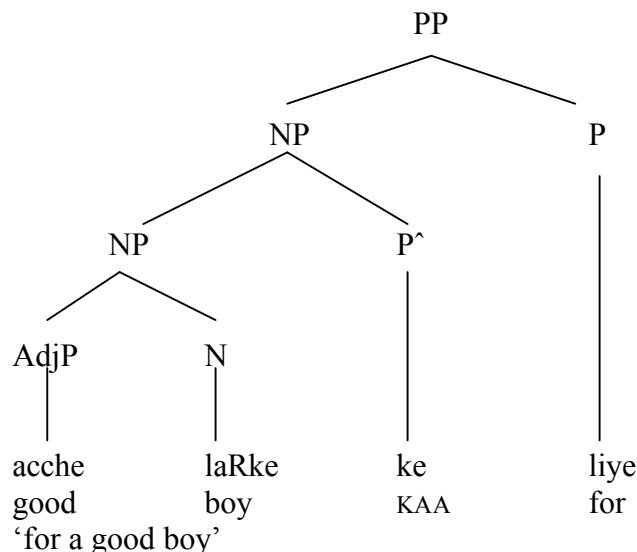
5. Proposal³

In the analysis I propose the inflectional paradigm of a noun includes the attribute [CASE: {NOMINATIVE, OBLIQUE, VOCATIVE}]. In addition Hindi appeals to a syntactic CASE attribute which plays a role in agreement and government phenomena and which is realized by the corresponding morphological attribute. In other words, an NP marked CASE NOM or CASE OBL, say, has that case realized by the appropriate inflected form of the head noun of that NP. In order for the analysis to work smoothly it is helpful to assume that the CASE attribute is an obligatory part of any complete f-structure corresponding to a complete NP in c-structure. In other words, CASE is an obligatory morphosyntactic category in Hindi.

The Layer II postpositional markers are clitics or phrasal affixes, taking the form of non-projecting words (Toivonen 2003), adjoined directly to the right edge of the NP which serves as their complement. Toivonen suggests that such non projecting words generally adjoin to a lexical head in syntactic representation, but there's no need to assume this and I shall propose that the Hindi postpositional clitics adjoin to the NP (or DP if you assume that Hindi has such a category). The category of the NP to which the postposition is adjoined will remain NP. In this respect, the NP=postposition complex is similar to a case-marked NP in languages with genuine case systems. The proposed analysis of the case postpositions is virtually identical to the analysis proposed by Sharma (2003) for the Hindi emphasis particle *hii* (though not to her analysis of the 'case' postpositions).

The analysis is illustrated in (10), where P[^] indicates a non-projecting category:

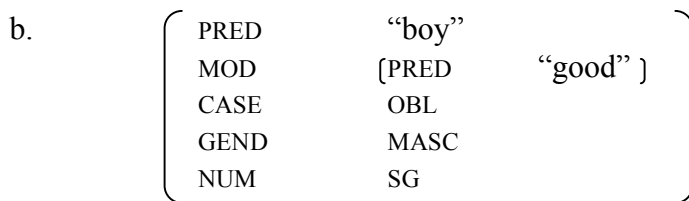
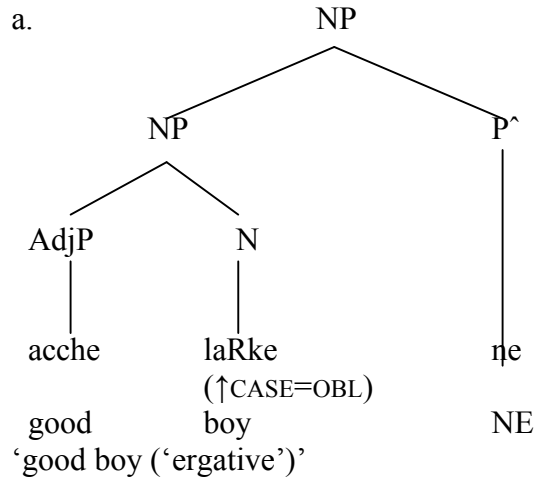
(10) Case clitic as non-projecting postposition



³ For a more detailed version of these proposals, extended to other Indo-Aryan languages and with a detailed and explicit account of the morphology-syntax mapping see Otaguro (forthcoming, chapter 5). That thesis also provides considerable further evidence against the proliferation of 'case' features in grammars of various types.

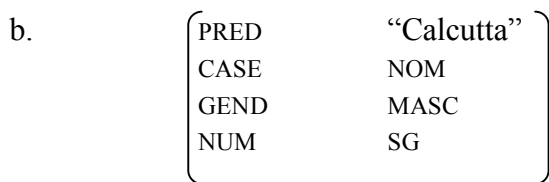
I am assuming that all NPs have an obligatory CASE attribute in their f-structures. The value of this CASE attribute comes from the inflected form of the head noun. As a lexical property, all non-projecting postpositions select the CASE OBL form of the NP. This means that the head noun bears morphological oblique case. Because the postpositions fail to project, the category of the phrase they form is no different from that of their host, and in particular the CASE value remains the same, that is CASE OBLIQUE. This is illustrated in (11):

(11) Case clitic as non-projecting postposition



We may contrast (11) with (12, 13) in which we see the noun *kalkataa* ‘Calcutta’ in its bare nominative form and in its oblique form (which could be used as a directional complement to a verb of motion):

(12) a. *kalkataa*: (↑CASE=NOM)



(13) a. kalkate: (↑CASE=OBL)

b.

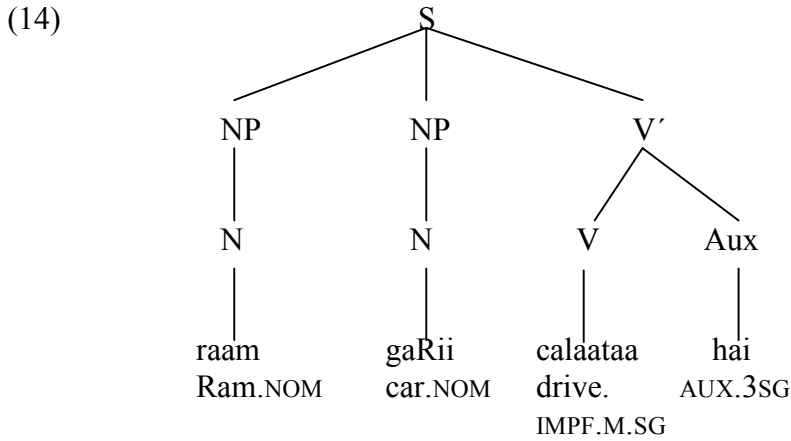
| | | | |
|---|------|------------|---|
| (| PRED | “Calcutta” |) |
| (| CASE | OBL |) |
| (| GEN | MASC |) |
| (| NUM | SG |) |

The proposed analysis permits us to unite modifier and predicate agreement in a natural fashion. The predicate agreement principles for lexical verb forms seek out an appropriate CASE NOM NP to trigger agreement (as in Mohanan’s formulation above). On the other hand, the modifier agreement principles operate over CASE {NOM, OBL} (we can assume that vocative case is syncretized with oblique case on the modifiers themselves). But notice that in both modifier agreement and predicate agreement, some head (adjective or verb) agrees in CASE NOM with either a head noun or a noun phrase. There is no prevarication over ambiguous case labels. Thus, in (9a) above, the adjective *acchaa* ‘good’ and the verb form *calaataa* ‘driving’ are both in the CASE NOM form and this is ultimately because the head noun *laRkaa* ‘boy’ is in the CASE NOM form. Likewise, in (9b), *acche laRke ne* ‘good boy’ is in the CASE OBL form (not CASE ERGATIVE!) and for that reason only modifier agreement can apply to that phrase.

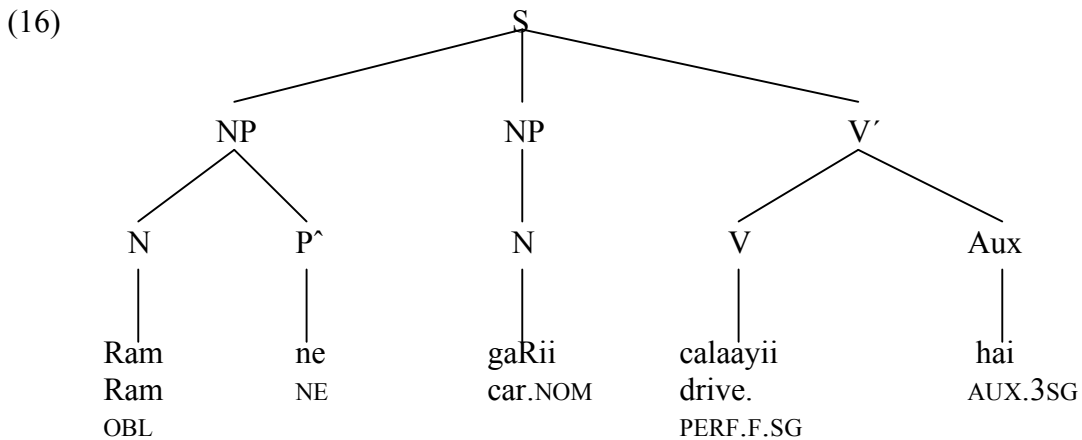
By appealing to Toivonen’s notion we have achieved our goal. The whole of the Hindi nominal system can now be given a simple, but unified, treatment. The selection of the postpositions themselves still has to be defined, however. Now, in the postpositions-as-cases analysis the postpositions project two sorts of information, one governing the grammatical functions themselves and the other a CASE label. We dispense with the CASE label for the postpositions in this analysis, since that label is completely superfluous. The postpositions themselves can be readily identified by virtue of their form. Thus, we may assume a feature, FORM, which defines the morphophonological shape of a lexical entry. In the case of *ne* we will have FORM NE, while for *ko* we will have FORM KO. Nothing more need be said. Where in previous analyses we might have postpositions realizing or constructing specific grammatical functions and supplying CASE labels, now they serve solely to realize the grammatical functions (and various semantico-syntactic properties of those functions). However, the postpositions do not define a CASE value at f-structure. That attribute is determined by the form of the head noun of the NP.

We capture the Layer I, II, III distinction categorially. The troublesome member of the triple is the Layer II set, the postpositions. These are distinct from true postpositions because they fail to project a PP node, but they are different from Layer I inflections because they themselves are words. In this way the non-projecting word plays the same role as the KP vs. NP/DP distinction in Butt and King (2004). However, because the postpositions are non-projecting we automatically have an account for why they fail to show the full panoply of X-bar syntax. One final point is that the *kaa* marker gives to an NP the agreement syntax of an adjective, while remaining an (oblique case marked) NP (This completely answers the objections of Payne 1995 to an analysis of *kaa* as an adjectival marker.)

The proposal is further illustrated in (14 - 17), where we see simplified c-structures for ‘Ram drives a car’, ‘Ram has driven a car’:



- (15)
- | | |
|------------------|--|
| <i>raam:</i> | (↑CASE)=NOM |
| <i>gaRii:</i> | (↑CASE)=NOM |
| <i>calaataa:</i> | agrees with <i>raam</i> as highest GF which is marked case nom |



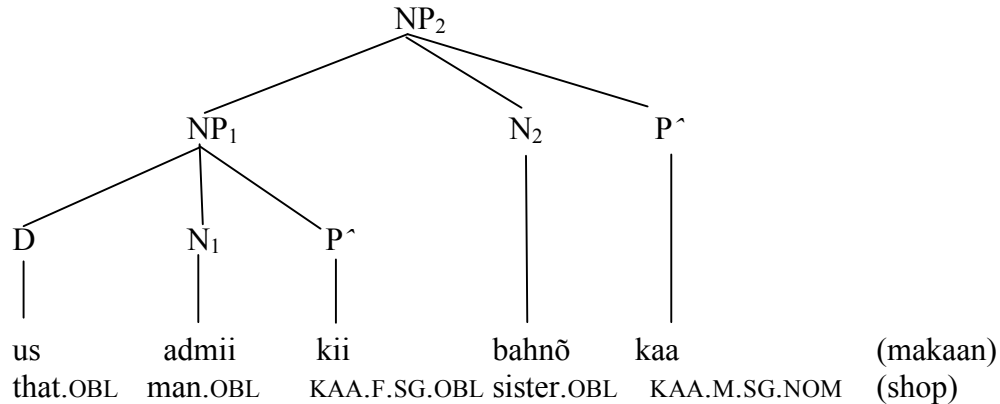
- (17)
- | | |
|--------------|---------------------------------------|
| <i>raam:</i> | (↑CASE)=OBL |
| <i>ne:</i> | (↑FORM)=NE |
| | ((SUBJ↑)OBJ) |
| | ((SUBJ↑)TENSE-ASP)= _c PERF |

The verb form *calaayii* agrees with *gaRii* as the sole grammatical function which bears nominative case. The annotations on *ne* state that it constructs a subject and that the f-structure containing that subject also contains an object. This is achieved by means of the inside-out designator, (SUBJ↑). This is interpreted to mean that f-structure corresponding to the mother of *ne*, that is, the f-structure corresponding to the NP *raam=ne*, is the value of a SUBJ attribute. Moreover, the f-structure containing that SUBJ attribute, namely, (SUBJ↑), itself contains an attribute OBJ. This is the way we capture the notion of ‘ergative case’ in the model of Nordlinger (1998). However, there is no requirement in her model

that what actually constructs a grammatical function has to be a case (as opposed, say, to an adposition). Finally, the annotation $((\text{SUBJ}\uparrow)\text{TENSE-ASP})=\text{PERF}$ constrains the clause to have a perfective aspect value. The constraint states that the f-structure containing the SUBJ attribute (that is the f-structure of the clause) also contains a TENSE-ASP attribute whose value is constrained to be PERF.⁴

In (18) we see a simplified tree for the recursively embedded possessor construction, together with relevant lexical entries (19):

(18) Possessor construction:



- (19) *kaa*: $(\uparrow\text{FORM})=\text{KAA}$
 $(\text{POSS}\uparrow)$
us agrees with N_1 (= NUM SG, GEN MASC, CASE OBL)
 NP_1 agrees with N_2 (and NP_2 agrees with ‘shop’)

Again, we make use of the notion of constructive case in order to ensure that the *kaa* postposition creates a POSSESSOR grammatical function by means of the inside-out designator (POSS \uparrow). This annotation says that the mother node of *kaa*, that is NP_1 , corresponds to an f-structure which is the value of a POSS attribute. This means that the *kaa*-marked NP is the possessor within NP_2 .

6. Inflecting pronominals

In this section I deal with one remaining objection to the postpositional analysis. As in many languages, pronouns in Hindi behave in a slightly different way from other nominals, in that they seem to have a distinct inflectional forms, corresponding to the *ko*-marked form. In addition the *ne* postposition selects a distinct form of some pronouns. I argue that this does not affect the overall analysis.

The 1st, 2nd person pronouns together with the demonstrative pronouns (which double as 3rd person pronouns), the interrogative pronouns *kyaa* ‘what’, *kaun* ‘who’ and the relative pronoun *jo* have a distinct ‘accusative/dative’ form. In addition, the non-personal

⁴ See Butt and King 2003 for *ne* marking on volitional intransitive subjects and other refinements.

pronominals sometimes have a special form of the oblique plural selected by the *ne* form ((20), Dymšits 1986a: 99f):

(20) Pronominal inflection

| | | | | | |
|-----------------------|-------------------|-------|----------------|-------|--|
| a. | Personal pronouns | | | | |
| | 1sg | 2sg | 1pl | 2pl | |
| Direct | māĩ | tuu | ham | tum | |
| Oblique | mujh | tujh | ham | tum | |
| Accusative/ Dative | mujhe | tujhe | hamẽ | tumhẽ | |
| b. | Other pronominals | | | | |
| | ‘this’ | | ‘that’ | | |
| | sg | pl | sg | pl | |
| Direct | yah | ye | vah | ve | |
| Oblique | is | in | us | un | |
| Accusative/ Dative | ise | inhẽ | use | unhẽ | |
| <i>ne</i> form | (is) | inhõ | (us) | unhõ | |
| | rel. pron. | | interrog. pron | | |
| | sg. | pl. | sg. | pl. | |
| Direct | jo | jo | kyaa/kaun | | |
| Oblique | jis | jin | kis | kin | |
| Accusative/ Dative | jise | jinhẽ | kise | kinhẽ | |
| <i>ne</i> form | (jis) | jinhõ | (kis) | kinhõ | |

These portmanteau accusative/dative forms are doublets in the sense that full forms are also possible with the expected postposition, *mujh ko*, *ham ko* and so on.

The special plural forms co-occurring with *ne* do not motivate any change to the case analysis of Hindi (though they require us to make decisions about whether all pronouns have two oblique cases which are syncretized by *in* in the singular). Indeed, the pronominal ‘second obliques’ are reminiscent of the Russian ‘second locative’, a special form of the prepositional case found with a hundred or so nouns and used exclusively with the prepositions *v* ‘in’ or *na* ‘on’ in their spatial use. A further point to note is that the 1st/2nd pronouns appear in their direct, nominative case form with the postposition *ne*. Again, this is an idiosyncrasy of morphology and does not bear on whether we need to treat the *ne* postposition as a case or not.

However, on the face of it the existence of the accusative/dative portmanteaus suggests that for ‘accusative/dative’ at least Beard’s Criterion is met: there are distinct forms for distinct word classes and so a generalization is missed if we fail to generalize over these forms and set up a CASE attribute (with values {NOMINATIVE, OBLIQUE,

ACCUSATIVE/DATIVE}). However, it would be premature to take this position. First, notice that treating the portmanteaus as separate case forms will offer justification solely for the ‘accusative/dative’ case form, not for the other Layer II postpositions. Moreover, instances of this sort of sporadic inflection are quite common cross-linguistically in languages for which it is very difficult to motivate a true case system.

We can think of the pronoun portmanteaus as an instance of what Haspelmath (2000) calls ‘anti-periphrasis’. This occurs when a normally periphrastic (multi-word) construction is expressed as a single word form, generally for a handful of common lexical items (often function words). Other examples of this sort of thing include inflection prepositions, such as French *du*, a portmanteau for *de* ‘of’ and *le* ‘masc. sg. definite article’, or German *zum*, a portmanteau for *zu* ‘to’ and *dem* ‘masc. sg. dative definite article’. Notice that the Hindi situation is rather different from the situation with English pronouns. In English pronouns have retained vestiges of a case system which has been completely lost in the rest of the language. This means that there is no periphrastic construction corresponding to object pronoun forms such as *him* or *us*. In this respect the Hindi system is easier to describe and analyse. Nonetheless, it’s worth bearing in mind that even in English the pronouns provide scant evidence for any kind of bona fide case system (Hudson 1995).

There is another reason for being wary of the Hindi pronominal portmanteau evidence. Some pronominals also have a distinctive ‘emphatic’ form derived from fusion with emphatic particle *hii* (the data in (21) are transcribed from Snell and Weightman 1989: 100; see also Dymšits 1986a: 110 and the discussion in Sharma 2003):

- (21) *mujh + hii* = *mujji*
 is + hii = *isii*
 ham + hii = *hamĩĩ*
 tum + hii = *tumhĩĩ*
 in + hii = *inhĩĩ*

The problem of portmanteau forms generally awaits a satisfactory solution even though it is difficult to find an inflecting language which doesn’t exhibit this phenomenon. In any event the problem of inflecting pronouns is hardly unique to Hindi. It has been argued (Spencer 1991: 383, Wescoat 2002) that English personal pronouns show inflection for tense/aspect/mood categories. The reduced auxiliary verb component of forms such as *she’ll*, *they’ve*, *I’m* show all the properties of being true affixes rather than simple clitics, which means that, morphologically speaking, such forms are inflected forms of the pronoun.

An important aspect of the Hindi pronoun portmanteaus, including the emphatic forms, is that they consist of a single word form which seems to occupy two adjacent ‘slots’ in syntactic structure. This property is also true of the Romance/German prepositional portmanteaus. A simple but effective treatment of such constructions has been offered by Wescoat (2002). He argues that languages sometimes exhibit ‘lexical sharing’. Modifying the traditional conception of lexical insertion somewhat he argues that portmanteaus

represent a deviation from the default, canonical mapping between word forms and syntactic terminals. Normally, syntactic terminals and word forms are in a one-one correspondence. However, Wescoat argues that portmanteaus prove that we must countenance the possibility that adjacent syntactic terminals, even if part of distinct constituents can be mapped to a single word form. Wescoat provides a formalization of this idea within LFG, demonstrating that the proposal can be incorporated relatively straightforwardly into the existing architecture.

The Hindi pronouns inflected for case or emphasis can therefore be regarded as portmanteau forms of the uninterrupted linear sequence of pronoun + postposition or pronoun + emphasis particle. More generally, the solution to the problem of the pronominals is a solution to the problem of portmanteau forms. It has nothing specifically to do with case and the pronoun forms provide motivation for extending the analysis of Hindi case beyond what is necessary for inflecting nouns generally.

7. Conclusions

The only justification for setting up a CASE attribute in Hindi is provided by the three Layer I inflections: nominative, oblique, and the almost universally neglected vocative. Formatives which have recently come to be treated as case markers, the Layer II clitic postpositions, do not justify setting up an additional attribute, any more than prepositions in English justify a CASE attribute. The Layer I inflections require appeal to a CASE attribute because their forms depend on the inflectional class and grammatical number of the noun. Thus, generalizations would be lost if we try to define their distribution solely in terms of their morphological forms. In the syntax a CASE attribute is required because modifiers show agreement for case and because Layer II postpositions select the oblique case form of the noun they combine with. The Layer II postpositions show none of these effects. If we wish to refer to the fact that a transitive subject is marked with the *ne* postposition all we need to do is to refer to the form of that postposition, and write a rule, constraint or equation which maps the relevant grammatical function to a word identified as FORM NE. Giving such forms an additional case label is completely superfluous.

The other respect in which CASE impacts on the morphosyntax of Hindi is in predicate agreement. Verb forms derived historically from participles show agreement with the highest ranking nominative NP. That is, the verb agrees with a nominative marked subject, and if there is no such subject but there is a nominative object the verb agrees with the object. If there are no subjects or objects in the nominative then the verb takes the default masculine singular form. If the Layer II postpositions are really cases then we have an uneasy tension between the set of case features required for modifier agreement (appealing to the inflectional nominative/oblique distinction) and the set of case features required for predicate agreement (which sets nominative NPs, lacking postpositions, from any NP combined with a postposition, including the transitive subject *ne* postposition). In effect, the term 'nominative' is being used in two distinct (but unacknowledged) senses. This, leads to needless complications, as I have shown.

By adopting Toivonen's (2003) notion of 'non-projecting word' we can easily reconcile the two agreement principles. The Layer II postpositions are syntactic terminals, but

morphologically they are phrasal affixes. This means that we can treat them as words which fail to project a phrase. They adjoin directly to the NP to which they apply, but since they fail to project, the categorial features of the host NP remain unchanged. In particular, this means that the case value of the NP will remain that of the head noun, namely, oblique. We can now introduce a simple constraint into the predicate agreement principles stating that predicates only agree with CASE NOM NPs. The ‘nominative’ feature value of both modifier agreement and predicate agreement thus denotes the same formal entity, and there is therefore no prevarication over the case labelling.

Analyses of Indo-Aryan languages which automatically label the Layer II postpositions as cases are guilty of introducing a totally redundant feature into the grammar, but the motivation for this is easy to understand. The Layer II postpositions are the markers par excellence of grammatical functions, and this is the function par excellence of traditional cases. Yet it seems odd to say that subjects and objects are regularly realized as postpositional phrases. Moreover, the feature ‘nominative’ seems to play an important role in the morphosyntax of agreement, so it seems necessary to set up a case attribute. The idea that the Layer II postpositions are categorially deficient, and fail to project a phrase solves all of these analytical problems and permits us to do justice to all aspects of the nominal morphosyntax of the Indo-Aryan language group.

References

- Beard, R. 1995. *Lexeme Morpheme Base Morphology*. Stony Brook, NY: SUNY Press.
- Butt, M. and T. H. King 2003. Case systems: beyond structural distinctions. In E. Brandner and H. Zinsmeister (eds), *New Perspectives on Case Theory*. Stanford University: CSLI, 53—87.
- Butt, M. and T. H. King 2004. The status of case. In V. Dayal & A. Mahajan (eds) *Clause Structure in South Asian Languages*, Springer Verlag.
- Dymšits, Z. M. 1986a. *Grammatika jazyka xindi 1. Pis'mennost', fonetika, morfologija, znamenatel'nye časti reči*. Moscow: Nauka.
- Dymšits, Z. M. 1986b. *Grammatika jazyka xindi 2. Morfologija, služebnye časti reči, sintaksis*. Moscow: Nauka.
- Haspelmath, M. 2000. Periphrasis. (Art. 68). In G. Booij, Ch. Lehmann and J. Mugdan (eds) *Morphology. An International Handbook on Inflection and Word-Formation. Volume 1*. Berlin: Walter de Gruyter, 654—664.
- Hudson, R. 1995. Does English really have case? *Journal of Linguistics* 31: 375—392.
- McGregor, R. S. 1995. *Outline of Hindi Grammar* [3rd edition]. Oxford: Oxford University Press.
- Masica, C. P. 1991. *The Indo-Aryan Languages*. Cambridge: Cambridge University Press.
- Mohanan, T. 1993. Case alternation on objects in Hindi. *South Asian Language Review* 3: 1—31.
- Mohanan, T. 1994a. *Argument Structure in Hindi*. Stanford: CSLI.
- Mohanan, T. 1994b. Case OCP: A constraint on word order in Hindi. In M. Butt, T. H. King and G. Ramchand (eds) *Theoretical Perspectives on Word Order in South Asian Languages*, 185—216.

- Nordlinger, R. 1998. *Constructive Case*. Stanford: CSLI.
- Otoguro, R. (forthcoming). *Morphosyntax of Case: A Theoretical Investigation of the Concept*. Unpublished PhD dissertation, University of Essex, Colchester, UK.
- Payne, J. 1995. Inflecting postpositions in Indic and Kashmiri. In F. Plank (ed.) *Double Case. Agreement by Suffixation*. Oxford: Oxford University Press, 283—298.
- Sharma, D. 2003. Nominals and constructive morphology in Hindi. In M. Butt & T. H. King (eds) *Nominals inside and outside*. Stanford: CSLI.
- Snell, R. and S. Weightman 1989. *Teach Yourself Hindi*. London: Hodder and Stoughton.
- Spencer, A. 1991. *Morphological Theory*. Oxford: Blackwells.
- Spencer, A. and R. Otoguro 2005. Limits to case – a critical survey of the notion. In M. Amberber & H. de Hoop (eds) *Competition and Variation in Natural Languages. The Case for Case*. Elsevier.
- Toivonen, I. 2003. *Non-Projecting Words. A Case Study of Swedish Particles*. Dordrecht: Kluwer.
- Wescoat, M. T. 2002. *On Lexical Sharing*. Unpublished PhD dissertation, Stanford University.
- Zograf, G. A. 1976 *Morfologičeskij stroj novyx indoarijskix jazykov*. Moskva: Nauka.

PRO-DROP IN NOMINAL POSSESSIVE
CONSTRUCTIONS

Jan Strunk

Sprachwissenschaftliches Institut
Ruhr-Universität Bochum

strunk@linguistics.rub.de

Proceedings of the LFG05 Conference
University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract

I provide LFG analyses for three nominal possessive constructions of modern Low Saxon, a less-studied West Germanic language closely related to Dutch and German. I argue that elegant synchronic analyses of these constructions can be given if it is assumed that they involve a phenomenon which is largely parallel to verbal pro-drop and which I accordingly call *nominal pro-drop* of the possessor. I corroborate this claim by pointing out parallels between verbal and nominal pro-drop in the use of overt pronouns for the subject and possessor respectively. I then extend the nominal pro-drop analysis also to cases of a “missing” possessum phrase and provide evidence against ellipsis accounts. I furthermore argue that my analysis is also suitable for the Low Saxon s-possessive construction. I conclude my paper by giving examples of similar constructions from almost all Germanic languages and also from genetically unrelated languages.

1 Introduction

1.1 Agreement, Pronoun Incorporation, and Pro-Drop in LFG

In many languages, arguments of a head are indexed by morphology on this head. In LFG, it is generally assumed that morphological material attached to a head can specify information that is projected into the grammatical functions of the indexed arguments of this head. The interaction of this morphological material with an overt syntactic expression of the indexed argument(s) determines what information is assumed to be provided by the head-marking. The following outline of this subject is based on Bresnan (2001, chapter 8).

Simple **agreement** morphology as in English subject-verb agreement puts restrictions on certain agreement features of the argument such as e.g. person and number. Thus, the English third person singular verb form *walks* can be used with the third person singular subject *Mary*; cf. (1); but not with a plural subject like *people*; cf. (2).

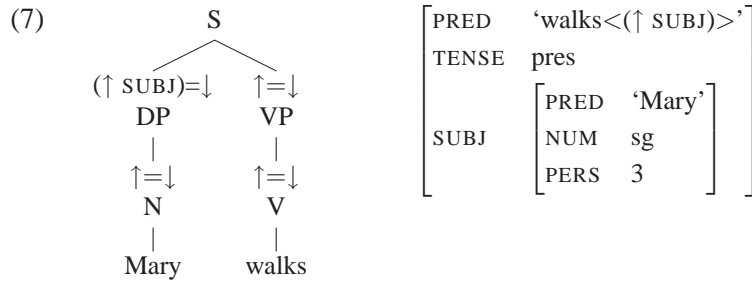
- (1) Mary walks. (2) * People walks. (3) * walks.

This agreement is modelled in LFG by assuming that the lexical entries of *Mary* in (4) and *people* in (5) contain agreement features and that the verbal head *walks* in (6) restricts the values of the agreement features of its subject by projecting information into the SUBJ function within its own f-structure.

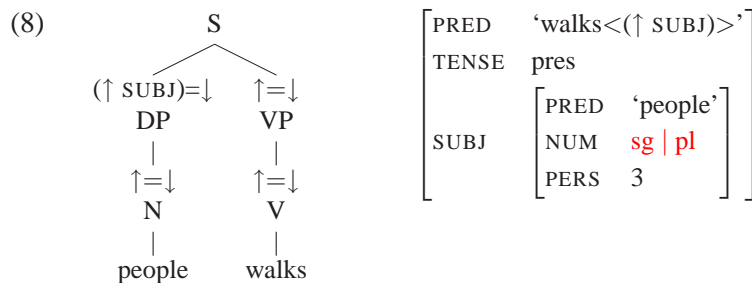
- | | | | | | | | |
|-----|------|---|-------------------|-----|--------|---|---------------------|
| (4) | Mary | N | (↑ PRED) = ‘Mary’ | (5) | people | N | (↑ PRED) = ‘people’ |
| | | | (↑ NUM) = sg | | | | (↑ NUM) = pl |
| | | | (↑ PERS) = 3 | | | | (↑ PERS) = 3 |
| | | | (↑ GEND) = f | | | | |

- (6) walks V (↑ PRED) = 'walk<(↑ SUBJ)>'
 (↑ TENSE) = pres
 (↑ SUBJ NUM) = sg
 (↑ SUBJ PERS) = 3

When *walks* is combined with the third person singular subject *Mary* in the c-structure shown in (7) the result is a well-formed f-structure because the agreement information specified by the head noun of the subject DP and that projected into the SUBJ function by the agreement affix on the verb do not differ.

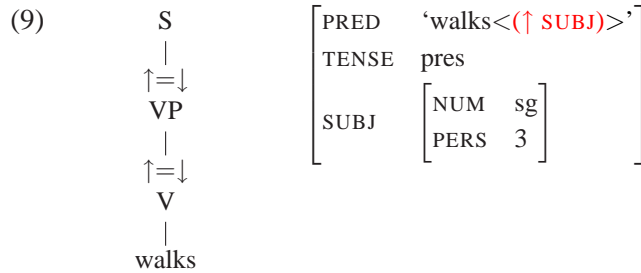


When *walks* is combined with a plural subject like *people* as in (8) the resulting f-structure is not well-formed because the value for the number feature of the subject projected from the head noun of the subject DP itself and the value projected from the agreement affix on the verb are in conflict which leads to a violation of the *uniqueness principle*.



The fact that the sentence *walks* without an overt subject DP in (3) is ungrammatical is modelled by assuming that the agreement affix on the verb only restricts the values of certain agreement features of its subject but does not provide any semantic content, i.e. no PRED feature, for its subject. The c-structure without an overt subject in (9) leads to an *incomplete* f-structure because the verbal head *walks* not only requires the presence of a SUBJ function but the *completeness principle* also demands that this function have semantic content.

To sum up, **agreement** means that a head puts restrictions on one of its argument functions by projecting agreement features into this function. However, simple **agreement** morphology does not provide any semantic content, i.e. no PRED feature, for this function.



In some languages, heads may appear with morphological material that allows them to occur without an overt complement phrase in which case the missing complement is interpreted pronominally. The Chicheŵa example in (10)¹ taken from Bresnan (2001, chapter 8), for example, contains a verb form with a subject affix and an object affix. The subject affix *zi-* agrees with the overt subject *njûchi* (“the bees”); the object affix *wá-* indicates the noun class of the object, but no overt object is present. Instead, the object affix gives rise to a pronominal interpretation for the object.

- (10) *Njûchi zi- ná- wá- lum -a.*
 10.bee 10.S- PST- 2.O- bite -FV
 “The bees bit **them**.”

In contrast to the English subject-verb agreement suffix *-s*, the Chicheŵa object affix *wá-* cannot co-occur with an overt realization of the argument that it indexes; cf. example (11) also taken from Bresnan (2001, chapter 8).²

- (11) * *Njûchi zi- ná- wá- lum -a a- lenje.*
 10.bee 10.S- PST- 2.O- bite -FV 2- hunter
 “The bees bit them the hunters.”

The object affix thus behaves like an ordinary syntactic object pronoun that has been incorporated into the verbal head. This phenomenon is therefore referred to as **pronoun incorporation**. In LFG, incorporated pronouns are modelled by assuming that they provide a pronominal PRED value for the argument function in question in addition to agreement information; cf. the lexical entry of the verb in (12) and the nominal entries in (13) and (14).

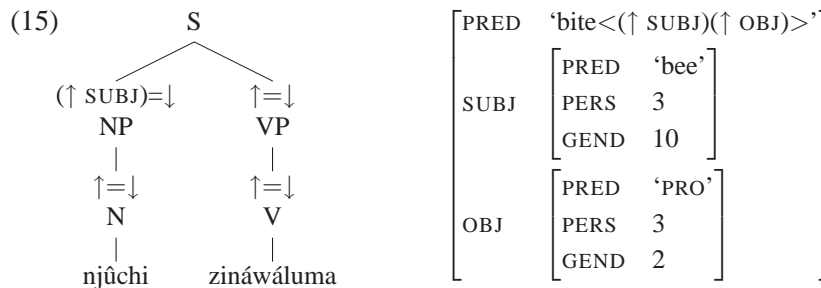
- (12) *zináwáluma V* (↑ PRED) = ‘bite<(↑ SUBJ)(↑ OBJ)>’
 (↑ SUBJ GEND) = 10
 (↑ OBJ PRED) = ‘PRO’ ← *pronominal PRED value*
 (↑ OBJ GEND) = 2

¹Abbreviations used in the glosses: ACC – accusative, FV – final vowel, LK – possessive linker, NOM – nominative, O – object, PL – plural, PST – past, S – subject, SG – singular. Numbers indicate noun classes in the Chicheŵa examples and person in Low Saxon examples.

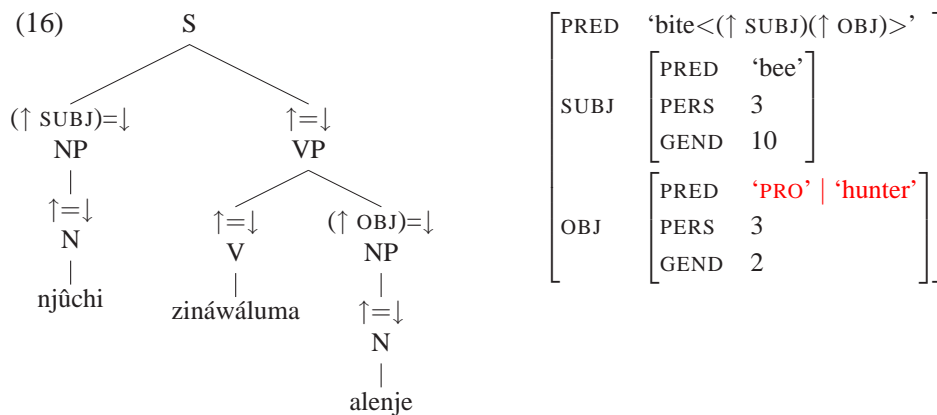
²Like its English translation, sentence (11) can be made grammatical by setting the NP *alenje* (“the hunters”) off intonationally. In this case, however, *alenje* would be a right-dislocated topic that is coreferent with the pronominal object affix contained in the verb and not itself the verb’s object.

- (13) njûchi N (↑ PRED) = ‘bee’ (14) alenje N (↑ PRED) = ‘hunter’
 (↑ PERS) = 3 (↑ PERS) = 3
 (↑ GEND) = 10 (↑ GEND) = 2

The pronominal PRED feature projected into the OBJ function by the object affix on the verb provides semantic content for the OBJ function and thus satisfies the *completeness principle* in a sentence without an overt object NP such as example (10); cf. the structure in (15).



However, if an overt object NP is present at the same time as the object affix the sentence is correctly ruled out because the PRED feature of the object head noun *alenje* and the PRED feature projected from the affix on the verbal head clash and violate the *uniqueness principle*; cf. (16).³



To sum up, **pronoun incorporation** means that an affix on a head can provide a pronominal interpretation for an argument of that head but cannot co-occur with an overt, syntactic realization of this argument.

³PRED is a semantic feature which means that its value can never be unified with anything else (cf. Bresnan 2001, p. 47): Not even two different ‘PRO’ values can unify, so that a Chicheŵa object affix cannot even co-occur with an overt object NP that is pronominal.

Agreement morphology that *can* co-occur with an overt exponent of the grammatical function that it indexes but is interpreted pronominally if no overt, syntactic complement phrase is present exhibits so-called **pro-drop** behavior. Pro-drop morphology functions like agreement morphology when an overt complement phrase is present; cf. the subject affix *zi-* in the Chicheŵa example (17);⁴ but can nevertheless provide a pronominal interpretation for a missing complement; cf. (18).

- (17) *Njũchi zi- ná- wá- lum -a.* (18) **Zi-** ná- wá- lum -a.
 10.bee **10.S-** PST- 2.O- bite -FV **10.S-** PST- 2.O- bite -FV
 “The bees bit them.” “**They** bit them.”

This behavior is standardly modelled in LFG by assuming that the agreement affix on the head provides an *optional* pronominal PRED value for the argument function; cf. the revised lexical entry of *zináwáluma* in (19).

- (19) *zináwáluma* V (↑ PRED) = ‘bite<(↑ SUBJ)(↑ OBJ)>’
 ((↑ SUBJ PRED) = ‘PRO’) ← *optional* PRED feature
 (↑ SUBJ GEND) = 10
 (↑ OBJ PRED) = ‘PRO’ ← *non-optional* PRED feature
 (↑ OBJ GEND) = 2

This lexical entry can be combined with an overt subject NP as in (15) because there is one solution in which the verbal head does not project a PRED feature for its subject, but the verb can also satisfy the *completeness principle* if no overt subject NP is present by projecting a PRED feature into its SUBJ function; cf. (20).

- (20)
- | | | | | | | |
|-------------------|------|------|-------------------------|------|-------|---|
| S | [| PRED | ‘bite<(↑ SUBJ)(↑ OBJ)>’ | | | |
| | | SUBJ | [| PRED | ‘PRO’ |] |
| ↑=↓ | | | PERS | 3 | | |
| VP | | | GEND | 10 | | |
| | | OBJ | [| PRED | ‘PRO’ |] |
| ↑=↓ | PERS | | 3 | | | |
| V | GEND | | 2 | | | |
| | | | | | | |
| <i>zináwáluma</i> | | | | | | |

To sum up, an affix on a head shows **pro-drop** behavior if it can act as agreement marking when the argument it indexes is overtly realized but can also provide a pronominal interpretation if no overt argument phrase is present.

1.2 The Low Saxon Language

Low Saxon is a West Germanic language spoken in northern Germany, the east of the Netherlands, and in emigrant communities throughout the world. It can be

⁴The pronominal interpretation of *zi-* vanishes completely when an overt subject is present. *Njũchi* is a real subject in example (17), not a left-dislocated topic.

considered a “major” minor language in that estimates of the number of speakers are sometimes as high as 10,000,000; cf. the Ethnologue.⁵ However, its survival is threatened because the language is often no longer passed on to children.

Typologically, Low Saxon is a typical West Germanic language with the unmarked word order SVO in main clauses and the order SOV in subordinate clauses. It shows verb-second behavior which means that only one constituent is allowed to appear in front of the finite verb in main clauses. Its case system has been eroded considerably in comparison e.g. with German or Icelandic and only nominative and accusative forms are distinguished.⁶ Low Saxon has three different genders: masculine, feminine, and neuter. Determiners and adjectives in nominal phrases have to agree with the head noun in number, gender, and case. Verbal pro-drop does not occur in the dialects of Low Saxon.

Most of the examples that I use to illustrate my points in the rest of the paper are authentic examples taken from a one million word corpus of Low Saxon that I built by manually harvesting the internet for Low Saxon texts.⁷ All invented examples are explicitly marked.

2 The Possessive Pronoun Construction

A pronominal possessor in Low Saxon is usually expressed by a possessive pronoun preceding a possessum NP; cf. examples (21)–(23). I will refer to this construction as the *possessive pronoun construction*.

- | | | |
|--|--|--|
| (21) <i>ehr Huus</i> her house “her house” | (22) <i>uns Vadder</i> our father “our father” | (23) <i>miene eajne Henj</i> my own hands “my own hands” |
|--|--|--|

The possessive pronoun occurs in the same syntactic position as the definite and indefinite articles, demonstratives, etc. and is in complementary distribution with them; cf. Strunk (2004, p. 40). I therefore conclude that the possessive pronouns are of category determiner and analyze them as a D co-head of the possessum NP; cf. also Dipper (2003) for German. The possessive pronoun agrees with the possessum NP in number, gender, and case; cf. examples (24) and (25).

- (24) *he geiht sien-en Weg*
he goes his-M.SG.ACC way.M.SG.ACC
“He goes his way.”

⁵www.ethnologue.com

⁶In fact, only pronouns and masculine singular nouns have preserved the distinction between nominative and accusative.

⁷There is neither a written nor a spoken standard variety of Low Saxon. Authors use their own dialectal forms and often idiosyncratic writing systems. I will not attempt any form of normalization of the examples I analyze but will always provide an interlinear gloss and an English translation.

- (25) * *he geht sien-e Weg*
 he goes his-F.SG/-PL way.M.SG.ACC

The stem of the possessive pronoun specifies the person, number, and gender of the possessor; cf. examples (21)–(23). The possessive pronoun thus has a kind of dual nature: It indexes both the possessor with the stem and the possessum with an agreement affix. The DP analysis, in which the possessive pronoun of category D is a co-head of the possessum NP, allows for a straightforward modelling of these agreement facts; cf. the lexical entry of a possessive pronoun in (26).

- (26) *ehr* D (↑ POSS PRED) = ‘PRO’
 (↑ POSS PERS) = 3
 (↑ POSS NUM) = sg
 (↑ POSS GEND) = f
 (↑ NUM) = sg
 (↑ GEND) = n
 (↑ CASE) = acc
- (27) *Gesicht* N (↑ PRED) = ‘face<(↑ POSS)>’
 (↑ PERS) = 3
 (↑ NUM) = sg
 (↑ GEND) = n
 (↑ CASE) = acc

The agreement information about the possessum is projected into the f-structure of the pronoun’s mother node, which is the same as that projected by the head noun of the possessum NP because possessive pronoun and possessum NP are co-heads. Thus, agreement with the possessum is enforced. The information about the possessor is projected into the grammatical function POSS(essor) in the mother’s f-structure; cf. (28).⁸

- (28)
- | | |
|--|---|
| $ \begin{array}{c} \text{DP} \\ \swarrow \quad \searrow \\ \uparrow = \downarrow \quad \uparrow = \downarrow \\ \text{D} \quad \quad \text{NP} \\ \quad \quad \\ \text{ehr} \quad \uparrow = \downarrow \\ \quad \quad \text{N} \\ \quad \quad \\ \quad \quad \text{Gesicht} \end{array} $ | $ \left[\begin{array}{l} \text{PRED} \quad \text{‘face}<(\uparrow \text{POSS})>’ \\ \text{PERS} \quad 3 \\ \text{NUM} \quad \text{sg} \\ \text{GEND} \quad \text{n} \\ \text{CASE} \quad \text{acc} \\ \\ \text{POSS} \quad \left[\begin{array}{l} \text{PRED} \quad \text{‘PRO’} \\ \text{PERS} \quad 3 \\ \text{NUM} \quad \text{sg} \\ \text{GEND} \quad \text{f} \end{array} \right] \end{array} \right] $ |
|--|---|

⁸The nature of the POSS function is still a subject of debate; cf. e.g. Laczko (1997) and Chisarik and Payne (2001). The question whether the POSS function is an argument or a non-argument function is largely orthogonal to the issues discussed in this paper. I will simply assume that POSS is an argument function and that all nouns can optionally be augmented by a lexical template to subcategorize for a POSS argument; cf. also Bresnan (2001, p. 169).

3 The Possessive Linker Construction

One possessive construction that is frequently used with non-pronominal possessor phrases in Low Saxon consists of a full possessor DP preposed to a possessive pronoun construction; cf. example (29).

- (29) *[[de'n Jung sien Vadder]*
the.M.SG.ACC boy.M.SG.ACC his.M.SG father.M.SG
“the boy’s father”

I will refer to this construction as the *possessive linker construction*. In this construction, the possessor DP has to stand in the accusative case. The possessor DP and the possessive pronoun in the possessive linker construction agree in number and gender just like a possessive pronoun in the possessive pronoun construction agrees with its antecedent; cf. (30) with a feminine possessor.

- (30) *[[Gerda ehr Mudder]*
Gerda.F.SG.ACC her.F.SG mother.F.SG
“Gerda’s mother”

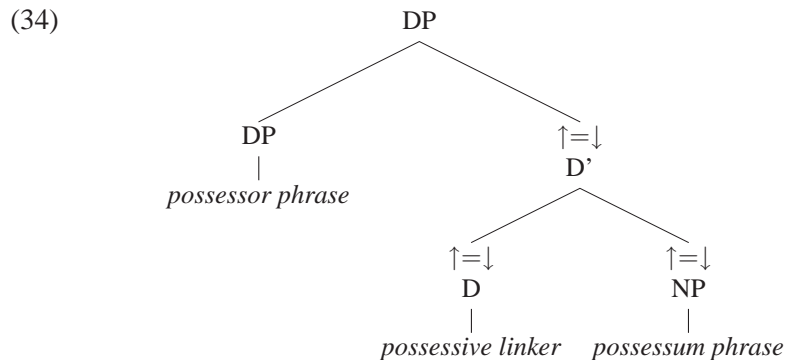
However, in order to analyze these examples as a nominal construction separate from the possessive pronoun construction, it has to be shown that the possessor DP, the possessive pronoun/linker, and the possessum NP form one constituent and that the possessor DP is not the usual antecedent of the possessive pronoun which occurs directly adjacent to it by chance. The evidence for this is very clear: First, as is shown in example (31), the whole construction can occur in front of the finite verb in a verb-second clause, where only one constituent is allowed. Second, when the possessor DP is a relative pronoun, the whole possessive linker construction is pied piped along to the front of the relative clause; cf. example (32). And third, a possessive linker construction can be recursively embedded in another possessive linker construction as possessor phrase; cf. example (33).

- (31) *[Wendtland sien Vadder] harr gor Fritz Reuter kennt.*
Wendtland his father had even Fritz Reuter known.
“Wendland’s father had even known Fritz Reuter.”

- (32) *ena [daem sien Shoobaunt] ekj nich faeich sie loos to moake*
one who his shoe string I not able am loose to make
“one whose shoe string I am not able to untie”

- (33) *[[Paul siene Sesta] aea Saen]*
Paul his sister her son
“Paul’s sister’s son”

These pieces of evidence and the fixed position of the possessor DP directly to the left of the possessive pronoun/linker suggest a c-structure for this construction in which the possessor phrase is located in the specifier of the whole DP; cf. figure (34). A structure like this has been proposed or discussed by a variety of authors for constructions similar to the Low Saxon possessive linker construction or the English s-possessive: Abney (1987), Delsing (1991), Taylor (1996), Norde (1997), Weerman and de Wit (1999), etc.



Once this structure is established, the next question is the nature of the relation between the possessor DP and the possessive pronoun/linker. Is this relation the same as the anaphoric relation of coreference between a possessive pronoun and its antecedent in the preceding context? Does the possessive pronoun in the possessive linker construction resume the referent introduced by possessor DP? An analysis along these lines is suggested for example by the name given to this type of construction by Norde (1997): *resumptive possessive pronoun construction*.

However, I would like to argue against the view that the possessive pronoun in the possessive linker construction functions as a resumptive pronoun. I have already shown that the possessive linker construction forms one constituent and thereby provided evidence against *DP-external resumption*, i.e. resumption understood as left dislocation of the possessor DP outside of the possessive construction and resumption by a possessive pronoun in an ordinary possessive pronoun construction. One further piece of evidence against *DP-external resumption* is the fact that a possessive linker construction can occur in the middle of a clause; cf. (35).

- (35) *De grugelige Bang' in [mudder ehr Ogen] seih ick noch hüt.*
 the terrible fear in mother her eyes see I still today
 “Even today I still see the terrible fear in mother’s eyes.”

One could also understand resumption as *DP-internal resumption*, i.e. the introduction of a referent by the possessor DP inside the possessive linker construction and subsequent resumption of this referent by the possessive pronoun. But although this account is harder to argue against because it is not entirely clear to

me what properties it would predict for the possessive linker construction, I still think that there is some evidence against it. First, the possessor phrase of the possessive linker construction can contain question words or negative possessive pronouns; cf. examples (36) and (37); which should be pragmatically odd if the possessive pronoun was a second act of reference to a discourse entity whose existence is negated or at least not asserted (see also Falk 2002); cf. the infelicitous English left-dislocation example in (38).

- (36) *[[wecke Geister] ehre Kinner]* (37) *[[niimms] siin Vadder]*
 whose minds their children nobody his father
 “the children of whose minds” “nobody’s father”

- (38) # Nobody_i, his_i father is nicer than mine.

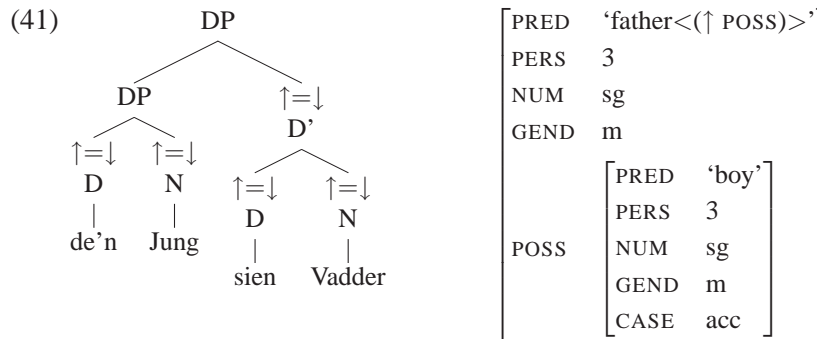
Second, although this is not particularly strong evidence, my informants also do not seem to perceive the possessive pronoun/linker as a second act of reference. Last but not least, note that the possessive pronoun/linker is always directly adjacent to the possessor DP and obligatorily “bound” by it. I would like to argue that even if there had been resumption in the beginning there would have been diachronic pressure to reanalyze the possessive pronoun as a mere possessive linker (or possessive marker) without anaphoric function, because its “antecedent” can always be found directly to the left with no need to perform anaphora resolution.

The alternative approach that I would like to propose is to regard the difference between the possessive pronoun construction with a pronominal interpretation of the possessor and the possessive linker construction with an overt possessor DP as a case of *nominal pro-drop*: When there is no overt possessor DP, the possessive pronoun/linker provides a pronominal interpretation for the possessor by projecting a pronominal PRED feature into the POSS function. It thus gives semantic content to this function and satisfies the *completeness principle*. When there is an overt possessor DP, the possessive pronoun/linker is no longer interpreted anaphorically but only agrees with the possessor in number and gender, i.e. it only projects agreement information into the POSS function but not a PRED feature. The only difference between the lexical entry of the possessive pronoun/linker in (39) and the one I proposed in (26), apart from different agreement information, is that the pronominal PRED feature for the POSS function has been made optional; cf. the standard account of verbal pro-drop in section 1.1.

- (39) sien D ((↑ POSS PRED) = ‘PRO’) ← *now optional*
 (↑ POSS PERS) = 3
 (↑ POSS NUM) = sg
 (↑ POSS GEND) = m
 (↑ POSS CASE) = acc
 (↑ NUM) = sg
 (↑ GEND) = m

After this revision of the lexical entry, an overt possessor DP can be combined with a possessive linker and a possessum NP to model the possessive linker phrase in example (40) without incurring a violation of the *uniqueness principle*; cf. (41).

- (40) *[[de'n Jung] sien Vadder]*
 the.M.SG.ACC boy.M.SG.ACC his.M.SG father.M.SG
 “the boy’s father”



However, so far I have left open how the information from the possessor DP is projected into the POSS function: I did not provide a functional annotation for the possessor DP node in (41). As the possessor DP itself is not specifically marked as possessor and it occupies a fixed position in c-structure, I assume that the DP specifier node should be annotated with an appropriate functional equation. The simplest possible annotation shown in (42) would license an ungrammatical example like (43) without a possessive linker.

- (42) DP \rightarrow DP D' (43) * *[[de'n Jung] Vadder]*
 (↑ POSS)=↓ ↑=↓ the boy father
 “the boy’s father”

The presence of the possessive pronoun/linker is crucial: It acts as possessive marking and establishes the possessive relation; cf. also Plank (1980). I therefore propose to add the equation in (44) to the lexical entries of all possessive pronouns/linkers and to use the alternative c-structure annotation in (45).⁹ The use of an overt possessor DP is now only allowed if a possessive linker is present that establishes the possessive relation by projecting the POSS MARKING feature and possibly also agreement information about the possessor. If there is no possessive linker that acts as possessive marking, the constraining equation in (45) will fail and the possessive construction is ruled out as ungrammatical.

⁹I originally used the implicational c-structure annotation $(\uparrow \text{POSS}) \Rightarrow (\uparrow \text{POSS})=\downarrow$, which yields two unconnected f-structures in case there is no possessive linker to establish the POSS function. I would like to thank Ron Kaplan for pointing out that the unconnectedness of an f-structure is not standardly taken to lead to ungrammaticality.

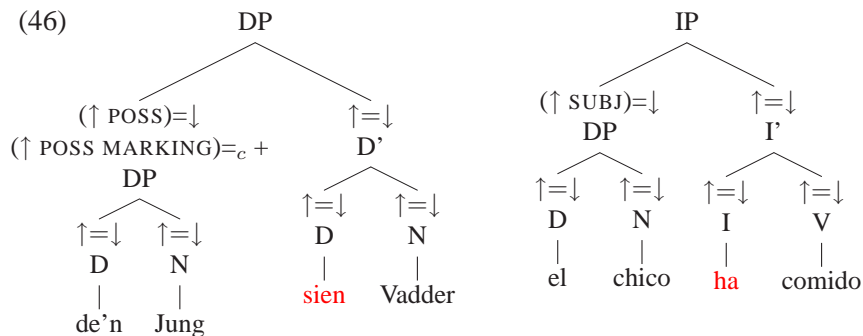
(44) sien D ...
 (↑ POSS MARKING)=+
 ...

(45) DP → DP D'
 (↑ POSS)=↓ ↑=↓
 (↑ POSS MARKING)=_c +

It may seem a little unintuitive at first to call the difference between the possessive pronoun construction and the possessive linker construction *nominal pro-drop* because according to the traditional view there is still a pronominal element present when there is no overt possessor DP, namely the possessive pronoun, while in “canonical” verbal pro-drop only an affix on the verb stem is used when no overt subject is present.¹⁰ This issue was also raised by an anonymous reviewer:

The terminology is slightly confusing since verbal pro-drop usually refers to a pronominal being “dropped”, but this is not the case here. The pronominal is there, but the PRED is dropped [...]

On closer look, however, my proposal is not unintuitive at all. First, I hope to have shown that what I call *nominal pro-drop* can elegantly be modelled using the same formal devices as standardly used in LFG to model verbal pro-drop. Second, the co-head as locus of agreement and pro-drop morphology is not as strange as it may seem: compare the Low Saxon possessive construction with the Spanish periphrastic perfect example, which means (“The boy has eaten.”), in (46). In the Spanish example, which could also be used without the overt subject *el chico* (“the boy”) and thus exhibits verbal pro-drop, the agreement and pro-drop information is also located on the perfective auxiliary *ha*, which is the co-head, and not on the non-finite verb *comido*.



And third, the possessive pronoun/linker is not directly comparable to a simple personal pronoun: It does not only refer to one discourse participant but contains

¹⁰I will also give a more “canonical” example of nominal pro-drop in which the possessor is expressed by an affix on the noun stem in section 7.

information on both possessor and possessum and also establishes the possessive relation. It is not so much an ordinary pronoun as a possessive marker. In the next section, I will indeed provide examples where the possessive linker construction is used with a pronominal possessor DP.

4 Comparing Verbal and Nominal Pro-Drop

In order to corroborate the plausibility of analyzing the Low Saxon possessive pronoun construction and possessive linker construction as a case of *nominal pro-drop*, I would like to point out parallels in the use of nominal and verbal pro-drop.

First, in case of anaphoric reference to a highly accessible referent, no overt subject DP is used in verbal pro-drop and no overt possessor DP is used in nominal pro-drop. Second, if one wants to express lexical content, in both verbal and nominal pro-drop, one has to use an overt subject or possessor DP respectively. The most interesting cases are the special contexts in which the use of overt pronominal subjects or pronominal possessors is possible. In order to see whether nominal pro-drop and verbal pro-drop put the same conditions on the use of overt pronominal subjects and possessors respectively, I devised a short questionnaire and did a small exploratory study with my informants. Specifically, I constructed some examples with contexts in which the use of an overt pronoun should be possible according to the literature on verbal pro-drop and asked them to evaluate whether it was natural to use overt pronouns in the possessive linker construction in these contexts.¹¹

Overt subject pronouns can be used in verbal pro-drop to convey contrastive focus; cf. Larson and Lujà (1989), Cameron (1992), Bresnan (2001), Amaral and Schwenter (2005), etc. The same seems to be true for nominal pro-drop in Low Saxon; cf. example (47).

- (47) *Ik heff all en moien Wogen, man [em sien Auto] is nog
I have already a nice car but **him** his car is still
veel beter.
much better.*

“I already have a very nice car, but **his** car is still much better.”

Overt subject pronouns are also used in coordination; cf. Larson and Lujà (1989). The same is possible in Low Saxon nominal pro-drop; cf. example (48).

- (48) *Dat sünd [[em un sien Broder] ehr Peer].
That are **him** and his Brother their horses
“Those are **his** and his brother’s horses.”*

¹¹All the examples used in this section are constructed examples that my informants judged to be “natural sounding”.

Overt subject pronouns are also used deictically for example while pointing at the intended referent. In the Low Saxon possessive linker constructions, overt pronouns can also be used in this function; cf. (49).

- (49) *Wokeen hört dei tou? – Och, dat is [em sien Wogen].*
 Who belongs that to? Well, that is **him** his car
 “Who does that one belong to? Well, that’s **his** car.”

Most importantly, an overt pronoun can be used in the Low Saxon possessive linker construction to refer to a referent that is currently not the most accessible; cf. the example in (50).

- (50) *Jan wull gern angeln gohn. He wull sien Fründ Hinnerk ok inloden.*
 “Jan wanted to go fishing. He wanted to invite his friend Hinnerk.”
Man [em sien Telefoon] wöör twei.
 “But **his** phone was broken.” (i.e. Hinnerk’s phone was broken)

The overt masculine accusative pronoun *em* makes clear that the intended referent for the possessor of the phone is *Hinnerk* and not *Jan*, which has been the subject of the preceding two sentences and therefore is the most accessible referent. If no overt pronoun had been used, *Jan* would have been interpreted as the possessor of the broken phone. This function of overt pronouns has been termed *switch reference* in the literature on verbal pro-drop; cf. e.g. Cameron (1992) and Dimitriadis (1996).

I thus conclude that the pragmatic conditions on the use of verbal and nominal pro-drop seem to be entirely parallel and that this lends further plausibility to my account of the possessive constructions in Low Saxon and my use of the term *nominal pro-drop*.

5 Pro-Drop of the Possessum

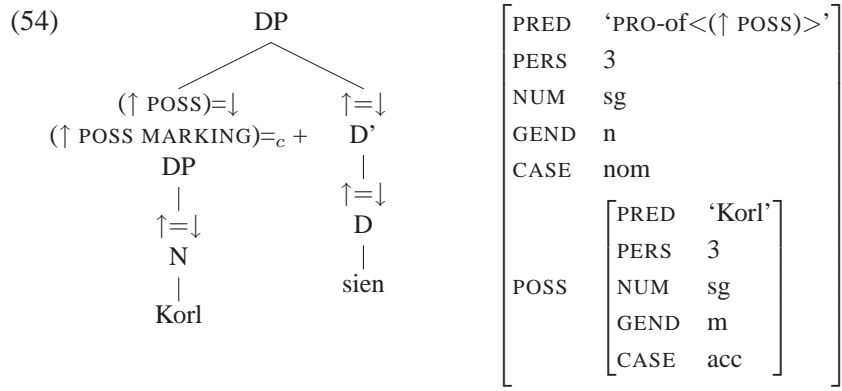
In the preceding two sections, I have established the existence of nominal pro-drop of the possessor in Low Saxon possessive constructions. I now want to argue that the dual nature of the possessive pronoun/linker also makes it plausible to analyze examples like (51) and (52) in which there is no overt possessum as cases of pro-drop of the possessum.

- (51) *säi läegt höör kop tegen [mîn]*
 she lays her head against mine
 “She leans her head against mine.”
- (52) *Mien Öller ... [Fritz sien] ... un [Korl sien] ...*
 my age Fritz his and Korl his
 “My age ... Fritz’s ... and Korl’s ...”

In the examples (51) and (52), no overt possessum NP is present in the bracketed possessive constructions (although one could have been used there) and the possessum is inferred from the context. This can be modelled by assuming that the possessive pronoun/linker does not only project agreement information about the possessum but in addition provides an optional pronominal PRED feature for the possessum in the same way as it optionally provides such a feature for the possessor. The only information that has to be added to the lexical entry of the possessive linker in (39) is shown in (53).

- (53) sien D ...
 ((↑ PRED) = 'PRO-of<(↑ POSS)>') ← *optional*
 ...

The possessive linker now optionally projects a pronominal semantic feature into the f-structure of the whole DP and the pronominal interpretation of the possessum is thus also modelled as a case of nominal pro-drop; cf. the structure in (54).



But why should one not treat cases of missing possessum NP as ellipsis? First of all, note that the possessive relation which is established by the possessive pronoun/linker always entails the existence of a possessum. Second, the possessive pronoun/linker has to contain information about the possessum anyway in order to model agreement. It is thus quite plausible to assume that the possibility of a pronominal interpretation for the possessum is a lexical fact stated in the lexical entry of the possessive pronoun/linker. Third, other determiners such as demonstratives can also be interpreted pronominally when they occur without a following NP and often it is not really possible to reconstruct what exactly could have been elided; cf. example (55).

- (55) *Dat kann he doch ni nich!*
 that can he though never
 "But he could never do that!"

The same is true for the possessive pronoun/linker; cf. example (56), in which the possessum is interpreted very abstractly as *possessions* or *belongings* but could be interpreted in a variety of ways.

- (56) *Hest du dien, de Anner sien ...*
 have you yours the other his
 “If you have yours and the other his, ...”

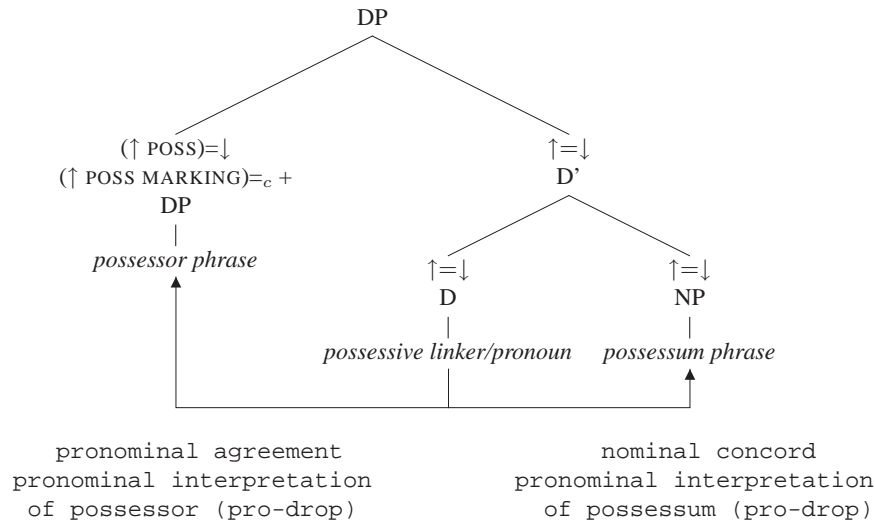
If demonstrative pronouns like the one in example (55) are treated as pronominal elements and no ellipsis is assumed, the same should apply to possessive pronouns/linkers like those in (56). Last but not least, there are forms of the possessive pronoun/linker in some dialects of Low Saxon that can be analyzed as incorporating a pronominal possessum because these forms can never occur with an overt possessum NP but always provide a pronominal interpretation for the possessum; cf. example (57) from the dialect of Groningen in the Netherlands.

- (57) *heur voader en mienent (*voader)*
 her father and mine father

As the standard analyses of pronoun incorporation and pro-drop are quite similar in LFG, the existence of pronoun incorporation of the possessum in Low Saxon is a further (theory-internal) argument for modelling missing possessums as pro-drop and not as some form of ellipsis.

To sum up, the structure of the Low Saxon possessive linker construction can be schematized as in figure (58).

- (58)



6 The S-Possessive Construction

Most dialects of Low Saxon use a third possessive construction, which I will call the *s-possessive construction*; cf. examples (59) and (60). This construction is

similar to the s-possessives in other Germanic languages, such as e.g. Dutch, German, Scandinavian, and also English. It is traditionally regarded as a possessive construction with a possessor phrase in genitive case. However, in Strunk (2004) I show that the invariant =s possessive marking, which always appears once in between the possessor DP and the possessum NP, behaves more like a clitic possessive linker than like case-marking morphology. Moreover, the =s clitic seems to occupy the same syntactic position as the possessive pronouns. Many other authors have come to similar conclusion regarding the s-possessive in other Germanic languages; cf. e.g. Janda (1980), Delsing (1991), Hudson (1995), Taylor (1996), Norde (1997), and Weerman and de Wit (1999), etc.

- (59) *[[höör ollen] =s hus]* (60) *[[Antje] =s Bröögam]*
her parents 's house Antje 's bridegroom
“her parents’ house” “Antje’s bridegroom”

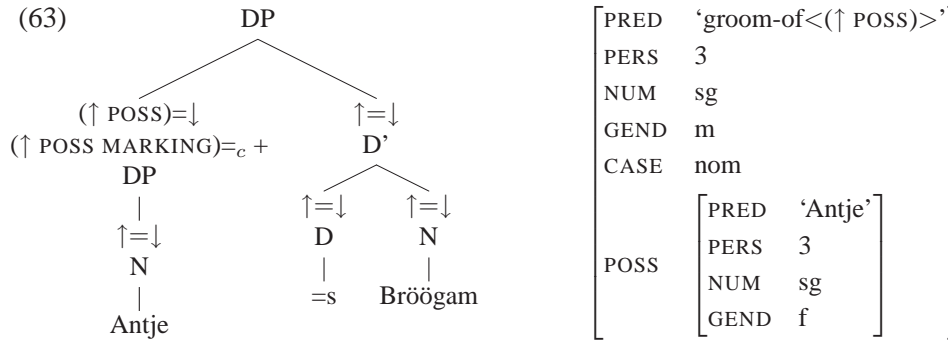
Because of the structural similarities between the s-possessive and the possessive linker construction, I would like to argue that the s-possessive in Low Saxon (and also in other Germanic languages) can be analyzed in a similar way. Like the possessive pronoun, the =s clitic functions as a possessive marker that establishes the possessive relation. However, in contrast to the possessive pronoun/linker, the =s morpheme is invariant and thus it does not project any agreement information, neither about the possessor nor about the possessum. The Low Saxon s-possessive always has to occur with an overt possessor DP, the possessum NP can be missing and is then interpreted pronominally; cf. example (61).

- (61) *Hinnerk =s Huss iss groote den Antje =s.*
Hinnerk 's house is bigger than Antje 's
“Hinnerk’s house is bigger than Antje’s.”

I therefore assume the lexical entry in (62) for the =s possessive linker. It acts as possessive marking so that the information from the possessor DP can be projected into the POSS function. It also contains an optional pronominal PRED feature to allow for pro-drop of the possessum.

- (62) =s D (↑ POSS MARKING) = +
((↑ PRED) = ‘PRO-of<(↑ POSS)>’) ← *optional*

With this lexical entry, examples like (59) and (60) but also examples with a missing possessum NP like (61) can be analyzed; cf. (63).



7 Similar Constructions in Other Languages

Many modern Germanic languages make use of a pronominal linker construction that should be amenable to the same kind of analysis as proposed for Low Saxon: Afrikaans (64), Dutch (65), Frisian (66), colloquial German, and Norwegian (67), and West Flemish.

- | | |
|--|--|
| <p>(64) <i>my moeder se huis</i> my mother LK house “my mother’s house”</p> <p>(66) <i>heit syn hynder</i> father his horse “father’s horse”</p> | <p>(65) <i>mijn moeder d'r auto</i> my mother her car “my mother’s car”</p> <p>(67) <i>Per sin bil</i> Per his car “Peter’s car”</p> |
|--|--|

Most Germanic languages also have an s-possessive construction: Dutch (68), English, Frisian (69), German (70), Swedish (71) (and the other Scandinavian languages), and maybe West Flemish (72).

- | | |
|---|--|
| <p>(68) <i>mijn moeder =s auto</i> my mother 's car</p> <p>(70) <i>Mutter =s Auto</i> mother 's car</p> | <p>(69) <i>ús buorman =s tún</i> our neighbor 's garden</p> <p>(71) <i>Per =s bil</i> Per 's car</p> <p>(72) <i>Marie-se boek</i> Mary’s book</p> |
|---|--|

Moreover, there are many languages in the world with similar possessive constructions that could be analyzed as cases of nominal pro-drop; cf. Koptjevskaja-Tamm (2001, p. 963) on so-called “possessor-doubling constructions”. The term *pro-drop* has also been used by Chisarik and Payne (2001) in connection with Hungarian possessive constructions and by Kathol (2001) in connection with possessives in Luiseño. Another such language is the Oceanic language Roviana

(Corston-Oliver 2002), which make use of an even more “canonical” version of nominal pro-drop in that pronominal possessors are expressed by affixes on the noun stem; cf. (73). A syntactic possessor phrase can then be combined with a noun inflected for “possessive agreement”; cf. example (74). The affix can be analyzed as establishing the possessive relation and optionally providing a pronominal PRED feature for the possessor in case there is no overt possessor phrase present.

| | |
|---------------------|--|
| (73) <i>tama-na</i> | (74) [<i>tama-na</i> [<i>tie</i> <i>hoi</i>]] |
| father-3.SG | father-3.SG person that |
| “his/her father” | “that person’s father” |

8 Conclusion

I hope to have shown that the analysis standardly assumed in LFG to model agreement, pronoun incorporation, and pro-drop behavior in the verbal domain can also be used to account for the behavior of possessive constructions in modern Low Saxon, in other Germanic languages, and in many other languages from around the world. Specifically, I have argued for a nominal pro-drop analysis for so-called possessor doubling phenomena and against a resumptive pronoun approach to such constructions. Furthermore, I have extended the pro-drop analysis also to cases of “missing” possessum phrases. My analyses show that the possessive pronouns in the Germanic languages are not simple pronouns but act as a possessive marker with a dual nature that contains information about both possessor and possessum.

I would like to thank Joan Bresnan, Reuben Epp, Reinhard F. Hahn, Nikolaus Himmelmann, Dan Jurafsky, Ron Kaplan, Judith Köhne, Jonny Meibohm, Eldo Neufeld, Friedrich W. Neumann, Anette Rosenbach, Helge Tietz, Tom Wasow, Holger Weigelt, and Shirley Wyatt and the audience at the LFG05 conference in Bergen!

References

- Abney, S. P. (1987). *The English Noun Phrase in its Sentential Aspect*. Ph. D. thesis, MIT, Cambridge.
- Amaral, P. M. and S. A. Schwenter (2005). Contrast and the (Non-) Occurrence of Subject Pronouns. Selected Proceedings of the 7th Hispanic Linguistics Symposium, pp. 116–127. Cascadilla Proceedings Project, Somerville.
- Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics 16. Blackwell, Malden/Oxford.
- Cameron, R. (1992). *Pronominal and Null Subject Variation in Spanish: Constraints, Dialects, and Functional Compensation*. Ph. D. thesis, University of Pennsylvania.

- Chisarik, E. and J. Payne (2001). Modelling Possessor Constructions in LFG: English and Hungarian. Proceedings of the LFG01 Conference, Hong Kong. CSLI Publications, Stanford. <http://csli-publications.stanford.edu/>.
- Corston-Oliver, S. (2002). Roviana. In J. Lynch, M. Ross, and T. Crowley (Eds.), *The Oceanic Languages*, pp. 467–497. Curzon Press, London.
- Delsing, L.-O. (1991). Om genitivens utveckling i fornsvenskan. In S.-G. Malmgren and B. Ralph (Eds.), *Studier i svensk språkhistoria* 2, Nordistica Gothoburgensia 14, pp. 12–30. ACTAG, Gothenburg.
- Dimitriadis, A. (1996). When Pro-Drop Languages Don't: Overt Pronominal Subjects and Pragmatic Inference. Proceedings of the 32th meeting of the Chicago Linguistic Society.
- Dipper, S. (2003). *Implementing and Documenting Large-Scale Grammars – German LFG*. Ph. D. thesis, University of Stuttgart.
- Falk, Y. N. (2002). Resumptive Pronouns in LFG. Proceedings of the LFG02 conference, Athens. CSLI Publications, Stanford. <http://csli-publications.stanford.edu/>.
- Hudson, R. (1995). Does English Really Have Case? *Journal of Linguistics* 31, 375–392.
- Janda, R. D. (1980). On the Decline of Declensional Systems: The Overall Loss of OE Nominal Case Inflections and the ME Reanalysis of -es as his. In E. C. Traugott, R. Labrum, and S. Shepherd (Eds.), *Papers from the 4th International Conference on Historical Linguistics*, pp. 243–253. John Benjamins, Amsterdam/Philadelphia.
- Koptjevskaja-Tamm, M. (2001). Adnominal Possession. In M. Haspelmath, E. König, W. Oesterreicher, and W. Raible (Eds.), *Language Typology and Language Universals*, HSK 20.2, pp. 960–970. Walter de Gruyter, Berlin.
- Laczkó, T. (1997). Action Nominalization and the Possessor Function within Hungarian and English Noun Phrases. *Acta Linguistica Hungarica* 44 (3-4), 413–475.
- Larson, R. and M. Lujà (1989). Emphatic Pronouns. <http://semlab5.sbs.sunysb.edu/~rlarson/emphpro.pdf>.
- Norde, M. (1997). *The History of the Genitive in Swedish. A Case Study in Degrammaticalization*. Ph. D. thesis, University of Amsterdam.
- Plank, F. (1980). Encoding Grammatical Relations: Acceptable and Unacceptable Non-Distinctness. In J. Fisiak (Ed.), *Historical Morphology*, Trends in Linguistics. Studies and Monographs 17, pp. 289–325. Mouton, The Hague.
- Strunk, J. (2004). Possessive Constructions in modern Low Saxon. Master's thesis, Stanford University. <http://www.linguistics.rub.de/~strunk/mathesis.pdf>.
- Taylor, J. R. (1996). *Possessives in English. An Exploration in Cognitive Grammar*. Clarendon Press, Oxford.
- Weerman, F. and P. de Wit (1999). The Decline of the Genitive in Dutch. *Linguistics* 37/6, 1155–1192.

**ENGLISH NONSYLLABIC AUXILIARY CONTRACTIONS:
AN ANALYSIS IN LFG WITH LEXICAL SHARING**

Michael T. Wescoat
University of California, Davis

Proceedings of the LFG05 Conference
University of Bergen
Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications
<http://csli-publications.stanford.edu/>

Abstract

English auxiliary contractions may reduce to varying degrees, sometimes becoming nonsyllabic, with only a consonant. Most nonsyllabic contractions exhibit behavior that suggests they are joined to the preceding form in the lexicon. Yet paradoxically they behave syntactically like a clitic group, formed from two distinct constituents. I conclude that these forms are *lexical clitics*. To model lexical clitics, I employ a mechanism called *lexical sharing*, allowing two or more atomic constituents to be instantiated by the same word. Combining lexical sharing with LFG provides a way to model functional constraints associated with nonsyllabic auxiliary contractions. I also show that lexical sharing provides an illuminating analysis of so-called second-word clitics, concluding that adding lexical sharing to LFG provides a useful component in the analysis of cliticization.

1 Introduction

This paper examines a problematic area of cliticization and considers how one might tackle it within the theory of Lexical-Functional Grammar (LFG). The issue revolves around a subset of English auxiliary contractions, specifically those which are most radically reduced, leaving only a consonant. These contractions do not form syllables unto themselves; therefore, I call them *nonsyllabic auxiliary contractions*. The behavior of some members of this class runs counter to the traditional thinking about clitics. Here I explain why these forms are challenging, and I offer an analysis that combines the tools traditionally made available in LFG with a mechanism that I call *lexical sharing* (Wescoat 2002). I go on to suggest that the incorporation of lexical sharing into an LFG proves useful for analyzing other types of clitic phenomena.

2 The traditional view of auxiliary contractions

English auxiliary contractions are routinely treated as clitics. Indeed, Zwicky and Pullum (1983) offer two auxiliary contractions, 's (*is* or *has*) and 've (*have*), as paradigmatic exemplars of clitics. More specifically, Zwicky treats English auxiliary contractions as members of the class of *simple clitics*, which comprises “cases where a free morpheme, when unaccented, may be phonologically subordinated to a neighboring word” (1977:5). For instance, 'll corresponds to the free form *will*, the contracted form arising only in places where the full form could have occurred:

- (1) a. I'll help.
- b. I **will** help.

The idea that auxiliary contractions are syntactically free yet phonologically bound is echoed in Di Sciullo and Williams's assumptions about the process by which such forms are derived:

The correct distribution for *I'll* is obviously arrived at in this way: *first*, independently determine the distribution of *I* and *will* according to syntax, and *then* weld the two together when they occur juxtaposed. Clearly, if this description is correct, then *I'll* is not a *syntactic atom* in any sense, because it is composed of syntactically accessible parts. So if *I'll* is a word at all, . . . it is a *phonological word*. (1987:107, emphasis added)

The term *syntactic atom* may be interpreted as referring to the smallest, indivisible units within the c(onstituent)-structure. Thus, we are told that *first* the syntax incorporates *I* and *will* into the c-structure as two autonomous syntactic atoms, *then*, in some postsyntactic readjustment, *I* and *will* are joined into *I'll*, which constitutes a word as far as the rules of phonology are concerned.

3 Lexicalist counteranalyses

Spencer (1991:383) notes that some auxiliary contractions seem to be incompatible with the traditional view of these forms as clitics. It has been recognized at least since the work of Sweet (1890:14–16) that auxiliary contractions may be reduced to varying degrees:¹

- (2) **aar** ‘are’: ə(r).
- æm** ‘am’: əm; m.
- hæd** ‘had’: həd; əd; d.
- hæv** ‘have’: həv; əv; v.
- hæz** ‘has’: həz; əz; z, s...
- iz** ‘is’: iz; z, s...
- wil** ‘will’: əl; l.
- wud** ‘would’: wəd; əd; d.

In the most extreme instances, all that remains is a single consonant that is realized as the final coda of the preceding word. As noted above, since these contractions do not form a syllable unto themselves, I call the single-consonant forms nonsyllabic auxiliary contractions. Spencer observes that this subclass of auxiliary contractions exhibits curious properties that make them more amenable to a lexicalist analysis.

3.1 Selection

One telling characteristic of nonsyllabic auxiliary contractions is their propensity to *select* pronouns and *wh*-words as the forms to which they attach:

- (3) a. **I’ll** help. [aɪl]
- b. **Ai’ll** help. [aɪ.l/*aɪl]²
- (4) a. **We’re** a big group. [wi:ɹ]
- b. The **Cree’re** a big group. [kɹi:ɹ/*kɹi:ɹ]
- (5) a. **They’ve** gone. [ðeɪv]
- b. They **may’ve** gone. [meɪ.əv/*meɪv]
- (6) a. **I’m** happy. [aɪm]
- b. **So am** I. [səʊ.m/*səʊm]
- (7) a. **How’ve** you been? [haʊv]
- b. The **Au’ve** been polled. [aʊ.əv/*aʊv]³

Note that ‘.’ symbolizes a syllable boundary and that the vertical line ‘|’ below a sonorant-consonant symbol indicates that the sound is being used vocally, as a syllable peak. With the nonpronoun, non-*wh*-words above, the nonsyllabic contractions are incompatible; instead, a less reduced contraction that contains its own syllable peak must be employed.

The property of selecting pronouns and *wh*-words is shared by some but not all nonsyllabic auxiliary contractions. Most notably, the nonsyllabic forms of ‘s (*is* or *has*) may attach to words of any category:

¹Sweet’s original transcriptions are retained in (2); ⟨aa⟩, ⟨i⟩, ⟨i̇⟩, and ⟨u⟩ correspond to ⟨a⟩, ⟨ɪ⟩, ⟨i̇⟩, and ⟨ʊ⟩ in the International Phonetic Alphabet.

²*Ai* is a Japanese given name.

³The Au are a people of Papua New Guinea.

- (8) a. **Pat's** gone. [pæts]
 b. **So's** John. [souz]

The behavior of 'd (*had* or *would*) is similar in some dialects. On the one hand, Spencer reports the judgments in (9) and (10), which suggest that nonsyllabic 'd selects pronouns for some speakers:

- (9) a. **She'd** seen it. [ʃi:d]
 b. **Lee'd** seen it. [li:əd/*li:d]
 (10) a. **I'd** have seen it. [aɪd]
 b. **Bligh'd** have seen it. [blaɪ.əd/*blaɪd] (1991:383)

Yet in other dialects, including my own and that described by Zwicky (1970:331–332), even the b. sentences above are compatible with nonsyllabic 'd.⁴ Thus, the nonsyllabic contractions of 's, and for some speakers 'd, pattern more like the corresponding syllabic contractions and will not be relevant to this discussion. Henceforth, I use the term *restrictive* nonsyllabic auxiliary contractions to refer to the nonsyllabic forms of 'll (*will*), 'm (*am*), 're (*are*), and 've (*have*) (along with 'd in dialects like that described by Spencer); these forms share the crucial properties on which this argument is based.

The fact that they select pronouns or *wh*-words as the forms to which they attach is significant for determining the morphological status of restrictive nonsyllabic auxiliary contractions. Zwicky and Pullum provide some well-known criteria for distinguishing clitics and affixes. The latter term should be interpreted here as describing morphemes that attach to stems in the derivational or inflectional morphology. One of their criteria is given in (11):

(11) **Zwicky and Pullum's criterion A**

Clitics can exhibit a low degree of selection with respect to their hosts, while affixes exhibit a high degree of selection with respect to their stems. (1983:503)

By criterion A, restrictive nonsyllabic auxiliary contractions seem rather affix-like.

3.2 Morphophonological idiosyncrasies

Another curious property of nonsyllabic auxiliary contractions concerns morphophonological idiosyncrasies in the form of pronouns to which they attach. The literature reveals some degree of dialect variation in this area, but for me,⁵ the following generalizations hold: *I* [aɪ] may be pronounced [ɑ], but only in association with 'll (*will*), yielding *I'll* [ɑl]; moreover, *you* [ju:] may become [jɔ], but only when followed by 're (*are*), resulting in *you're* [jɔr]:⁶

- (12) *I'll* [aɪl/ɑl] *I'm* [am/*ɑm] *I've* [av/*ɑv]
you'll [ju:l/*jɔl] *you're* [ju:ɹ/jɔr] *you've* [ju:v/*jɔv]

This, by the way, is not a fast-speech phenomenon; *I'll* [ɑl] and *you're* [jɔr] may be heavily stressed and elongated.

Pronouns to which nonsyllabic auxiliary contractions attach undergo another process which is somewhat more regular. Zwicky (1970:330) describes this with a phonological rule called *Phonetic*

⁴Coincidentally, Bloch indirectly corroborates the existence of the latter sort of dialect; in discussing the phonology of "Midwestern English," he states "*pod* . . . is phonetically identical with *pa'd*" (1941:283–284).

⁵My thanks go to the students of Linguistics 121, Spring 2005, at the University of California, Davis, for discussing and corroborating these judgments.

⁶Sweet also reports this idiosyncratic pronunciation of *you're*, which he renders as "jɔə, jɔr" (1890:25).

Laxing, which causes vowels that are long and tense (i.e., with advanced tongue root, [+ATR]) to become short and lax (i.e., [-ATR]). For me, Phonetic Laxing is most clearly applicable when the following contraction consists of just a liquid, as in 'll (*will*) or 're (*are*); with other nonsyllabic auxiliary contractions the rule appears not to apply unless the pronoun is *you*:⁷

| | | | | | | | | | | | | |
|------|-------------|-------|----------------|-------|----------------|-------|----------------|--------|---------------|--------|--------------|--------|
| (13) | <i>we</i> | [wi:] | <i>we'll</i> | [wɪl] | <i>we're</i> | [wɪr] | <i>we've</i> | *[wɪv] | <i>we'd</i> | *[wɪd] | | |
| | <i>you</i> | [ju:] | <i>you'll</i> | [jʊl] | <i>you're</i> | [jʊr] | <i>you've</i> | [jʊv] | <i>you'd</i> | [jʊd] | | |
| | <i>he</i> | [hi:] | <i>he'll</i> | [hɪl] | | | | | <i>he'd</i> | *[hɪd] | <i>he's</i> | *[hɪz] |
| | <i>she</i> | [ʃi:] | <i>she'll</i> | [ʃɪl] | | | | | <i>she'd</i> | *[ʃɪd] | <i>she's</i> | *[ʃɪz] |
| | <i>they</i> | [ðeɪ] | <i>they'll</i> | [ðɛl] | <i>they're</i> | [ðɛr] | <i>they've</i> | *[ðɛv] | <i>they'd</i> | *[ðɛd] | | |

I find that all of the licit forms in (13) may be stressed and elongated; thus, Phonetic Laxing is not a fast-speech phenomenon. Significantly, this rule has a highly restricted range of application, operating only in vowel-final pronouns to which nonsyllabic auxiliary contractions are attached; witness *we'll* [wi:l/wɪl] vs. *wheel* [wi:l/*wɪl].⁹

Here another of Zwicky and Pullum's criteria for distinguishing clitics and affixes comes into play:

(14) **Zwicky and Pullum's criterion C**

Morphophonological idiosyncrasies are more characteristic of affixed words than of clitic groups. (1983:504)

In light of criterion C, since the above [ai/a] (*I*) and [ju:/jɔ] (*you*) allomorphies and the highly constrained rule of Phonetic Laxing are morphophonological idiosyncrasies triggered by the attachment of nonsyllabic auxiliary contractions, these forms are once again revealed to be affix-like.

3.3 A lexical source for restrictive nonsyllabic auxiliary contractions

In lexicalist theories, selection and morphophonological idiosyncrasies like those described above are lexical matters. One may therefore follow Spencer (1991:383) in assuming some version of (15):

(15) **Lexical source hypothesis**

The nonsyllabic contractions of *am*, *are*, *have*, and *will* (and for some speakers, *had* and *would*) are attached to pronouns and *wh*-words *in the lexicon*.

Motivated by this hypothesis, some researchers have analyzed restrictive nonsyllabic auxiliary contractions not as clitics but as suffixes that attach to a stem to form affixed words. Sadler (1998) treats pairings of pronouns with nonsyllabic auxiliary contractions as *tense-marked pronouns* (D), as in (16a). Bender and Sag's (2001) HPSG analysis *incorporates* the pronoun into the auxiliary

⁷Plainly there is variability among speakers here; for instance, Zwicky (1970:330) indicates that Phonetic Laxing does not occur with 's (*is* or *has*), but he seems to accept it with other auxiliary contractions.

⁸Sweet's data also show the effect of Phonetic Laxing in this form; *they're* is rendered as "ðɛər" (1890:25).

⁹Even more drastic reductions of the form of the pronouns are possible. When the nonsyllabic auxiliary consists of an obstruent, the pronoun may be realized with a central vowel, as in *she'd* [ʃɪd] or *they've* [ðɛv]; however, it appears to me that these occur strictly in fast speech. When the auxiliary consists of a sonorant, it may be vocalized, in which case it takes over as the syllable peak of the pronoun-auxiliary unit, as in *you're* [jɹ] and *they'll* [ð]. Interestingly, I find that the forms with vocalized sonorants can be stressed and elongated, though *he'll* [hɪ] may be an exception to this generalization. I will not consider these data further here, beyond noting that these phenomena provide yet more evidence of phonological processes that apply only in pronoun+auxiliary combinations.

(V), which in turn combines with a base-form VP to yield a saturated sentence requiring no subject NP, as in (16b):



4 Problems with affixed-word analyses

Approaches that treat the combination of a pronoun or *wh*-word with a restrictive nonsyllabic auxiliary contraction as an affixed word encounter difficulties when it comes to their predictions about the syntax.

First note that coordination fails to apply to the hypothesized affixed word, as seen in (17):

(17) ***You're** [jɔɪ] and **I'm** [am] helping.

Consider (17) in the light of another of Zwicky and Pullum's criteria, shown in (18):

(18) **Zwicky and Pullum's criterion E**

Syntactic rules can affect affixed words, but cannot affect clitic groups. (1983:504)

Since the rule of coordination cannot combine *you're* [jɔɪ] with *I'm* [am], by criterion E, these forms behave more like clitic groups than affixed words. Consider the reasoning underlying this conclusion. The affixed-word analyses sketched in (16) would predict that *you're* [jɔɪ] and *I'm* [am] are syntactic atoms, in the sense in which Di Sciullo and Williams (1987) employ this term, designating the smallest parts of a c-structure. As syntactic atoms, *you're* [jɔɪ] and *I'm* [am] would be constituents, so one is left to wonder why they would not undergo coordination. In contrast, clitic groups are assumed to comprise multiple syntactic atoms, which are phonologically bound. If *you're* [jɔɪ] and *I'm* [am] are each composed of a pronoun D and an auxiliary I, which would not together constitute a constituent, then there is every reason to expect coordination to fail in (17).

Another problem with the affixed-word analyses concerns I' coordination, as in (19):

(19) **I'll** [aɪl/əl] be there on Sunday and [_I am looking forward to seeing you]

Here the future tense of *I'll* [aɪl/əl] takes scope only over the left-hand conjunct. Moreover, the right-hand conjunct, headed by the tensed auxiliary *am*, needs a first-person, singular subject; this subject is in fact shared by both conjuncts. These observations are easily handled if one assumes that *I'll* [aɪl/əl] in (19) corresponds to a clitic group comprising two syntactic atoms, a first-person, singular D and a future-tense I. The D lies outside of the coordinate structure, taking scope over both conjuncts, while the I is inside of the left-hand conjunct, which corresponds, appropriately enough, to the perceived scope of the future tense. However, if *I'll* [aɪl/əl] were an affixed word, and thus a syntactic atom, one would have to choose between two equally unpalatable analyses. The left-hand conjunct would either contain the presumed syntactic atom *I'll* [aɪl/əl] or lack it. In the former case, the left-hand conjunct would be the whole clause *I'll be there on Sunday*; in the latter, it would be the phrase *be there on Sunday*, headed by uninflected *be*. Neither candidate

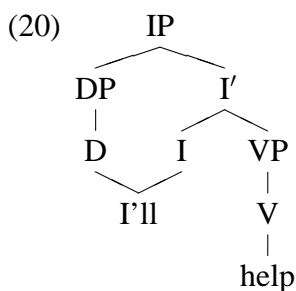
is plausible as the co-conjunct of an I' headed by the present-tense auxiliary *am*. Given these observations, *I'll* [aɪl/aɪ] once again seems more like a clitic group than an affixed word.

5 Toward lexical sharing

A paradox emerges here. By the lexical source hypothesis in (15), restrictive nonsyllabic auxiliary contractions are attached to pronouns and *wh*-words in the lexicon, as are affixes. Yet these pairings behave syntactically like clitic groups, and it is a widely held assumption that “All cliticization . . . follows syntax” (Zwicky & Pullum 1983:504). Thus, the conclusion seems so far to be that the derivation of restrictive nonsyllabic auxiliary contractions takes place in two parts of the grammar that lexicalist theories strive to keep scrupulously separate. To resolve this problem, one of the foregoing assumptions must be abandoned; I propose to explore the hypothesis that *not* all cliticization follows syntax. More specifically, I claim that restrictive nonsyllabic auxiliary contractions are instances of *lexical cliticization* of the sort advanced by Booij and Rubach (1987:36) to describe Polish preterite clitics. Moreover, I propose to treat lexical cliticization as an instance of a phenomenon that I call *lexical sharing* (Wescoat 2002), in which two or more syntactic atoms share a single word as their *lexical exponent*.

5.1 How *not* to model lexical sharing

The capacity to associate one word with two syntactic atoms is something that must be carefully and precisely implemented in the theory of c-structure. In fact, the traditional model of c-structure is ill-equipped for the task. Consider the form *I'll* [aɪl/aɪ], which, as an instance of lexical sharing, would need to be associated with two syntactic atoms, a D and an I. The traditional way to represent the fact that a word *w* is associated with a syntactic atom of category X within the c-structure model is to have the tree include a terminal node labeled *w* immediately dominated by a preterminal node labeled X. Following this practice, one might propose a terminal node labeled *I'll* immediately dominated by one preterminal node labeled D and another labeled I:



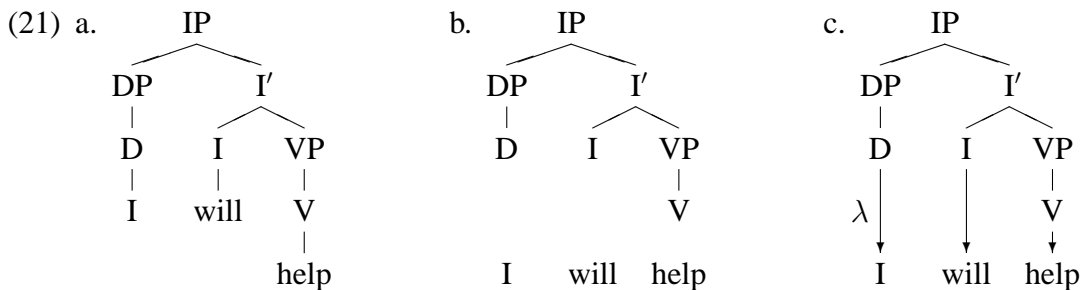
However, a structure like (20) is not a well-formed c-structure tree. In graph-theoretic terms, a c-structure is defined as a *directed tree*; such a structure has a *root* node from which every other node in the tree is reachable by *exactly one* directed path (Thulasiraman & Swamy 1992:106). There is a directed path from a ‘source’ node to a ‘goal’ node precisely when there is a sequence of nodes beginning with the source and ending with goal, such that each node other than the goal immediately dominates the next node in the sequence. In (20), IP is the only node from which there are directed paths leading to every other node; however, *I'll* may be reached from IP via two distinct paths, one going through DP and D, and another going through I' and I. Consequently, (20) is ill-formed, and some modification of the traditional c-structure model is required in order to allow for lexical sharing.

5.2 Freeing words from domination

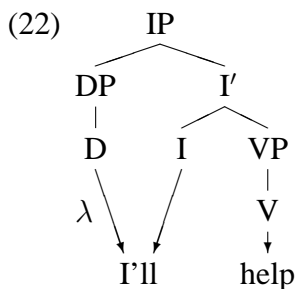
The single factor that prevents the traditional c-structure model from being able to represent lexical sharing is its dependence on immediate domination to convey the fact that a word corresponds to a syntactic atom of a particular category. To overcome this difficulty, I propose to remove words from the domination relation. To that end, I exploit LFG’s notion of a grammatical architecture based on “parallel structures flexibly related by correspondence mappings” (Bresnan 2001:43). Words will be removed from c-structure and set off in a structure of their own. The correspondence mapping that relates these two structures will thus be independent of the relation of domination that holds among nodes *within* c-structure. Consequently, the constraints on c-structure that undermined the first attempt at modeling lexical sharing will no longer be an issue.

The proposal may be visualized in three simple steps:

1. As a familiar conceptual starting point, begin with a traditional c-structure like (21a).
2. Sever the words from the tree, and arrange them in a separate, linearly ordered representation, called *l(lexical)-structure*, as in (21b). The terminal nodes of the new c-structure are the former preterminals. The new terminals represent syntactic atoms, which I henceforth describe as *atomic constituents*, to emphasize that these are elements of *constituent-structure* and formally on a par with complex constituents like those represented by DP, IP, I', etc.
3. Establish a correspondence mapping, λ , which relates each atomic constituent in c-structure to a word in l-structure, which one may call the atomic constituent’s lexical exponent. If a word w is the lexical exponent of an atomic constituent X, one may alternatively express that fact by saying that w *instantiates* X or by using functional notation: $\lambda(X) = w$. The correspondence mapping λ may be diagrammed with arrows, as in (21c).



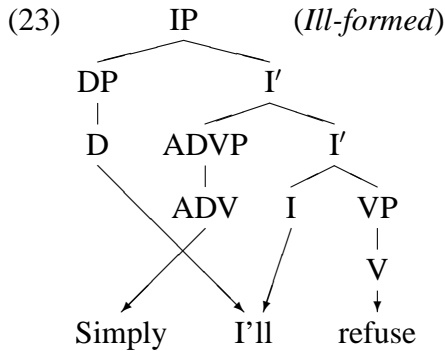
Since the correspondence mapping λ is distinct from domination, it is free to map D and I to distinct words, as in (21c), or to map them to *the same word*, as in (22):



In the latter case, multiple atomic constituents share a common lexical exponent, whence the name ‘lexical sharing.’

5.3 Homomorphic lexical integrity

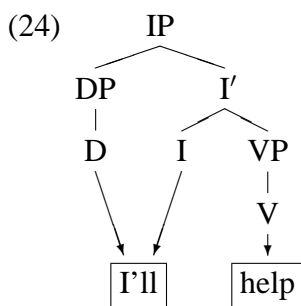
There is a very important constraint on the new correspondence mapping λ : It must be *order-preserving*. Notice that the terminals of the c-structure tree, which represent atomic constituents, are linearly ordered, as are the words in l-structure. For λ to be order-preserving, given two atomic constituents, X and Y, if X precedes Y, then $\lambda(Y)$ may not precede $\lambda(X)$. This condition may be easily appreciated in simple graphic terms: The arrows representing λ may *never* cross. Thus, for instance, the correspondence mapping between the c- and l-structures in (22) is *not* countenanced by the present theory:



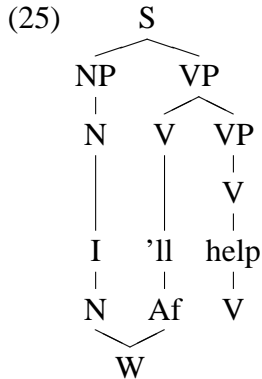
Given the requirement that λ be order-preserving, it follows that analyses framed within the lexical-sharing approach will have a property that I call *homomorphic lexical integrity*: Atomic constituents that share a lexical exponent will always be *adjacent*. The name of this property is derived from the fact that in the jargon of lattice theory, λ turns out to be a *homomorphism*. Notice, that in the illicit (23), D and I, which share *I'll*, are not next to each other in the linear ordering of atomic constituents; in contrast, in (22), which is allowed by the theory, the D and I that share *I'll* are side-by-side. The property described here constitutes a variant of Bresnan's (2001:92) notion of lexical integrity: Expressed in the terms of the present study, Bresnan's version would amount to saying that λ must be a one-to-one mapping, and therefore an *isomorphism*. This sort of isomorphic lexical integrity would of course render lexical sharing impossible; thus, I opt for the homomorphic variety, which makes an interestingly strong statement about the integrity of words without undermining lexical sharing.

5.4 The separation of syntax and morphology

Consider lexical sharing from the perspective of Lapointe's Generalized Lexical Hypothesis: "No syntactic rule can refer to elements of morphological structure" (1985:8). The crucial observation to make here is that the correspondence mapping λ maps atomic constituents to *unanalyzed words*:



Thus, under lexical sharing, the syntax is not privy to information about a word's internal composition. Contrast this with Autolexical Syntax (Sadock 1991: especially 52–53), which links X^0 -labeled nodes to *morphemes*, the latter being grouped into words in another tier of a multi-tiered model of grammar:

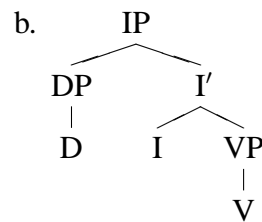


The fact that lexical sharing deals with words and not morphemes will prove advantageous in analyzing other clitic phenomena.

5.5 A rule formalism for lexical sharing

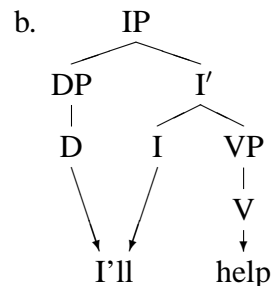
I provide a rule formalism for describing lexical sharing. To describe constituent structure, I use a normal, context-free phrase-structure grammar. Since these rules are concerned with c-structure, they contain only syntactic category symbols. The rules in (26a) admit the c-structure in (26b):

- (26) a. $IP \rightarrow DP I'$
 $I' \rightarrow I VP$
 $DP \rightarrow D$
 $VP \rightarrow V$



To determine l-structure and the correspondence mapping λ , I add *lexical-instantiation rules*, distinguished with a *leftward-pointing arrow*, as seen in (27a). The left-hand side of a lexical-instantiation rule is a word, and the right-hand side is a *sequence* of one or more syntactic categories. A lexical-instantiation rule of the form $w \leftarrow X_1 \cdots X_n$ allows the word w to appear in l-structure, provided λ maps n adjacent terminal nodes of the c-structure to w , and those terminal nodes are labeled, in order, X_1, \dots, X_n . Thus, the lexical-instantiation rules in (27a) associate the c-structure in (26b) with the l-structure *I'll help* via the correspondence mapping λ displayed in (27b):

- (27) a. $I'll \leftarrow D I$
 $help \leftarrow V$



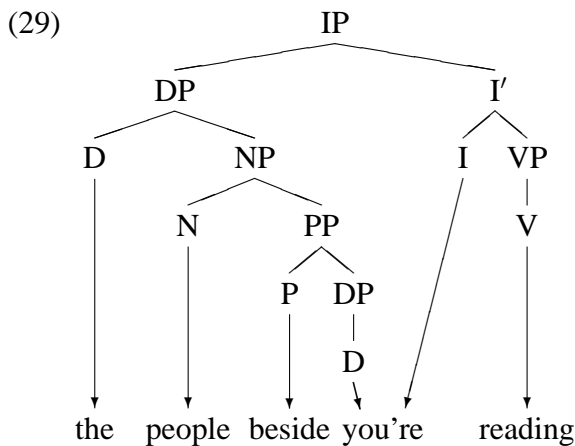
Thus, one may easily write lexical instantiation rules that implement the lexical-sharing analysis of nonsyllabic auxiliary contractions, as it has been described up to this point.

6 Incorporating lexical sharing into LFG

A further idiosyncrasy in the behavior of nonsyllabic auxiliary contractions becomes apparent in the following array of data:

- (28) a. **You're** reading. [ju:ɪ/jʊɪ/jɔɪ]
 b. The people beside **you're** reading. [ju:ɪ/*ju:ɪ/*jʊɪ/*jɔɪ]
 c. The people who helped **you're** kind. [ju:ɪ/*ju:ɪ/*jʊɪ/*jɔɪ]

When the auxiliary contraction is nonsyllabic, only (28a) is grammatical. However, if one gives *you're* [ju:ɪ/jʊɪ/jɔɪ] the lexical-instantiation rule *you're* ← D I, it ought to be possible to derive even the ill-formed sentences in (28), assigning (28b) the structure in in (29), for instance:



Some means is needed to constrain nonsyllabic auxiliary contractions to occur only in structures like (28a); a natural solution to this problem emerges if one combines lexical sharing with the over-arching grammatical theory of LFG.

6.1 Nonsyllabic auxiliary contractions and discourse functions

Faced with data like (28), Zwicky offers the following comment, framed in terms of his rule of *Auxiliary Reduction*:

The correct generalization is that Auxiliary Reduction applies to *will*, *have*, *am*, and *are* only after one of a small set of pronominal forms. . . , and then only when these NPs are immediately dominated by S. It may be significant that this S is always the one to which the auxiliary belongs (where a node X is said to *belong to* an S if that S is the lowest S dominating X). (1970:332)

Here S and NP may be interpreted as IP and DP, and one may readily infer that the DP in question is the *specifier* of the IP, i.e. [_{IP} DP. . .]. LFG proposes universal principles for associating grammatical functions with elements of c-structure, including this one, which is relevant to *functional categories* like IP: “Specifiers of functional categories are the grammaticalized *discourse functions*” (Bresnan 2001:102, emphasis added). The discourse functions recognized within LFG are *topic*, *focus*, and *subject*. Assume that in (28), *you're* [ju:ɪ/jʊɪ/jɔɪ] instantiates a D and an I; then

in the grammatical (28a), the D heads the subject of the clause headed by the I. The same cannot be said, however, of the ill-formed (28b) and (28c). Also, in (30), if *how've* [haʊv] instantiates a ADV and a C, then the former is the focus of the clause headed by the latter:¹⁰

(30) **How've** you been? [haʊv]

Thus, one may reformulate Zwicky generalization in the terms of this study as follows:

(31) **Functional identification hypothesis**

When a restrictive nonsyllabic auxiliary contraction is lexically cliticized to a host of category X, the result is an instance of lexical sharing instantiating two atomic constituents, an X along with an I or C, and the X is constrained to bear a discourse function, subject or focus, with respect to the I/C.

6.2 Lexical sharing and f-structure

To incorporate the functional identification hypothesis from (31) into this analysis, it is necessary to integrate lexical sharing into LFG. This may be accomplished with the following three steps:

1. Revise the correspondence mapping φ , which traditionally relates c-structure to f-structure, in order to have it map from *both c- and l-structure* to f-structure.
2. Provide a new metavariable \Downarrow , meaning $\varphi(\lambda(*))$.¹¹ Here * represents the c-structure node with which the annotation containing the metavariable is associated; it is convenient to paraphrase * by employing first-person pronouns. Then, $\lambda(*)$ represents ‘my lexical exponent,’ and $\varphi(\lambda(*))$ may consequently be read as ‘my lexical exponent’s f-structure.’
3. Furnish the right-hand sides of lexical instantiation rules with annotations, which will then be associated with the c-structure terminals instantiated by the word on the rule’s left-hand side.

With the foregoing changes in place, one may annotate the right-hand sides of the lexical-instantiation rules in (27a) above, yielding (32):

(32) *I'll* [aɪl/aɪ] ← D I
(\Downarrow PRED) = ‘PRO’ (\Downarrow TNS) = FUT
(\Downarrow SUBJ) = _c \Downarrow \Downarrow = \Downarrow
help ← V
(\Downarrow PRED) = ‘HELP(\Downarrow SUBJ)’
 \Downarrow = \Downarrow

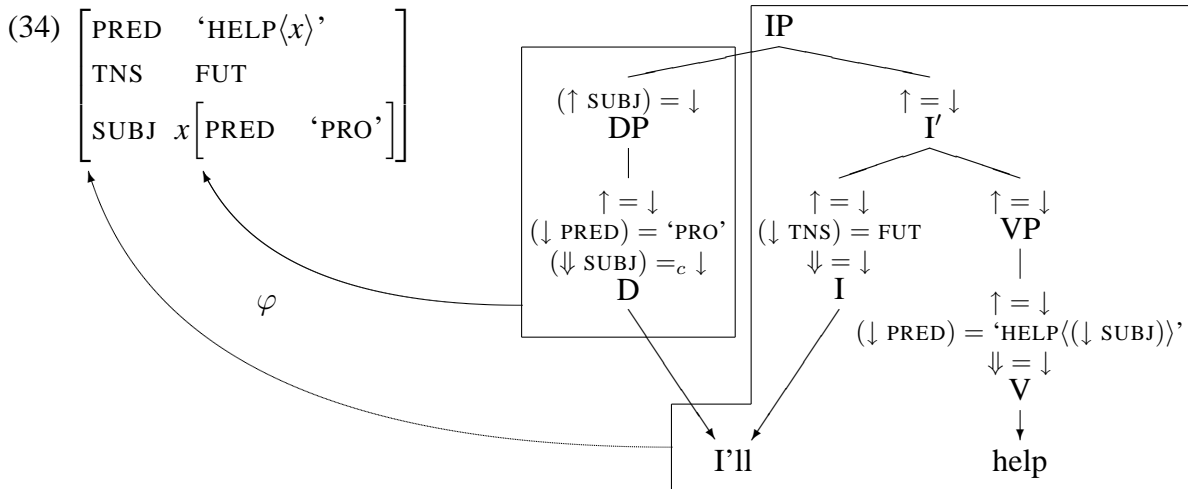
Universal principles of structure-function association (Bresnan 2001:102) provide the phrase-structure rules in (26a) above with the annotations seen in (33):

¹⁰I assume that inverted auxiliary verbs are in the head position of CP but that they constitute *extended heads* of their clauses, in the sense of Bresnan (2001:132).

¹¹Rather, I should say that this is a new usage for an old metavariable. The symbol \Downarrow was used in early LFG for describing long-distance dependencies (Kaplan & Bresnan 1995 [1982]:82–113). However, with the advent of LFG analyses of long-distance dependencies based on functional uncertainty (Kaplan & Zaenen 1995 [1989]), this older use of \Downarrow seems to have been abandoned. I therefore assume that \Downarrow is available for recycling.

- (33) $IP \rightarrow DP \quad I'$
 $(\uparrow \text{SUBJ}) = \downarrow \quad \uparrow = \downarrow$
 $I' \rightarrow I \quad VP$
 $\uparrow = \downarrow \quad \uparrow = \downarrow$
 $DP \rightarrow D$
 $\uparrow = \downarrow$
 $VP \rightarrow V$
 $\uparrow = \downarrow$

The rules in (32) and (33) then give rise to the c-, l-, and f-structures in (34):



Three points should be emphasized in connection with (34). First, the terminal nodes of the c-structure, D, I, and V, receive annotations both from the lexical-instantiation rules in (32) and from the phrase-structure rules in (33). The second matter concerns the correspondence mapping φ from c- and l-structure to f-structure; φ maps DP and D to the smaller f-structure labeled x , and all other elements of c- and l-structure to the larger f-structure. In particular, the annotation $\downarrow = \downarrow$ on I is responsible for equating the f-structures of I' and I; the $\downarrow = \downarrow$ on V similarly equates the f-structures of $help$ and V. The final point concerns the annotation $(\downarrow \text{SUBJ}) =_c \downarrow$ on D; the use of $'=_c'$ indicates that this is a *constraining equation*. Once the f-structure in (34) is created by the various *defining equations* expressed with $'=,'$ the constraining equation $(\downarrow \text{SUBJ}) =_c \downarrow$ checks that the f-structure associated with D is the SUBJ of the f-structure associated with I' , which also happens to be the f-structure for I.

One may now see how the functional identification hypothesis from (31) is implemented by the lexical-instantiation rule for I' [a1/a1] in (35):

- (35) $I' [a1/a1] \leftarrow D \quad I$ [repeated from (32)]
 $(\downarrow \text{PRED}) = \text{'PRO'}$ $(\downarrow \text{TNS}) = \text{FUT}$
 $(\downarrow \text{SUBJ}) =_c \downarrow \quad \downarrow = \downarrow$

Though associated with different atomic constituents, the annotations $(\downarrow \text{SUBJ}) =_c \downarrow$ and $\downarrow = \downarrow$ interact through the \downarrow metavariable, which refers to the f-structure of the shared lexical exponent I' . Together, these annotations ensure that the f-structure for D must turn out to bear the discourse function SUBJ with respect to the f-structure for I, just as required by (31). Similar comments hold for examples in which a restrictive nonsyllabic auxiliary contraction is lexically cliticized to a *wh*-word, as in (30). There, a minimally different lexical-instantiation rule along the lines of (36)

might be employed:

$$(36) \text{ how've [haʊv] } \leftarrow \text{ ADV} \quad \text{C}$$

$$\begin{array}{ll} (\downarrow \text{ PRED}) = \text{'HOW'} & (\downarrow \text{ TNS}) = \text{PRES} \\ (\downarrow \text{ FOCUS}) =_c \downarrow & (\downarrow \text{ ASP}) = \text{PERF} \\ & \downarrow = \downarrow \end{array}$$

6.3 More on coordination

There is yet another interesting idiosyncrasy associated with restrictive nonsyllabic auxiliary contractions. They do not attach to coordinated hosts:

$$(37) \text{ She and I'll help. [aɪ.l/*aɪl/*ɑl]}$$

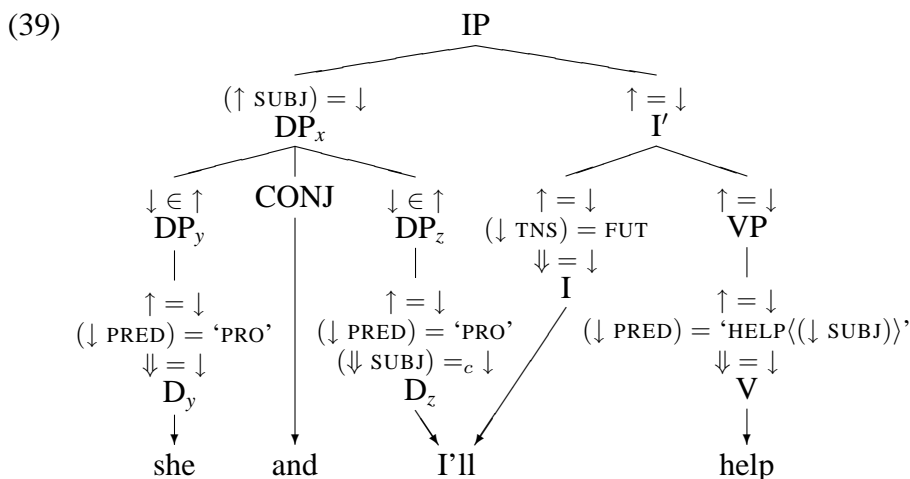
Felicitously, an explanation for this behavior is already at hand, thanks to the foregoing LFG implementation of the functional identification hypothesis. The analysis offered here closely parallels that presented by Sadler (1998), though the underlying technical details differ.

To model the coordinate structure in (37), one might follow Dalrymple and Kaplan (2000) in employing an annotated phrase-structure rule not unlike (38):¹²

$$(38) \text{ DP} \rightarrow \text{DP CONJ DP}$$

$$\begin{array}{ccc} & \downarrow \in \uparrow & \downarrow \in \uparrow \\ & \text{CONJ} & \text{DP} \end{array}$$

With this rule, one may construct the c- and l-structures in (39):



Using the defining equations in (39), one may build the f-structure in (40):

$$(40) \left[\begin{array}{ll} \text{PRED} & \text{'HELP}\langle x \rangle\text{' } \\ \text{TNS} & \text{FUT} \\ \text{SUBJ} & x \left\{ \begin{array}{l} y \left[\text{PRED} \text{'PRO'} \right] \\ z \left[\text{PRED} \text{'PRO'} \right] \end{array} \right\} \end{array} \right]$$

The elements of (40) labeled *x*, *y*, and *z* are the f-structure correlates of the c-structure nodes bearing the same labels as subscripts in (39). Now consider the constraining equation $(\downarrow \text{ SUBJ}) =_c \downarrow$ on

¹²Dalrymple and Kaplan's (2000) analysis of coordination introduces several nuances which I ignore here, since they have no effect on the present argument. See Dalrymple and Kaplan's article for more details.

D_z ; it requires that z be the SUBJ of the f-structure corresponding to *I'll*, which turns out to be the unlabeled f-structure in (40). Of course, the SUBJ of the unlabeled f-structure is x rather than z , so the constraining equation is not satisfied, the f-structure is deemed *inconsistent*, and the string in (37) is consequently ruled out.

7 Beyond nonsyllabic auxiliary contractions

I have focused on restrictive nonsyllabic auxiliary contractions because such forms seem to *require* a lexical-sharing analysis. However, nonrestrictive auxiliary contractions, whether syllabic or nonsyllabic, are no less *compatible* with lexical sharing. The foregoing discussion leads to conclusions of the following sort: There is a lexical process that attaches nonsyllabic 'll [l] (*will*) to a host, yielding a lexical-sharing structure; the host must be a pronoun or *wh*-word, the attachment of 'll [l] triggers morphophonological idiosyncrasies, and functional restrictions are involved. If one accepts that such conclusions are necessary, then it probably makes sense to analyze other auxiliary contractions with statements along these lines: There is a lexical process that attaches 's [z/s/əz] (*is* or *has*) to a host, yielding a lexical-sharing structure; the host may be anything, the attachment of 's [z/s/əz] triggers no morphophonological idiosyncrasies, and no functional restrictions are involved. The lack of morphophonological and functional intricacies in no way undermines a lexical-sharing analysis.

Beyond auxiliary contractions, lexical sharing is useful as a tool for treating various other clitic phenomena. In general, simple clitics, in Zwicky's (1977) terminology, may be candidates for such an analysis. Recall that simple clitics are characterized as unstressed versions of free words which become phonologically dependent on a neighbor; by positing a single word that instantiates a sequence of adjacent atomic constituents, lexical sharing accords well with this sort of phenomenon, as suggested by the foregoing analysis of English auxiliary contractions. In contrast to simple clitics, Zwicky also posits *special clitics*, which include forms with a special syntax that situates them in places where one would not expect to find corresponding non-clitics. This is illustrated by Romance clitic pronouns, as in these French examples:

- (41) a. Je **lui** prêterai un livre.
 I to.him will.lend a book
 'I will lend a book to him.'
- b. Je prêterai un livre à **Jean**.
 to
 'I will lend a book to Jean.'

Clitic phenomena of this sort are not compatible with a lexical-sharing analysis, since nonadjacent parts of c-structure seem to be involved; rather, an approach that posits distinct c-structures that map to similar f-structures, as proposed by Grimshaw (1982), is more appropriate for capturing the relationship between the sentences in (41). Zwicky posits a third class of clitic phenomena, containing what he calls *bound words*; these are forms that are "semantically associated with an entire constituent while being phonologically attached to one word of this constituent" (1977:6). Lexical sharing provides an interesting approach to members of this class, such as English possessive 's, which I discuss elsewhere (Wescoat 2002:30–36). Among Zwicky's bound words is a subtype that particularly illustrates the utility of lexical sharing, with its strict separation of syntax and morphology. The forms in question are known as *second-word clitics*.

One of the usual examples of a second-word clitic is the Latin *-que* ‘and.’ The place of *-que* in a coordinate structure is not in between conjuncts, but rather attached to the end of the final conjunct’s first word:

- (42) a. [NP boni pueri] [NP pulchrae**que** puellae]
good boys pretty-and girls
 ‘good boys and pretty girls’
 b. *boni pueri**que** pulchrae puellae (Sadock 1991:63)

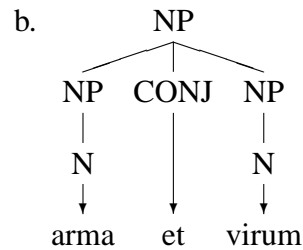
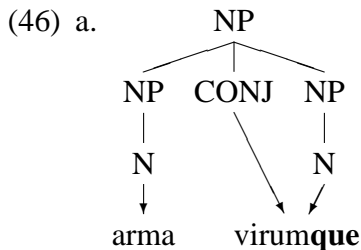
The term ‘second-word clitic’ arises from the assumption that *-que*, though phonologically bound, is a word unto itself, which occurs second-in-line within the conjunct. However, *-que* arguably forms a word with its host, since the rules of accent placement apply to the host and *-que* as a single unit:

- (43) a. vírum
 ‘the man (ACC)’
 b. virúm**que**
 ‘and the man’ (Zwicky 1977:30)

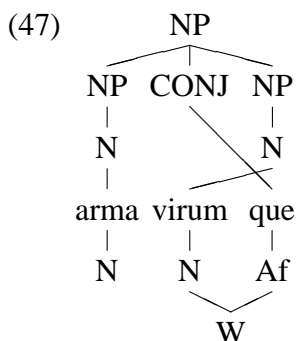
If one assumes a lexical-sharing treatment in which *virumque* instantiates two atomic constituents, a CONJ followed by an N, as specified by (44), one might give (45) the analysis in (46a), which shares the c-structure of (46b), formed with the free conjunction *et* ‘and’:

- (44) *virumque* ← CONJ N

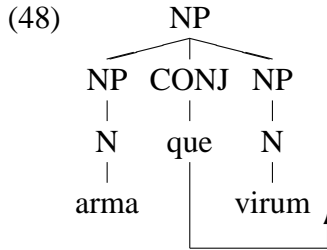
- (45) arma virum**que**
arms man-and
 ‘arms and the man’



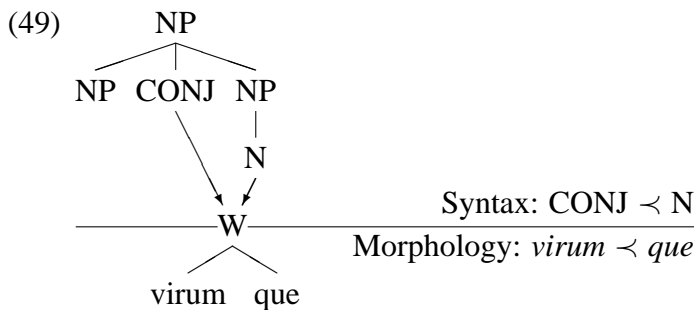
The lexical-sharing analysis of *-que* compares favorably with other approaches. Using the multi-tiered model of Autolexical Syntax, Sadock (1991:63–64) proposes a structure in which the association lines between the syntactic tier and the morphological tier are crossed, as in (47).



In order to capture the fact that *-que* is in ‘second’ position and not third, fourth, etc., Sadock proposes a theory of morphosyntactic mismatches in order to limit the degree of crossover between tiers. Halpern (1995) suggests including a procedure called *Prosodic Inversion* in the mapping from the syntax to prosodic structure; this would move the clitic to the opposite side of its host:



However, to ensure that the clitic is in ‘second’ position, Halpern stipulates that *Prosodic Inversion* allows movement over just one phonological word (1995:63). Whereas the foregoing theories require special measures to prevent *-que* from being placed too far to the right, the same effect follows without stipulation from the lexical-sharing analysis. The only word in the final conjunct that is able to act as host for *-que* is the leftmost one; this allows CONJ and the atomic constituent to its immediate right to share the word bearing *-que* as their lexical exponent, and if this state of affairs does not obtain, a violation of homomorphic lexical integrity will result. Recall that the correspondence mapping λ relates atomic constituents to unanalyzed words; thus, the role of the syntax in situating *-que* is strictly limited to placing it somewhere inside of the first word of the conjunct. Beyond that, the position of *-que* within the word is independently determined by the morphology; since *-que* is a suffix, it will occur at the word’s right edge. This distribution of labor between the syntax and the semantics is schematized in (49):



In this manner, *-que* winds up at the right edge of the first word of the final conjunct, and this accounts for the perception that *-que* is in ‘second-word’ position. Thus, lexical sharing acquits itself rather well in the analysis of so-called second-word clitics.

Elsewhere I offer some suggestions about how lexical sharing might fit into a more comprehensive theory of clitics (Wescoat 2002:57–64). Essentially, exploiting the sort of capabilities illustrated in the discussion of Latin *-que*, I propose lexical sharing as a candidate to take over the role played by *Prosodic Inversion* within the overarching theory devised by Halpern (1995).

8 Conclusion

In sum, lexical sharing affords an analysis that successfully captures the characteristics of restrictive nonsyllabic auxiliary contractions as lexical clitics. Moreover, it shows promise for explaining other cliticization phenomena, such as the pseudo-movement of ‘second-word’ clitics. I hasten to

point out, however, that lexical sharing is not limited in its application to the analysis of clitics. For instance, in addition to the topics mentioned in the foregoing discussion, I have employed lexical sharing in the treatment of English pronominal determiners (e.g. *These are good*), Romance preposition+determiner compounds (e.g. French *au*), and Hindi noun incorporation (Wescoat 2002). Lexical sharing has also been applied to Korean copular constructions by Kim et al. (2004) and by Kim and Sells (2005). Lexical sharing affords analyses of these phenomena in a tractable and straightforward formalism. For a grammar composed of phrase-structure rules and lexical-instantiation rules of the sort outlined here, the problem of recognizing whether or not a string is generated by the grammar may be solved in time proportional to the cube of the string's length (Wescoat 2002). This compares quite favorably with many of the mechanisms employed in modern theories of grammar, so by the objective measure of computability, lexical sharing is a relatively simple grammatical tool. Moreover, lexical sharing integrates nicely with LFG, in a manner that allows one to express the functional constraints at work in restrictive nonsyllabic auxiliary contractions without sacrificing the clarity of the rule formalism. The availability of such simple analyses suggests that lexical sharing may prove to be a useful component in the LFG explanation of cliticization.

References

- Baltin, Mark R., & Anthony S. Kroch, eds. (1989) *Alternative conceptions of phrase structure*. University of Chicago Press.
- Bender, Emily, & Ivan A. Sag (2001) 'Incorporating contracted auxiliaries in English'. In Ronnie Cann, Claire Grover, & Philip Miller, eds., *Grammatical interfaces in HPSG*. Stanford: CSLI Publications, 1–15.
- Bloch, Bernard (1941) 'Phonemic overlapping'. *American speech* 16, 278–284.
- Booij, Geert, & Jerzy Rubach (1987) 'Postcyclic versus postlexical rules in Lexical Phonology'. *Linguistic inquiry* 18, 1–44.
- Bresnan, Joan W., ed. (1982) *The mental representation of grammatical relations*. MIT Press series on cognitive theory and mental representation. Cambridge, MA: MIT Press.
- (2001) *Lexical-functional syntax*. Oxford: Blackwell.
- Dalrymple, Mary, & Ronald M. Kaplan (2000) 'Feature indeterminacy and feature resolution'. *Language* 76, 759–798.
- Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III, & Annie Zaenen, eds. (1995) *Formal issues in Lexical-Functional Grammar*. Stanford: CSLI Publications.
- Di Sciullo, Anna-Maria, & Edwin Williams (1987) *On the definition of word*. Cambridge, MA: MIT Press.
- Grimshaw, Jane (1982) 'On the lexical representation of Romance reflexive clitics'. In Bresnan (1982:87–148).
- Halpern, Aaron L. (1995) *On the placement and morphology of clitics*. Stanford: CSLI Publications.
- Kaplan, Ronald M., & Joan W. Bresnan (1995 [1982]) 'Lexical-Functional Grammar: A formal system for grammatical representation'. In Dalrymple et al. (1995:29–130). [First appeared in Bresnan (1982:173–281).]
- Kaplan, Ronald M., & Annie Zaenen (1995 [1989]) 'Long-distance dependencies, constituent structure, and functional uncertainty'. In Dalrymple et al. (1995:137–165). [First appeared in

- Baltin and Kroch (1989:17–42).]
- Kim, Jong-Bok, & Peter Sells (2005) *Copy constructions and their interaction with the copula in Korean*. [Paper presented at HPSG 2005, Lisbon]
- Kim, Jong-Bok, Peter Sells, & Michael T. Wescoat (2004) ‘Korean copular constructions: A lexical sharing approach’. In M. Endo Hudson, Sun-Ah Jun, & Peter Sells, eds., *Japanese/Korean linguistics 13*. Stanford, CA: CSLI Publications.
- Lapointe, Steven G. (1985) *A theory of grammatical agreement*. New York: Garland. [Ph.D. dissertation, University of Massachusetts, Amherst, 1980.]
- Sadler, Louisa (1998) ‘English auxiliaries as tense inflections’. *Essex research reports in linguistics* 24, 1–16.
- Sadock, Jerrold M. (1991) *Autolexical syntax: A theory of parallel grammatical representations*. Studies in contemporary linguistics. Chicago: University of Chicago Press.
- Spencer, Andrew (1991) *Morphological theory*. Oxford: Blackwell.
- Sweet, Henry (1890) *A primer of spoken English*. Oxford: Clarendon.
- Thulasiraman, K., & M. N. S. Swamy (1992) *Graphs: Theory and algorithms*. New York: John Wiley & Sons.
- Wescoat, Michael T. (2002) *On lexical sharing*. Ph.D. dissertation, Stanford University.
- Zwicky, Arnold M. (1970) ‘Auxiliary reduction in English’. *Linguistic inquiry* 1, 323–336.
(1977) *On clitics*. Bloomington: Indiana University Linguistics Club.
- Zwicky, Arnold M., & Geoffrey K. Pullum (1983) ‘Cliticization vs. inflection: English *n’t*’. *Language* 59, 502–513.

ON PIVOTS AND SUBJECTS IN GEORGIAN
Thomas R. Wier

University of Chicago

Proceedings of the LFG05 Conference

University of Bergen

Miriam Butt and Tracy Holloway King (Editors)

2005

CSLI Publications

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

Abstract.

Properties of cross-clausal coreference have long been taken as criteria of grammatical relations (e.g. Chomsky 1965 and before). In this paper I will cast doubt on that proposition by identifying one language, Georgian, in which pivot phenomena are sensitive to argument structural information and not case or position or some other outward sign of grammatical status. These pivot facts will clearly distinguish two approaches to grammatical relations in LFG, Manning (1996) and Falk (forthcoming). In closing, I will speculate about the formal origin of the typology of this pivot.

1. Introduction

In this paper¹ I will be looking at questions of how different theories formalize properties of morphosyntax that traditionally go under the label of grammatical relations, specifically looking at evidence from pivot data in Georgian that would help decide whether all properties that have traditionally been assigned to ‘subjects’ can be reducible to a single theoretical construct, or may alternatively arise from completely separate processes. These issues are important, because grammatical relations having primitive status or not is one of the primary features that distinguish multistratal theories of grammar such as Minimalism and its predecessors. Thus, if evidence can be brought to bear that grammatical relations are not primitives, or are not primitives in the standard way as currently conceived, then that might cast doubt on the monostratal theories or would require rethinking their premises.

To narrow down this subject, I will be focusing on two different lexicalist attempts at accounting for the properties of pivots: that of Manning (1996) and that of Falk (1998) and Falk (to appear). Both of these theories assume monostratality, but account for the behavior of pivots in different ways, which we will shortly see. After reviewing the claims made by these theories, I will be introducing new facts from Georgian which I will claim crucially distinguish between the two, and discuss their implications for syntactic theory in general.

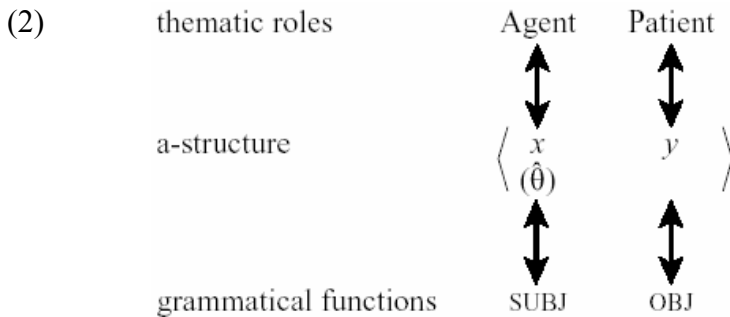
2. Manning 1996.

Because monostratal theories cannot appeal to a kind of isomorphism between syntactic structure and semantic or thematic structure, they must have recourse to something else. LFG, HPSG and some other recent theories such as that of Jackendoff assume that the properties that have been traditionally and rather theory-neutrally associated with subjects, such as control of reflexivization or control of pivots are localized in an SUBJ function, however they choose to formalize this. In LFG, these are typically captured by an alignment between two hierarchies: the *thematic hierarchy* of agents, beneficiaries, etc. and a separate *functional hierarchy* of subcategorized arguments. You can see this in (1a) and (1b).

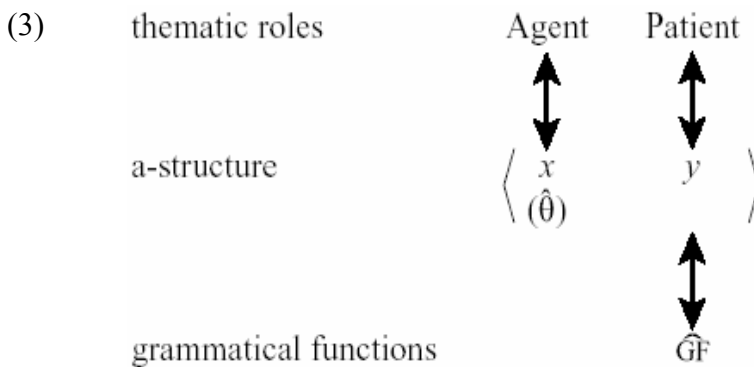
¹ I would like to thank Dr. Gela Tevzadze for his unstinting and patient help in responding to linguists’ obscure questions. I would also like to thank Ilya Yakubovich and Adam Cooper for their comments and critiques of this work. All errors are of course my own.

- (1) a. Thematic hierarchy: Agent > Beneficiary > Experiencer > Instrument > Patient, etc.
 b. Functional hierarchy: SUBJ > OBJ > OBJ_θ > OBL_θ

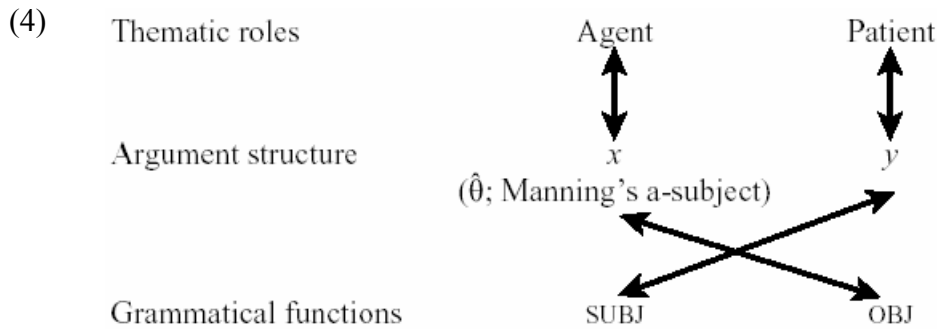
In general, these two hierarchies will always align, so that the most prominent member of the f-structure hierarchy, the SUBJ, will also be the most prominent member of the thematic hierarchy, the agent, as you can see in the diagram in (2).



And as you can see in (3), any deviation between these two must be the direct result of some lexical process that forces it, so that, for example, in the case of a passive sentence, the subject aligns with the patient because of some specific morphology motivating that.



Importantly, for Manning not all phenomena will make reference to both hierarchies. Some properties of the sentence, such as the subcategorization requirements, relativization, topicalization and pivots, will make reference only to the functional hierarchy, while binding, control and the imperative addressee will make reference only to a-structure. Manning is particularly concerned with ergativity, which he explains as a mismatch, or misalignment, between the two hierarchies, as you can see in (4), where the a-structure element assigned the agent thematic role is actually syntactically an OBJECT, while the a-structure patient is treated as the syntactic SUBJECT. This is important, because it predicts that syntactically ergative languages should separate out these two sets of phenomena and they should never mix.



What is more important for our purposes is that for any given language, whether it be a nominative-accusative one like English or a syntactically ergative one like Dyrbal or Samoan, there can only be one pivot for any interclausal patterns of coreference. As we will see, this will clearly distinguish Manning's view from that of Falk, which I will now proceed to describe, since there do exist languages with multiple pivots.

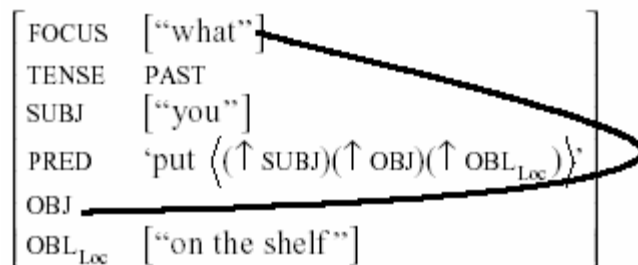
3. Falk

For Falk, in contrast, pivots are not assigned by any kind of mismatch between the separate modules, but rather arise because pivots constitute a separate entity in f-structure. In LFG, certain functions in f-structure are obligatory because of subcategorization requirements, such as SUBJs or OBJs if any. Other functions in f-structure, such as discourse notions like TOPIC and FOCUS or ADJuncts are not subcategorized, but must be licensed by the Extended Coherence² condition which functions effectively in some respects like the Projection Principle in generativist terms. Falk holds that there is another PIV function which, like TOPIC and FOCUS is an overlay function which must be anaphorically bound to some subcategorized argument.

So, for example, in a wh-construction like in (5) "What did you put on the shelf?", the wh-word is in focus, and is bound to the OBJ function in the f-structure matrix. What concerns us here is not the surface position of the wh-word, but the fact that the formal existence of the FOCUS function depends on something already licensed by the syntax.

(5) a. What did you put on the shelf?

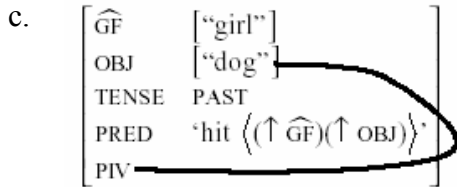
b.



² Defined thus: 'All functions in an f-structure must be incorporated into the semantics. Argument functions are subject to the Coherence condition. Overlay functions must be identified with arguments or adjuncts. Adjuncts must be in f-structures containing preds.' (Falk 2001: 64)

(6) Samoan (Falk p. 60)

- a. $(\uparrow\text{OBJ}) \Rightarrow (\uparrow\text{PIV}) = (\uparrow\text{OBJ})$
 b. Sā fasi le maile e le teine.
 PAST hit ART dog ERG ART girl
 ‘The girl hit the dog.’



As you can see in (6a), Falk believes that the PIV function works similarly. In the case of this Samoan example, a lexical default rule simply identifies the PIV function with this particular function. Since not all verbs have OBJects, any intransitive coordinated with the transitive verb in (6b) will naturally assign the same entity to the pivot function, or otherwise there would be a feature clash and the sentence would be ruled ungrammatical. It is also important that this is defined lexically, because it predicts that languages may vary on precisely this point of how and when they assign pivots. Most languages may assign the pivot to the SUBJ function of any verb, transitive or intransitive, but other languages differ in this respect.

Falk formalizes this as the *Pivot Condition* in (7). (7) is translated roughly as follows: “A path inward through f-structure into another predicate-argument domain [that is, into a subordinate clause of some kind] or sideways into a coordinate f-structure must terminate in the function PIV.”

(7) THE PIVOT CONDITION:

In a functional designation of the form $(\uparrow\dots\alpha\dots\beta)$ where:

$$\alpha \quad \text{or } (\varphi(<^* \dots \beta)) \text{ or } (\varphi(>^* \dots \beta)), \text{ if } \beta \neq \emptyset, \beta = \text{PIV}$$

(\rightarrow PRED ARG1)

4. Georgian

Now that we have looked at the formal debate within LFG circles, I would like to bring in some new Georgian data that may shed light on the debate. So, as you can see in the data in Table 1 and in (8), Georgian has a complicated split system featuring splits not just in tense but also among different classes of verbs. There are three series which indicate combinations of tense, aspect and/or modality: the present/future and aorist being more or less straightforward present and past tenses, respectively, while the perfect series is rather a kind of modal evidential form which broadly implicates but not necessarily entails an event having occurred in the past³. Importantly, the two classes of intransitive verbs, the second and the third, behave differently with respect to case and verbal agreement. The second conjugation consists of unaccusative verbs, most of whose

³ The semantics of the perfect series are complex; see Wier (ms.) for more information on the nexus of semantic and pragmatic phenomena it evinces.

members take patient subjects, while the third conjugation consists of unergative verbs most of whose members take agent subjects.

TABLE 1. Case assignment across verbal conjugations and tense-aspect series

| Series / Conj. | 1 st Conj. | 2 nd Conj. | 3 rd Conj. | 4 th Conj. |
|---------------------------|--|-----------------------|-----------------------|---|
| <i>Present/Future</i> | Nom _{AG} – Dat _{PAT} – Dat _{GOAL} | Nom _{PAT} | Nom _{AG} | Dat _{EXP} – Nom _{PAT} |
| <i>Aorist</i> | Erg _{AG} -Nom _{PAT} – Dat _{GOAL} | Nom _{PAT} | Erg _{AG} | Dat _{EXP} – Nom _{PAT} |
| <i>Perfect Evidential</i> | Dat _{AG} – Nom _{PAT} - - <i>tvis</i> _{GOAL} | Nom _{PAT} | Dat _{AG} | Dat _{EXP} – Nom _{PAT} |

- (8) a. Ivane da Ketevan-i c'eril-s Mariam-s da-u-mal-av-en
 John.NOM and Ketevan-NOM letter-DAT Mary-DAT PVB-3DAT-hide-TH-3PL
 'John and Ketevan are hiding the letters from Mary.'
- b. Ivane-m da Ketevan-ma c'eril-i Mariam-s da-u-mal-es
 John-ERG and Ketevan-ERG letter-NOM Mary-DAT PVB-3DAT-hide-3PLAOR
 'John and Ketevan hid the letters from Mary.'
- c. Ivane-s da Ketevan-s c'eril-i Mariam-is-a-tvis da-u-mal-eb-i-a
 John-DAT and Ketevan-DAT letter- NOM Mary-GEN-EXT-for PVB-3DAT-hide-TH-3NOM-3SG/PL
 'John has apparently hidden the letters from Mary.'

The interesting fact is that although such classes in Georgian have always been challenging morphologically, they have always appeared to be more straightforward syntactically, at least with respect to the topic under discussion today, as in (9) and (10). Thus, when we coordinate a transitive verb of the first conjugation with an unergative verb as in (9a, c) and (10a, c) or an unaccusative verb as in (9b, d) or (10b, d), in either the present or the aorist series, we get in both cases a fairly unremarkable nominative-accusative pivot where the A argument and the S argument obligatorily corefer. Below each morphological gloss is a case-frame. If coreference tracks case as in, say, Yidiny (Comrie 1981), then the pattern of data in examples (9) and (10) should differ, in that (9) should have a nominative-accusative pivot, while (10) should have an ergative-absolutive pivot, given that the intransitive verbs without overt argument realization are in the imperfective, which patterns along with the Present/Future series, and not the Aorist, in terms of case-assignment.

(9) Present/Future series:

- a. Ivane_i Mariams_j xed-av-s da pro_{i*} t'ir-i-s. (Intr. = 3rd Conj.)
 John-NOM Mary-DAT see-TH-3SGS and cry-TH-3SGS
NOM DAT NOM

'John sees Mary, and (John/*Mary) is crying.' (S_a/A)

- b. Ivane_i Mariams_j xed-av-s da pro_{i*} ga-c'itl-d-eb-a. (Intr.=2nd Conj.)
 John-NOM Mary-DAT see-TH-3SGS and PRVB-red-INGR-TH-3SGS
NOM DAT NOM

'John sees Mary, and (John/*Mary) blushes.' (S_o/A)

- c. Ivane_i Mariams xedavs rodesac pro_{i*} t'ir-i-s
 John-NOM Mary-DAT see-TH-3SGS when cry-TH-3SGS
NOM DAT NOM
 'John sees Mary, when (John/*Mary) is crying.' (S_a/A)

d. Ivane_i Mariams xedavs rodesac pro_i*_j ga-c'itl-d-eb-a
 John-NOM Mary-DAT see-TH-3SGS when PRVB-red-INGR-TH-3SGS
NOM DAT NOM

‘John sees Mary, when (John/*Mary) blushes.’ (S_o/A)

(10) Aorist series:

a. Maia-m_i Eduard-i_j nax-a da pro_i*_j t'ir-od-a
 Maia-ERG Eduard-NOM see-TH-3SGS and cry-IMPf-3SGS
ERG NOM NOM

‘Maia saw Eduard, and (Maia/*Eduard) cried.’ (S_a/A)

b. Maia-m_i Eduard-i_j nax-a da pro_i*_j ga-c'itl-d-eb-od-a.
 Maia-ERG Eduard-NOM see-TH-3SGS and PRVB-red-INGR-TH-IMPf-3SGS
ERG NOM NOM

‘Maia saw Eduard, and (Maia/*Eduard) was blushing.’ (S_o/A)

c. Maia-m_i Eduard-i_j nax-a rodesac pro_i*_j t'ir-od-a
 Maia-ERG Eduard-NOM see-TH-3SGS when cry-IMPf-3SGS
ERG NOM NOM

‘Maia saw Eduard when (Maia/*Eduard) was crying.’ (S_a/A)

d. Maia-m_i Eduard-i_j nax-a rodesac pro_i*_j ga-c'itl-d-eb-od-a.
 Maia-ERG Eduard-NOM see-TH-3SGS when PVB-red-INGR-TH-IMPf-3SGS
ERG NOM NOM

‘Maia saw Eduard when (Maia/*Eduard) was blushing.’ (S_o/A)

In fact, as (10) shows, coreference cannot directly track case-marking, given that the nonovert arguments of the intransitive verbs should receive nominative case, and yet must corefer with the previous verb’s subject, which takes ergative case. So, it would seem based on these facts that coreference must track the grammatical relations whatever case those relations in fact receive. However, what has not been realized is that the perfect series appears to behave differently, and most importantly for us, it falls right along the split that we see in the case and agreement morphology. That is, on the one hand, if we coordinate a transitive verb with an unergative intransitive verb in the perfect series, marked with dative case, we get the same obligatory nominative-accusative pivot, as in (11a) and (11c).

(11) Perfect series:

a. Tamaz-s_i Zurab-i_j u-nax-i-a da pro_i*_j u-t'ir-i-a.
 Tamaz-DAT Zurab-NOM 3IO-see-TH-3SgO and 3IO-cry-th-3Sg
DAT NOM DAT

‘Tamaz has (apparently) seen Zurab, and (Tamaz/*Zurab) has cried.’ (S_a/A)

b. Tamaz-*s_i* Zurab-*j* u-nax-i-a da pro_{*ij*} ga-c'itl-eb-ul-a.
 Tamaz-DAT Zurab-NOM 3IO-see-TH-3SGO and PRVB-red-TH-PF-3SGS
DAT NOM NOM

‘Tamaz has (apparently) seen Zurab, and (Tamaz/Zurab) has blushed.’ !!! (S_o/A or O)

c. Tamaz-*s_i* Zurab-*j* u-nax-i-a rodesac pro_{*ij*} u-t'ir-i-a.
 Tamaz-DAT Zurab-NOM 3IO-see-TH-3SGO when 3IO-cry-th-3Sg
DAT NOM DAT

‘Tamaz has (apparently) seen Zurab, when (Tamaz/*Zurab) has cried.’ (S_a/A)

d. Tamaz-*s_i* Zurab-*j* u-nax-i-a rodesac pro_{*ij*} ga-c'itl-eb-ul-a.
 Tamaz-DAT Zurab-NOM 3IO-see-TH-3SGO when PRVB-red-TH-PF-3SGS
DAT NOM NOM

‘Tamaz has (apparently) seen Zurab, when (Tamaz/Zurab) has blushed.’ !!! (S_o/A or O)

On the other hand, if we coordinate a transitive verb with an unaccusative verb, marked with nominative case as they are in all series, we get an optional fluid-S pivot, where the intransitive’s argument can corefer to either the agent or the patient of the transitive verb, as in (11b) and (11d). I have tested these with a number of other unaccusative verbs, so I don’t think these are pragmatic accidents.

The formal analysis is as in (12) below. This is a rather complicated lexical default rule, but it basically means in the special case that an unaccusative verb is coordinated with a transitive in the perfect series, the pivot may optionally be assigned to either argument of the transitive. As a kind of elsewhere condition, we would always expect a nominative/accusative pivot, that should work itself out as a natural consequence of the harmonic alignment of the two hierarchies I mentioned in (1a) and (1b).

(12)

| | | | | |
|--------|-------------------------------|------|---------------------|--------------------|
| f-str: | (↑TENSE = PERF) ^ ∃(↑OBJ) ^ (| α | or (φ(<* ... SUBJ)) | or (φ(*> ... SUBJ) |
| | (→ PRED ARG1 SUBJ) | | ↑ | ↑ |
| | | ↓ | ↓ | ↓ |
| a-str: | | [−r] | [−r] | [−r] |

⇒ (↑PIV) = (↑SUBJ) ∨ (↑OBJ)

5. Constructions and Pivots.

Languages with multiple pivots are not new; Dixon (1994), e.g., notes them for Chukchee (from Nedjalkov 1979), Greenlandic Eskimo (Woodbury 1975), Yupik (Payne 1982), Tongan, and Yidin^y (Comrie 1981)⁴. What distinguishes these languages from the Georgian facts is that to my knowledge for all these languages, the pivot patterns along with some property of the clausal structure, such as having an S/A pivot in matrix clauses but an S/O pivot in subordinate clauses, as in some languages of the Pacific, or closely tracks morphological case marking, as in Yidin^y.

⁴ It may also have been the case for Hurrian (Ilya Yakubovich, p.c.)

The Georgian facts, however, seem not to be dependent on such higher levels of syntactic organization. The same pivot pattern occurs in both matrix and subordinate clause types (as in (9)-(11) above), and the pivot clearly does not follow case marking. If, say, it followed nominative case, we would expect an S/A pivot in the first three conjugations in the present (and not in the dative constructions of the fourth conjugation), a split-S pivot in the aorist, and a different split-S pivot in the perfect. What we actually have is an S/A pivot in two tenses, but a fluid-S in the perfect.

So, why, or whence, does this strange configuration arise? One possibility is that we might ultimately find a diachronic explanation in the pivot properties of Kartvelian languages. This explanation, however, is fraught with difficulties, as there are obviously no longer any speakers of proto-Kartvelian around to make inquiries about which pivots are possible in which contexts, and I am unaware of any attempts to reconstruct such by the Comparative Method. It seems then that we are stuck with a lexical analysis, where specifications for particular pivots can even become lexicalized properties of everyday lexical items, in this case tracking argument structural properties of those items. This proposal is not entirely new. In fact, Falk (ms.) proposes precisely this kind of lexical analysis for Tagalog “voice” morphology, which in his analysis contribute PIV features linked to particular grammatical functions (q.v.).

But if pivots may be lexically determined like this, as they seem to need to be, what then constrains the typology of pivots? Can just any syntactic⁵ pivot pattern occur? If so, why is it the case that so large a number of languages have pivots that target the SUBJ function, producing languages that may be purely morphologically ergative, say, but syntactically accusative? In answer, it may be pointed out that there are many properties of human languages that do not seem to be readily explainable in terms of broader generalizations, “syntactic nuts” in Culicover (1999)’s words. All languages have idioms, for example, which are not semantically compositional, but it is not as frequently recognized or admitted that many languages have syntactically or morphologically idiosyncratic idiomatic constructions as well. In German, e.g., one can say colloquially/dialectally “Butter bei die Fische” (literally “butter by the fish”) meaning something like “okay, now let’s carry out our proposal”, but, as is well known, in the productive parts of German grammar, the preposition *bei* obligatorily takes the dative case in all contexts, which here would imply *den Fischen*, not accusative *die Fische* as in the idiom. These simple exceptions to the rule must be learned at some level, as they cannot be predicted based on anything else speakers may have learned in their environment, and they certainly are not innate.

It may be the case that the Georgian pivot properties are similar, in that speakers have latched onto the very clear sensitivity of Georgian grammar to the underlying argument structure of lexical items (which surfaces most directly in the aorist series), and contrary to the rest of the grammar isolated this one tense for special treatment⁶. Culicover (1999, p. 194-232) has some extended discussion that may provide some insight. He is here broadly concerned with a range of extraction phenomena which do not clearly align with the typological generalizations about extraction hierarchies provided by Keenan and Comrie (1977). He cites Chung and Seiter (1980), for example, on the ergative Polynesian language Rennellese which, contrary to the Keenan-Comrie

⁵ I am excluding pragmatically determined pivots for purposes of this discussion.

⁶ For possible semantic motivations, see Aronson 1990, and Wier Ms.

hierarchy, prefers to have pronouns in long-distance dependencies where you would expect a gap in subject position, but has gaps lower down on the hierarchy. His analysis is nuanced, and a full discussion is well beyond the scope of this paper. But in explaining the Rennellese extraction exceptionality, he crucially makes reference to Hawkins (1994)'s Complexity Metric, in which the constructional correspondence rules between syntactic and conceptual structure can be ranked in terms of the number of constituents required to formulate those very correspondence rules. The significance of this for Culicover is great, in that the Hawkins Complexity Metric provides an objective basis for integrating the formal properties of a grammatical theory such as his into our understanding of language acquisition. As Culicover says, "[I]ndividuating structures has important consequences. For CAL [the Conservative-Attentive Learner], the acquisition of wh-movement for subjects will not entail the possibility of wh-movement of direct objects in the absence of positive evidence about the latter" (205).

Importantly for our understanding of the Georgian pivot properties, since as long as the formal rules can be discretely defined, they can also be measured and ranked in complexity. In this respect, the reason more languages do not have lexically specified pivots such as Georgian, or Tagalog in Falk's analysis, is that the very pivot rules that we deduce for them are so complex that the Conservative Attentive Learner will not readily pick rules like (12) up unless there is positive evidence to the contrary. It may not be the case that all exceptions to linguistic generalizations can be explained away like this, but it may go far in that direction, and bring about a more concrete understanding of human language as a mental and social phenomenon.

Works Cited

- Aronson, H. 1990. *Georgian: A Reading Grammar*. Columbus, OH: Slavica.
- Chomsky, Noam. 1965. *Syntactic Structures*. Mouton.
- Chung, S., and Seiter, W. 1980. 'The History of Raising and Relativization in Polynesian'. *Language* 56:622-38.
- Comrie, B.. 1981. *Language Universals and Linguistic Typology*. 1st Ed.
- Culicover, P.. 1999. *Syntactic Nuts*. Oxford: Oxford University Press.
- Falk, Yehuda. 1998. 'On Pivots and Subjects.' LFG Conference 1998.
- Falk, Yehuda. 2001. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Stanford, CA: CSLI Publications.
- Falk, Yehuda. Ms. *Explaining Subjecthood*.
<<http://pluto.mscc.huji.ac.il/~msyfalk/SubjectBook.pdf>>
- Hawkins, J. A. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.
- Keenan, E., and Comrie, B. 1977. 'Noun Phrase Accessibility and Universal Grammar'. *Linguistic Inquiry*, 8:63-99.
- Manning, C. 1996. *Ergativity: Argument Structure and Grammatical Relations*. Stanford, CA: CSLI Publications.
- Nedjalkov, V. 1979. 'Degrees of ergativity in Chukchee. Ergativity: Towards a theory of grammatical relations', ed. by Frans Plank, 241-262. New York/London: Academic Press
- Payne, T. 1982. 'Role and Reference related subject properties and ergativity inYup'ik Eskimo and Tagalog'. *Studies in Language* 6.75-106.
- Wier, T. Ms. 'The Semantics of the Georgian Perfect'. Contact: trwier@uchicago.edu.
- Woodbury, A. C. 1975. *Ergativity of Grammatical Processes: A Study of Greenlandic Eskimo*. M. A. thesis. University of Chicago.

Contact information:

Thomas Wier
Dept. of Linguistics
University of Chicago
1010 E. 59th St.
Chicago, IL 60637

Email: <trwier@uchicago.edu>