Proceedings of the LFG 02 Conference

National Technical University of Athens

Editors: Miriam Butt and Tracy Holloway King

2002

CSLI Publications

ISSN 1098-6782

The Proceedings of the LFG'02 Conference

National Technical University of Athens

Editors: Miriam Butt and Tracy Holloway King ☐
2002 CSLI Publications
ISSN 1098-6782

Editors' Note

The program committee for LFG'02 were Rachel Nordlinger and Jonas Kuhn. We would like to thank them 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. This is especially true for Yanis Maistros, Stella Markantonatou, Marina Vassiliou, and the local organizing committee, who put together a superb conference!

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 via their e-mail addresses.

This pdf file contains all of the papers submitted to the LFG02 proceedings. Use the view bookmarks option to navigate between papers.

Table of Contents

The Syntax of Preverbal Particles and Adjunction in Irish

Ash Asudeh
Stanford University

Proceedings of the LFG02 Conference
National Technical University of Athens, Athens
Miriam Butt and Tracy Holloway King (Editors)

2002
CSLI Publications
http://csli-publications.stanford.edu/

Abstract It is shown that five apparently irreconcilable claims about the clausal syntax of Irish can be reconciled in a natural, base-generated LFG analysis that builds on the standard LFG theory of endocentricity and coheads/extended heads, the LFG projection architecture, and Toivonen's (2001) work on non-projecting categories and c-structure adjunction. The analysis also builds on McCloskey's (1996) analysis of Irish adjunction, but does not posit complementizer lowering. The principal theoretical consequences of the analysis are 1) the reconciliation of the five claims, in particular a synthesis of McCloskey's position that the Irish preverbal particles are complementizers and Sells's (1984) position that they are head-adjoined to the verb, 2) the elaboration of Toivonen's (2001) theory of c-structure adjunction, 3) correct predictions about not only adjunction to matrix and subordinate clauses, but also adjunction to appositives.

1 Introduction[1]

The goal of this paper is to reconcile the apparently irreconcilable claims about the clausal syntax of Irish listed in (1). Claims 1 and 2 have been proposed in the literature as universals and claims 3 to 5 have been proposed as specific to the syntax of Irish.

(1)
1. Adjunction to a lexically selected phrase is prohibited (universal).
   (Chomsky 1986)
2. Phrasal categories are endocentrically headed (universal).
   (Jackendoff 1977)
3. The preverbal particles *go*, *aL*, and *aN* in Irish are complementizers.[2]
   (McCloskey 1979, to appear)
4. The preverbal particles are head-adjoined to the finite verb.
   (Sells 1984)
5. The order of a subordinate clause with an adjoined adverbial phrase is *Adverbial Particle V S O* (*not* Particle Adverbial V S O).
   (McCloskey 1996)

There are three principal theoretical consequences of the proposed analysis. First, there is the reconciliation of claims 1 to 5 itself. In particular, a synthesis is achieved between Sells's position (the particles are head adjuncts) and McCloskey's position (the particles are complementizers). Second, the class of adjunction structures will be further restricted, building on recent work by Toivonen (2001). Third, there are further consequences for the theory of adjunction and c-structure. Specifically, the impossibility of adjunction to appositives is derived, while allowing the possibility of adjunction to matrix clauses.

In section 2, I lay out why the claims in (1) present various problems when taken together. I go on to show how an extension of Toivonen's (2001) theory of c-structure adjunction (section 3.1) and use of LFG's projection architecture (section 3.2) and theory of endocentricity (section 3.3) solves these problems, building on transformational work by McCloskey (1996).

[1]I would like to thank Joan Bresnan, Mary Dalrymple, and Ida Toivonen for valuable comments, criticism, and discussion. I owe special thanks to Jim McCloskey and Peter Sells for their generosity in sharing their expertise on this topic. Any errors are my own. This research was funded in part by SSHRC 752-98-0424.

[2]Although *aL* and *aN* are both phonetically realized as a schwa, McCloskey (1979) argues convincingly that they should be treated as separate but homophonous morphemes. One part of his argument is that *aL* and *aN* induce differing mutations on following words: *aL* induces lenition and *aN* induces nasalization. This difference in mutation-triggering is indicated by the *L* and *N*.

2 The Problem: Five Conflicting Claims

It seems at first that of the claims in (1) only claims 3 and 4 are necessarily conflicting. However, I will show in this section that these five claims taken together yield a mess of contradictions that can be fairly intricate at certain points.

The first claim is:

(2) **Claim 1**
Adjunction to a lexically selected phrase is prohibited. **(Universal)**

This is based on the following principle proposed by Chomsky (1986:6), which McCloskey (1996:57) calls the *Adjunction Prohibition*:

(3) Adjunction to a phrase s-selected by a lexical head is ungrammatical.

As noted, claim 1 is postulated as a universal condition on adjunction. This claim applies to lexically selected nominals as well as lexically selected clauses, but in this paper I will concentrate on the latter.

McCloskey (1996) notes that claim 1 accounts for the ungrammaticality of sentences like the following:

(4) a. *[$_{CP}$ When she moved to the city [$_{CP}$ that she could actually get a job]] was amazing.
(McCloskey 1996:57, (21a))

b. *It was amazing [$_{CP}$ when she moved to the city [$_{CP}$ that she could actually get a job]].
(McCloskey 1996:57, (22a))

c. *After [$_{IP}$ last year [$_{IP}$ she resigned]], she moved to Paris.
(McCloskey 1996:58, (26))

The CP *that she could actually get a job* is a sentential subject in (4a) and the complement of an adjective in (4b). The ungrammaticality of adjoining the adverbial *wh*-phrase *when she moved to the city* to this clause is explained by (2), since in both cases the clause is lexically selected. Similarly, in (4c) the IP *she resigned* is the lexically selected complement to *after*; (2) prohibits adjunction of the adverbial NP *last year* to this lexically selected clause.

The importance of the phrase "lexically selected" in (2) is further illustrated by the following variants of (4a) and (4b) McCloskey (1996:57 ,(21b–c) and (22b–c)):

(5) a. [$_{CP}$ That [$_{IP}$ [$_{IP}$ she could actually get a job] when she moved to the city]] was amazing.

b. [$_{CP}$ That [$_{IP}$ when she moved to the city [$_{IP}$ she could actually get a job]]] was amazing.

(6) a. It was amazing [$_{CP}$ that [$_{IP}$ [$_{IP}$ she could actually get a job] when she moved to the city]].

b. It was amazing [$_{CP}$ that [$_{IP}$ when she moved to the city [$_{IP}$ she could actually get a job]]].

There is a crucial difference between the sentences in (5) and (6) and (4a) and (4b) respectively. In the ungrammatical cases, adjunction is to CP, which is lexically selected. This adjunction is ruled out by (2). In the grammatical cases, adjunction is to the IP complement of C. Since C is not a lexical head, the IP complement of C is not lexically selected, and there is no violation of (2). These cases also contrast with (4c), in which there was ungrammatical adjunction to an IP. The difference is that the IP in (4c) is the complement of the lexical head *after* and is therefore a lexically selected clause, to which adjunction is prohibited by (2).

The prohibition against adjunction to a lexically selected clause also holds for Irish. McCloskey (1996:64–65) notes that adjunction of an adverbial to a *wh*-complement is ungrammatical:[3]

[3]I have added some phrase structural annotations to McCloskey's examples for the sake of exposition.

(7) a. * Ní bhfuair siad amach ariamh [$_{CP}$ an bhliain sin [$_{CP}$ cé a bhí ag goid a gcuid
NEG found they out ever that-year who COMP was steal PROG their
móna]].
turf
They never found out who was stealing their turf that year.
(McCloskey 1996:65, (45))

b. * Níor thuig mé [$_{CP}$ roimh an Nollaig [$_{CP}$ cé chomh gnóitheach is a bheadh
NEG-PAST understand I before Christmas how busy as COMP be.COND
siad]].
they
I didn't realize how busy they would be before Christmas.
(McCloskey 1996:65, (46))

The *wh*-complements in (7) are lexically selected by *bhfuair siad* ('found out') and *thuig* ('understand/realize'). Therefore (2) predicts the ungrammaticality of adjunction to these CPs.

The examples in (7) show that it is false to simply assume that (2) does not hold for Irish, yet (2) is seemingly contradicted by other Irish examples, where there is apparent adjunction to lexically selected clauses:

(8) a. Deiridís [$_{CP}$ an chéad Nollaig eile [$_{CP}$ go dtiocfadh sé aníos
they-used-to-say the first Christmas other COMP would-come he up]].
They used to say that next Christmas he would come up.
(From *Bhí Mo Lá Agam*, by Ger Ó Cíobháin, as cited by McCloskey (1996:59, (30)))

b. Is dóiche [$_{CP}$ faoi cheann cúpla lá [$_{CP}$ go bhféadfaí imeacht
COP.PRES probable at-the-end-of couple day COMP could.IMPERS leave.[− FIN]
It's probable that in a few days it would be possible to leave.
(From *Bhí Mo Lá Agam*, by Ger Ó Cíobháin, as cited by McCloskey (1996:59, (31)))

c. chun isteacht duit [$_{CP}$ [$_{CP}$ nuair a bhíos thall ar an tamhnach] go bhfaca mé
to tell.[-FIN] to-you when COMP I-was over on the slope COMP saw I
ceann de do chuid beithíoch]
one of your portion cattle.GEN
to tell you that when I was over on the hillside, I saw one of your cattle
(From *An Leacht Nár Tógadh*, by Séamas Ó Conghaile, as cited by McCloskey (1996:60, (33)))

If we assume that the preverbal particles mark the left edge of the CP (McCloskey 1979, Sells 1984), then these sentences have the structure indicated and seem to be straightforward violations of (2).

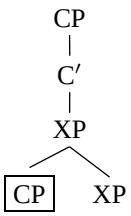
Another puzzle for (2) is appositives, which are not lexically selected, but nevertheless cannot be adjoined to:

(9) *Her prediction, when she moved to the city that her social life would improve, was false.

Given (9), it seems that (2) is not general enough. I will consider appositives again in section 3.1, but for now I wish to focus on the Irish problem raised by (8) and McCloskey's (1996) solution to it.

Based on the premise that claim 1 is universal and prior, McCloskey (1996) argues that these adverbials are in fact not adjoined as indicated in (8), but are rather adjoined *inside* a subcategorized CP, as in the following structure (where the adjoined CP is boxed):

(10)



The outermost CP is the lexically selected clause. Since the adjoined CP is adjoining to an XP inside this selected clause, there is no adjunction to a lexically selected clause and there is no violation of (2).

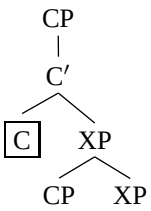
It is a normal assumption of X-bar theory that a maximal projection like the outermost CP in (10) must have a head, a C in this case. This brings us to claim 2:

(11) **Claim 2**
Phrasal categories are endocentrically headed. **(Universal)**

The presence of a CP requires the presence of a C projecting the CP. If the adverbial modifier in (8) is adjoined inside a subcategorized CP, there must be a complementizer dominated by and projecting the CP in question.

This suggests expanding (10) as follows (where the head of the CP in question is boxed):

(12)

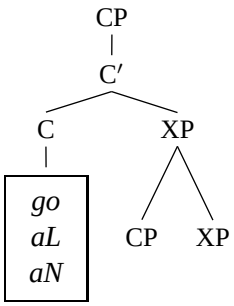


The natural next question to ask is: what is the morphological realization of C in (12)? McCloskey (1996) argues that the C in question is the preverbal particle. This brings us to claim 3:

(13) **Claim 3**
The preverbal particles *go*, *aL*, and *aN* are complementizers. **(Irish, theoretical)**

In other words, the C in (12) expands as follows:

(14)



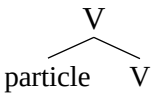
There is independent motivation for assuming that the preverbal particles are complementizers (McCloskey 1979, 1990). First, the particles in question are generally left-peripheral. Second, the particles are sensitive to extraction phenomena in the clause they introduce, famously registering (roughly) whether

an unbounded dependency that passes through their clause terminates in a gap (registered by *aL*) or a resumptive pronoun (registered by *aN*).[4] Third, the particles are sensitive to tense and negation in the clauses they introduce, indicating morphologically whether the clause is past or non-past and whether it is negated.[5] The fact that the particles are sensitive to the presence or absence of arguments and how the argument's extraction site is registered and their sensitivity to tense and negation indicates that the particles are part of the extended verbal functional domain. This coupled with their left-peripheral position argues in favour of treating them as complementizers.

However, Sells (1984) argues that the particles are not complementizers, making claim 4 instead:

(15) **Claim 4**
The preverbal particles are head-adjoined to the finite verb. **(Irish, theoretical)**

In particular, he proposes that the preverbal particles are base-generated as adjuncts to the verbal head:

(16)


As adjuncts to V, the preverbal particles are still within the verbal domain. In fact, they are part of the core verbal domain, rather than the extended functional domain of the verb that complementizers appear in. The evidence that McCloskey gives for the complementizer status of the preverbal particles (that they are left-peripheral, register extraction phenomena, and register tense and negation information) is therefore compatible with Sells's position that they are head-adjoined to the verb.

Two pieces of evidence that Sells (1984) presents for his position is that 1) no material can separate the particle from the verb, and 2) in VP coordination structures the particle must occur in each conjunct:

(17) a. an fear aL cheannaionn agus aL dhíolann tithe
the man [ptc] buys and [ptc] sells houses
the man that buys and sells houses
(Sells 1984:131, (25a))

b. *an fear aL cheannaionn agus d(h)íolann tithe
the man [ptc] buys and sells houses
(Sells 1984:131, (25b))

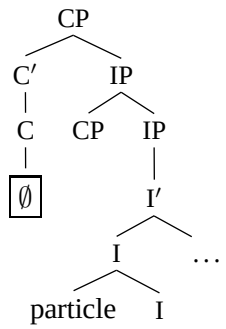
This leads to a somewhat complicated situation. Claim 4 as Sells presents it, i.e. with the structure (16), contradicts claim 3. However, it seems desirable to maintain McCloskey's (1996) structure in (10), which preserved the universal about adjunction (claim 1). If we also wish to maintain claim 2 as a matter of X-bar theory, then we have two choices.

The first choice is that there is a null complementizer heading the CP, as in (18). Note that I have updated Sells's proposal, reflecting the argument that finite verbs in Irish occupy I, not V (Chung and McCloskey 1987, McCloskey 1996).

[4]The choice between *aL* and *aN* is actually extremely complex — especially when the unbounded dependency passes through more than one clause — as discussed in McCloskey (1979) and in more detail in McCloskey (to appear).

[5]See McCloskey (1979:11) for the full morphological paradigm.

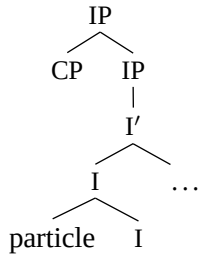
(18)



The problem with this proposal is that it proposes a null element. Not only is this undesirable from an LFG-specific ontological perspective, it even seems undesirable from a transformational perspective, since there is an overt element that is arguably a complementizer, the preverbal particle.

The second choice, again accepting the argument that finite verbs occupy I, is to make the selected clause an IP, not a CP, effectively peeling away the CP layer in (18):

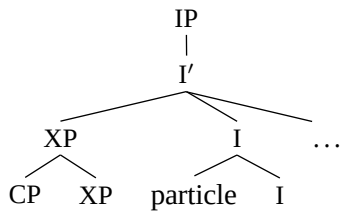
(19)



The problem is that this structure runs equally afoul of claim 1, because there is adjunction to a lexically selected IP.

A possible solution to the latter problem is to make a position for adjunction inside the IP, adopting McCloskey's (1996) strategy. Presumably this position would be a functional projection, as there are no candidate lexical projections, and we would get a stucture like the following:

(20)



However, we must ask ourselves what the functional projection XP in (20) could be. The only functional projections that are standardly part of LFG ontologies are CPs, DPs, and IPs, but none of these are appropriate. Also, to get the correct word order *Adverbial Particle V S O*, XP must be empty; otherwise it is wrongly predicted that some functional element can intervene between the adjoined adverbial (CP in this case) and the particle.[6] This solution therefore introduces an unmotivated functional projection that is independently problematic.

At this point we seem to be rather stuck. It seems that there is no way to simultaneously maintain claim 1 about universal adjunction possibilities, McCloskey's (1996) proposed structure for Irish clausal adjunction,

[6]The adverbial could also be right-adjoined to XP, but this does not solve either of these problems: XP is still unmotivated and we would wrongly predict the possibility of some functional element to the left of the adverbial.

the widely-accepted contention that finite verbs in Irish occupy I, and Sells's (1984) well-motivated claim that the preverbal particles in Irish are in fact base-generated as head adjuncts to the verb. It seems that we may have to abandon Sells's proposal and retreat to claim 3, McCloskey's claim that the preverbal particles are complementizers.

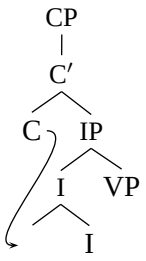
However, things are not even this easy, due to an empirical observation about Irish word order (McCloskey 1996), which constitutes claim 5:

(21) **Claim 5**
The order of a subordinate clause with an adjoined adverbial phrase is *Adverbial Particle V S O*. **(Irish, empirical)**

This is in fact true of all the Irish sentences presented above. However, if the particles are complementizers, as per claim 3, and adjunction takes place *inside* the lexically selected CP, as per McCloskey's (1996) proposal in (10), then we would expect the order *Particle Adverbial V S O*. This is *not* what is observed, which is *Adverbial Particle V S O*.

To resolve this contradiction, McCloskey (1996) proposes that the C is lowered and adjoins to the verb, which occupies I:

(22)



There are various problems with this proposal. First, from an LFG perspective, a lowering analysis, or in fact any movement-based analysis, is certainly not desirable. Second, even in a transformational theory such as the one McCloskey (1996) adopts, lowering is potentially problematic. The principal problem is that the analysis makes false predictions about possible landing sites. There can be indefinitely many CPs adjoined to the IP hosting the proper landing site for the lowered C. However, each of these CPs will presumably also contain an IP. The challenge is to prevent lowering into one of these IPs and to guarantee lowering only into the proper IP; this would require unmotivated stipulations regarding landing sites.[7] A second problem is that minimality requirements on head movement (e.g., the Head Movement Constraint and its descendants; Travis 1984) would need to be adapted to work in the lowering direction to make sure that this kind of head movement is not a violation.

2.1 Summary

This is indeed a tangled web of conflicting assumptions, but in outline the problem is simple. To maintain claim 1 (no adjunction to a selected clause), McCloskey (1996) argues that what looks like adjunction to a CP is adjunction *under* a CP. Claims 1 and 2 (endocentricity) naturally lead to claim 3 (the Irish preverbal particles are complementizers), which is sufficient to satisfy endocentric headedness of a CP. Claim 4 (the preverbal particles are affixes) explains certain Irish data well, but is seemingly incompatible with claims

[7] It may be that a phase-based analysis within the assumptions of the Minimalist Program (Chomsky 1995, 2001) provides a motivated solution to this problem (McCloskey, p.c.). CPs in this framework are phases that are closed upon completion and all but their left-peripheral position is closed to the operation MOVE. Under these assumptions, one might be able to derive the fact that one cannot adjoin to intervening, adjoined CPs since these would be complete and impenetrable to movement. However, it should be noted that lowering operations in general are eschewed in Minimalism.

1–3, unless problematic assumptions are made. Retreating to claim 3 and maintaining the structure proposed by McCloskey to maintain claim 1 is not possible, since there is a clash between this proposed structure and claim 5 (the observation about Irish word order). McCloskey's (1996) solution to this clash is problematic from an LFG perspective, and perhaps even from a theory-internal transformational perspective.

3 The Solution

I will show that these contradictions can be resolved in a natural, base-generated LFG analysis that builds on the standard LFG projection architecture (Kaplan 1995) and theory of coheads/extended heads (Bresnan 2001, among others) and Toivonen's (2001) work on non-projecting categories and c-structure adjunction.

3.1 A Theory of Adjunction

Toivonen (2001) extends and modifies the theory of X-bar structure for LFG presented in Bresnan (2001). She proposes that there is a fundamental distinction between *projecting* and *non-projecting* categories and that X′ and XP level categories can only dominate projecting categories. I will write non-projecting preterminal categories as $\hat{\text{X}}$, using the circumflex accent (ˆ) to indicate iconically that these categories have a "roof" and cannot project any further.[8] Projecting preterminal categories will be written as X^0. Note that X′ is also a projecting category, but it is not a projecting *preterminal*, as it does not dominate a terminal node.

Toivonen (2001:59) also assumes that the class of admissible adjunction structures is restricted by the following generalization:

(23) **Adjunction Identity**
Same [X-bar level] adjoins to same.

In the context of Toivonen's (2001) system, the force of this generalization is that the only permissible adjunction structures involve adjunction of a maximal projection to a maximal projection or adjunction of a non-projecting preterminal to a projecting preterminal:

(24) $\text{XP} \rightarrow \text{XP}, \text{YP}^*$

(25) $\text{X}^0 \rightarrow \text{X}^0, \hat{\text{X}}^*$

This differs from Bresnan's (2001:102–103, 121) theory which allows X′ adjunction and disallows X^0 adjunction (i.e., head adjunction). Note that Toivonen (2001) allows for the possibility of multiple flat adjunction (hence the Kleene star annotation on the adjoining element).

The annotation for the adjunction target is unsurprisingly ↑=↓ (Bresnan 2001:102–103; Toivonen 2001:58–66); since the essential purpose of adjunction is to divide one c-structural category into two segments, it makes sense that the two segments should map to the same f-structure. The versions of (24) and (25) with the adjunction target annotated are as follows:

(26) $\text{XP} \rightarrow \underset{\uparrow=\downarrow}{\text{XP}}, \text{YP}^*$

(27) $\text{X}^0 \rightarrow \underset{\uparrow=\downarrow}{\text{X}^0}, \hat{\text{X}}^*$

[8] This notation departs from (Toivonen 2001), where non-projecting categories are written as X, and projecting categories as X^0. Although this is true to the letter of X-bar theory, it is potentially confusing, as X^0 is often abbreviated as X.

I leave aside the annotation of YP (see Bresnan 2001:102–103), as it is not really relevant to the discussion at hand. The annotation of the adjoined element in all the examples in this paper will be $\downarrow \in (\uparrow \text{ ADJ})$, which is in any case a permissible annotation for YP in Bresnan's (2001) theory of structure-function mappings.

We can capture the universal prohibition against adjunction to a selected phrase (claim 1) as a further refinement of the adjunction structure (26). An initial attempt to do this is shown in (28). The adjunction site XP is annotated with a negative inside-out constraining equation, $\neg(\text{GF} \uparrow)$, which states that the f-structure of the XP must not serve as a grammatical function.

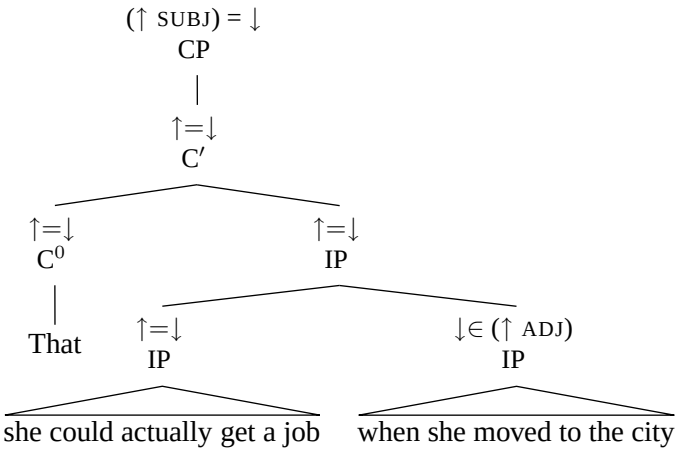
(28) $\text{XP} \longrightarrow \underset{\neg(\text{GF} \uparrow)}{\text{XP}}, \text{YP}^*$

This will not work for functional coheads, though. In particular, it will not work for C^0 and I^0.

The problem is illustrated by example (5a), repeated here along with a partial specification of its annotated constituent-structure:

(5a) [$_{\text{CP}}$ That [$_{\text{IP}}$ [$_{\text{IP}}$ she could actually get a job] when she moved to the city]] was amazing.

(29)



The adjunction site is the IP on the lower left. Although this IP is not *annotated* with a GF, since it is a cohead of C^0 it will receive the grammatical function SUBJ. This is evident if we follow the $\uparrow=\downarrow$ head paths starting at this IP. Since the IP does have a GF, (28) will erroneously rule that (5a) is ungrammatical. The same problem occurs with the left adjunction variant (5b) and the examples in (6).

A tempting move is to adjust rule (28) so that it refers specifically to CPs, since the problematic case of adjunction in both English and Irish involved CP adjunction:

(4a) *[$_{\text{CP}}$ When she moved to the city [$_{\text{CP}}$ that she could actually get a job]] was amazing.

(7a) * Ní bhfuair siad amach ariamh [$_{\text{CP}}$ an bhliain sin [$_{\text{CP}}$ cé a bhí ag goid a gcuid móna]].
NEG found they out ever that-year who COMP was steal PROG their turf
They never found out who was stealing their turf that year.

The CP-specific version of (28) would look like this:

(30) $\text{CP} \longrightarrow \underset{\neg(\text{GF} \uparrow)}{\text{CP}}, \text{YP}^*$

This rule allows adjunction to a CP only if it does not have a grammatical function.

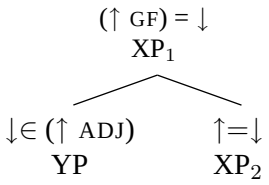
This will indeed rule out (4a) and (7a) but allow (5) and (6). The problem is that there is in fact IP-adjunction that needs to be barred too, as we saw in (4c):

(4c) *After [$_{\text{IP}}$ last year [$_{\text{IP}}$ she resigned]], she moved to Paris.

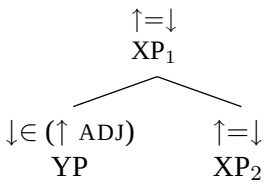
The CP-specific rule (30) will erroneously admit this sentence, since the adjunction site is an IP. We need to capture the distinction between adjoining to an IP that happens to bear a grammatical function because it is a cohead to C^0 and one that bears a grammatical function in its own right. Otherwise we cannot distinguish between grammatical IP-adjunction cases like those in (5) and (6) and the ungrammatical cases like (4c).

What we need is an adjunction rule that distinguishes between structures like (31), which should be disallowed, and (32), which should be allowed. Notice that I am not using the indices *1* and *2* to indicate identity, but just as convenient labels to make subsequent reference to the parts of the XP easier.

(31)



(32)



The crucial difference between the two structures is that XP_1 is annotated with a GF in (31) and with the $\uparrow{=}\downarrow$ head path in (32). However, there is no way within LFG's formal theory to refer to occurrences of annotations on c-structure nodes.

Yet we can take advantage of the $\uparrow{=}\downarrow$ head path in the licit structure (32). A consequence of the two occurrences of the $\uparrow{=}\downarrow$ path in (32) is that the f-structure of XP_1's mother will be identical to the f-structure of XP_2. In other words, the f-structure of the adjunction target XP_2 is identical to the f-structure of its grandmother. This is not true in (31), where the f-structure of XP_1's mother is the f-structure of the predicate for which XP_1's f-structure is a GF.

Therefore, what is needed is a way of referring to a node *x*'s grandmother's f-structure so that the usual kind of f-structure equality can be stated between the grandmother's f-structure and *x*'s f-structure. To accomplish this, we use the standard function M (Kaplan 1995:10), which maps from a node to its mother:

(33) $\text{M}: \text{N} \rightarrow \text{N}$

For example, the $\uparrow$ metavariable for the f-structure of a node's mother is defined as follows, where '$*$' indicates the current node (Dalrymple 2001):

(34) $\uparrow \equiv \phi(\text{M}(*))$

Since M is function from nodes to nodes, we can apply it recursively. Thus, we can define a metavariable $\uparrow^2$ for the f-structure of a node's grandmother as follows:

(35) $\uparrow^2 \equiv \phi(\text{M}(\text{M}(*)))$

Using the metavariable $\uparrow^2$, we can capture the necessary distinction if we further annotate the adjunction rule (26) as follows:

(36) $\text{XP} \longrightarrow \underset{\substack{\uparrow=\downarrow \\ \{\neg(\text{GF}\ \uparrow)\ |\ \uparrow^2 = \downarrow\}}}{\text{XP}}\ ,\ \text{YP}^*$

We will see shortly that independent parts of LFG theory will do the rest of the work.

Let us first take a brief digression to consider a matter of descriptive power, particularly locality. A possible objection to the grandmother metavariable is that it is nonlocal in nature, despite the general desirability of keeping syntactic relations strictly local. There are two responses to this. The first is that the metavariable is in fact local, in the sense of not being global: it requires reference to a c-structure node that is a bounded distance away from the node that is decorated by the metavariable. Second, and more interestingly, it is precisely in the case of adjunction that we would expect the grandmother relation to be relevant. The reason is that adjunction splits one category into two parts. The resulting two c-structure categories are then in some sense really the same category and should have the same mother. Suppose there is an XP, call it $\mathrm{XP}_{d(aughter)}$, that has the annotation $\uparrow{=}\downarrow$. Let us call this XP's mother $\mathrm{YP}_{m(other)}$. When XP_d is split by adjunction, into an upper part (XP_{d1}) and a lower part (XP_{d2}), then both parts should identify their f-structural information with that of YP_m, since the unsplit category XP_d identifies its f-structure with that of YP_m, via the $\uparrow{=}\downarrow$ annotation. The fact that XP_{d2} has the same f-structure as YP_m is indirectly captured by annotating XP_{d2} and XP_{d1} with $\uparrow{=}\downarrow$; it is directly captured by annotating XP_{d2} with $\uparrow^2 = \downarrow$. Since the two parts of the adjunction should map to the same f-structure, XP_{d2} is also annotated $\uparrow{=}\downarrow$.[9]

We can now return to a consideration of (36) and how it captures the correct pattern of data. Consider first structure (31), which represents the kind of adjunction that should be blocked. Since the upper XP (XP_1) bears the annotation ($\uparrow$ GF) = $\downarrow$ its f-structure must be the argument of some predicate that selects for GF in order to satisfy Coherence (Kaplan and Bresnan 1982, Bresnan 2001), which requires that every GF be designated by a PRED. The f-structures of XP_1 and the adjunction target XP_2 are identified as being the same by the $\uparrow{=}\downarrow$ equation on the adjunction target XP_2. A schematic specification of the resulting f-structure is:

(37)
$$\textit{xp}_1\textit{'s mother}\begin{bmatrix}\text{PRED} & \text{`...GF...'}\\ \text{GF} & \textit{xp}_1, \textit{xp}_2\begin{bmatrix}\text{PRED} & \text{`...'}\end{bmatrix}\end{bmatrix}$$

The f-structure corresponding to XP_2 has a GF; therefore $\neg$(GF $\uparrow$) is false and $\uparrow^2 = \downarrow$ must hold. This means that the f-structure of XP_1's mother, i.e. the outermost f-structure in (37), will be equated with the f-structure for XP_2:

(38)
$$\textit{xp}_1\textit{'s mother}\begin{bmatrix}\text{PRED} & \text{`...GF...'}\\ \text{GF} & \textit{xp}_1, \textit{xp}_2\begin{bmatrix}\text{PRED} & \text{`...'}\end{bmatrix}\end{bmatrix}\!\!\Big\}$$

This results in a functional uniqueness violation for the semantic feature PRED; thus, structures like (31) are blocked. The adjunction rule (36) prevents adjunction to lexically selected phrases, no matter their category, maintaining the claim 1.

Consider next the structure (32), which represents the valid kind of adjunction further articulated in (29). XP_1 in (32) is identifying its f-structure with that of its mother and XP_2 is identifying its f-structure with that of its own mother, which is XP_1. Therefore XP_2's f-structure is independently asserted to be the same as its grandmother's f-structure and $\uparrow^2 = \downarrow$ does no further work or harm. Although the left disjunct $\neg$(GF $\uparrow$) is false of XP_2 in (29) (i.e., the IP that is the adjunction target has a GF as discussed above), the right disjunct is true and the structure is licensed. The adjunction rule (36) does not prevent adjunction to an XP contained in a lexically selected phrase.

A case that we have not considered so far is adjunction to matrix CPs, which should be allowed:

(39) When she moved to the city, where did she live?

[9]The $\uparrow^2 = \downarrow$ annotation can also be added to the righthand X^0 in the head-adjunction rule (27). An X^0 head will always have the $\uparrow{=}\downarrow$ annotation; therefore the mother of X^0 with no adjunction and the grandmother of the lower X^0 in an adjunction structure will always be the same. The reader can check this against the head adjunction structures shown in (44) and (47).

The adjunction rule (36) does not block matrix adjunction: since the root clause is not a selected phrase, the negative constraining equation $\neg(\text{GF} \uparrow)$ is satisfied.

Thus, the universal barring adjunction to a lexically selected phrase, claim 1, is maintained in (36) by extending further the theory of adjunction presented by Bresnan (2001) and modified by Toivonen (2001). In fact, (36) is slightly more general than claim 1. As noted in section 2, the universal has nothing to say about the badness of adjunction to an appositive as in (9), since appositives are not lexically selected.

(9) *Her prediction, when she moved to the city that her social life would improve, was false.

However, the appositive does have a GF in LFG: ADJUNCT. This feature is set-valued (see Dalrymple 2001:153–158 and references therein); the appositive in (9) occurs in an f-structure that can be schematically represented as:

(40) $\left[\text{ADJUNCT} \quad \left\{\left[\text{“that her social life would improve”}\right]\right\}\right]$

Thus, it seems that the crucial concept for claim 1 should not be whether the adjunction site is lexically selected, but rather whether it bears a grammatical function, even a non-selected function like ADJUNCT.

The set-valued nature of ADJUNCT necessitates a slight notational modification to (36), such that it does not matter if the inside-out path that is checking for a grammatical function passes through a set or not:

(41) $\text{XP} \longrightarrow \underset{\substack{\uparrow=\downarrow \\ \{\ \neg(\text{GF}\ (\in)\uparrow)\ \mid\ \uparrow^2 = \downarrow\}}}{\text{XP}}\ ,\ \text{YP}^*$

Notice that the optionality of the path through the set ($\in$) means that the left disjunct is equivalent to the negated disjunction $\neg[(\text{GF} \uparrow) \vee (\text{GF} \in \uparrow)]$. This in turn is equivalent to the conjunction $\neg(\text{GF} \uparrow) \wedge \neg(\text{GF} \in \uparrow)$ (by DeMorgan's Law). Therefore, in order for the left disjunct in (41) to be satisified, the f-structure corresponding to the adjunction site cannot be either the value of a GF or a member of a set that is the value of a GF.

The equation $\neg(\text{GF}\ (\in) \uparrow)$ is not satisfied in the f-structure for (9), since the appositive is a member of an ADJUNCT set, as shown in (40); $\uparrow^2 = \downarrow$ cannot be satisfied either, for essentially the same reasons as discussed for (31) (i.e., the f-structure reentrancy introduced results in a functional uniqueness violation). The LFG theory of adjunction presented here not only preserves claim 1, it goes further by correctly blocking adjunction to appositives.

3.2 Irish Complementizers as Head-adjoined Verbal Particles

In the previous section I built on Toivonen's (2001) theory of adjunction, which is in turned based on Bresnan (2001). In this section I will show how Toivonen's (2001) distinction between projecting and non-projecting heads can be used to synthesize McCloskey's claim that the Irish preverbal particles *go*, *aL*, *aN* and their morphological alternants are complementizers and Sells's claim that they are head adjuncts.

The synthesis is achieved by treating the particles as *non-projecting* complementizers. This is demonstrated in the following lexical entry for one of the realizations of the complementizer *go*, which has the non-projecting category $\hat{\text{C}}$, rather than the projecting category C^0:[10]

(42) *goN* $\hat{\text{C}}$ $(\uparrow \text{TENSE}) \neq \text{past}$
$(\uparrow \text{MOOD}) = \text{affirmative}$

[10]The affirmative, non-past *go* induces the nasalization mutation, hence it is written *goN* (McCloskey 1979:11).

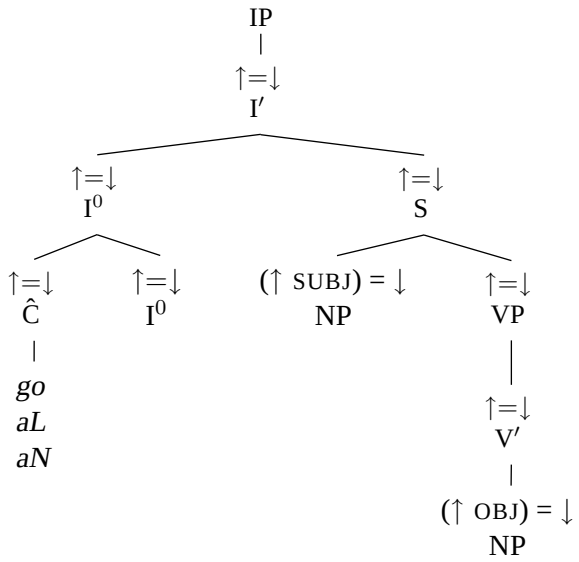
The particles are head-adjoined to finite verbs base-generated in I by the c-structure rule (43), an instantiation of Toivonen's (2001) head-adjunction rule (see (27) above).

(43) $\mathrm{I}^0 \longrightarrow \underset{\uparrow=\downarrow}{\hat{\mathrm{C}}} \quad \underset{\uparrow=\downarrow}{\mathrm{I}^0}$

This maintains Chung and McCloskey's (1987) claim that Irish finite verbs occupy I, but with no movement from V to I. The assignment of the category I to finite verbs is normal practice in LFG (cf. King's (1995) analysis of Russian finite verbs and the analysis of Welsh in Bresnan 2001:127–131).

The resulting IP structure for Irish will be:[11]

(44)



If we adjoin an adverbial to the left of this IP, we get the correct word order, *Adverbial Particle V S O*, as per claim 5. This is akin to McCloskey's (1996) solution of lowering the complementizer to adjoin to I^0, but everything is base-generated and there is no lowering.

A problem remains, though: how is adjunction to this IP possible if it is a lexically selected clause (i.e., a COMP)? I will adopt McCloskey's (1996) solution of shielding the IP inside a CP. The next question is where this CP comes from, as the Irish complementizer is a non-projecting head and does not project a CP. The problem of projecting a CP is solved by further annotating the head-adjunction rule that adjoins $\hat{\mathrm{C}}$ to I^0:

(45) $\mathrm{I}^0 \longrightarrow \underset{\substack{\uparrow=\downarrow \\ \mathrm{CP} \in \mathrm{CAT}(\uparrow)}}{\hat{\mathrm{C}}} \quad \underset{\uparrow=\downarrow}{\mathrm{I}^0}$

The rule uses the CAT operator defined in (10) (Kaplan and Maxwell 1996:93–94; Dalrymple 2001:171):

(46) $\mathrm{CAT(f)} = \{c \,|\, \exists n(n \in \phi^{-1}(f) \wedge \lambda(n) = c)\}$

CAT(f), where f is an f-structure node, returns the set of category labels of c-structure nodes that map to f, using the labeling function λ (Kaplan 1995). Thus, rule (45) states that there is a CP in the set of category labels of c-structure nodes that map to the particle-verb complex's f-structure. A CP is projected, but by the adjunction rule that forms the particle-verb complex, not by the complementizer itself, which cannot project.

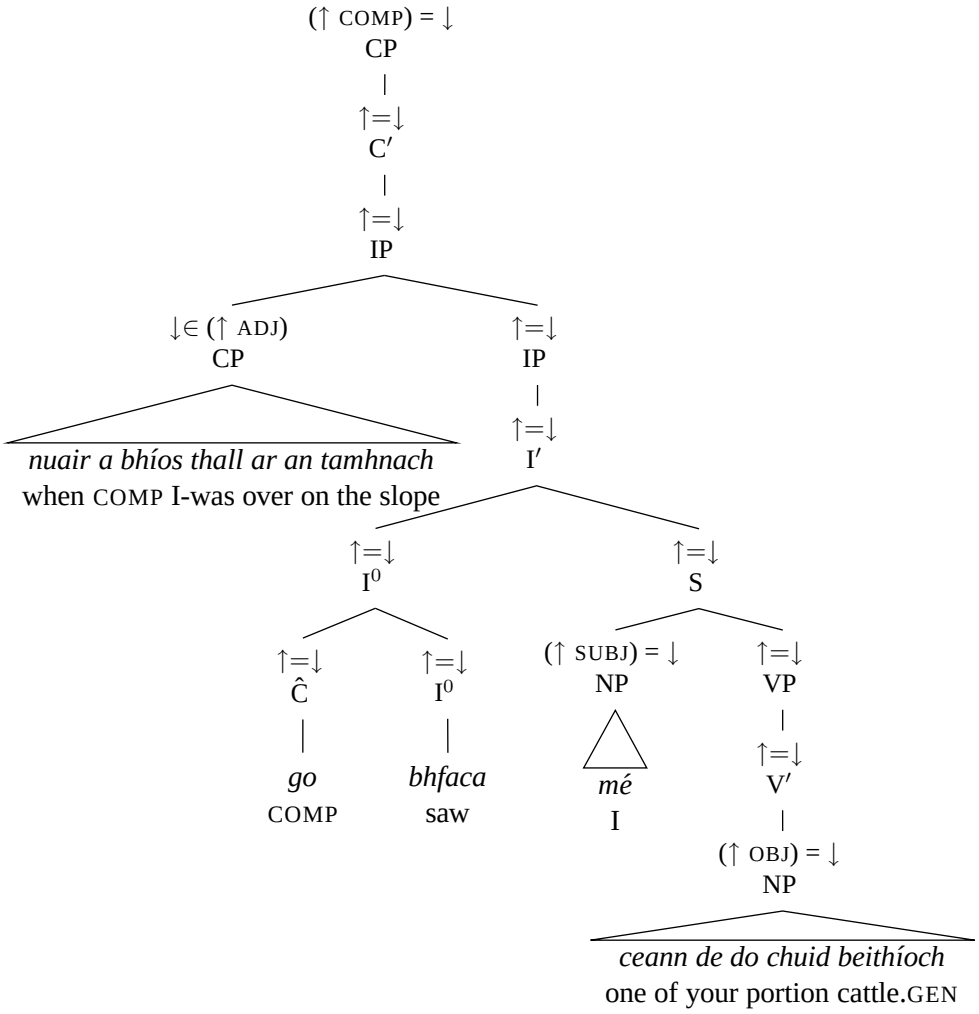
Let us see how all of this comes together in the analysis of example (8c), repeated below, which demonstrated the possibility of adjunction to the left of the particle, yielding the word order *Adverbial Particle V S O*:

[11] The small clause analysis of S dominating NP and VP is also motivated by Chung and McCloskey (1987).

(8c) chun isteacht duit [$_{CP}$[$_{CP}$ nuair a bhíos thall ar an tamhnach] go bhfaca mé ceann de do chuid beithíoch]
to tell.[-FIN] to-you when COMP I-was over on the slope COMP saw I one of your portion cattle.GEN
to tell you that when I was over on the hillside, I saw one of your cattle

The structure yielded for the CP complement of *isteacht* ('tell') in this example, according to the particle-verb adjunction rule (45), the phrasal adjunction rule (41), and lexical selection, is shown in (47):

(47)



The adjunction rule (45) that creates the particle-verb complex requires that a CP be in the set of categories corresponding to the f-structure node of the complementizer $\hat{C}$. This CP is assigned the GF COMP by the lexical item that selects it as a complement. The requirements of the XP adjunction rule (41) are such that adjunction to this CP is impossible, but adjunction to the IP is possible, for the reasons discussed in section 3.1 above.

The analysis has so far achieved a base-generated synthesis of McCloskey position that the Irish preverbal particles are complementizers (claim 3) and Sells position that the Irish particles are head-adjoined to the finite verb (claim 4). This has been done using Toivonen's (2001) theory of X-bar structure and adjunction such that the universal prohibiting adjunction to a selected phrase (claim 1) is maintained and such that the correct and surprising word order *Adverbial Particle V S O* (claim 5) is attained.

3.3 Endocentricity

The last remaining consideration is claim 2, the universal requirement that phrases be endocentrically headed. We saw in section 2 that this was a problem for claim 4, that the complementizers are head adjuncts, if we wish to maintain McCloskey's (1996) solution to the problem of adjunction to selected clauses in Irish. The problem is illustrated in (47): the CP that shields the IP adjunction does not dominate a C^0 in its maximal projection, seemingly violating endocentricity. In fact, c-structure (47) contains an apparent further violation of endocentricity, since the VP does not dominate a V^0.

However, the independently-motivated LFG theory of endocentricity and heads (Bresnan 2001:ch. 7) allows structures such as (47). The statement of endocentricity in this theory is as follows (Bresnan 2001:134):

(48) **Endocentricity:** Every lexical category has an extended head.

Bresnan (2001:132) defines *extended head* as in (49), based on previous work by Zaenen and Kaplan (1995) and Bresnan (2000):

(49) **Definition of Extended Head**: Given a c-structure containing nodes $\mathcal{N}$, $\mathcal{C}$, and c- to f-structure correspondence mapping ϕ, $\mathcal{N}$ is an **extended head** of $\mathcal{C}$ if $\mathcal{N}$ is the minimal node in $\phi^{-1}(\phi(\mathcal{C}))$ that c-commands $\mathcal{C}$ without dominating $\mathcal{C}$.

The force of this definition is to define the notion of head partly in terms of f-structure and partly in terms of c-structure, since these are the two syntactic projections (or levels) in LFG. Although, the definition is somewhat complicated, its basic import is that a c-structural head X^0 of an XP is defined as its extended head if such an X^0 is present;[12] otherwise the immediately c-commanding c-structure node that is on the same $\uparrow{=}\downarrow$ head path as the XP serves as its extended head.

We can now see that the VP in (47) satisfies (48), because its extended head is the upper I^0, which hosts the particle-verb complex. But what about the CP? Since CP is not a lexical category, it does not need even an extended head, according to (48). The capacity for CPs to lack heads is motivated by Bresnan (2001:133, (15a–c)), based on examples such as the following:

(50) a. I wonder [$_{\mathrm{CP}}$ [$_{\mathrm{C}}$ if] [$_{\mathrm{IP}}$ I am tall enough]].
b. I wonder [$_{\mathrm{CP}}$ [$_{\mathrm{AP}}$ how tall] [$_{\mathrm{IP}}$ I am]].
c. *I wonder [$_{\mathrm{CP}}$ [$_{\mathrm{AP}}$ how tall] [$_{\mathrm{C}}$ if] [$_{\mathrm{IP}}$ I am]].

Bresnan (2001) argues that the correct generalization is that either the interrogative complementizer of CP (*if*) or the specifier of CP (*how tall*) is present, but not both. In the former case the CP is headless. Bresnan (2001:133) notes that there is evidence that the *wh*-phrase cannot be an alternative realization of the head of CP, since the *wh*-phrase licenses ellipsis of the IP (*She's tall. I wonder how tall.*), but the complementizer does not, even when heavily stressed (** They say she'll do it, but I wonder IF.*). Thus, CPs in general can lack heads, hence the formulation of endocentricity in (48). There is nothing exceptional about the CP in (47).

4 Conclusion

I have shown in this paper that the apparently irreconcilable claims 1–5 can be reconciled in a natural, base-generated LFG analysis that builds on the standard LFG theory of endocentricity and coheads/extended heads (Bresnan 2001, among others), the LFG projection architecture (Kaplan 1995), and Toivonen's (2001)

[12] Note that Bresnan (2001:142, fn. 11) defines c-command such that "[a node] c-commands itself and all of the nodes dominated by its mother (including that mother)".

work on non-projecting categories and c-structure adjunction. The analysis built on McCloskey's (1996) analysis of Irish adjunction, but does not posit complementizer lowering. The principal theoretical consequences of the analysis are 1) the reconciliation of claims 1–5, in particular the synthesis of McCloskey's position that the Irish preverbal particles are complementizers and Sells position that they are head-adjoined to the verb; 2) the extension of Toivonen's (2001) theory of c-structure adjunction; 3) the correct prediction that adjunction to appositives is impossible, while also disallowing adjunction to lexically selected clauses and allowing adjunction to matrix clauses.

References

Bresnan, Joan (2000). Optimal syntax. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, syntax, and acquisition*, (pp. 334–385). Oxford: Oxford University Press.

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

Chomsky, Noam (1986). *Barriers*. Cambridge, MA: MIT Press.

Chomsky, Noam (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam (2001). Derivation by phase. In Michael Kenstowicz (ed.), *Ken Hale: A life in language*, (pp. 1–50). Cambridge, MA: MIT Press.

Chung, Sandra, and James McCloskey (1987). Government, barriers and small clauses in Modern Irish. *Linguistic Inquiry*, *18*, 173–237.

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

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

Jackendoff, Ray (1977). *X′ syntax: A study of phrase structure*. Cambridge, MA: MIT Press.

Kaplan, Ronald M. (1995). The formal architecture of Lexical-Functional Grammar. In Dalrymple et al. (1995), (pp. 7–27).

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*, (pp. 173–281). Cambridge, MA: MIT Press.

Kaplan, Ronald M., and John T. Maxwell, III (1996). *LFG grammar writer's workbench*. Tech. rep., PARC, Palo Alto, CA. URL `ftp://ftp.parc.xerox.com/pub/lfg/lfgmanual.ps`.

King, Tracy Holloway (1995). *Configuring topic and focus in Russian*. Stanford, CA: CSLI Publications.

McCloskey, James (1979). *Transformational syntax and model theoretic semantics: A case-study in Modern Irish*. Dordrecht: Reidel.

McCloskey, James (1990). Resumptive pronouns, $\overline{\text{A}}$-binding and levels of representation in Irish. In Randall Hendrick (ed.), *Syntax of the Modern Celtic languages*, vol. 23 of *Syntax and Semantics*, (pp. 199–248). San Diego, CA: Academic Press.

McCloskey, James (1996). On the scope of verb movement in Irish. *Natural Language and Linguistic Theory, 14*, 47–104.

McCloskey, James (to appear). Resumption, successive cyclicity, and the locality of operations. In Samuel Epstein and T. Daniel Seeley (eds.), *Prospects for derivational explanation*. Oxford: Blackwell.

Sells, Peter (1984). *Syntax and semantics of resumptive pronouns*. Ph.D. thesis, University of Massachusetts, Amherst.

Toivonen, Ida (2001). *The phrase structure of non-projecting words*. Ph.D. thesis, Stanford University.

Travis, Lisa (1984). *Parameters and effects of word order variation*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Zaenen, Annie, and Ronald M. Kaplan (1995). Formal devices for linguistic generalizations: West Germanic word order in LFG. In Dalrymple et al. (1995), (pp. 215–239).

Coordination and Parallelism in Glue Semantics: Integrating Discourse Cohesion and the Element Constraint

Ash Asudeh and Richard Crouch
Stanford University and PARC

Proceedings of the LFG02 Conference
National Technical University of Athens, Athens
Miriam Butt and Tracy Holloway King (Editors)

2002
CSLI Publications
http://csli-publications.stanford.edu/

Abstract We present initial work on a theory of coordination and parallelism in Glue Semantics (GLUE; Dalrymple 1999, 2001). We will explore points of convergence and divergence between our approach to coordination and similar Categorial Grammar (CG) approaches. We also compare our approach to a previous GLUE approach to coordination (Kehler et al. 1995, 1999) and argue that our approach is superior on the grounds that it preserves a very strong notion of resource-sensitivity (Dalrymple et al. 1993). We conclude by discussing parallelism in connection with the Coordinate Structure Constraint (CSC; Ross 1967). The CSC is a putatively robust condition on extraction which has been argued to be a feature of the CG approach to coordination and of other related approaches. It is standardly assumed to have two parts, the Conjunct Constraint and the Element Constraint (Grosu 1973). The Conjunct Constraint is quite robust, but the Element Constraint has been challenged repeatedly, most recently by Kehler (2002), who argues that the CSC is not a syntactic condition, but rather follows from conditions on discourse coherence and parallelism. We discuss a constraint language on the structure of GLUE derivations, and show how Kehler's theory of discourse cohesion can be related to parallelism in such derivations.

1 Introduction[1]

This paper presents an account of the semantics of coordination, framed within the theory of Glue Semantics (GLUE Dalrymple 1999, 2001). We compare this account to related work in Categorial Grammar (CG) (Steedman 1985, Emms 1990, Carpenter 1997), and an earlier GLUE approach to coordination (Kehler et al. 1995, 1999). We discuss parallelism in connection with the Coordinate Structure Constraint (CSC; Ross 1967), which is standardly assumed to have two parts: the Conjunct Constraint and the Element Constraint (Grosu 1973). Kehler (2002) discusses a set of exceptions to the Element Constraint and argues that it is not a syntactic condition, but follows from conditions on discourse coherence and parallelism. By treating GLUE derivations as first-class semantic representations on which interesting parallelism relations can be defined (Crouch 1999, Asudeh and Crouch to appear), we show how our account of coordination is able to deal with violations of the Element Constraint.

2 A Brief Overview of Glue Semantics

GLUE embodies a treatment of "semantic interpretation as deduction", similar to "parsing as deduction" in Categorial Grammar. GLUE identifies two separate logics in semantic interpretation: a meaning logic for expressing the target semantic representation; and a GLUE logic which specifies how chunks of meaning are deductively assembled. A variety of options are open for the meaning logic (IL, DRT, etc.). But the natural choice for the GLUE logic is a restricted fragment of propositional Linear Logic (Girard 1987). First, the resource-sensitivity of linear logic closely reflects that of natural language (Dalrymple et al. 1993). Second, the existence of a Curry-Howard isomorphism for the GLUE fragment of linear logic both renders it suitable for driving the deductive assembly of meanings, and also provides GLUE derivations with non-trivial identity criteria. We enlarge on these points below.

2.1 A Brief Overview of Linear Logic

Linear logic is resource-sensitive. Unlike traditional logics, linear logic derivations literally consume their premises in order to draw conclusions. This can be illustrated by the following contrastive patterns of inference ($\multimap$ is linear implication, and $\otimes$ is multiplicative conjunction):

[1]We would like to acknowledge Mary Dalrymple, Chris Potts, Mark Steedman and Ida Toivonen for comments on and criticism and discussion of various incarnations of these ideas. We would also like to acknowledge ourselves for all remaining errors. Asudeh is funded in part by SSHRC 752-98-0424.

	Traditional	Linear
Duplication	$a, a \rightarrow b \vdash b$ $a, a \rightarrow b \vdash b \wedge a$ *(a duplicated)*	$a, a \multimap b \vdash b$ $a, a \multimap b \nvdash b \otimes a$ *(No duplication of a)*
Deletion	$a, b \vdash a \wedge b$ $a, b \vdash b$ *(a deleted)*	$a, b \vdash a \otimes b$ $a, b \nvdash b$ *(No deletion of a)*

This ensures that each premise has to be used once, and exactly once, in a derivation; premises can be neither deleted nor duplicated.[2]

The same pattern of strict resource accounting is to be found in natural language: the contribution of each word and phrase must be used once and exactly once in the analysis of a sentence (Dalrymple et al. 1993, 1999b, Asudeh in progress). One cannot freely delete or duplicate the contributions that words make. As we will see in section 3, coordination provides a *prima facie* counterexample to this strict resource accounting. However, we will show how a strictly resourced account can be given.

The Curry-Howard Isomorphism (Howard 1980) pairs logical derivations with terms in the lambda-calculus. In particular, the proof rule of Implication Elimination (or *modus ponens*) corresponds to an operation of function application, while Implication Introduction (or hypothetical reasoning) corresonds to λ-abstraction:

(1) **Implication Elimination**

$$\frac{P : a \multimap b \qquad Q : a}{P(Q) : b} \multimap_{\mathcal{E}}$$

Implication Introduction

$$\frac{\begin{array}{c}[X : a]^i \\ \vdots \\ \phi : b\end{array}}{\lambda X.\phi : a \multimap b} \multimap_{\mathcal{I},i}$$

For Implication Elimination, if P is the λ-term labelling the derivation of $a \multimap b$, and Q is the term labelling the derivation of a, then $P(Q)$ is the term labelling the resultant derivation of b. The implication $a \multimap b$ is thus a function, P, which when applied to an argument Q of type a returns a result $P(Q)$ of type b. For Implication Introduction, suppose that assuming an arbitrary a, with a variable X labelling its unknown derivation, allows us to obtain a derivation ϕ of b. We can then discharge the assumption to get an implication from arguments of type a to results of type b, where the function corresponding to the implication is $\lambda X.\phi$.

Within the setting of GLUE, the λ-terms labelling linear logic formulas will be expressions from the meaning logic. Derivations will assemble these meanings by means of the function application and λ-abstraction operations defined by the Curry-Howard Isomorphism.

The following two derivations of $a, a \multimap b \vdash b$ show the interaction of the proof rules and λ-terms:

(2) (a)
$$\frac{A : a \quad P : a \multimap b}{P(A) : b} \multimap_{\mathcal{E}}$$

(b)
$$\frac{A : a \qquad \dfrac{\dfrac{[X : a]^1 \qquad P : a \multimap b}{P(X) : b} \multimap_{\mathcal{E}}}{\lambda X.P(X) : a \multimap b} \multimap_{\mathcal{I},1}}{(\lambda X.P(X))(A) : b} \multimap_{\mathcal{E}}$$

Derivation (2b) introduces an unnecessary detour. By applying $a \multimap b$ to an assumption of a and then immediately discharging the assumption, we rather pointlessly derive $a \multimap b$ from $a \multimap b$. We then apply this to the premise a to conclude b. Derivation (2a) is a more straightforward way of achieving the same result. Interestingly, the λ-terms of the two derivations are equivalent given η/β-reduction: $(\lambda X.P(X))(A) \Rightarrow P(A)$. This is non-accidental. The equivalence of the terms shows that the two derivations in (2a) and (2b) correspond to the same underlying proof.

The Curry-Howard Isomorphism thus induces non-trivial identity criteria for proofs, such that (2a) and (2b) denote the same proof, despite their surface differences. These identity criteria also match

[2]This strict resource accounting can locally be turned off by means of linear logic's ! modality, where $!a$ means that a is no longer resourced, and can be duplicated or deleted at will. However, we do not include ! in our GLUE fragment of linear logic; see section 3.4 below.

those induced by proof normalization (Prawitz 1965), which provides a set of rules (3) for expunging unnecessary detours from derivations

(3) $$\dfrac{\dfrac{A \quad \dfrac{\begin{array}{c}[A]^i\\ \vdots\\ B\end{array}}{A \multimap B}\multimap_{\mathcal{I},i}}{B}} \quad \stackrel{\beta}{\Longrightarrow} \quad \begin{array}{c} A\\ \vdots\\ B\end{array}$$

The presence of such identity criteria (following Quine's dictum of "no entity without identity") allows us to regard proofs as first-class objects. In particular, normal form derivations (with all detours expunged) can be viewed as canonical representations for their underlying proofs. This opens the possibility of viewing normal form GLUE derivations as first-class levels of representation in semantic theory. In section 5 we will argue that GLUE proofs form an important level of representation in guaging semantic parallelism.

2.2 Examples of GLUE Derivations

In GLUE, meaning constructors for semantic composition are obtained from lexical items instantiated in particular syntactic structures. Each constructor has the form $\mathcal{M} : G$, where $\mathcal{M}$ is a term from some meaning language (e.g., IL, DRT, etc.) and G is a formula of propositional linear logic (Dalrymple et al. 1999a). The goal of a GLUE derivation is to consume all the lexical premises to produce a single conclusion stating the meaning of the sentence. Semantic ambiguity (e.g., scope ambiguity) results when there are alternative derivations from the same set of premises. The Curry-Howard Isomorphism combines lexical meanings in parallel with the structure of the linear logic deduction to build meaning terms.

In this paper we will assume an LFG syntax (see Dalrymple 2001, among others), with one important caveat to which we return in section 5.2, and a very generic predicate calculus, for the sake of exposition. Given these, consider (4) and its lexical items (5):

(4) John saw Fred.

(5)

John N	*Fred* N	*saw* V
$(\uparrow \text{PRED}) = \text{'John'}$	$(\uparrow \text{PRED}) = \text{'Fred'}$	$(\uparrow \text{PRED}) = \text{'see'}$
$john : \uparrow_{\sigma_e}$	$fred : \uparrow_{\sigma_e}$	$see : (\uparrow \text{OBJ})_{\sigma_e} \multimap (\uparrow \text{SUBJ})_{\sigma_e} \multimap \uparrow_{\sigma_t}$

The second line of each entry is its GLUE meaning constructor. The σ subscripts in the GLUE constructors are functions that map syntactic phrases onto their corresponding semantic resources. These resources are typed: e for entity, t for truth-value. The resources are denoted by atomic linear logic propositions. (We will suppress the σ and type subscripts on the linear logic atoms where convenient).

Parsing (4) both constructs the f(unctional)-structure (6) and instantiates the lexical entries so that the $\uparrow$ metavariables refer to nodes within (6), instantiating the lexical premises as in (7).

(6) $$s\begin{bmatrix} \text{PRED} & \text{'see'} \\ \text{SUBJ} & j\left[\text{PRED} \quad \text{'John'}\right] \\ \text{OBJ} & f\left[\text{PRED} \quad \text{'Fred'}\right] \end{bmatrix}$$

(7) $$\begin{array}{l} john : j_{\sigma_e} \\ fred : f_{\sigma_e} \\ see : f_{\sigma_e} \multimap j_{\sigma_e} \multimap s_{\sigma_t} \end{array}$$

The formulas in (7) are used as premises in a linear logic derivation. This must consume all the premises to produce a single conclusion stating the meaning paired with the head resource of the sentence (s_σ). In this case the derivation is straightforward: the three premises combine through two instance of implication elimination, which is functional application in the meaning language:

(8)
$$\dfrac{\dfrac{\dfrac{\mathit{fred} : f_e \qquad \mathit{see} : f_e \multimap j_e \multimap s_t}{\mathit{see}(\mathit{fred}) : j_e \multimap s_t}\multimap_{\mathcal{E}} \qquad \mathit{john} : j_e}{\mathit{see}(\mathit{fred})(\mathit{john}) : s_t}\multimap_{\mathcal{E}}}{\mathit{see}(\mathit{john},\ \mathit{fred}) : s_t}\ \mathit{Notational\ convention}$$

The second example will illustrate quantification. It also shows that quantifier scope ambiguity is handled in the GLUE derivations, without positing an ambiguous syntactic representation. Consider the following sentence, f-structure, and instantiated lexical items:

(9) Everyone found at least one gremlin.

(10)
$$f\begin{bmatrix}\text{PRED} & \text{'find'} \\ \text{SUBJ} & e\begin{bmatrix}\text{PRED} & \text{'everyone'}\end{bmatrix} \\ \text{OBJ} & g\begin{bmatrix}\text{PRED} & \text{'gremlin'} \\ \text{SPEC} & \begin{bmatrix}\text{PRED} & \text{'at least one'}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

(11)
$$\begin{array}{l}\lambda P.\mathit{every}(\mathit{person},\ P) : (e_e \multimap X_t) \multimap X_t \\ \lambda u, v.\mathit{find}(v, u) : g_e \multimap e_e \multimap f_t \\ \lambda Q.\textsc{alo}(\mathit{gremlin},\ Q) : (g_e \multimap Y_t) \multimap Y_t\end{array}$$

The meaning terms $\lambda P.\mathit{every}(\mathit{person},\ P)$ and $\lambda Q.\textsc{alo}(\mathit{gremlin},\ Q)$ are standard generalized quantifier expressions, which we will henceforth abbreviate as EO and ALOG. Reading the types from the linear logic formulas, it can be see that both are of the (familiar) semantic type $\langle\langle e, t\rangle, t\rangle$. The upper case variables, X_t and Y_t, range over arbitrary type t atomic resources that the quantifiers could take as their scope. Essentially, the two quantifiers can apply to any type t clause that depends on the meaning of the subject e_e or the object g_e, and discharge this dependency by scoping the quantitifier.

From the three premises in (11), there are two distinct derivations f_t. Both have the same initial derivation (12), producing the semantic resource f_t dependent on both e_e and g_e. The derivations then fork, depending on which of these dependencies are discharged first via scoping a quantifier. ((13) for surface scope and (14) for inverse scope).

(12)
$$\dfrac{[x : e]^2 \qquad \dfrac{[y : g]^1 \qquad \lambda u, v.\mathit{find}(v, u) : g \multimap e \multimap f}{\lambda v.\mathit{find}(v, y) : e \multimap f}\multimap_{\mathcal{E}}}{\mathit{find}(x, y) : f}\multimap_{\mathcal{E}}$$

(13)
$$\dfrac{\dfrac{\textsc{eo} : (e \multimap X) \multimap X \qquad \dfrac{\dfrac{\textsc{alog} : (g \multimap Y) \multimap Y \qquad \dfrac{\mathit{find}(x, y) : f}{\lambda y.\mathit{find}(x, y) : g \multimap f}\multimap_{\mathcal{I},1}}{\textsc{alog}(\lambda y.\mathit{find}(x, y)) : f}\multimap_{\mathcal{E}Y=f} \qquad \lambda x.\textsc{alog}(\lambda y.\mathit{speak}(x, y)) : e \multimap f}\multimap_{\mathcal{I},2}}{\textsc{eo}(\lambda x.\textsc{alog}(\lambda y.\mathit{find}(x, y))) : f}\multimap_{\mathcal{E}X=f}$$

(14)
$$\dfrac{\dfrac{\textsc{alog} : (g \multimap Y) \multimap Y \qquad \dfrac{\dfrac{\textsc{eo} : (e \multimap X) \multimap X \qquad \dfrac{\mathit{find}(x, y) : f}{\lambda x.\mathit{find}(x, y) : e \multimap f}\multimap_{\mathcal{I},2}}{\textsc{eo}(\lambda x.\mathit{find}(x, y)) : f}\multimap_{\mathcal{E}X=f} \qquad \lambda y.\textsc{eo}(\lambda x.\mathit{find}(x, y)) : e \multimap s}\multimap_{\mathcal{I},1}}{\textsc{alog}(\lambda y.\textsc{eo}(\lambda x.\mathit{find}(x, y))) : f}\multimap_{\mathcal{E}Y=f}$$

Note that the two scopings for the sentence arise solely from alternative linear logic derivations from premises to conclusion, using standard rules of inference. No syntactic ambiguity needs to be posited. Moreover, no special assumptions need to be made about the meaning terms. Glue Semantics enforces a modular separation between the GLUE language and the target meaning language, so that the range of possible GLUE derivations is determined solely by the linear logic formulas expressing relations between

semantic resources, and is entirely independent of the meaning terms associated with these resources. This modularity will be important when we come to define a level of semantic parallelism that abstracts away from differences in meaning (section 4.2).

3 Coordination in Glue Semantics

In this section we will look at two approaches to coordination in Glue Semantics (see also Dalrymple 2001:361–387). First we present our approach, which treats coordination essentially as modification and bears similarities to coordination in Montague Grammar, Generalized Phrase Structure Grammar (GPSG), Combinatory Categorial Grammar, and Type-Logical Grammar (Montague 1973, Gazdar 1980, 1981, 1982, Partee and Rooth 1983, Gazdar et al. 1985, Keenan and Faltz 1985, Steedman 1985, 1989, 1990, 2000, Emms 1990, Carpenter 1997). Second, we present and contrast a GLUE approach pursued by Kehler et al. (1995, 1999). Anticipating somewhat, the key difference between the two GLUE approaches is that the coordination as modification approach employs a resource management strategy to handle apparently conflicting resource consumption requirements, while the latter approach instead chooses to relax resource sensitivity.

3.1 An Example of VP Coordination

Let us start by considering a simple VP-coordination with three conjuncts:[3]

(15) Kim slept, dreamt and laughed.

(16)
$$f\left[\begin{array}{l}\text{CONJ}\quad \text{and}\\ \left\{\begin{array}{l} s\left[\begin{array}{ll}\text{PRED} & \text{`sleep'}\\ \text{SUBJ} & k\left[\text{PRED}\quad \text{`Kim'}\right]\end{array}\right]\\ d\left[\begin{array}{ll}\text{PRED} & \text{`dream'}\\ \text{SUBJ} & \end{array}\right]\\ l\left[\begin{array}{ll}\text{PRED} & \text{`laugh'}\\ \text{SUBJ} & \end{array}\right]\end{array}\right\}\end{array}\right]$$

(17) Target semantics = $sleep(kim) \wedge dream(kim) \wedge laugh(kim)$

The target semantic representation suggests that coordination poses a problem for the strict resource accounting inherent in GLUE. The single word *Kim* appears to make a threefold semantic contribution, as the subject argument to each of the three coordinated VPs. Spelling this out in terms of instantiated GLUE constructors, we would expect to obtain three ordinary intransitive verb premises but just one subject premise

(18)

kim	$kim : k$
slept	$sleep : k \multimap s$
dreamt	$dream : k \multimap d$
laughed	$laugh : k \multimap l$

How can a GLUE derivation match the one producer of the semantic resource k with the three consumers of k?

Our solution to this apparent resource mismatch is a treatment of coordination that consumes multiple dependencies on shared arguments to produce single dependencies on the shared arguments. Let

[3] We follow the usual convention of referring to the elements that are coordinated as conjuncts, even though the logical coordination relation is not limited to conjunction.

us add the following additional constructors, contributed by the coordination itself (For expository purposes, we here show these additional constructors in a simplified form. Their full form, (23), and how they are obtained is discussed below):

(19)
$$\begin{array}{l} \lambda P.P : (k \multimap l) \multimap (k \multimap f) \\ \lambda P, Q, x.P(x) \wedge Q(x) : (k \multimap d) \multimap (k \multimap f) \multimap (k \multimap f) \\ \lambda P, Q, x.P(x) \wedge Q(x) : (k \multimap s) \multimap (k \multimap f) \multimap (k \multimap f) \end{array}$$

The first constructor consumes the final conjunct VP ($k \multimap l$), and turns it into the 'seed' meaning for the coordinated VP ($k \multimap f$). This consumes one dependency on the argument k to produce another. The other two constructors each consume one of the remaining conjuncts, and turn them into modifiers of the seed VP meaning ($(k \multimap f) \multimap (k \multimap f)$). Each consumes one dependency on the shared argument k to modify the existing seed dependency on k. The GLUE derivation for the sentence proceeds as follows

(20)
$$\dfrac{\dfrac{\dfrac{\dfrac{\begin{array}{l}\mathit{laugh}: \\ k \multimap l\end{array} \quad \begin{array}{l}\lambda P.P: \\ (k \multimap l) \multimap (k \multimap f)\end{array}}{\begin{array}{l}\mathit{laugh}: \\ k \multimap f\end{array}} \quad \dfrac{\begin{array}{l}\mathit{dream}: \\ k \multimap d\end{array} \quad \begin{array}{l}\lambda P \lambda Q \lambda x.[P(x) \wedge Q(x)]: \\ (k \multimap d) \multimap ((k \multimap f) \multimap (k \multimap f))\end{array}}{\begin{array}{l}\lambda Q \lambda x.[\mathit{dream}(x) \wedge Q(x)]: \\ (k \multimap f) \multimap (k \multimap f)\end{array}}}{\begin{array}{l}\lambda x.[\mathit{dream}(x) \wedge \mathit{laugh}(x)]: \\ k \multimap f\end{array}} \quad \dfrac{\begin{array}{l}\mathit{sleep}: \\ k \multimap s\end{array} \quad \begin{array}{l}\lambda P \lambda Q \lambda x.[P(x) \wedge Q(x)]: \\ (k \multimap s) \multimap ((k \multimap f) \multimap (k \multimap f))\end{array}}{\begin{array}{l}\lambda Q \lambda x.[\mathit{sleep}(x) \wedge Q(x)]: \\ (k \multimap f) \multimap (k \multimap f)\end{array}}}{\dfrac{\begin{array}{l}\lambda x.[\mathit{sleep}(x) \wedge (\lambda x.[\mathit{dream}(x) \wedge \mathit{laugh}(x)])(x)]: \\ k \multimap f\end{array}}{\begin{array}{l}\lambda x.[\mathit{sleep}(x) \wedge [\mathit{dream}(x) \wedge \mathit{laugh}(x)]]: \\ k \multimap f\end{array}}} \quad \begin{array}{l}\mathit{kim}: \\ k\end{array}}{\begin{array}{l}[\mathit{sleep}(\mathit{kim}) \wedge [\mathit{dream}(\mathit{kim}) \wedge \mathit{laugh}(\mathit{kim})]]: \\ f\end{array}}$$

The derivation first consumes the three $k \multimap l, k \multimap s$ and $k \multimap d$ VPs to produce one $k \multimap f$ coordinate VP. Only then is the single k resource consumed.

We now address the question of where the additional GLUE constructors in (19) come from. In doing so, we will also generalize the constructors in (19) so that they apply to coordinators besides boolean conjunction.

In addition to purely lexical GLUE premises, we assume that c-structure rules can sometimes also introduce non-lexical, constructional GLUE premises (see also Dalrymple 2001). Such is the case with the rule for VP coordination (21), with the lexical entry for the conjunct *and* shown in (22).[4]

(21)
$$\mathrm{VP} \rightarrow \begin{array}{l} \quad\quad\quad\quad \mathrm{VP}^{+} \\ \quad\quad\quad\quad \downarrow \in \uparrow \\ \lambda P, Q, C, x.\ C(P(x), Q(C, x)): \\ \ [((\in \uparrow)_\sigma \, \mathrm{CREL}) \multimap (\uparrow \mathrm{SUBJ})_\sigma \multimap \uparrow_\sigma] \\ \ \multimap [((\in \uparrow)_\sigma \, \mathrm{CREL}) \multimap (\uparrow \mathrm{SUBJ})_\sigma \multimap (\in \uparrow)_\sigma] \\ \ \multimap [((\in \uparrow)_\sigma \, \mathrm{CREL}) \multimap (\uparrow \mathrm{SUBJ})_\sigma \multimap (\in \uparrow)_\sigma] \end{array} \quad \begin{array}{c} \mathrm{Conj} \\ \uparrow = \downarrow \end{array} \quad \begin{array}{l} \quad\quad\quad\quad \mathrm{VP} \\ \quad\quad\quad\quad \downarrow \in \uparrow \\ \lambda P, C.\ P: \\ \ [(\uparrow \mathrm{SUBJ})_\sigma \multimap \uparrow_\sigma \\ \ \multimap [((\in \uparrow)_\sigma \, \mathrm{CREL}) \multimap (\uparrow \mathrm{SUBJ})_\sigma \multimap (\in \uparrow)_\sigma \end{array}$$

(22) Lexical entry: *and* Conj
$(\uparrow \mathrm{CONJ}) = \text{'and'}$
$\mathit{and} : (\uparrow_\sigma \mathrm{COORD\text{-}REL})$

The c-structure rules assigns an additional GLUE premise to each VP conjunct. The final VP is the seed for the coordination. The meaning of the final VP, $(\uparrow \mathrm{SUBJ})_\sigma \multimap \uparrow_\sigma$, is consumed to produce an initial meaning for the coordinate VP, $(\uparrow \mathrm{SUBJ})_\sigma \multimap (\in \uparrow)_\sigma$.[5] This initial meaning is also *vacuously* dependent on the meaning of the coordinator, $((\in \uparrow)_\sigma \, \mathrm{CREL})$, provided by the lexical entry for the word *and*. Here, CREL (an abbreviation for COORDINATION-RELATION) is a feature in s(emantic)-structure, similar to other s-structure features like VAR(IABLE) and RESTR(ICTION) (for more details see Dalrymple 2001).

Non-final VP conjuncts each induce a constructor that consumes the conjunct VP meaning to produce a modifier of coordinate VP meaning. The meaning term for the constructor, $\lambda P, Q, C, x.\ C(P(x), Q(C, x)$ perhaps needs explanation. P represents the meaning of the conjunct

[4]We follow standard LFG practice in assuming that the Kleene plus (+) applies not just to the category, but also to its annotations.

[5]The coordinate f-structure is refered to by means of $(\in \uparrow)$, since the conjunct $\uparrow$ is contained within a set argument of the coordinate f-structure (Kaplan and Maxwell 1988).

VP, and Q the meaning of the seed, coordinate VP. C represents the meaning of the conjunction, and x the meaning of the shared subject. The constructor abstracts over all of these meanings, which is to say we are creating a function that takes them as arguments. The result of the function applies P to the subject x, and Q to both x and the conjunction C. The application of Q to C ensures that the value of the conjunction is threaded down throughout the coordinate VP. Finally, both the conjunct clause meaning, $P(x)$, and the coordinate clause meaning, $Q(C, x)$, are conjoined by the conjunction C.

Applying these rules in the setting of (16) (where c abbreviates the semantic resource $(f_\sigma \text{ CREL})$ and f abbreviates the semantic resource for the entire coordinate structure), we obtain the following premises:

(23)
1. kim $kim : k$
2. slept $sleep : k \multimap s$
3. dreamt $dream : k \multimap d$
4. and $and : c$
5. laughed $laugh : k \multimap l$
6. $\lambda P, C.P : (k \multimap l) \multimap (c \multimap k \multimap f)$
7. $\lambda P, Q, C, x.\ C(P(x), Q(C, x)) : (k \multimap d) \multimap (c \multimap k \multimap f) \multimap (c \multimap k \multimap f)$
8. $\lambda P, Q, C, x.\ C(P(x), Q(C, x)) : (k \multimap s) \multimap (c \multimap k \multimap f) \multimap (c \multimap k \multimap f)$

Going through the derivation step by step, we first apply 6 to 5 to get the seed meaning for the coordination, 9.

(24) 9. $\lambda C.\ laugh : (c \multimap k \multimap f)$

We also apply 7 to 3 to obtain the seed modifier 10:

(25) 10. $\lambda Q, C, x.\ C(dream(x), Q(C, x)) : (c \multimap k \multimap f) \multimap (c \multimap k \multimap f)$

We then modify 9 by 10 to give 11

(26) 11. $\lambda C, x.\ C(dream(x),\ laugh(x)) : (c \multimap k \multimap f)$

Applying 8 to 2 gives another seed modifier, 12

(27) 12. $\lambda Q, C, x.\ C(sleep(x), Q(C, x)) : (c \multimap k \multimap f) \multimap (c \multimap k \multimap f)$

Applying 12 to 11 gives

(28) 13. $\lambda C, x.\ c(sleep(x),\ C(dream(x),\ laugh(x)) : (c \multimap k \multimap f)$

Applying 13 to 4 and then 1 finally yields the conclusion

(29) $and(sleep(kim),\ and(dream(kim),\ laugh(kim)) : f$

Note how both the meaning of the subject and the meaning of the conjunction appear at several places in the final meaning term, without duplication of the corresponding GLUE resources.

In this section, we have deliberately chosen an example with three conjuncts. The recursive nature of our analysis (turn the rightmost conjunct into a seed meaning for the coordination, and all other conjuncts into modifiers of the coordination) ensures that we can deal with coordinations with any number of conjuncts. This addresses a criticism by Dalrymple (2001:379) that a previous version of the coordination-as-modification approach was insufficiently general.

Another criticism made by Dalrymple (2001) is that there are non-coordination cases where resource sharing is an issue, such as the following example, which she cites from Hudson (1976):

(30) Citizens who support, paraded against politicians who oppose, two trade bills.

Although there is no coordination in this right-node raising case, the resource sharing issue is similar to the VP-coordination case we have been examining, as the verbs *support* and *paraded* share an object, *two trade bills*. The rule for right-node raising would handle the resource management of the VPs in essentially the same manner as the coordination rule, but without making the semantic contribution of coordination.

3.2 A General Semantic Schema for Coordination?

Our method for dealing with VP coordination readily generalizes to the coordination of any phrases with zero or more shared arguments (Kaplan and Maxwell 1988). It is tempting, therefore, to attempt a general schema for the semantics of coordination, along the lines of

(31)
$$\mathrm{X} \rightarrow \begin{array}{c} \mathrm{X}^{+} \\ \downarrow \in \uparrow \end{array} \quad \begin{array}{l} \lambda P, Q, C, \vec{x}.\ C(P(\vec{x}), Q(C, \vec{x})) : \\ [((\in \uparrow)_{\sigma}\ \mathrm{CREL}) \multimap \langle \vec{\alpha_{\multimap}} \rangle_n \multimap \uparrow_{\sigma}] \\ \quad \multimap [((\in \uparrow)_{\sigma}\ \mathrm{CREL}) \multimap \langle \vec{\alpha_{\multimap}} \rangle_n \multimap (\in \uparrow)_{\sigma}] \\ \quad \multimap [((\in \uparrow)_{\sigma}\ \mathrm{CREL}) \multimap \langle \vec{\alpha_{\multimap}} \rangle_n \multimap (\in \uparrow)_{\sigma}] \end{array} \qquad \begin{array}{c} \mathrm{Conj} \\ \uparrow = \downarrow \end{array} \qquad \begin{array}{c} \mathrm{X} \\ \downarrow \in \uparrow \end{array} \quad \begin{array}{l} \lambda P, C.\ P : \\ [\langle \vec{\alpha_{\multimap}} \rangle_n \multimap \uparrow_{\sigma} \\ \quad \multimap [((\in \uparrow)_{\sigma}\ \mathrm{CREL}) \multimap \langle \vec{\alpha_{\multimap}} \rangle_n \multimap (\in \uparrow)_{\sigma}] \end{array}$$

where X specifies the category level of the coordination, $\langle \vec{\alpha_{\multimap}} \rangle_n$ represent a sequence of n implications over the shared argument resources of the conjuncts, and $\vec{x}$ is the corresponding sequence of meaning variables. The value of n is set by the level of the coordination: for sentential coordination, where there are no shared arguments, $n = 0$; for VP coordination $n = 1$, etc.

While is may be possible to define such a schematic c-structure rule and semantics for coordination, it is unclear that this is desirable. Let us leave aside the question of whether a single schema is syntactically possible. It is implausible to demand that all levels of coordination must have exactly the same semantics. For example, purely boolean coordination, of the kind outlined above, may be what is required for VPs. But for NPs, we might wish to adopt a semantics where the seed NP conjunct additionally introduces a group entity, and where the conjunct NPs quantify over elements of the group, e.g.

(32) Kim and/or a dog left
$\exists X.group(X) \wedge$
$and/or(member(kim, X),\ \exists d.\ dog(d) \wedge member(d, X))$
$\wedge\ leave(X)$

This is not a purely boolean coordination. However, the conjunctive word *and/or* does have a purely boolean semantics within the coordination. One would not be able to combine these two styles of semantics for different coordination levels under a single c-structure schema for coordination.

Instead, it is preferable to assume one c-structure coordination rule per category. The syntactic portion of each c-structure rule will indeed always have the form of the syntactic portion of (31), as shown in (33), where *X* can be any category of the language, including the categories for partial constituents, such as *x*-VP, defined by Maxwell and Manning (1996) in their treatment of nonconstituent coordination and coordination of unlikes.

(33)
$$\mathrm{X} \rightarrow \begin{array}{c} \mathrm{X}^{+} \\ \downarrow \in \uparrow \end{array} \qquad \begin{array}{c} \mathrm{CONJ} \\ \uparrow = \downarrow \end{array} \qquad \begin{array}{c} \mathrm{X} \\ \downarrow \in \uparrow \end{array}$$

However, the semantics of the each rule can be tuned to suit the particular category in question.

Significant linguistic generalizations can be achieved by means of macros[6] encoding common patterns of analysis. Thus one can encode our semantic analysis of shared arguments in one macro, and invoke this in slightly different settings in various specific c-structure coordination rules.

3.3 A Brief Comparison to Coordination in Categorial Grammar

Our approach bears a lot of similarity to various Categorial Grammar (CG) approaches to coordination, including approaches in both Combinatory Categorial Grammar (CCG) and Type-Logical Grammar (TLG) (Steedman 1985, Emms 1990, Carpenter 1997). The boolean coordinator we introduced in (22)

[6] A macro is a device widely used in programming, and in computational grammars, where repeatedly used chunks of code/rules are written once, possibly parameterized, in a single place. Macro calls in the grammar are expanded out to be replaced by the rule chunks. If (33) were defined as a macro, parameterized by the category X, then multiple calls to the macro for different categories (S, VP, NP, etc.) would expand out to different instances of the coordination rule.

and the schema we use for coordination in (21) are similar to Emms's (1990) polymorphic generalization of Steedman's (1985) work, which is also adopted for Type-Logical Grammar by Carpenter (1997). We also noted that our syntactic schema is compatible with the previous LFG work on nonconstituent coordination and coordination of unlikes by Maxwell and Manning (1996).

The crucial distinction between our approach and the CCG/TLG approaches is that GLUE assumes a level of syntax that is separate from the level of semantic composition. Another interesting difference between the GLUE approach and categorial approaches is that the latter have tended to assume binary coordination.[7] The independent level of syntax in GLUE allows for constructional premises associated with c-structure rules, permitting a straightforward analysis of n-ary coordinations, such as *Kim slept, dreamt and laughed*. On a binary approach, one is forced to treat coordination syncategorematically or to treat the comma in written language or some phonetic cue in spoken language as an additional lexical conjunction. However, this is empirically unmotivated, since the comma is merely an orthographic deviceand there is no clear phonetic correlate of the "coordinating" comma in normal, connected speech.[8]

3.4 The Paths-as-resources Approach to Coordination

Kehler et al. (1995, 1999) offer a different solution to the problem of resource sharing caused by reentrancy in f-structures, of which coordination is just one example. They propose a modification to GLUE in which *paths* through f-structures contribute resources, rather than f-structure *nodes* contributing resources, as we have been assuming and as is assumed in most work on Glue Semantics (see Dalrymple 1999, 2001 and references therein). For example, even if there is only one occurence of the subject in a VP-coordination f-structure, there are as many paths leading to the shared subject as there are heads subcategorizing for it. Thus, there will in fact be as many subject resources contributed as there are verbs requiring subjects, solving the resource sharing problem. Beside solving this problem, Kehler et al. (1995, 1999) show how their account gets correct results for the interaction of coordination with intensional verbs and with right-node raising. Dalrymple (2001:377-378) additionally notes that the Kehler et al. (1995, 1999) approach correctly forces a shared quantified subject in VP-coordination to have wide scope. For example, in *Someone laughed and cried* it is the same person doing the laughing and crying (Partee 1970).

However, there are empirical and theoretical objections to the Kehler et al. approach. First, as Dalrymple (2001:378) points out, we do not want the resource duplication offered by this approach in other cases with shared arguments at f-structure, such as in raising and possibly control (Asudeh 2000, 2002) and in unbounded dependencies involving sharing of an argument function with TOPIC or FOCUS. Second, their approach makes crucial use of the *of course* or *bang* modality (!) of linear logic. This modality turns off resource accounting for any formula it takes as its argument.[9] The problem with using this modality is that it undermines the potentially powerful explanation of natural language resource sensitivity offered by GLUE and the potential for simplifying or eliminating several principles and generalizations offered in the literature, such as Full Interpretation, the Theta Criterion, Completeness and Coherence, and possibly others (for discussion see Asudeh in progress).

By contrast, we stick to the multiplicative fragment of linear logic without bang, thus preserving a strict notion of resource sensitivity. Furthermore, the coordination-as-modification approach solves the resource sharing problem introduced by structure sharing while maintaining the usual GLUE notion of resources being contributed by f-structure nodes rather than paths. Thus, our approach does not run into resource duplication problems with raising, control, or unbounded dependencies. Lastly, our

[7] There are at least two possible exceptions to this. The first is Morrill's (1994) proposal to give coordinators the schematic form $(X+\backslash X/X)$, where X+ expands into one or more categories of type X. The second is the proposal by Steedman (1989:210–212). Steedman (1989) does not explicitly discuss a solution to n-ary coordination, but it is clear from Steedman (1990:fn. 9) that he means the generalization of coordination in Steedman (1990) and the syncategorematic treatment in Steedman (1989) to extend to such cases. However, syncategorematic treatments of lexically-realized elements are generally disfavoured in CG, and this is abandoned in Steedman (2000).

[8] Orthographic devices may be indicative of some linguistically-relevant factor, but this is unreliable. For example, no linguist would agree that a good test for whether something is a German noun is whether it is written capitalized or not.

[9] The bang modality is used when adding the rules for Weakening and Contraction and is therefore useful for showing relations between linear logic and classical logics.

approach also achieves correct results for wide-scope quantified subjects and for coordinations involving intensional verbs.[10]

4 Semantic Parallelism

In the next section we turn to parallelism in coordination, and in particular to examples of non-parallelism discussed by Kehler (2002). In this section we lay some groundwork by describing how semantic parallelism, cast as parallelism in GLUE derivations, can be measured.

The starting point is the observation made in section 2.1 that, for certain logical systems including linear logic, derivations have non-trivial identity criteria. These are sufficient to make proofs first class objects in logical theory, so that it is interesting not only to study *what* is proved, but also *how* it is proved.[11] These identity criteria also allow us to view normal forms derivations as canonical representations of underlying proofs.

Within a linguistic setting, this means that normal form GLUE derivations can be genuine objects in semantic theory, reifying the syntax-semantics interface to show *how* meanings are constructed, abstracting away from details of *what* meanings are constructed. This in turn enables one to compare derivations of different meanings for parallel structures.[12]

4.1 A Level of Semantic Representation

Logical formulas are traditionally not regarded as a genuine level of semantic representation (Montague 1970), as they generally have no non-trivial identity criteria other than through model theoretic semantics. For example, the two apparently distinct scopings of (34)

(34) Every man saw every woman

a. $\forall x.\, man(x) \rightarrow \forall y.\, woman(y) \rightarrow see(x, y)$

b. $\forall y.\, woman(y) \rightarrow \forall x.\, man(x) \rightarrow see(x, y)$

are model theoretically equivalent. Other than by exploiting arbitrary properties of the logical notation, there is no semantic basis for distinguishing these two formulas. Yet the urge to state generalizations over a level of semantic representation is very strong.

In this respect, GLUE contrasts with various closely related frameworks, such as Montague Grammar, GPSG, CCG, and TLG. Montague Semantics, as in Montague's own work (Montague 1973) or as in the variant espoused in GPSG, uses a level of purely syntactic representation that is systematically translated into a semantic formulae. However, there is no true level of semantic representation beyond the model theory. On the other hand we have Categorial Grammar (CCG and Type-Logical Grammar). Categorial derivations do have identity criteria that are distinct from the model theory used to interpret the semantics, but there is no separation of syntax and semantics. There is no level of syntactic representation that is distinct from the syntax of the type theory. Thus, Montague Grammar has no separate level of semantic representation and Categorial Grammar has no separate level of syntactic representation.

Glue Semantics, by contrast, posits both a level of syntactic representation and a separate level of semantic representation. There is flexibility in the choice of both levels, however. The syntactic frameworks GLUE has been defined for include Lexical Functional Grammar (Dalrymple 1999, 2001), Lexicalized Tree Adjoining Grammar (Frank and van Genabith 2001), Head-driven Phrase Structure Grammar (Asudeh and Crouch 2002), Categorial Grammar (Asudeh and Crouch 2001), and Context

[10] The Kehler et al. (1995, 1999) account assumes a Montogovian treatment of intensional verbs. While our account does work for such a treatment, one of us has argued in separate work (Condoravdi et al. 2001) that the Montogovian approach is flawed, based on problems with existence predicates embedded under predicates such as *prevent*. Our account also works for the concept-based treatment of intensional verbs offered by Condoravdi et al. (2001).

[11] This strand within proof theory was in fact a key motivation behind the development of linear logic (Girard et al. 1989).

[12] Asher et al. (1997, 2001) offer an alternative theory of semantic parallelism, cast in Segmented Discourse Representation Theory; however, they do not provide identity criteria for their representations (for futher discussion, see Asudeh and Crouch to appear).

Free Grammars (Asudeh and Crouch 2001). The semantic framework can be any logic for semantics that supports the lambda calculus, such as Intensional Logic, Discourse Representation Theory (Dalrymple et al. 1999c), and Underspecified DRT (Crouch and van Genabith 1999, van Genabith and Crouch 1999). GLUE derivations are similar to categorial derivations and likewise possess identity criteria distinct from model theory, but they differ in being solely semantic, since there is a separate level of syntactic representation.

4.2 Defining Semantic Parallelism

In order to show that two normal form GLUE derivations are parallel, we need to establish that there is a homomorphism (a structure preserving map) between them. Given two structured objects G and H, with structral relations R_G and R_H holding between elements of G and H respectively, a homomorphism from G to H is a mapping $\mathcal{F}$ such that

(35) If $R_G(x, y)$ then $R_H(\mathcal{F}(x), \mathcal{F}(y))$

We need to decide on two things to measure semantic parallelism: (1) what kind mapping is f and between what kinds of objects in the derivations, and (2) what kind of structural relation, R, should be preserved.

We will define f in terms of the σ projection from f-structures to semantic structures (Dalrymple 2001). Suppose that we wish to compare the derivations arising from two elements of f-structure, p and q. Then in the first instance, we want $\mathcal{F}(p_\sigma) = q_\sigma$: that is, the semantic resources for these two elements should be made parallel. And then, recursively, we want $\mathcal{F}((p\ \text{SUBJ})_\sigma) = (q\ \text{SUBJ})_\sigma$, $\mathcal{F}((p\ \text{COMP})_\sigma) = (q\ \text{COMP})_\sigma$, $\mathcal{F}((p\ \text{COMP SUBJ})_\sigma) = (q\ \text{COMP SUBJ})_\sigma$, etc., for all cases where matching f-structure paths from the roots p and q exist. In other words, the $\mathcal{F}$ mapping pairs (atomic) semantic resources for syntactically matching elements.

In cases where there are mismatched syntactic elements (i.e. when there is a path in one f-structure, but no corresponding path in the other), then the $\mathcal{F}$ mapping is undefined and filters out unmatched elements from an assessment of parallelism. This allows us to compare derivations for sentences like *Every boy saw a girl* and *Every young man saw a woman*, where the extra adjective *young* is filtered out of the comparison.[13]

The relation R needs to be defined over atomic resources / linear logic propositions, since f maps atomic resources to atomic resources. Such a relation was defined in Crouch and van Genabith (1999). In a normal form GLUE derivation, one can identify the last point of occurrence of atomic semantic resources. These indicate the points in the derivation at which the corresponding syntactic elements make their final semantic contributions. An ordering, $\prec$, over these final semantic contributions provides a high level description of the topology of the derivation; as shown in Crouch and van Genabith (1999), this ordering can be used to express scope relations.

As an example of the resource ordering relation $\prec$, re-consider the following two derivations showing the alternative scopings of (12):

(36)

$$\dfrac{\dfrac{[e]^2 \qquad \dfrac{[g]^1 \qquad g \multimap e \multimap f}{e \multimap f}}{f} \qquad (e \multimap X) \multimap X}{\dfrac{\dfrac{f}{g \multimap f} \qquad (g \multimap Y) \multimap Y}{\underline{f}}}$$

$$\dfrac{\dfrac{[e]^2 \qquad \dfrac{[g]^1 \qquad g \multimap e \multimap f}{e \multimap f}}{f} \qquad (g \multimap Y) \multimap Y}{\dfrac{\dfrac{f}{e \multimap f} \qquad (e \multimap X) \multimap X}{\underline{f}}}$$

The final occurrences of the subject resource e, object resource g and sentential resource $\mathcal{F}$ are shown underlined. The tree structure of the derivations imposes a partial ordering over these final occurrences: $e \prec g \prec f$ (subject outscopes object) and $g \prec e \prec f$ (object outscopes subject) respectively.

[13] Here we are assuming an obvious notion of a syntactic match: subjects match subjects, objects match object, etc. Other ways of matching syntactic elements may be empirically motivated. In some cases we may want looser matches, e.g. objects can match obliques when no matching objects are present. We leave these, and many other questions, open.

In summary, to show that two derivations are semantically parallel, we need to do three things. First establish a pairing, $\mathcal{F}$, between atomic resources on the basis of their grammatical roles. Second, compute the $\prec$ ordering of these resources in the two derivations. Finally, ensure that precedence orderings coincide, so that if $a \prec b$ in derivation 1, then $\mathcal{F}(a) \prec \mathcal{F}(b)$ in derivation 2.

4.3 Scope parallelism in coordination

We now look at scope parallelism in coordination. The following example shows that the preferred reading is that in which the parallel quantifiers have parallel scopes, even if this goes against the general scope preferences of particular quantifiers:

(37) [**Context**: The animals really misbehaved last night.]
Every dog ate a bun and a cat gnawed each table leg.

The semantics of *eat* make the first conjunct plausible only with surface scope, i.e. *every* $\succ$ *a*. The parallelism between the conjuncts makes surface scope preferred in the second conjunct, too, despite the fact that *each* normally prefers wide scope.

(38)

$$\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{d} \multimap \mathbf{b} \multimap \mathbf{e} \quad [\mathbf{d}]^1}{\mathbf{b} \multimap \mathbf{e} \quad (\mathbf{b} \multimap \mathbf{X}) \multimap \mathbf{X}}}{\mathbf{e}}}{\mathbf{d} \multimap \mathbf{e} \qquad (\mathbf{d} \multimap \mathbf{Y}) \multimap \mathbf{Y}} \qquad \dfrac{\dfrac{\dfrac{\mathbf{c} \multimap \mathbf{l} \multimap \mathbf{g} \quad [\mathbf{c}]^2}{\mathbf{l} \multimap \mathbf{g} \quad (\mathbf{l} \multimap \mathbf{X}) \multimap \mathbf{X}}}{\mathbf{g}}}{\mathbf{c} \multimap \mathbf{g} \qquad (\mathbf{c} \multimap \mathbf{Y}) \multimap \mathbf{Y}}}{\dfrac{\mathbf{e} \qquad e \multimap (c \multimap f) \multimap (c \multimap f)}{(c \multimap f) \multimap (c \multimap f)}} \qquad \dfrac{\mathbf{g} \qquad g \multimap (c \multimap f)}{(c \multimap f)}}{\dfrac{c \qquad\qquad (c \multimap f)}{f}}$$

Computing parallelism proceeds as follows. First, we are comparing the derivation for clause g with that of clause e. We therefore concentrate on the sub-derivations (shown in bold) terminating at g and e. Within these, we have a pairing of subject and object resources, such that $\mathcal{F}(d) = c$ and $\mathcal{F}(b) = l$. For the two sub-derivations we have the resource ordering $b \prec d \prec e$ and $l \prec c \prec g$, i.e. $\mathcal{F}(b) \prec \mathcal{F}(d) \prec \mathcal{F}(e)$. Thus this derivation preserves semantic parallelism between the two conjuncts.

5 Violations to the Element Constraint

We have presented a theory of coordination as modification in Glue Semantics and shown how GLUE derivations can be used as a real level of semantic representation, which can be used in defining semantic parallelism. Next we will bring these two strands together and show how parallelism in GLUE can be interfaced with Kehler's (2002) theory of discourse parallelism to deal with violations of the Element Constraint (Grosu 1972, 1973), a subpart of the Coordinate Structure Constraint (Ross 1967). We conclude this section by comparing our results with the treatment of the Element Constraint in Montague Grammar, CCG, and TLG. We argue that these other theories are either too strict, allowing no exceptions to the Element Constraint, or too permissive, not capturing the Element Constraint at all.

5.1 The Coordinate Structure Constraint

The Coordinate Structure Constraint (CSC), one of Ross's (1967) island constraints, reads as follows:[14]

(39) **The Coordinate Structure Constraint**
In a coordinate structure, no conjunct may be moved, nor may any element contained in a conjunct be moved out of that conjunct.

[14] As the initial work on the CSC was done in Transformational Grammar, (39)–(41) make reference to movement, which does not make sense in non-transformational theories such as LFG. The constraints should be read as making appropriate restrictions on unbounded dependencies, no matter how these are dealt with.

Grosu (1972, 1973) subsequently pointed out that there are two parts to this constraint and that there are tests for distinguishing them. The parts are:

(40) **The Conjunct Constraint**: No conjunct of a coordinate structure may be moved.

(41) **The Element Constraint**: No element in a conjunct of a coordinate structure may be moved.

The key distinction between the two parts of the CSC, for our purposes, is that there are exceptions to the Element Constraint, but not to the Conjunct Constraint. Ross himself noticed certain exceptions with asymmetric coordination; Grosu points out that these are exceptions to the Element Constraint, but not to the Conjunct constraint:

- Element Constraint violations

  (42) I went to the store and bought some whiskey.

  a. This is the whiskey which I went to the store and bought.

  b. This is the store which I went to and bought some whiskey.

- No corresponding Conjunct Constraint violations

  (43) John is looking forward to going to the store and buying some whiskey.

  a. *What John is looking forward to and buying some whiskey is going to the store.

  b. *What John is looking forward to going to the store and is buying some whiskey.

The Conjunct Constraint violations should be compared to across-the-board (ATB) extraction. Ross noticed that the CSC is violable if an element is extracted from all conjuncts, or across the board:

(44) What John is looking forward to and excited about is buying some whiskey.

We will return to the Conjunct Constraint in section 5.5.

Kehler (2002) is the latest in a long line of literature that argues that the Element Constraint has principled exceptions (see Kehler (2002) for references). He notes that there are three main classes of exception to the Element Constraint and uses his theory of discourse coherence relations to explain these cases. His key insight is that *if discourse parallelism holds then the Element Constraint holds*. In other words, if the discourse coherence relation governing the conjuncts is *Parallel*, then there is either no movement out of conjuncts or there is across-the-board (i.e., parallel) movement.

Kehler's coherence relations (those relevant here) are defined as follows, with consequences for extraction possibilities as indicated:[15]

(45) **Parallel**: Infer $p(a_1, a_2, \ldots)$ from the assertion of S_1 and $p(b_1, b_2, \ldots)$ from the assertion of S_2, where for some property vector $\vec{q}$, $q_i(a_i)$ and $q_i(b_i)$ for all i.

Any extraction must be across the board

a. *What book did John buy and read the magazine?

b. What book did John buy and read?

(46) **Cause-Effect**

1. **Violated Expectation**: Infer P from the assertion of S_1 and Q from the assertion of S_2, where normally $P \rightarrow \neg Q$.
2. **Result**: Infer P from the assertion of S_1 and Q from the assertion of S_2, where normally $P \rightarrow Q$.

[15] S_1 and S_2 are respectively the first and second clauses being compared for discourse coherence; p_1 is a relation holding over a set of entities $a_1 \ldots a_n$ from S_1 and p_2 is a relation holding over a corresponding set of entities $b_1 \ldots b_n$ from S_2; q_i is a common or contrasting property for the i^{th} arguments (a_i and b_i), and the set of such properties is $\vec{q}$ (see Kehler 2002:ch. 2).

Extraction possible from initial (primary) clause

a. How much can you drink and still stay sober?

b. That's the stuff that the guys in the Caucasus drink and live to be a hundred.

c. That's the kind of firecracker that I set off and scared the neighbors.

(47) **Contiguity**

1. **Occasion (i)**
   Infer a change of state for a system of entities from S_1, inferring the final state for this system from S_2.
2. **Occasion (ii)**
   Infer a change of state for a system of entities from S_2, inferring the initial state for this system from S_1.

Extraction not necessary from "scene-setting" or "supporting" clauses

a. Here's the whiskey which I went to the store and bought.

Kehler (2002) thus gives a theory of discourse coherence that classifies and explains the exceptions to the Element Constraint systematically. In the next section we relate his discourse theory to Glue Semantics by relating proof parallelism to discourse parallelism.

5.2 Discourse Coherence and Derivational Parallelism

Kehler (2002) provides us with a theory of discourse coherence relations that explains the seeming exceptions to the Element Constraint. His theory essentially states that if discourse parallelism holds, then the Element Constraint holds. We have outlined a proof-theoretic notion of parallelism in Glue Semantics, as well as a GLUE theory of coordination as modification. Now we would like to relate Kehler's discourse theory to GLUE. Our central claim is that *if discourse parallelism holds, then proof parallelism holds*. The implication is read more usefully from right to left, and we name it the *Discourse-Proof Relation*, which sounds fancy but is just meant for us to be able to refer to it more easily:

(48) **Discourse-Proof Relation (DPR)**
If there is no proof parallelism then there is no discourse parallelism.

In other words, if parallelism does not hold in the GLUE proof for a coordination, the discourse relation cannot be parallelism. If the discourse relation is not parallelism, then it must be some other discourse relation; *if another discourse relation is compatible with the coordination, then the coordination is licensed, otherwise it is not licensed*. This allows the following understanding of the element constraint:

(49) **The Element Constraint**: when an element is extracted from some but not all conjuncts in a coordinate structure, there is no proof parallelism, and therefore no discourse parallelism.

With (48) and (49) in hand, we can now look at how our theory handles exceptions to the Element Constraint by appealing to Kehler's (2002) theory where proof parallelism does not hold. We need to consider examples like (45a), where the Element Constraint is violated and ungrammaticality results, the across-the-board case (45b), and the cases in (46) and (47) where the Element Constraint is violated but no ungrammaticality results.

Before turning to the exposition of the relevant examples we need to make clear one important caveat. Throughout this paper we have been assuming an LFG syntax. In the standard LFG treatment of unbounded dependencies and coordination (Kaplan and Maxwell 1988, Dalrymple 2001), a strict version of the Element Constraint is upheld, allowing no unbounded dependencies that terminate in some but not all of the conjuncts of a coordination. This is due to the way that inside-out functional uncertainties (which are used to handle unbounded dependencies) interact with sets. Grammatical functions

are standardly defined as distributive features (Dalrymple and Kaplan 2000, Dalrymple 2001) and an inside-out path written in terms of grammatical functions will therefore distribute to all members of a coordination set. One solution is to make grammatical functions nondistributive features instead. We cannot explore the consequences of such a solution here. However, as we stressed above, GLUE can be coupled with a variety of syntactic frameworks and as such this contingent fact about LFG does not impact our GLUE account. The account could be paired with another syntactic theory instead; the only desideratum would be that the syntax have coordination rules like (33), which capture the Conjunct Constraint, as we will see in section 5.5 below.

5.3 The Element Constraint and Derivational Parallelism

5.3.1 Non-Parallel Extraction

Consider (45a), repeated as (50) below, which violates the Element Constraint, since the object of *buy* has been extracted, but the object of *read* has not. The (perfectly valid) GLUE derivation for this example is shown in (51), with the resource f corresponding to the coordination and all other resources named mnemonically as usual.

(50) *What book did John buy and read the magazine?

(51)
$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{w}\multimap\mathbf{j}\multimap\mathbf{b}\quad[\mathbf{w}]^1}{\mathbf{j}\multimap\underline{\mathbf{b}}}\quad(j\multimap b)\multimap(c\multimap j\multimap f)\multimap(c\multimap j\multimap f)}{(c\multimap j\multimap f)\multimap(c\multimap j\multimap f)}\quad\dfrac{\dfrac{\dfrac{\mathbf{m}\multimap\mathbf{j}\multimap\mathbf{r}\quad[\mathbf{m}]^2}{\mathbf{j}\multimap\mathbf{r}}\quad[\mathbf{j}]^3}{\mathbf{r}}\,{}^{\multimap_{\mathcal{I},2}}\quad\dfrac{(\mathbf{m}\multimap\mathbf{Y})\multimap\mathbf{Y}}{Y=r}}{\mathbf{r}}\,{}^{\multimap_{\mathcal{I},3}}\quad\dfrac{\mathbf{j}\multimap\underline{\mathbf{r}}\qquad(j\multimap r)\multimap(c\multimap j\multimap f)}{(c\multimap j\multimap f)}}{c\qquad(c\multimap j\multimap f)}}{\dfrac{\underline{j}\qquad j\multimap f}{f}}\,{}^{\multimap_{\mathcal{I},1}}\quad\dfrac{\underline{w}\multimap f\qquad(w\multimap X)\multimap X}{X=f}}{f}$$

The sub-derivations for the two conjunct resources, b and r, are shown in bold, and terminate at the conclusions $j \multimap b$ and $j \multimap r$ respectively. The mapping of subjects to subjects, objects to objects, etc., gives the following pairings of resources:

(52) $\mathcal{F}(b) = r,\ \mathcal{F}(j) = j,\ \mathcal{F}(w) = m$

The final occurrences of each of these resources are shown in red and underlined; these induce the non-parallel orderings

(53) $b \prec j \prec w$
$m \prec r \prec j$ (i.e. not $\mathcal{F}(b) \prec \mathcal{F}(j) \prec \mathcal{F}(w)$)

Thus, although the GLUE derivation is logically valid, it does not preserve parallelism between the conjuncts.

According to the DPR, in (48) above, since there is no proof parallelism there is no discourse parallelism. Therefore, the discourse relation cannot be *Parallel*. However, none of the other discourse relations in (46) and (47) are compatible with (50) either. The *Cause-Effect* relations *Violated Expectation* and *Result* cannot relate the conjuncts in (50), since there is no implicational relationship between the assertion of *John bought x* and *John read the magazine*. The two different kinds of *Contiguity* relation, *Occasion (i)* and *Occasion (ii)* also do not hold: *John bought x* does not set the scene for or support *John read the magazine* (more specifically, there is no change of state from S_1 to S_2, or vice versa).

5.3.2 Across-the-Board Extraction

The situation for (50) contrasts with the across-the-board extraction case (45b) and with the cases in (46a–c) and (47a). In the latter cases, there is no proof parallelism either, and therefore no discourse parallelism, but the relevant discourse relation licenses each case. As for across-the-board extraction, this does lead to derivational parallelism, as shown in proof (55) for example (54):

(54) What book did John buy and read?

(55)
$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{w}\multimap\mathbf{j}\multimap\underline{\mathbf{b}} \quad (w\multimap j\multimap b)\multimap(c\multimap w\multimap j\multimap f)\multimap(c\multimap w\multimap j\multimap f)}{(c\multimap w\multimap j\multimap f)\multimap(c\multimap w\multimap j\multimap f)} \quad \dfrac{\mathbf{w}\multimap\mathbf{j}\multimap\underline{\mathbf{r}} \quad (w\multimap j\multimap r)\multimap(c\multimap w\multimap j\multimap f)}{c\multimap w\multimap j\multimap f}}{c \qquad c\multimap w\multimap j\multimap f} \quad [w]^1}{w\multimap j\multimap f}}{\underline{j} \qquad j\multimap f}}{f}}{\dfrac{\underline{w}\multimap f \quad \multimap_{\mathcal{I},1} \quad (w\multimap X)\multimap X}{f}\; X=f}$$

For this derivation we have the resource mappings and the parallel orderings

(56) $\mathcal{F}(b)=r,\ \ \mathcal{F}(j)=j,\ \ \mathcal{F}(w)=w$

(57) $b\prec j\prec w,\quad r\prec j\prec w$

Across the board extraction preserves derivational parallelism.

5.4 Syntactic and Semantic Parallelism

Lack of discourse parallelism as per Kehler's (2002) theory is indeed reflected by lack of GLUE derivational parallelism. However, one could arguably also point to a more obvious lack of syntactic parallelism in these cases. That is, there will be a difference between the syntactic structure corresponding to a conjunct with no element extracted and the structure for a conjunct with an extracted element. In LFG terms, for example, the extracted GF will be shared with a discourse function, TOPIC or FOCUS, while the parallel unextracted GF will not.

We feel that there are still compelling reasons for preferring the GLUE parallelism account of the Element Constraint to an alternative account based on purely syntactic parallelism. First, GLUE is framework-independent and can be combined with a variety of syntactic frameworks, as noted above. This makes the GLUE treatment more general than a syntactic treatment cast in some specific framework. For example, the grammatical functions used in LFG do not have clear correspondences in most other syntactic frameworks. Second, proof parallelism in GLUE generalizes to scope parallelism, as we showed above, and to ellipsis parallelism, as we have shown elsewhere (Asudeh and Crouch to appear). Thus, the GLUE parallelism theory has potential as a general theory of parallelism, unlike a theory of syntactic parallelism devised just for coordination. Third, the close relationship between GLUE, Combinatory Categorial Grammar, and Type-Logical Grammar allows us to compare the GLUE account of the Element Constraint to the standing of this constraint in these other theories. We turn to this comparison now.

5.5 The Goldilocks Effect

We have noted that a desirable situation is one where the Conjunct Constraint holds generally, but where the Element Constraint is allowed certain systematic exceptions governed by discourse parallelism. In our theory the Conjunct Constraint follows from the syntactic rule for forming coordinate structures, shown for VP-coordination in (21) above. The general form of the syntactic portion of the rule, which was shown in (33) above, is repeated here:

(58)
$$\begin{array}{lccc} \mathrm{X}\rightarrow & \mathrm{X}^{+} & \textsc{conj} & \mathrm{X} \\ & \downarrow\ \in\ \uparrow & \uparrow=\downarrow & \downarrow\ \in\ \uparrow \end{array}$$

As the category-specific rules for coordination will each be coordinating one or more conjuncts on the left with a conjunct on the right, it does not license structures that are missing conjuncts. This is not new; a similar rule is also a feature of Generalized Phrase Structure Grammar (Gazdar et al. 1985) and Combinatory Categorial Grammar (Steedman 1985), as well as the earlier account of Dougherty (1970). The rule is also fairly neutral with respect to syntactic formalism: most syntactic approaches could readily accommodate such a rule.

However, the Element Constraint does not *necessarily* follow from this rule. If the categories that the rule conjoins register extraction (as in GPSG, with slashes, or CCG, with functors), then the strict version of the Element Constraint does follow, because we would be coordinating a slashed category, for example, with an unslashed one. This is undesirable, since it is too strong for the Element Constraint to hold in general; exceptions must be allowed.

By contrast, the Element Constraint does not hold at all in Type-Logical Grammar, without further stipulation. If a conjunct has an extracted element, it is always possible to reduce the categories of all the other conjuncts in the coordination to the category of the conjunct with the missing element. This is done by hypothetical reasoning. The assumptions can be discharged after the coordination has been carried out. In fact, the same thing is true of our GLUE theory of coordination as modification if it is not related to a discourse theory like Kehler's (2002). Given the proof-theoretic similarity between TLG and GLUE, it is quite feasible that our parallelism account could be ported to TLG. However, as it stands TLG does not capture the Element Constraint at all. This is too weak, as not all exceptions to the Element Constraint result in grammatical outputs and there are generalizations about the grammatical exceptions that would be missed by allowing the Element Constraint to fail in general.

Thus, we have what we call the Goldilocks Effect. One brand of Categorial Grammar (CCG), as well as Montague Grammar and GPSG, is too strong: it allows no exceptions to the Element Constraint. The other brand of Categorial Grammar (TLG) is too weak: all exceptions to the Element Constraint are permitted. Perhaps Glue Semantics is just right.

6 Conclusion

We set out do three things in this paper. The first was to develop the GLUE account of coordination as modification, which was accomplished in section 3. The resulting theory not only preserves resource sensitivity, it is also able to deal with n-ary coordination and does not overgenerate in the manner of certain Categorial Grammar treatments, since GLUE preserves a notion of constituency through its pairing with a syntactic theory. The second goal was to present a theory of proof parallelism and to show how this theory extends to coordination; this was presented in section 4. The third goal, which we ended with in section 5, was to interface the theory of proof parallelism with Kehler's (2002) theory of discourse coherence, in order to capture the Element Constraint in general while allowing a systematic class of exceptions.

References

Asher, Nicholas, Daniel Hardt, and Joan Busquets (1997). Discourse parallelism, scope, and ellipsis. In Aaron Lawson (ed.), *Proceedings of SALT VII*.

Asher, Nicholas, Daniel Hardt, and Joan Busquets (2001). Discoure parallelism, ellipsis, and ambiguity. *Journal of Semantics, 18*.

Asudeh, Ash (2000). Functional identity and resource-sensitivity in control. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG00 conference*, Stanford, CA. CSLI Publications.

Asudeh, Ash (2002). A resource-sensitive semantics for equi and raising. In David Beaver, Stefan Kaufmann, Brady Clark, and Luis Casillas (eds.), *The construction of meaning*. Stanford, CA: CSLI Publications.

Asudeh, Ash (in progress). *Resumption as resource management*. Ph.D. thesis, Stanford University, Stanford, CA.

Asudeh, Ash, and Richard Crouch (2001). Glue semantics: A general theory of meaning composition. Talk given at Stanford Semantics Fest 2, March 16, 2001.

Asudeh, Ash, and Richard Crouch (2002). Glue semantics for HPSG. In Frank van Eynde, Lars Hellan, and Dorothee Beermann (eds.), *Proceedings of the 8th international HPSG conference*, Stanford, CA. CSLI Publications.

Asudeh, Ash, and Richard Crouch (to appear). Derivational parallelism and ellipsis parallelism. In Line Mikkelsen and Christopher Potts (eds.), *WCCFL 21 proceedings*, (pp. 1–14). Somerville, MA: Cascadilla Press.

Carpenter, Robert (1997). *Type-logical semantics*. Cambridge, MA: MIT Press.

Condoravdi, Cleo, Richard Crouch, Martin van den Berg, John Everett, Valeria de Paiva, Reinhard Stolle, and Daniel Bobrow (2001). Preventing existence. In *Proceedings of Formal Ontologies in Information Systems (FOIS)*, Maine.

Crouch, Richard (1999). Ellipsis and glue languages. In Shalom Lappin and Elabbas Benmamoun (eds.), *Fragments: Studies in ellipsis and gapping*. Oxford: Oxford University Press.

Crouch, Richard, and Josef van Genabith (1999). Context change, underspecification, and the structure of glue language derivations. In Dalrymple (1999), (pp. 117–189).

Dalrymple, Mary (ed.) (1999). *Semantics and syntax in Lexical Functional Grammar: The resource logic approach*. Cambridge, MA: MIT Press.

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

Dalrymple, Mary, Vaneet Gupta, John Lamping, and Vijay Saraswat (1999a). Relating resource-based semantics to categorial semantics. In Dalrymple (1999), (pp. 261–280).

Dalrymple, Mary, and Ron Kaplan (2000). Feature indeterminacy and feature resolution in description-based syntax. *Language*, *74*, 759–798.

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

Dalrymple, Mary, John Lamping, Fernando Pereira, and Vijay Saraswat (1999b). Overview and introduction. In Dalrymple (1999), (pp. 1–38).

Dalrymple, Mary, John Lamping, Fernando Pereira, and Vijay Saraswat (1999c). Quantification, anaphora, and intensionality. In Dalrymple (1999), (pp. 39–89).

Dalrymple, Mary, John Lamping, and Vijay Saraswat (1993). LFG semantics via constraints. In *Proceedings of the sixth meeting of the European ACL*, (pp. 97–105), University of Utrecht. European Chapter of the Association for Computational Linguistics.

Dougherty, Ray (1970). A grammar of coordinate conjoined structures. *Language*, *46*, 850–898.

Emms, Martin (1990). Polymorphic quantifiers. In *Studies in Categorial Grammar*, no. 5 in Edinburgh Working Papers in Cognitive Science, (pp. 65–111). Edinburgh: Centre for Cognitive Science.

Frank, Anette, and Josef van Genabith (2001). LL-based semantics construction for LTAG — and what it teaches us about the relation between LFG and LTAG. In *Proceedings of the LFG01 conference*. CSLI Publications.

Gazdar, Gerald (1980). A cross-categorial semantics for coordination. *Linguistics and Philosophy, 3*, 407–410.

Gazdar, Gerald (1981). Unbounded dependencies and coordinate structure. *Linguistic Inquiry, 12*, 155–184.

Gazdar, Gerald (1982). Phrase structure grammar. In Pauline Jacobson and Geoffrey K. Pullum (eds.), *The nature of syntactic representation*, vol. 15 of *Synthese Language Library*, (pp. 131–186). Dordrecht: Reidel.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, and Ivan A. Sag (1985). *Generalized phrase structure grammar*. Cambridge, MA: Harvard University Press.

Girard, Jean-Yves (1987). Linear logic. *Theoretical Computer Science, 50*, 1–102.

Girard, Jean-Yves, Paul Taylor, and Yves Lafont (1989). *Proofs and types*. Cambridge: Cambridge University Press.

Grosu, Alexander (1972). *The strategic content of island constraints*. Ph.D. thesis, Ohio State University, Columbus, OH. Reprinted as Working Papers in Linguistics no. 13, OSU.

Grosu, Alexander (1973). On the nonunitary nature of the Coordinate Structure Constraint. *Linguistic Inquiry, 4*, 88–92.

Howard, William A. (1980). The formulae-as-types notion of construction. In Jonathan P. Seldin and J. Roger Hindley (eds.), *To H.B. Curry: Essays on combinatory logic, lambda calculus and formalism*, (pp. 479–490). London: Academic press. Often cited as Howard (1969).

Hudson, Richard (1976). Conjunction reduction, gapping, and right-node raising. *Language, 52*, 535–562.

Kaplan, Ronald M., and John T. Maxwell, III (1988). Constituent coordination in Lexical-Functional Grammar. In *Proceedings of COLING-88*, vol. 1, Budapest. Reprinted in Dalrymple et al. (1995:199–214).

Keenan, Edward L., and Leonard M. Faltz (1985). *Boolean semantics for natural language*. Dordrecht: Reidel.

Kehler, Andrew (2002). *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI Publications.

Kehler, Andrew, Mary Dalrymple, John Lamping, and Vijay Saraswat (1995). The semantics of resource-sharing in Lexical-Functional Grammar. In *Proceedings of the 1995 meeting of the European Chapter of the Association for Computational Linguistics*, Dublin, Ireland.

Kehler, Andrew, Mary Dalrymple, John Lamping, and Vijay Saraswat (1999). Resource sharing in glue language semantics. In Dalrymple (1999), (pp. 191–208).

Maxwell, John T., III, and Christopher D. Manning (1996). A theory of non-constituent coordination based on finite-state rules. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG96 conference*, Stanford, CA. CSLI Publications.

Montague, Richard (1970). English as a formal language. In Bruno Visentini et al. (eds.), *Linguaggi nella societá e nella tecnica*, (pp. 189–224). Milan: Edizioni di Communità. Reprinted in Montague (1974:188–221).

Montague, Richard (1973). The proper treatment of quantification in ordinary English. In Jakko Hintikka, Julian Moravcsik, and Patrick Suppes (eds.), *Approaches to language*, (pp. 221–242). Dordrecht: D. Reidel. Reprinted in Montague (1974:247–270).

Montague, Richard (1974). *Formal philosophy: Selected papers of Richard Montague*. New Haven: Yale University Press. Edited and with an introduction by Richmond H. Thomason.

Morrill, Glyn (1994). *Type Logical Grammar*. Dordrecht: Kluwer.

Partee, Barbara (1970). Negation, conjunction, and quantifiers: Syntax vs. semantics. *Foundations of Language*, *6*, 153–165.

Partee, Barbara, and Mats Rooth (1983). Generalized conjunction and type ambiguity. In Rainer Bäuerle, Christoph Schwarze, and Arnim von Stechow (eds.), *Meaning, use, and interpretation of language*. Berlin: Walter De Gruyter.

Prawitz, Dag (1965). *Natural deduction: A proof-theoretical study*. Stockholm: Almquist and Wiksell.

Ross, John R. (1967). *Constraints on variables in syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

Steedman, Mark (1985). Dependency and coordination in the grammar of Dutch and English. *Language*, *61*, 523–568.

Steedman, Mark (1989). Constituency and coordination in a combinatory grammar. In Mark Baltin and Anthony Kroch (eds.), *Alternative conceptions of phrase structure*, (pp. 201–231). Chicago: University of Chicago Press.

Steedman, Mark (1990). Gapping as constituent coordination. *Linguistics and Philosophy*, *13*, 207–263.

Steedman, Mark (2000). *The syntactic process*. Cambridge, MA: MIT Press.

van Genabith, Josef, and Richard Crouch (1999). How to glue a donkey to an f-structure. In Harry Bunt and Reinhard Muskens (eds.), *Computing meaning*, vol. 1. Dordrecht: Kluwer.

VP-Chaining in Oriya

Dorothee A.Beermann and Lars Hellan

NTNU, Trondheim, Norway

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

http:/csli-publications.stanford.edu/

Abstract

VP-chaining in Oriya (an Indo-Aryan language spoken in Orissa, India) is argued to be left-recursive VP- adjunction in c-structure with a corresponding f-structure displaying recursive embedding under the attribute ADJUNCT (for 'chain-adjunct'). Under VP-chaining 'subject sharing' is obligatory and analyzable in terms of functional control ('token sharing'), 'object sharing' is not only optional but in addition, does not require case or GF uniformity, and can skip an intervening intransitive verb. It is therefore argued that 'object sharing' is a form of anaphoric control ('reference sharing'). Passivization under VP-chaining needs to apply to all verbs in the VP-chain. The uniformity of diathesis pattern is accounted for via the adoption of an LFG architecture (from Butt, Dalrymple and Frank 1997) according to which c-structure maps directly to a (rgument) structure, and a-structure in turn to f-structure. This analysis allows us to 'drive' all the aspects of VP-chaining mentioned (apart from object sharing) from an annotation of c-structure rules.

VP-Chaining in Oriya*

Introduction

It is often suggested that Serial Verb Constructions (SVC) originate from paratactic constructions (e.g., Payne (1985), Foley and Olson (1985), Andrews & Manning (1999)). The relation between coordination and verb serialization seems particularly obvious for VP-chaining in the Indo-Aryan language Oriya, where a series of VPs describes a series of consecutive events. However, VP-chains in Oriya possess properties unlike those of coordination. First of all, under VP-chaining, there is an asymmetry in the morphology of the verbs involved, not expected to appear in coordination: A final finite verb combines with a series of dependent verb-forms, as in (1) (here and throughout, we highlight the verbs of the sequence by underscore):

(1) Raajaa maachha-Te kiN-i keLaa-i bhaaj-i khaa-il-aa
Raajaa fish-a buy-dM clean-dM fry-dM eat- PAST 3rd, sg
'Having bought, cleaned and fried a fish, Raajaa ate it.'

The verb forms *kiNi*, *KeLaai* and *bhaaji* are here marked as non-finite, dependent verbs by the suffix *-i* (glossed as 'd(ependent) M(arker)'),[1] while the last verb *khaailaa* is finite and marked for tense and agreement.

Secondly, NP-extraction out of one of the conjuncts is grammatical, violating the across-the-board constraint (otherwise valid in Oriya; together with the other island constraints). In (2) the NP *maachha* ('fish') has been extracted out of the first VP headed by the verb *bhaaj* ('fry').

(2)
sei maachha-Ti-ku mun bhaaj-i bhaata raandh-i bhaata o maachha khaa-il-i
that fish-Def-Acc I fry-dM rice cook-dM rice and fish eat-Past1st,sg
'It was that fish I fried, then (I) cooked the rice and ate the rice and the fish.'

Finally, unlike under coordination, the non-finite VPs can be fronted together, leaving the tensed verb behind, as in (3):

* The authors would like to thank Kalyanamalini Sahoo with whom they published an earlier paper on argument sharing, also based on Oriya (Beermann et.al. 2001). Kalyani is a native speaker of Oriya, and without her help also this paper would not have been possible. We further would like to thank the participants of the Spring 2001 seminar on 'Verb Serialization' at the Linguistic Department at Stanford University, and in particular Joan Bresnan, who took the time to discuss with us an early version of this paper. We also would like to thank Miriam Butt for discussion and Mary Dalrymple for carefully reading through this paper. The discussions we had have provided us with new and interesting insights on control, referential binding and complex verbal constructions. Not all of these ideas could be integrated in the present version of this paper. Errors and misunderstandings found in the text are the authors' responsibility alone.

[1] The suffix *-i* is in form identical to the perfective marker. That these two morphemes are different in meaning can be seen by the following example taken from Sahoo (2001):

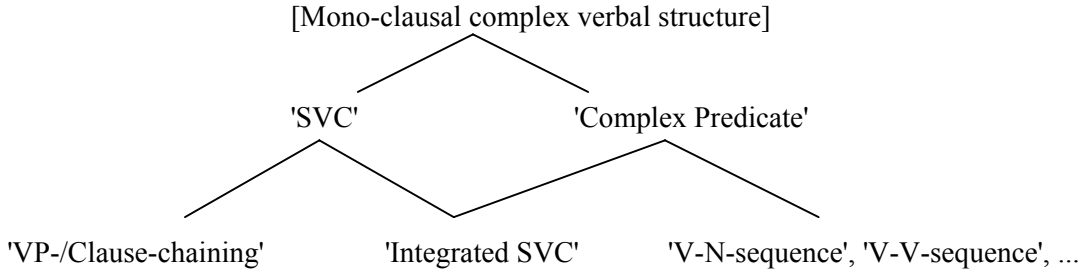
(i) se has-i has-i kathaa kah-u-th-il-aa
he laugh-dM laugh-dM tale talk-PROG-PAST-3rd,sg
'He was laughing and talking.'

The verb 'laugh' is duplicated, to express the progressive aspect, while the dM *–i* , pertinent to VP-chaining, is maintained.

(3) haata dho-i bhaata khaa-i mun skul-ku ga-li
hand wash rice eat I school-PP go- PAST 1st sg
'Having washed my hand and eaten rice, I went to school.'

VP-chaining in Oriya resembles in many respects what has been called 'Clause-chaining' in the context of the West-African languages (e.g., Osam 1994). For Akan in particular, Osam contrasts clause-chaining with 'Integrated SVCs', a dichotomy also recognized for many other languages, and properties of which are summarized by Kroeger (2001). Both constructions are commonly categorized as SVCs, and we follow this convention here; at the same time, 'Integrated SVCs' seem to fall under the notion 'Complex Predicates' as commonly used, which suggests a rough typology of notions as depicted in figure 1:

Figure 1. *Rough typology of notions referring to mono-clausal complex verbal structures*



The variant of SVCs that has received the main attention in the LFG literature is one instantiating 'Integrated SVCs'. This type is discussed by Niño (1997) and Bodomo(1997) for Dagaare. Niño, as summarized by Sells (2001), suggests that multiple verbal predicate constructions are analyzable in terms of multiple c-structure exponents of one complex predicate nucleus. The idea is that a nucleus of the type 'pred$_1$- …-pred$_n$' is formed by a process called 'composition of predicates' (cf. Alsina (1993, 1997), Butt (1993, 1995, 1997)). (4) gives a Dagaare example of an Integrated SVC , taken from Bodomo (1997), and the f-structure in figure 2 illustrates the basic idea of predicate composition:

(4) 0 da zo-ro gɛ-rɛ wuo-ro la haane.
3sg PAST run-IMPERF go-IMPERF collect-IMPERF FACT berries
He/she was always running there collecting berries.[2]

[2] Different from at least one of the meanings suggested by the English gloss of (4) given by Bodomo(1997) where she collects berries after having run, the f-structure shown in figure 2 suggests as PRED value a 'runningly' performed collecting event of berries. This is also the meaning suggested by the gloss given by Niño (1997) for the same example. The verbs *zo* and *gaa* are interpreted as semantically modifying the collecting event.

Figure 2. *'Predicate Composition' illustrated with an example from Dagaare (Bodomo (1997))*

f-structure of (4):

$$\begin{bmatrix} \text{PRED} & \text{'zo- gaa- wuo'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{NUMBER} & \text{SG} \\ \text{PERSON} & 3 \end{bmatrix} \\ \text{OBJ} & [\text{PRED 'haana'}] \\ \text{TENSE} & \text{past} \\ \text{ASPECT} & \text{prog} \end{bmatrix}$$

Building on data from Oriya, the intention of this paper is to clearer distinguish VP-chaining/Clause chaining from Integrated SVCs/complex predicate formation, and in particular to present an analysis which allows multiple verbal predicates to enter into a mono-clausal structure without necessarily forming a complex predicate nucleus.

Following T. Mohanan (1997) in her analysis of Hindi, we will assume that Oriya VP-chains involve a combination of independent predicates, associated with a single-headed mono-clausal c-structure. Employing the formalism suggested by Butt, Dalrymple and Frank (1997) and earlier work by Kaplan (1995), we will argue for a representation of VP-chains that is factored out over three parallel representations. We would like to argue that, at a-structure and f-structure, VP-chains build a structure of independent predicates, which corresponds at c-structure to a series of adjuncts that modify a single finite verbal head. Following Butt et al., a–structure is directly projected from c-structure nodes via a so-called α - function which allows us to directly construct diathesis alternations from morpho-syntactic information.

The patterns of argument sharing that are characteristic of VP-chains, will be represented at f-structure. For subject sharing we will have to extend the notion of 'functional control', from a lexically-induced concept (Bresnan 2001) to a constructionally-induced concept. Object sharing, on the other hand, will be described in terms of anaphoric control (Bresnan 1982a, 2001, Dalrymple 2001), thus accounting for the different properties of subject and object sharing.

Figure 3 schematically indicates the c-structure adjunction configuration assumed, and figure 4 the corresponding representations in a- and f-structure. Again following Butt et al., we describe the correspondence between a-structure and f-structure in terms of the λ-function:

Figure 3. *Schematic c-structure representing VP-Chains as a recursive process of left adjunction.*

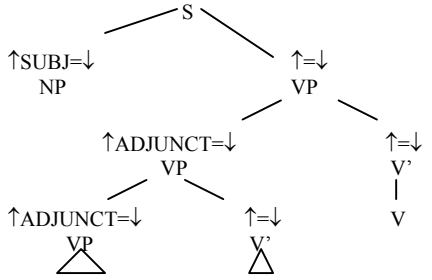
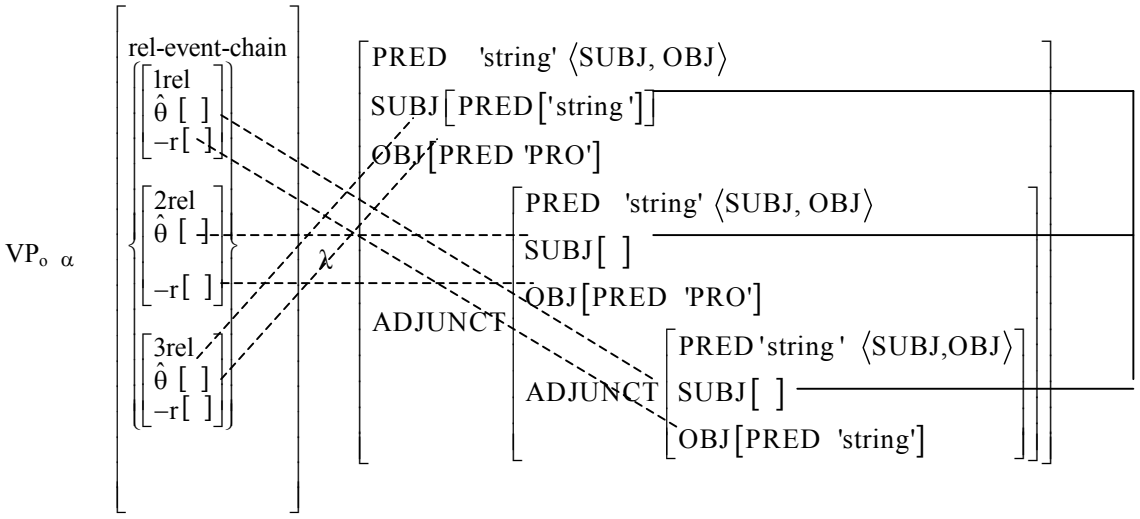


Figure 4. *Schematic representation of a VP-chain of three transitive verbs with subject- and object- sharing.*



(A-structure will reflect not only the argument structures associated with each predication, but also the temporal relationships between the events expressed by the predications. In figure 4, the label 'rel-event-chain' refers to this aspect of the representation, and as a notational convenience, we use the label '1rel' on the predication representing the temporally first event, '2rel' the temporally second, and so forth. See (20) below for an approximation to a more adequate representation of this feature of the semantics).

Details of these proposals will be developed in the following. We first provide the most crucial data.

2. Main data

Subject sharing

Subject sharing is obligatory in VP-chaining. Moreover the serialization of two predicates, one requiring a dative subject, the other a nominative subject, is ungrammatical, as the contrast between the grammatical *coordination* in (5) and the ungrammatical *serialization* in (6) shows:

(5) mote jara he-uchh-i aau mun ousadha khaa-u-chh-i
I-DAT fever happen-PROG-AUX-1st,sg and I medicine eat-PROG-AUX-1st,sg
' I am suffering from fever and I eat medicine.'

(6) * mote jara ho-i ousadha khaa-il-i.
I fever happen-dM medicine eat-PAST-1st,sg.
'Having had fever I took medicine.'

Object sharing

Different from subject sharing, object sharing under VP-chaining is optional. That is to say that each verb in the sequence maintains its independent object domain, as illustrated in (3), repeated, (7) and (8) below. In the VP-chain (3) none of the objects are shared. In (7) the verb

de ('give') is preceded by both of its objects, of which the indirect object (given in italic) is a shared object relative to the following two verbs. The first one of those following verbs, the verb *ne* ('take'), is separately modified by a PP, while the final verb *khuaa* ('feed') selects its direct object independently.

(3) **haata** dho-i **bhaata** khaa-i mun skul-ku ga-l-i
hand wash rice eat I school-PP go- PAST 1st sg
'Having washed my hand and eaten rice, I went to school.'

(7)
mun ***gariba pilaaTi-ku*** **lugaa** de-i **hotellku** ne-i **piThaa** khuaa-il-i
I poor child-DAT cloth give-dM hotel take-dM pancake feed-PAST-1st,sg
'Having given the child clothes, taking him to a hotel, I fed him pancake.'

(8) illustrates that adverbs need not take scope over the whole sentence, but may well modify chained verbs individually:[3]

(8) bilei-Taa aakhi pichhuLaake **maachhabhajaa-Taa** ne-i
cat-the an eyeblink's time fish cutlet-the take-dM
baaDipaTa-ku jaa-i bhaari majaare khaa-il-aa
backyard go-dM happily eat-PAST-3rd,sg
'Having taken the fish cutlet in an eyeblink's time, the cat went to the backyard and ate it happily.'

In short, in the construction at hand, each verb occurs inside the VP structure that it is normally associated with, possibly sharing one or two of its objects[4] with other verbs in the VP-chain.[5] Important is that object sharing can obtain even when the shared object realizes different grammatical functions relative to the verbs in the sharing sequence. The shared argument 'child' in (7) is an indirect object relative to the first and third verb, but a direct object relative to the second. In (9) below, the shared object *ghaa* ('wound') is a direct object relative to the first verb *dho* ('wash') and as such it is suffixed by the objective case marker *–ku*. Relative to the second verb *lage* ('apply') it serves in an oblique function. It indicates the location to which

---

[3] When they do take scope over the whole sentence, adverbs may occur either sentence initially, as in (i), or in a sentence internal position, as in (ii):
(i)
kaali raati-re mun maachha-Te kiN-i keLaa-i bhaaj-i khaa-il-i
yesterday night-PP I fish-a buy-dM clean-dM fry-dM eat- PAST 1st sg
'Last night, having bought, cleaned and fried a fish, I ate it.'
(ii)
mun tarakarri-Taa-ku aaji sakaaLe frizru baahaara kar-i garam kar-i khaa-il-i.
I curry-DEF-CASE today morning fridge out do-dM hot do-dM eat-PAST-1st,sg
'This moring, having taken the curry out of the fridge, I heated it and ate it.'

[4] An example of two shared objects is given in (i):
(i) mun mak-ku$_i$ saaDi-Taa dekhaa-i de-li
I mother sari-the show-dM, give-Past 1st,sg
'Having shown my mother the sari and gave it to somebody else/her.'

[5] Object sharing is not necessarily sequence initial as the following example shows:
(i) subaasa sakaaLu uTh-i jaLakhiaa khaa-i bilaru jaa-i TamaaTo toL-i bikk-il-aa
Subas morning-PP wake up breakfast eat farm go tomato pluck sell-PAST-3rd,sg
'Having got up in the morning, Subas had breakfast, went to the farm, plucked tomato, and sold them.'
For discussion see Beermann, Sahoo and Hellan (2001)

the medicine is applied. If overtly realized it would therefore be marked by the post-position -*re*.

(9)
mun ghaa-Thaa-ku dho-i ousadha lage-i byaandaze ka-li.
I wound-DEF-CASE wash-dM medicine apply-dM bandage do-PAST-1st,sg

In summary, shared objects can be functionally and morphologically distinct, while shared subjects need to be also morphologically identical.

Finally, object arguments can be shared across intransitive verbs, as illustrated by (8) above, while, in most cases, they cannot be shared across a transitive verb, as indicated in (10) below:

(10) #mun aambaTaa ne-i bhaata khaa-i kaaT-i khaa-il-i
I mango-the take-dM rice eat-dM cut-dM eat- PAST 1st sg
'Having taken the mango, I ate rice, then cut the mango and ate it.' [intended meaning]

Passivization
An intriguing property of VP-chains is that under passivization, *all* verbs of the sequence have to passivize. Since passivization does generally not apply to intransitives, 'chain passivization' becomes impossible when one of the verbs in the sequence is intransitive; hence, e.g., (8) cannot be passivized. (11) illustrates a grammatical passivization:

(11) maachha-Ti bhaj-aa jaa-i khi-aa-ga-laa
fish-the fry-PRTP go-Dm eat-PRTP-go-PAST.3rd
'Having been fried, the fish was eaten

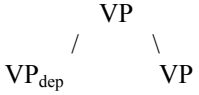
Notice that the passive is expressed by adding the 'light-verb' *jibaa* 'to go'. When associated with the finite verb, *jibaa* is realized as a suffix that is followed by the number/person inflection, while when associated with the dependent verb forms, it is perceived (and written) as an independent word suffixed by the dependent marker, a difference that we will ignore in the following.

3. The Analysis

3.1. The ADJUNCT status of the dependent VP

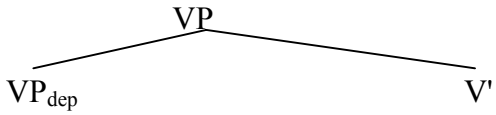
As indicated in figure 3, the series of VPs is conceived as a binary left-branching structure, with the dependent VP in each case adjoined to the matrix V', giving the constellation in figure 5 as the recursive minimal tree configuration in c-structure:[6]

[6] Other types of adjuncts in the verbal projection may conceivably also occur in the position of VP_{dep}, but we offer no considerations on this here. A technical reason why we don't want, e.g., the constellation



is to forestall structures where VP_{deps} originate under the rightmost VP - we assume strict left-recursiveness of the VP-chain construction.

Figure 5.



Through annotation as indicated in the phrase structure rule in (12), the functional status of VP_{dep} is that of a *chain-adjunct* relative to the head V' (for convenience, we here and throughout write only ADJUNCT as the f-structure attribute label, since no other types of adjuncts will be discussed):

(12) VP -> VP_{dep} (↑ADJUNCT=↓) V' (↑=↓)

Chain-adjuncts are headed by dependent verbs. A dependent verb is one of the many lexically induced verb forms of Oriya (for a discussion of Oriya verb morphology, see Sahoo 2001). A skeletal lexical entry for dependent verbs is given in figure 6:

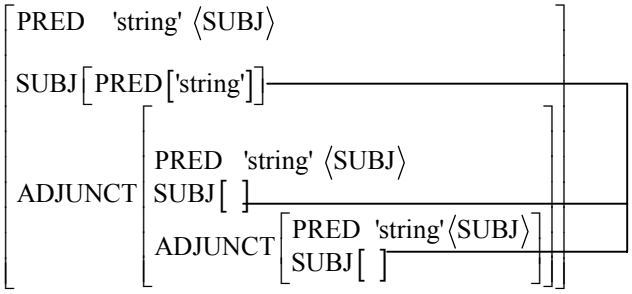
Figure 6.

Phrase Structure: VP_{dep} → ... V_{dep}

Lexicon: ‘string’: V (↑PRED) = ’string’
- i : (↑V-FORM) = dep

Through recursion of (12), the embeddings of chain-adjuncts inside of other chain-adjuncts is obtained. The c-structure induced by (12) corresponds to the schematic f-structure (showing one recursion) illustrated in figure 7:

Figure 7.



3.2. Subject sharing

Expanding phrase structure rule (12), we now can induce subject sharing as follows:

(13) VP -> VP (↑SUBJECT = ↓SUBJECT) V' (↑=↓)
.....

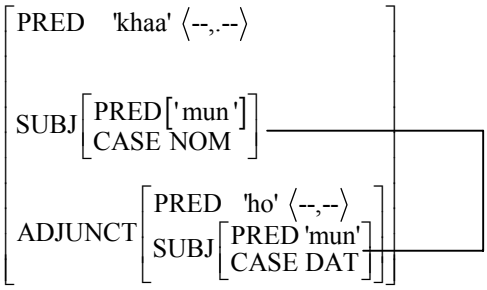
The control equation in (13) establishes identity of the subject of the matrix clause and the subject of the chain-adjunct, that is to say that just *one* syntactic item serves as the carrier of the argument realization of the 'subject-prone' argument of both verbs (and thus, by the recursion of (13), all verbs in the VP sequence). In the SVC literature, this is often referred to as 'token-sharing' (i.e., many verbs sharing one token item for a certain argument realization), as opposed to 'reference sharing', which is the situation where many syntactic items are co-referential.

In the LFG literature, 'token-sharing' corresponds to 'functional control'. An example is the case of English raising constructions, where the same item is an argument relative to both the matrix verb (e.g. 'seem' or 'believe') and the SUBJ of an XCOMP. Since the same f-structure function obtains relative to both matrix and embedded verb (i.e., SUBJ), morphological restrictions imposed by either of the verbs need to be observed by both. A crucial difference between English-type functional control and VP-chaining type functional control is that in the English case, functional control is induced lexically, being tied to the selection of a predicate complement or XCOMP by the matrix verb (see for further discussion Bresnan (2001)); under VP-chaining, in contrast, the verbs do not stand in a complementation relation, but instead form independent predication nuclei. We therefore have to introduce the control equation given in (13) as a constructional constraint.

Given structure-sharing of the subject function, the ungrammaticality of examples such as (6), repeated, follows from general constraints on unification, as indicated in figure 8:

(6) * mote jara ho-i ousadha khaa-il-i.
I fever happen medicine eat-PAST-1st,sg.
'Having had fever I took medicine.'

Figure 8. *Partial f-structure for (6):*



3.3. Object sharing

Suppose one were considering a rule for object sharing that corresponds to the functional control pattern that we have discussed for subject sharing. We would then attempt to apply the pattern in (13) to object sharing, as indicated in (14):

(14) VP -> VP V'
↑OBJECT = ↓OBJECT ↑=↓
.....

Such a rule, inducing token-sharing or functional control for objects, would *not* be adequate for VP-chaining, for the following reasons:

a) As seen earlier, not all VPs in a VP-chaining need to have object(s) (cf. (8));
b) as seen in (3) and (7), it is not required that transitive verbs in consecutive VPs share their object;
c) as seen in (8), a sharing effect can 'skip' an intermediate intransitive verb;
d) even when there is sharing of objects between two VPs, the argument-realization functions of the two verbs may realize the argument in question under distinct attributes, such as *direct* vs. *indirect* object, or object vs. object of preposition, as in (9).

It thus seems problematic to construe object sharing in these constructions as functional control and thus align it with subject sharing. Instead, we suggest construing object sharing as *reference sharing* or, following standard terminology, *anaphoric control*. Under anaphoric control, the 'covert' items are null-pronominals, of the type found in 'pro-drop' constructions, represented as PRED 'PRO' in figure 4. This construal is corroborated by two circumstances. The first is that the 'pro-drop' is pervasively attested in Oriya (Sahoo 2001) and related languages (e.g., Hindi (T.Mohanan 1994)). Secondly, pronominals may appear as part of the object sharing pattern, as illustrated in (15):

(15)

mun	maachha-Te$_i$	bhaaj-i	**taaku$_i$**	baaDipaTaku	ne-i	(*taaku)	khaai-li
I	fish-a	fry-dM.	it-ACC	backyard	take-dM	it-ACC	eat-Past 1st,sg

Such occurrences of overt pronominals are restricted in various ways. (15) shows that an overt pronominal may appear only *once* relative to the sharing of an object referent. Even so, the open pronoun in (15) it is clearly disfavored (Sahoo p.c.). Moreover, other types of occurrences of open bound pronouns are restricted to where they resolve possible ambiguity, such as in (16) below:

(16)

mun	mak-ku$_i$	saaDi-Taa	dekhaa-i	kaani	silei kar-i	(taaku$_i$)	de-li
I	mother	sari-the	show-dM,	seam	stich do-dM	her	give-Past 1st,sg

'Having shown my mother the sari, I stiched its seam and gave it to somebody else (her).'

In (16), a non-overt second reference to 'mother' may obtain, but cannot be controlled by the indirect object in the first VP-adjunct, due to the intervention of a transitive chain-adjunct. To enforce co-reference with the remote indirect object, an open pronoun has to appear. Such cases of pronominal disambiguation underline the pronominal nature of the object sharing phenomenon in VP-chaining.[7]

Notice that (16) seems to violate the constraints on object sharing across transitive verbs illustrated in (10). The direct object relative to *de* ('give') is understood as 'sari', although the complex verb *silei ka-i* ('stitch') takes *kaani* ('seam') as its direct object. Crucially, (16) doesn't mean that the speaker gave the seam. What enables 'sari' here to be the shared object is that it is the understood possessor of the intervening potential antecedent 'seam'; that is, (16) is only grammatical under the interpretation where it is the seam of the shown sari that has been stitched. Thus, 'sari' here can serve as shared object across a VP containing a second possible shared object, because it is understood in a whole-part/possessor-possessee relationship to this

[7] A question naturally arising at this point is then, given the presence of null-anaphora as something which needs to be addressed by the grammar of Oriya anyway, whether anything more specific needs to be added to that account with regard to VP-chaining. Constraints like those mentioned in connection with (15) might suggest that it does, but discussing this goes far beyond the scope of this paper.

intervening object and thus 'kept on stack' as the antecedent for the non-overt pronominal object of the verb *de*.

In sum, the occurrence of *taaku* in (16) indicates that although open pronouns are disfavored, they appear in VP-chains for disambiguation purposes to exclude arbitrary anaphoric control. Secondly, shared objects can survive intervening possible antecedents if referentially kept available. Both of these observations consolidate the general point namely that object sharing is a matter of null-pronominal anaphora, as indicated in figure 4.

It is beyond the scope of this paper to fully spell-out the mechanism of anaphoric object sharing in VP-chains. As a typological conjecture, however, it may be suggested that a possible characteristic of VP-chaining, as opposed to 'Integrated SVCs', is that object sharing in VP-chaining is a matter of anaphoric control, and not token sharing, whereas for Integrated SVCs, object sharing may be construed as structure sharing (alias functional control, or 'token-sharing').[8]

4. Uniformity of diathesis marking

We finally turn to passivization under VP-chaining. Recall that all verbs in the VP-chain need to passivize, and that intransitive verbs do not passivize; hence, passivization cannot apply across intransitive verbs. This makes the passive of (8) ungrammatical. Passivization marginally applies to shared objects that fulfill different object functions relative to the individual verbs. The passivization of (7), shown below as (17), therefore seems only marginally acceptable:

(17) *?? pilaaTi-ku* lugaa di-aa-jaa-i hotell-ku ni-aa-jaa-i
child-DAT cloth give-PRTP go-dM hotel take-PRTP-go-dM
piThaa khu-aa- ga-l-aa
pita feed-PRTP-go-PAST-3sg
'The child having gotten cloth, been taken to a hotel, was fed pita.'

Notice that in (17), the passivized NP *pilaa* ('child') is marked for objective case by the morpheme *-ku*. Constructions with an objective case marked subject seem to correspond to the passivization of secondary objects.[9] For (17) this means that *pilaa*, although the direct object relative to the verb *ni* ('take'), is passivized as secondary object, a function that it holds only relative to the first and the last verb of the VP-chain.

Here we will concentrate on the basic passivization pattern, illustrated in (18) and (19) below:

(18) *'active'*
mun maachha-Te bhaaj-i khaa-il-i
I fish-a fry-dM eat- PAST 1st sg
'Having fried the fish, I ate it.'

(19) *'passive'*
maachha-Ti bhaj-aa jaa-i khi-aa-ga-laa
fish-the fry-PRTP-go-dM eat-PRTP-go-PAST.3rd
'Having been fried, the fish was eaten.'

[8] See, e.g., Agyeman (2002).
[9] These are, according to Sahoo (p.c.), traditionally referred to as 'periphrastic passives'.

In order to illustrate the constraints that will allow us to state the properties of VP-chain-passivization, we adopt now a somewhat more elaborated correspondence architecture than employed so far. Following Butt et al., we will use the following notation:

Argument structure of mother node: $\hat{*}\alpha$
Argument structure of current node: $*\alpha$

Functional structure of mother node: $\hat{*}\alpha\lambda$
Functional structure of current node: $*\alpha\lambda$

The *ϕ-function* which defines the relationship between c-structure nodes and f-structures is in Butt et al defined as the composition of the α- and λ-function. Since we do not want to take a stand as to whether all information in c-structure relevant for the construction of f-structure is actually preserved across a-structure, we leave open the possibility that an independent function mediates between c- and f-structure. Its result will have to unify with the output of the α- and λ-function; we call it (likewise) ϕ, and represent it using the standard arrow notation.

We accommodate (18) through a set of annotated phrase structure rules. (In accordance with what was just said, we use the standard up/down arrows for the c-structure-to-f-structure correspondence function.)

Figure 9. A*nnotated phrase structure rules and lexicon accommodating (18)*

PS (1) S → NP (↑ SUBJ =↓) VP (↑=↓)

PS (2)

VP → VP_{dep} (↑ SUBJ =↓ SUBJ) V' (↑=↓)

$\lambda(\hat{*}\alpha\ \hat{\theta}) = SUBJ \wedge \lambda(*\alpha\ \hat{\theta}) = SUBJ \quad \vee$
$\lambda(\hat{*}\alpha\ \hat{\theta}) = NULL \wedge \lambda(*\alpha\ \hat{\theta}) = NULL$

PS (3) V' → V (↑=↓)

Lexicon

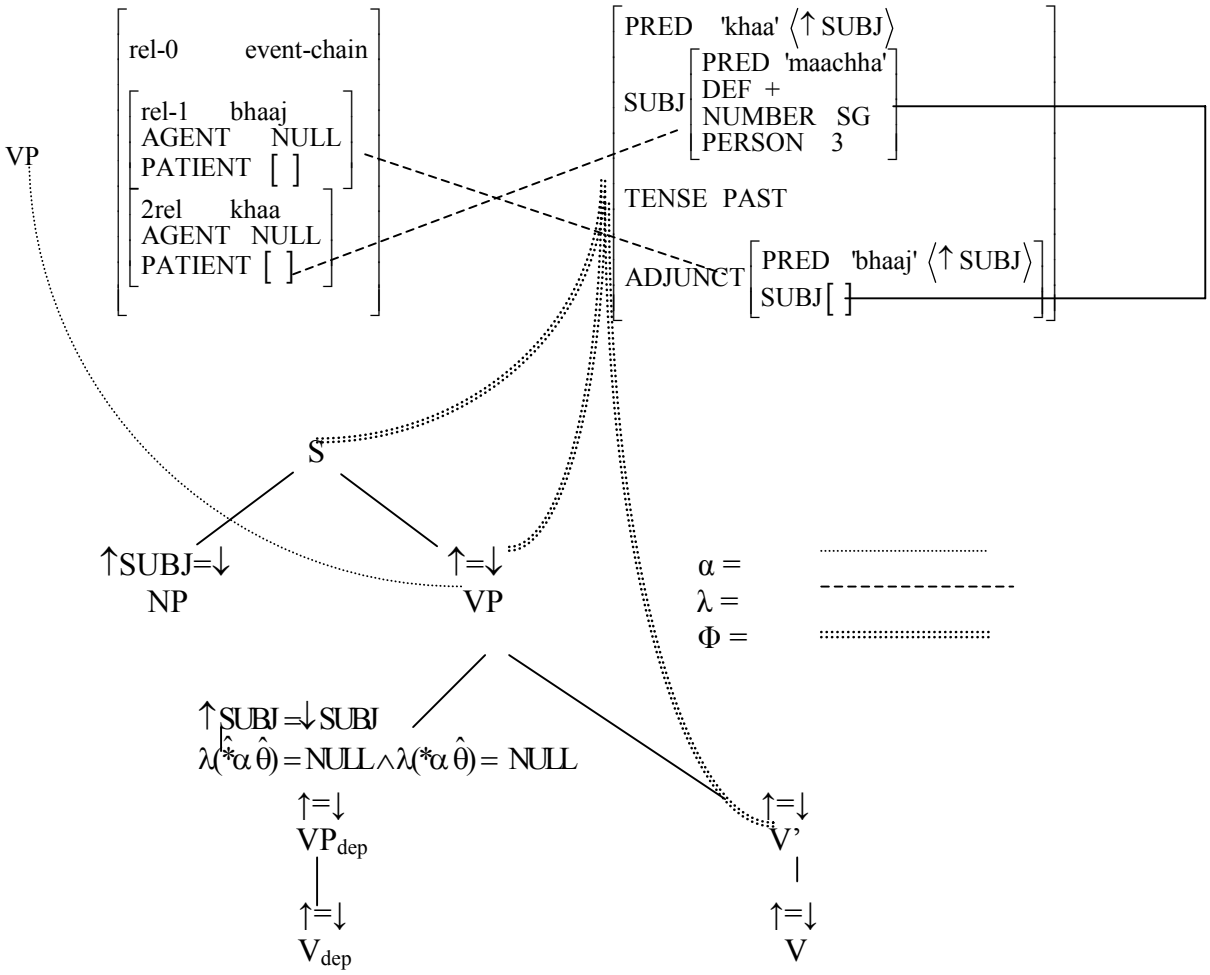
L(1): khi-aa-ga-laa: V	(↑ PRED) = 'khaa' (↑ TENSE) = PAST (↑ SUBJ PERSON) = 3 (↑ SUBJ NUMBER) = SG * α [$\hat{\theta}$]) = NULL λ($\hat{*}$ α [-r]) = SUBJ
L(2): bhaj-aa jaa-i: V	(↑ PRED) = 'bhaaj' (*α [$\hat{\theta}$]) = NULL λ($\hat{*}$ α [-r]) = SUBJ (↑ V-FORM) = dep
L(3): maachha-Ti	(↑ PRED) = 'maachha' (↑ DEF) = +

Crucial is here the rule PS(2), which demands, for each generation of a chain-VP, that this VP have the same diathesis value as the mother VP. Diathesis value is represented as follows: in active voice, the highest thematic role (^θ) is realized as SUBJECT, in passive voice, it is realized as NULL (or 'oblique', an option we leave out here, for perspicuity). The rule PS(2) thus states that either both VPs have a SUBJECT realization of ^θ, or they both have a NULL realization of it. Such an equivalence between diathesis value and ^θ-realization seems defensible in the case of Oriya, and possibly universally.

Phrase Structure Rules (1)-(3) and lexical entries (1)-(3) in figure 9, along with a tacit assumption to the effect that a [-r] PATIENT is realized as SUBJ in passive, generate the overall correspondence architecture shown in figure 10:

Figure 10. *Correspondence architecture for example (18):*

maachha-Ti bhaj-aa jaa-i khi-aa-ga-laa
fish-the fry-PRTP-go-dM eat-PRTP-go-PAST.3rd
'Having been fried, the fish was eaten



Notice that we do not exclude asymmetrical passivization, as exemplified in (17). While the suppression of the highest thematic role is enforced constructionally, we have left the mapping to subject of a passive verb open.

Another specification left out in the above is one representing the temporal order between the events/situations described by the VPs. What is needed is the construction of a set of precedence specifications for *pairs of VPs/situations*, schematically as follows (again, as specifications *added* to those stated so far for the constellation in question):

(20) VP -> VP V'

$*\alpha$ SIT-INDEX = n

$\widehat{*\alpha}$ SIT-INDEX = m

$\widehat{*\alpha}$ SIT-INDEX = $*\alpha$ SIT-INDEX

$\widehat{*\alpha}$ PRECEDENCE-SET {... , n<m}

....

That is, going up the tree from left, at each syntactic 'chain-juncture' point, the event/situation expressed by the left daughter VP is entered as temporally preceding ('<') the event expressed by the matrix V', this specification being added to a set accumulated as one goes from one matrix verb to a next matrix verb up.

While details of the latter representation will need to be further developed, it instantiates a common feature of the analyses of the various phenomena treated (apart from object sharing), namely that they 'drive' all aspects of VP-chaining from annotations of the c-structure rule inducing VP-embedding (here distributed over (12), (13), PS(2) in figure 9, and (20)), in tandem with lexical specifications as induced by these annotations.

5. Possible distinguishing properties of VP-chaining as opposed to Integrated SVCs

Although our focus has been exclusively on VP-chaining, and exclusively on how they are manifest in Oriya, it is not unreasonable at the end to offer some speculations as to how VP-chaining may stand apart from the heterogeneous family of construction types that we have tentatively labelled 'Integrated SVCs' (ISVCs). Among the properties of VP-chaining now described, at least three of them may seem good candidates for serving as factors distinguishing VP-chaining from ISVCs, not only in Oriya, but perhaps cross-linguistically:

(i) VP-chains have recursive embeddings of adjuncts in f-structure, whereas in accordance with the view exemplified in figure 2, ISVCs presumably have a flat f-structure.

(ii) While both types display subject sharing as token sharing, object sharing in VP-chains is a matter of anaphoric control, whereas in ISVCs, it may well be a matter of functional control, possibly induced by a schema like (14).

(iii) Semantically, the VPs in VP-chaining express distinct situations, ordered by temporal precedence (or status as 'prior given'). Although the semantics of ISVCs is not uniform, often it may be possible to see the verbs as situationally interleaving, describing the same situation but with each verb specifying a distinct aspect of it. (In this respect, a PRED value representation like 'zo-gaa-wuo' in figure (4) may perhaps be interpreted as an *event-type unification* of the event types expressed by 'zo', 'gaa' and 'wuo' - cf. Butt (1997) for a proposal in this direction.)

In addition to the features of VP-chaining constructions just mentioned, the 'passivize-all-or-no-verbs' phenomenon addressed in section 4 has been seen as lending itself to an architecture of an LFG grammar where c-structure projects to a-structure and a-structure to f-structure, in accordance with a proposal by Butt et. al. (1997). How this phenomenon, and our proposed way of dealing with it, places itself in a typological perspective, is not a matter we can comment further on here.

References

Agyeman, N.A. 2002. Serial Verb Constructions in Akan. M.Phil. thesis, NTNU, Trondheim.

Alsina, A. 1993. Predicate Composition: A theory of Syntactic Function Alternations, Doctoral Dissertation, Stanford University.

Alsina, A. 1997. A Theory of Complex Predicates: Evidence from Causatives in Bantu and Romance. In Alsina, A. et.al. (eds).

Alsina, A., J. Bresnan, and P. Sells (eds). 1997. *Complex Predicates.* CSLI Publications.

Andrews, A. & C. Manning 1999. *Complex Predicates and Information Spreading in LFG.* Stanford Monographs. CSLI Publications.

Beermann, D., K. Sahoo, and L. Hellan . 2001. What is 'Argument Sharing'? A case study on argument sharing under VP-serialization in Oriya. ms, NTNU.

Bodomo, A. 1997. Paths and Pathfinders: Exploring the Syntax and Semantics of Complex Verbal Predicates in Dagaare and Other Languages. Doctoral dissertation, NTNU, Trondheim.

Bresnan, J. 1982a The passive in lexical theory. In Bresnan 1982c.

Bresnan, J. 1982c. *The Mental Representation of Gammatical Relations.* MIT Press.

Bresnan, J. 2001. *Lexical Functional Syntax.* Blackwell.

Butt, M.J. 1993. The Structure of Complex Predicates in Urdu. Doctoral dissertation, Stanford University.

Butt, M. 1995. *The Structure of Complex Predicates in Urdu*. CSLI Publications.

Butt, M. 1997. Complex Predicates in Urdu. In Alsina et. al. (eds).

Butt, M., M. Dalrymple, and A. Frank 1997. An architecture of linking theory in LFG. In *Proceedings of the LFG97 conference*, UC, San Diego.

Dalrymple, M. 2001. *Lexical Funciional Grammar.* Academic Press.

Foley. J. and S. Olson 1985. Clausehood and Verb Serialization. In J. Nichols & A. Woodbury (eds) *Grammar Inside and Outside the Clause*. Cambridge University Press.

Kaplan, R.M. 1995. The Formal Architecture of LFG. In Dalrymple, M., R.M. Kaplan, J.T. Maxwell, and A. Zaenen (eds) *Formal Issues in Lexical Functional Grammar*. CSLI Publications.

Kroeger, P. 2001. Analyzing Syntax: a lexical functional approach. manuscript.

Mohanan, T. 1994. *Argument Structure in Hindi*. CSLI Publications, Stanford.

Mohanan, T. 1997. Multidimensionality of Representation: NV Complex predicates in Hindi. In Alsina, A. et.al. (eds).

Niño, M-E. 1997. The multiple expression of inflectional information and grammatical architecture. In F.Corblin, Godard, D. & J-M. Marandin (eds) *Empirical Issues in Formal Syntax and Semantics*. Bern, Peter Lang.

Osam, E.K.1994. From Serial Verbs to Prepositions – The road between. In *Sprachtypologie und Universalienforschung*, 47,1.

Payne, J.P.1985. Complex Phrases and Complex Sentences. In T.Shopen (ed) *Language Typology and Syntactic Description.* Volume II.

Sahoo, K. 2001. Oriya Verb Morpholgy and Complex Verb Constructions. Doctoral dissertation. NTNU, Trondheim.

Sells, P. 2000. Syntactic Information and its Morphological Expression. manuscript, Stanford University.

CONSTRAINT SYMMETRY IN OPTIMALITY THEORETIC SYNTAX

George Aaron Broadwell
University at Albany, State University of New York

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

http//csli-publications.stanford.edu/

1 Introduction

As Sells (2001a) has noted, few of the directional constraints thus far motivated in Optimality Theoretic Lexical Functional Grammar seem truly symmetrical in their application. It is clear that there are languages that attempt to align interrogatives, foci, and topics with the left edge of a clause (WH-L, FOC-L, TOP-L) and languages that attempt to align the head of a phrase with the left edge of that phrase (HEAD-L) (Choi 1999, Lee 2001, Morimoto 2001). It is less clear that WH-R, FOC-R, or TOP-R is well motivated.[1] Although the existence of HEAD-R might seem necessary, Sells (2001a) has argued that the effects of this constraint can be better captured with an asymmetrical constraint SPINE-R. The question naturally arises—are all syntactic constraints fundamentally asymmetric or are there constraints that operate symmetrically, with both left and right variants?

This paper pursues a constraint-based account of consistency effects. I'll use *consistency* as a general label for whatever constrains branching on the non-recursive side of a phrase. In particular, it appears that consistency constraints operate symmetrically, affecting right branches in some languages and left branches in others.

2 Zapotec pied-piping with inversion

San Dionicio Ocotepec Zapotec (hereafter SDZ) is an Otomanguean language spoken in Oaxaca, Mexico.[2] The basic word order of this language is VSO, with head-initial NPs and PPs:

1)	Ù-dííny	Juáàny	bèh'cw=rè'	cùn	yààg.
	com-hit	Juan	dog=that	with	stick

[1] However, see Van Valin and LaPolla (1997) for arguments in favor of a right-edge focus in some languages.

[2] SDZ is an Otomanguean language spoken in San Dionicio Ocotepec, Oaxaca, Mexico by 2,000 - 3,000 people. I thank Farrell Ackerman, Peter Austin, Chris Barker, Lee Bickmore, Cheryl Black, Joan Bresnan, Yehuda Falk, Ed Keer, Pamela Munro, Jerrold Sadock, Peter Sells, Yuching Tseng, and Robert Van Valin for useful discussion of this material. Special thanks to Luisa Martínez, who provided all the SDZ data. Most of the material discussed here is published as Broadwell (2001); earlier versions of the analysis appear as Broadwell (1999a,b).

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> = /ʒ/ before a vowel and /ʃ/ before a consonant, <xh> = /ʃ/, <dx> = /ʤ/, <ch> = /ʧ/, <c> = /k/ before back vowels, <qu> = /k/ before front vowels, <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.

'Juan hit that dog with a stick'.

However, there are some contexts in which this head-initial order gives way to a head final order. Suppose we question the object of the PP:

2) ¿Xhíí cùn ù-dííny Juààny bèh'cw?
what with com-hit Juan dog

'What did Juan hit the dog with?'

Now the PP is head final:

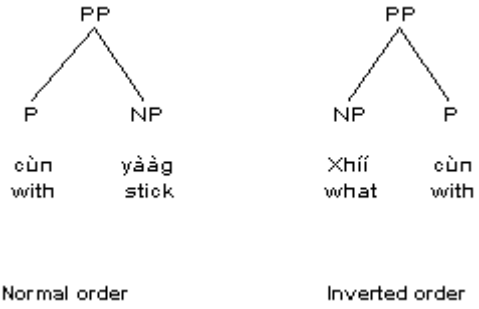


Figure 1 Normal and inverted orders for PP

Inverted, head-final orders are found in some other contexts as well, as we'll see in the following sections. Smith Stark (1988) was the first to draw attention to this pattern, and labelled it 'pied-piping with inversion' (PPI).

SDZ shows PPI is found with PPs, NPs, and QPs. For reasons of space, inversion in NP and QP doesn't receive an explicit analysis in this paper. See Broadwell (2001) for a fuller account.

3 Pied-piping with inversion and constraint interaction

Let's look at how these results can be obtained in an optimality-theoretic approach to syntax. We can express the idea that in the unmarked order heads precede both their complements and specifiers through constraints of the following sort:

3) Head ≺ Spec

A head must precede its specifier.

4) Head ≺ Comp

A head must precede its complement.

My claim is that the PPI phenomenon arises from the interaction of these ordering principles with another constraint that forces the wh-word to appear at the left edge of CP.

SDZ has obligatory wh-movement as shown by the following examples:

5) a. ¿Túú ù-dííny Juààny cùn yààg?
what com-hit Juan with stick

'What (anim.) did Juan hit with a stick?'[3]

b. *¿Ù-dííny Juààny túú cùn yààg? *wh-in-situ*
com-hit Juan what with stick

Wh-phrases appear in [Spec,CP]. The fact that wh-movement is obligatory suggests that SDZ shows the effects of a constraint like the following:

6) Align (IntF, L, CP, L) = Wh-L

Align the left edge of an interrogative focus phrase with the left edge of CP.[4]

4 Wh-L and PPs

Pied-piping with inversion is found with the objects of most prepositions.[5]

7) ¿Xhíí cùn ù-dííny Juààny bèh'cw? ✓*PPI*
what with com-hit Juan dog

'What did Juan hit the dog with?'

8) ¿Xhíí dèjts zúú bèh'cw? ✓*PPI*
what behind lie dog

'What is the dog behind?'

Though my consultant reports a preference for the inverted form, the uninverted form is also acceptable for these prepositions:

9) ¿Cùn xhíí ù-dííny Juààny bèh'cw? ✓*PP without inversion*
with what com-hit Juan dog

---

[3] SDZ uses the wh-words *xhíí* 'what, which' for inanimates and *túú* 'who, what, which' for animates (both people and animals). I've glossed the examples with the appropriate English wh-word.

[4] In what follows below, I have assumed that Wh-L is interpreted in a gradient manner, so that each word that intervenes between the wh-element and the left edge of CP triggers an additional violation.

[5] As discussed in Broadwell (2001), there is a small set of preposition, mostly those borrowed from Spanish, which fail to invert. For reasons of time, I will not discuss these cases here.

'With what did Juan hit the dog?'

10) ¿Dèjts xhíí zúú bèh'cw? ✓*PP without inversion*
behind what lie dog

'Behind what is the dog?'

Most prepositions show this pattern of optional inversion.

Preposition stranding is completely disallowed:

11) *¿Xhíí ù-dííny Juààny bèh'cw cùn? **P-stranding*
what com-hit Juan dog with

(What did Juan hit the dog with?)

The ungrammaticality of these examples seems to be due to an undominated constraint in SDZ that forbids preposition stranding. For our current purposes, the constraint can be stated as follows:[6]

12) *Prep-Strand

The object of a preposition must form a c-str constituent with the preposition.

The optionality of inversion for most prepositions indicates that the constraints Wh-L and Head◁Comp are equally ranked.

		*Prep-Strand	Wh-L	Head < Comp
a.	¿Xhíí zúú bèh'cw dèjts? (What lies dog behind?)	*!		
b.	☞¿Dèjts xhíí zúú bèh'cw? (Behind what lies dog ?)		*	
c.	☞¿Xhíí dèjts zúú bèh'cw? (What behind lies dog?)			*

5 Consistency

The constraints posited above correctly account for simple cases of pied-piping with inversion. But what about cases where we expect two flips? These are unexpectedly bad.

[6] For a more careful formulation of this constraint, see Broadwell (2001).

13) *¿ Xhíí yààg cùn ù-dííny Juààny bèh'cw?
which stick with com-hit Juan dog

'With which stick did Juan hit the dog?'

With our current constraints, the structure that we would predict would be as follows:

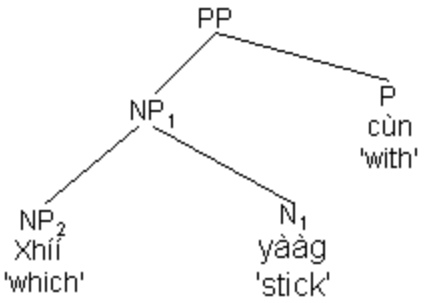


Figure 2 Double inversion with PP

I'll call this the 'double inversion' structure because it involves two inversion of the normal head-initial structure – one in the PP and the second in NP_1.

According to the argument so far, the tableau for such cases ought to look as follows:[7]

		Wh-L	Head <Comp
a.	¿Cùn yààg xhíí ù-dííny Juáàny bèh'cw? (with stick which hit Juan dog)	**	
b. [inversion of NP]	☞¿Cùn xhíí yààg ù-dííny Juáàny bèh'cw? (with which stick hit Juan dog)	*	
c. [double inversion]	☞¿Xhíí yààg cùn ù-dííny Juáàny bèh'cw? (which stick with hit Juan dog)		*
d. [inversion of PP]	¿Yààg xhíí cùn ù-dííny Juáàny bèh'cw? (stick which with hit Juan dog)	*	*

[7] For the sake of clarity, I have omitted from this tableau the penalty associated with violating the Head ≺ Spec constraint, since the Specifier *xhíí* 'which' precedes the head *yààg* 'stick'. Since the relevant candidates all violate the constraint to the same degree, this constraint doesn't play a crucial role in distinguishing the candidates under discussion.

The unexpectedly bad candidate is (c).

I suggest that the double inversion examples are bad due to the interaction of another constraint, Consistency. For all the cases under consideration, the double inversion configuration is ungrammatical. Consider the phrase structure tree that results from double inversion:

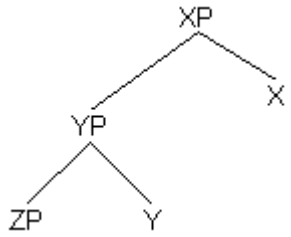


Figure 3 Prohibited double inversion structure

I want to argue that a tree of this sort violates a constraint on permissable branching types, modifying an idea originally due to Longobardi (1991). We can state the restriction as follows:

14) Consistency (San Dionicio Ocotepec Zapotec)

If a phrase XP is right-headed, then a constituent which is located on a left branch of XP may not branch.

There is an additional kind of evidence for this restricion, found in 'vexation interrogatives'. SDZ has expressions of the following sort:

15) ¿Xhíí chìngáàd ù-dàù Juáàny?
what the:hell com-eat Juan

'What the hell did Juan eat?'

16) ¿Túú chìngáàd ù-tò' x-còch-á'?
who the:hell com-sell p-car-1s

'Who the hell sold my car?'

These have the form of a Zapotec interrogative followed by the borrowed 'vexifier' *chìngáàd* (<Span. *chingada)*. I've glossed and translated this using 'the hell' as an approximate English equivalent.

When a vexation interrogative is the object of a preposition, we also see lack of pied-piping with inversion:

17) *¿Túú chìngáàd cùn ù-tò'=èhby x-còch-á'? [double inversion]
who the:hell with com-sell=3s p-car-1s

(Who the hell did he sell my car to?)

Assume that the relevant portion of the tree in the ungrammatical (17) looks like this:

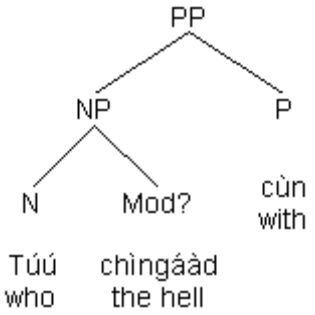


Figure 4 Inversion with a vexation interrogative

Then the problem seems to be that this PP is right-headed, but the complement is branching in violation of the Consistency condition.

In terms of the optimality-theoretic treatment of pied-piping with inversion, we want to say that the Consistency constraint dominates Wh-L. Consider again the case of PPs with branching objects.

18) *¿ Xhíí yàág cùn ù-dííny Juàány bèh'cw?
which stick with com-hit Juan dog

'With which stick did Juan hit the dog?'

If we add Consistency to the tableau, then the double inversion candidate is correctly ruled out.

		Consistency	Wh-L	Head < Comp
a.	¿Cùn yàág xhíí ù-dííny Juáány bèh'cw? (with stick which hit Juan dog)		**	
b.	☞¿Cùn xhíí yàág ù-dííny Juáány bèh'cw? (with which stick hit Juan dog)		*	
c. [double inversion]	¿Xhíí yàág cùn ù-dííny Juáány bèh'cw? (which stick with hit Juan dog)	*!		*
d.	¿Yàág xhíí cùn ù-dííny Juáány bèh'cw? (stick which with hit Juan dog)		*	*

Other Mesoamerican languages that allow PPI frequently disallow it in cases where a Consistency violation would obtain, so there is strong cross-linguistic support for a constraint of this sort.[8]

6 Amharic[9]

[8] For reasons of space, the data from other languages have been omitted from this version of the paper. A fuller version is available at http://www.albany.edu/anthro/fac/broadwell.htm

[9] Amharic is a Semitic language spoken by approximately 20 million people in Ethiopia. The transliteration used here follows the system of Appleyard (1995). ä is /ə/, ï is /ɨ/, and the apostrophe represents glottalization. I thank Daniel Clough, Kelly Moore, Sharon Rose, Yuching Tseng, and the members of the 2001 field methods class for their suggestions on the analysis of Amharic. Special thanks to Tejitu Molla and Daniel Wolde-Giorgis, who provided all the Amharic data not otherwise

Amharic (Semitic, Ethiopia) is a predominantly head-final, left-branching language. Sentences are verb final and noun phrases are noun-final:

19) [[Yïh t'ïllïk' säw]$_{NP}$ t'ïru näw.]$_{S}$
this big man good is:3ms

'This big man is good.'

The only exception to this branching pattern is found with prepositional phrases and complementizers.

6.1 The variety of adpositional phrases

Amharic has prepositions, postpositions, and circumpositions.[10] Some examples of prepositions:

20) [**wädä** bet-u]$_{PP}$
toward house-def

'toward the house'

21) [**bä**=bet-u]$_{PP}$ 'in the house'
in=house-def

As the last example shows, monosyllabic prepositions are proclitic on the following word.

Postpositions are shown in the following examples:

22) Bet-u [č'akka-w **dar**] näw.
house-def woods-def edge is

'The house is at the edge of the woods.' Leslau 1995:648

23) [Tärara-w **lay**] bïzu zaf allä.
mountain-def on many tree exist

'There are many trees on the mountain.' Leslau 1995:619

Circumfixes are shown in the following examples:

24) [**Ï**-bet **wïst'**] gäbba-hu-iñ.
in-house in enter-1s-1s

'I entered (into) the house.'

The possessive preposition is the proclitic *yä*, which attaches to the possessor. This preposition is normally obligatory.

---

attributed.

[10] I follow the traditional grammars of Amharic (Leslau 1995) in recognizing prepositions, postpositions, and circumpositions in the language. Some recent GB-oriented work on Amharic syntax (Tremblay and Kabbaj 1989, Halefom 1994) treats the prepositions as case markers. So far as I can see, this doesn't provide any explanation of the *yä*-deletion facts.

25) a.) [[yä=Yohannïs]$_{PP}$ bet]$_{NP}$ b.)*Yohannïs bet
of=John house John house

'John's house'

In particular, the subset of phrase structure rules that we are interested is as follows:[11]

NP ➔ PP Adj N
PP ➔ Prep NP

Prepositional phrases appear to be left-headed, right-branching structures. They impose a peculiar restriction on their complements, which we can state as follows:

26) Consistency (Amharic)

If a phrase XP is left-headed, then a constituent located on a right branch of XP may not contain a branching phrase YP.

As a result of this restriction, the object of an Amharic preposition can be a simple noun, or a noun with an adjective or determiner:

27) [wädä tïllïk'-u bet]$_{PP}$
toward big-def house

'toward the big house'

However, the object cannot contain a prepositional phrase, since it that would be a right-branching phrase.

28) *wädä [[yä=Yohannïs]$_{PP}$ bet]]$_{NP}$
toward of=John house

('toward John's house')

In such a case, the lower preposition, *yä*, must be deleted:

29) wädä [[Yohannïs]$_{PP}$ bet]]$_{NP}$
toward John house

'toward John's house'

The same restriction is not found for the object of a postpositional phrase. In this case, the preposition *yä* is optionally deleted:

30) [[(Yä=)Yohannïs t'äräp'p'eza]$_{NP}$ lay]$_{PP}$ bïzu sahïn-occ all-u.
(of=) John table on many plate-pl exist-3pl

'There are many plates on John's table.'

Circumfixes show the same pattern as prepositions—the *yä* is obligatorily deleted. In the following example, *bä... atäggäb* means 'near'.

[11] I will assume that all nodes are optional due to the principle of economy of expression (Bresnan 2001:91ff). Hence the following PS-rules don't include any parentheses.

31) Wïšša-w bä=Daniel bet atäggäb näw.
dog-def near=Daniel house near is

'The dog is near Daniel's house.'

*Wïšša-w bä=yä=Daniel bet atäggäb näw.
dog-def near=of=Daniel house near is

'The dog is near Daniel's house.'

There is also a purely postpositional variant of this adposition, which is *atäggäb* 'near'. Note that with the postpositional variant, the *yä* is again optional.

32) Wïšša-w (yä)=Daniel bet atäggäb näw.
dog-def (of)=Daniel house near is

'The dog is near Daniel's house.'

When not the object of an adposition, deletion of *yä* is bad:

33) Yä=Yohannïs bet tïllïk' näw.
of=John house big is

'Our house is big.'

*Yohannïs bet tïllïk' näw.
John house big is

6.2 An OT approach to *yä* deletion

Let us consider the phrase structure of an ungrammatical example like the following, repeated from above:

34) *wädä [[yä=Yohannïs]$_{PP}$ bet]]$_{NP}$
toward of=John house

('toward John's house')

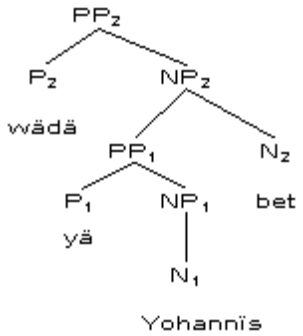


Figure 5 PP with object containing a branching phrase

This structure violates Consistency because PP_2 is a left-headed phrase, NP_2 is located on a right branch

of this phrase, and NP_2 contains PP_1, which is branching. The grammatical alternative deletes P_1.

However, recall that deletion of *yä* is ungrammatical outside the context of a PP:

35) Yä=Yohannïs bet tïllïk' näw.
of=John house big is

'John's house is big.'

*Yohannïs bet tïllïk' näw.
John house big is

Therefore there must be a penalty associated with the deletion of the preposition *yä*. We might call it Max(Prep):

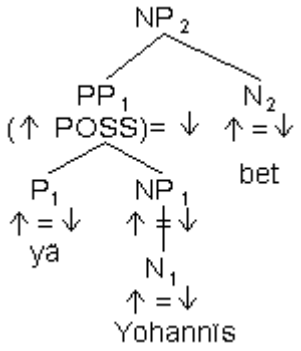
36) Max(Prep) (preliminary)

Do not delete a preposition.

However, this may be too broad. So far as I know, *yä* is the only Amharic preposition that ever deletes. We might understand that as a consequence of the fact that the f-str corresponding to a possessive structure contains no element corresponding to *yä*.

The f-structure and annotated c-structure for an Amharic possessive should look approximately as follows:[12]

$$\begin{bmatrix} \text{PRED} & \text{'house'} \\ \text{POSS} & [\text{PRED 'John'}] \end{bmatrix}$$



That is, P_1 here is purely 'functional' preposition, as opposed to other lexical prepositions. The situation is analogous to the categories CP and DP in English, as discussed by Bresnan (2001:134). Since Comp and Det are functional heads, they may be absent from the phrases they head, in constrast to lexical heads. Bresnan argues that the endocentricity requirement of English applies only to phrases headed by lexical categories.

Bresnan's (2001) endocentricity is stated categorically, rather than in terms of a violable constraint. Applied to Amharic prepositional phrases, it works best to think of an Endocentric-PP constraint with stronger and weaker versions.

37) Endocentric-PP(Lexical) Endo-PP(Lex)
A lexical PP must contain a head.

38) Endocentric-PP(Functional) Endo-PP(Fun)

[12] The structure given here is parallel to that proposed by Chisarek and Payne (2001) for English possessives like *the daughter of the king*, though I have retained the function POSS, rather than choose between SUBJ and NCOMP.

A functional PP must contain a head.

The ranking we want is Endo-PP(Lex), Consistency ≫ Endo-PP(Fun). This will produce results like the following:

$\begin{bmatrix}\text{PRED} & \text{'toward(OBJ)'} \\ \text{OBJ} & \begin{bmatrix}\text{PRED} & \text{'house'} \\ \text{POSS} & [\text{PRED 'John'}]\end{bmatrix}\end{bmatrix}$		Endo-PP(Lex)	Consis-tency	Endo-PP (Fun)
a	wädä yä=Yohannïs bet toward of=John house		*!	
b	☞wädä Yohannïs bet toward John house			*
c	yä=Yohannïs bet of=John house	*!		

Contrast this with the situation where we have a postposition instead of a preposition. Recall that in this case, deletion of *yä* is optional.

39) (Yä=)Yohannïs t'äräp'p'eza lay bïzu sahï n-očč all-u.
(of=)John table on many plate-pl exist-3pl

'There are many plates on John's table.'

$\begin{bmatrix}\text{PRED} & \text{'on (OBJ)'} \\ \text{OBJ} & \begin{bmatrix}\text{PRED} & \text{'table'} \\ \text{POSS} & [\text{PRED 'John'}]\end{bmatrix}\end{bmatrix}$		Endo-PP(Lex)	Consis-tency	Endo-PP(Fun)
a	☞yä=Yohannïs t'äräp'p'eza lay of=John table on			
b	☹ Yohnnïs t'äräp'p'eza lay John table on			*
c	yä=Yohannïs t'äräp'p'eza of=John table	*!		

Now the problem is that these constraints predict that only the fully faithful candidate (a) should be optimal. However, Amharic speakers also accept the (b) candidate.

It appears that there is some additional constraint violation in the (a) candidate that degrades

its grammaticality and makes (b) equivalent in grammaticality. It is not clear what the best formulation of this constraint should be. I tentatively propose the following:

40) Consistency (categorial)
If a phrase XP contains a phrase YP, and X and Y belong to the same syntactic category, then X and Y must both align with the same side of the phrase.

The intuition behind this constraint is that languages prefer consistently right- or left-headed phrases, especially when a phrase XP contains a phrase YP of the same categorial type. So if one PP contains another, both should be consistently prepositional or postpositional.

If we add this constraint to the tableau, weighted equally with Endo-PP (Fun), then we get the following result:

[PRED 'on(OBJ)' OBJ [PRED 'table' POSS [PRED 'John']]]		Endo-PP(Lex)	Consistency	Endo-PP(Fun)	Consist (Cat)
a	☞yä=Yohannïs t'äräp'p'eza lay of=John table on				*
b	☞Yohannïs t'äräp'p'eza lay John table on			*	
c	yä=Yohannïs t'äräp'p'eza of=John table	*!			

Now the (a) and (b) candidate both incur one violation, and we correctly predict that both are acceptable.

7 Consistency effects in English and the typology of consistency

7.1 Prenominal adjectives

English also shows effects of the Consistency constraint. There is a well-known restriction on adjective phrases in English—a prenominal adjective phrase may be composed of an adjective or an adjective plus a modifier:

41) a proud mother
a very proud mother

But the adjective phrase cannot contain another phrase.

42) *a proud of her son mother

The Italian facts are exactly the same, as discussed by Longobardi (1991).

We can state the English restriction in the same terms as the constraint found in Amharic:

43) Consistency (English AdjP)

If a phrase XP is located on a left branch, then it may not contain a branching phrase YP.

7.2 Consistency and sluicing

A less well known example involves PPs in sluicing contexts, as discussed by Merchant (to appear):

44) John was talking, but I don't know ?with who(m)/[**who with**].

Merchant labels the inversion of PPs that occurs in sluicing contexts *swiping* (for sluiced wh-word inversion with prepositions in Northern Germanic).

Merchant notes that for such inversion to occur, the wh-word must be 'minimal'. Consider the following contrasts:

45) John was talking, but I don't know who with.
*John was talking, but I don't know which guy with.

46) They were arguing; God only knows what about.
*They were arguing; God only knows what problem about.

Merchant's solution is to suggest that the wh-word incorporates (undergoes head movement) into the preposition in PF. Since this is head movement, only a single wh-word may move; a wh-phrase may not.[13]

Instead, I would suggest that it may be more productive to analyze the English PP inversion in the same way as the Zapotec inversion – the wh-word is still the complement of the P head, but exceptionally the complement precedes the head in this case. In this case, the ungrammatical examples in (76) and (77) above are violations of Consistency, and we may use the same variant of the constraint seen in Zapotec.[14]

47) Consistency (English swiping)

If an XP is right-headed, then a constituent which located on a left branch of XP may not branch.

Why should inversion occur in this context in English? Merchant suggests that the inversion is prosodically driven. Modifying his argument slightly, note that sluicing generally involves coordinated sentences, with focal stress on the final element of both.

48) John was TALKING, but
a.) ✓I don't know [who WITH]
b.) * I don't know [WHO with]

c.) ? I don't know [with WHO]
d.) * I don't know [WITH who]

Examples where the focal stress falls earlier than the final word are bad (b, d). The example where focal stress falls on WHO is also slighly marked, and this seems related to the fact that focal stress indicates

---

[13] But this is an odd sort of incorporation, considered from a cross-linguistic perspective. Note, for example, that multiple word vexation interrogatives may invert in swiping contexts:

They were arguing, but I don't know what on earth about.

In general, incorporation should only affect a N^0. However, it seems questionable to treat items like *what on earth* as a single N. Vexation interrogatives do behave in certain respects like a single lexical item, but lexical items may be larger than a single lexical head.

[14] However, Zapotec inversion disallows vexation interrogatives, while they seem relatively good in English swiping. It is possible that the difference is be due to the syntax or lexical status of the vexation interrogatives in the two languages. Alternately, the constraints may need to to be further differentiated in some way.

contrastive focus, while wh-word show *interrogative* focus. It appears to be difficult or impossible for an interrogative to bear focal stress:

49) JOHN will pick up Mary, but SUE will pick up Ralph.
*? JOHN will pick up Mary, but WHO will pick up Ralph?

We might call this constraint *FocStr (Interrogative). If *FocStr (Interrogative) is equally ranked with Head ≺ Complement, then the optionality of swiping can be seen as choice between two alternatives—one places focal stress on the interrogative and the other uses a marked word order..

	*FocStr (Int)	Head < Comp
☞... but I don't know [who WITH]		*
☞... but I don't know [with WHO]	*	

However, Consistency outranks both of these constraints, making inversion unavailable for branching wh-phrases:

	Consistency	*FocStr (Int)	Head < Comp
... but I don't know [which person WITH]	*		*
☞... but I don't know [with WHICH PERSON]		*	

7.3 A typology of Consistency constraints

Consider again the Consistency constraints we have identified so far:

50) Consistency (San Dionicio Ocotepec Zapotec)

If a phrase XP is right-headed, then a constituent which located on a left branch of XP may not branch.

51) Consistency (Amharic)

If a phrase XP is left-headed, then a constituent located on a right branch of XP may not contain a branching phrase YP.

52) Consistency (English swiping)

If an XP is right-headed, then a constituent which located on a left branch of XP may not branch.

53) Consistency (English AdjP)

If a phrase XP is located on a left branch, then it may not contain a branching phrase YP.

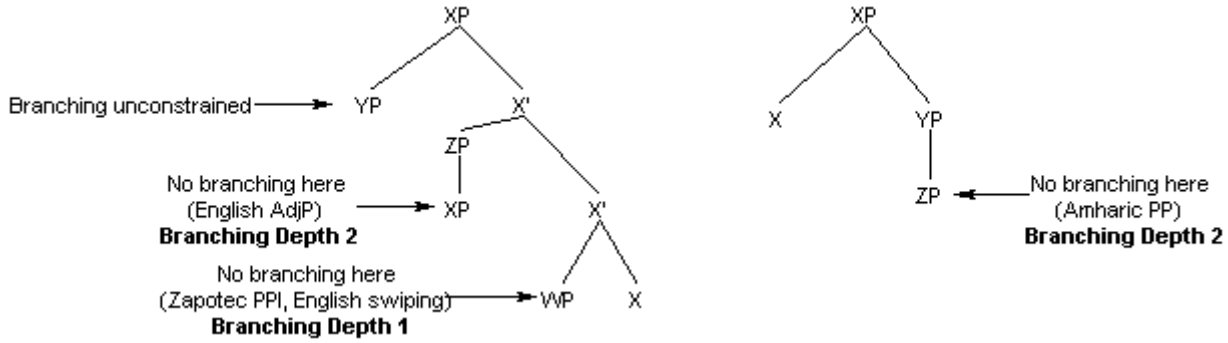
Why should the constraints work in this way? The crucial intuition is that languages restrict what may

appear on the non-recursive side of a phrase. Restrictions on the size or branching of constituents on the non-recursive side of the phrase appear to be quite common cross-linguistically. These seem to be of two types—1) either the constituent cannot branch, or 2) the constituent cannot contain a branching phrase.

It will be helpful to introduce the term *maximal branching depth* here. If a language allows no constituent at all in a position, then the maximal branching depth for this constituent is zero. If the language allows a non branching constituent, but not a branching constituent, then its maximal branching depth for this position is one. And a language which allows a branching constituent, but not one that contains a second branching constituent has a maximal branching depth of two.

In English, at least, the selection of maximal branching depth seems to be related to the depth of embedding. A constituent on the non-recursive side is most heavily restricted when it is a sister to the head; it has a maximal branching depth of one. A constituent is less heavily restricted when an adjunct to X', and has a maximal branching depth of two. When constituents appear in the specifier position they are (apparently) not restricted at all.

The correlation between maximal branching depth and position in the tree Consistency constraints might be schematized in the following way:



Although the correlation between depth of embedding and the degree to which branching is constrainted is intriguing, it doesn't seem to be possible to completely predict the form of the Consistency constraint from the complement/adjunct/specifier status of a constituent. In our data, for example, Amharic allows 2 deep complements, while Zapotec and English allow only 1 deep complements.

So far, we have interpreted the constraint Head ≺ Comp constraint categorially; either a head precedes its complement or it does not. If we interpret it gradiently, it may be possible to do away with a distinct Consistency constraint in favor of a revised ordering constraint. Let us think of the ordering constraints in terms of their inverses:

54) *Comp ≺ Head (*n*)
A complement (of branching depth *n)* must not precede its head

55) *Head ≺ Comp (*n*)
A head must not precede a complement (of branching depth *n*)

When $n = 0$, then these constraints are equivalent to the categorial versions seen before. When $n > 0$, they restrict successively larger and more complex types of constituents. We would expect larger branching depths to outrank smaller branching depths. Inverted or 'flipped' word orders result when some constraint is interpolated among the ordering constraints.

The constraint ranking relevant to English swiping would be as follows:

56) *Comp ≺ Head (2) ≫ *Comp ≺ Head (1) ≫ *FocStr (Interrog) ≫ *Comp ≺ Head (0) ≫ Wh-L

The constraint ranking for Zapotec would be

57) *Comp ≺ Head (2) ≫ *Comp ≺ Head (1) ≫ Wh-L≫ *Comp ≺ Head (0) ≫ *FocStr

(Interrog)

The constraint ranking for Amharic would be

58) *Head ≺ Comp (2) ≫ Align (P, L, PP, L) ≫ *Head ≺ Comp (1) ≫ *Head ≺ Comp (0) ≫ Wh-L, *FocStr (Interrog)

Many languages appear to be completely consistent in their branching order (e.g. Biblical Hebrew, Japanese). In such languages the *Comp ≺ Head (or *Head ≺Comp) constraints would be undominated.

The relevant ranking for English AdjP is less clear. Is there a parallel *Adjunct ≺ Head constraint? As Dryer (1988) has shown, statistically there is no clear correlation between the order of Adj and N and the OV/VO distinction (contra Greenberg 1966). I will leave this problem for future research.

8 References

Black, Cheryl. 1994. *Quiegolani Zapotec syntax*. PhD thesis. University of California, Santa Cruz.

Bresnan, Joan. 1998. Optimal syntax. Ms. Stanford University. (Available at http://www-lfg.stanford.edu/lfg/bresnan/pt3.ps.)

Bresnan, Joan. 2000. *Lexical functional syntax*. Blackwell.

Broadwell, George Aaron. 1999a. Focus alignment and optimal order in Zapotec. *Proceedings of the 35th Chicago Linguistics Society.* (Also available at http://www.albany.edu/anthro/fac/broadwell.htm)

Broadwell, George Aaron. 1999b. The interaction of focus and constituent order in San Dionicio Ocotepec Zapotec. *Proceedings of the LFG 99 conference,* ed. by Miriam Butt and Tracy Holloway King. CSLI Publications. *http//www-csli.stanford.edu/publications/*

Broadwell, George Aaron. 2000. On the phonological conditioning of clitic placement in Zapotec. in *Proceedings of the Workshop on Structure and Constituency in the Languages of the Americas.* Toronto: University of Toronto Department of Linguistics.

Broadwell, George Aaron. 2001. Optimal order and pied-piping in San Dionisio Zapotec, in Sells (2001b).

Chisarek, Erika and John Payne. 2001. Modelling possessor constructions in English and Hungarian: An LFG approach. Proceedings of the LFG 01 conference. Butt, Miriam and Tracy Holloway King, eds. http://www.csli-publications.stanford.edu.

Choi, Hye-Won. 1999. *Optimizing structure in context.* Stanford: CSLI.

Dryer, Matthew S. 1988. Object-Verb Order and Adjective-Noun Order: Dispelling a Myth. *Lingua* 74: 185-217.

Gazdar, Gerald; Ewan Klein; Geoffrey Pullum; and Ivan Sag. 1985. *Generalized phrase structure grammar*. Cambridge, Ma: Harvard.

Greenberg, Joseph. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, Joseph (ed.) *Universals of language.* Cambridge, Ma: MIT Press, pp. 73-116.

Halefom, Girma. 1994. *The syntax of functional categories: A study of Amharic*. Université du Québec à Montreal, Ph.D. thesis.

Kayne, Richard. 1994. *The antisymmetry of syntax* Cambridge, MA: MIT Press.

Lee, Hanjung. 2001. Markedness and word order freezing. in Sells (2001b).

Leslau, Wolf. 1995. *Reference grammar of Amharic*. Wiesbaden: Harrasowitz.

Longobardi, Giuseppe. 1991. Extraction from NP and the proper notion of head government. In A. Giorgi and G. Longobardi, eds. *The syntax of noun phrases.* Cambridge: Cambridge University Press, pp 57-112.

Merchant, Jason. to appear. Swiping in Germanic. In Jan-Wouter Zwart and Werner Abraham (eds.), Studies in Comparative Germanic Syntax.Amsterdam: John Benjamins.

Morimoto, Yukiko. 2001. Verb raising and phrase structure variation in OT. in Sells (2001b).

Pollard, Carl and Ivan Sag. 1987. *Information-based syntax and semantics.* Stanford:CSLI.

Sells, Peter 2001a. *Structure, alignment, and optimality in Swedish.* Stanford: CSLI.

Sells, Peter. ed. 2001b. *Formal and empirical issues in Optimality Theoretic syntax.* Stanford:CSLI

Publications.
Smith Stark, Thomas. 1988. 'Pied-piping' con inversion en preguntas parciales. Ms. Centro de estudios lingüísticos y literarios, Colegio de México y Seminario de lenguas indígenas.
Tremblay, Mireille and Ouadia Kabbaj. 1989. The internal structure of PPs in Amharic. in John Hutchison and Victor Manfredi, eds. *Current approaches to African linguistics* 7:185-197
Van Valin, Robert D. and Randy LaPolla. 1997. *Syntax: Structure,meaning and function.* Cambridge.

Parsing with PCFGs and Automatic F-Structure Annotation

Aoife Cahill Mairéad McCarthy Josef van Genabith Andy Way
Computer Applications, Dublin City University, Ireland
{ acahill, mccarthy, josef, away }@computing.dcu.ie

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

Abstract

The development of large coverage, rich unification- (constraint-) based grammar resources is very time consuming, expensive and requires lots of linguistic expertise. In this paper we report initial results on a new methodology that attempts to partially automate the development of substantial parts of large coverage, rich unification-(constraint-) based grammar resources. The method is based on a treebank resource (in our case Penn-II) and an automatic f-structure annotation algorithm that annotates treebank trees with proto-f-structure information. Based on these, we present two parsing architectures: in our pipeline architecture we firstextract a PCFG from the treebank following the method of [Charniak, 1993; Charniak, 1996], use the PCFG to parse new text, automatically annotate the resulting trees with our f-structure annotation algorithm and generate proto-f-structures. By contrast, in the integrated architecture we firstautomatically annotate the treebank trees with f-structure information and then extract an annotated PCFG (A-PCFG) from the treebank. We then use the A-PCFG to parse new text to generate proto-f-structures. Currently our best parsers achieve more than 81% f-score on the 2400 trees in section 23 of the Penn-II treebank and more than 60% f-score on gold-standard proto-f-structures for 105 randomly selected trees from section 23.

1 Introduction

The development of large coverage, rich unification-(constraint-) based grammar resources is very time consuming, expensive and requires considerable linguistic expertise [Butt et al, 1999; Riezler et al, 2002].

In this paper we report initial results on a new methodology that attempts to partially automate the development of substantial parts of large coverage, rich unification-(constraint-) based grammar resources.

A large number of researchers (cf. [Charniak, 1996; Collins, 1999; Hockenmaier and Steedman, 2002]) have developed parsing systems based on the Penn-II [Marcus et al, 1994] treebank resource but to the best of our knowledge, to date, none of them have attempted to semi-automatically derive large coverage, rich unification(constraint) grammar resources.

Our method is based on the Penn-II treebank resource and an automatic Lexical-Functional Grammar [Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001] f-structure annotation algorithm that annotates treebank trees with proto-f-structure information.

Based on these, we present two parsing architectures: in our *pipeline* architecture we firstextract a PCFG from the treebank following the method of [Charniak, 1993; Charniak, 1996], use the PCFG to parse new text, automatically annotate

the resulting trees with our f-structure annotation algorithm and generate proto-f-structures. By contrast, in the *integrated* architecture we first automatically annotate the treebank trees with f-structure information and then extract an annotated PCFG (A-PCFG) from the treebank. We then use the A-PCFG to parse new text to generate proto-f-structures.

The paper is structured as follows: first, we briefly describe the automatic f-structure annotation algorithm and proto-f-structure representations. Second, we introduce our two parsing architectures: the pipeline model and the integrated model. Third, we evaluate our approaches quantitatively and qualitatively in terms of f-structure fragmentation and precision and recall results against trees and manually encoded gold standard f-structures. We also compare our results with a PCFG resulting from a parent transformation as discussed in [Johnson, 1999] and we conduct a number of thresholding grammar compaction experiments [Krotov et al, 1998]. We briefly compare our approach with that of [Riezler et al, 2002]. Fourth, we outline further work on "proper" (rather than "proto-") f-structures and finally we summarise and conclude.

2 An Automatic F-Structure Annotation Algorithm

In this section we describe our automatic f-structure annotation algorithm by means of an example. Consider the following Penn-II treebank tree:

```
(S (NP-SBJ (DT The)
           (NN investment)
           (NN community))
   (, ,)
   (PP (IN for)
       (NP (CD one)))
   (, ,)
   (VP (VBZ has)
       (VP (VBN been)
           (VP (VBG anticipating)
               (NP (DT a)
                   (JJ speedy)
                   (NN resolution)))))
   (. .))
```

In LFG this tree would be associated with an f-structure of the following form representing abstract syntactic information approximating to predicate-argument-modifier(or: rich dependency-) structure:[1]

[1]Note that our f-structures are more hierarchical with XCOMP functions for temporal and aspectual auxiliaries than e.g. the f-structures given in [Butt et al, 1999]. The repetition of the subject NP

```
subj : spec : det : pred : the
       headmod : 1 : num : sg
                     pers : 3
                     pred : investment
       num : sg
       pers : 3
       pred : community
adjunct : 1 : obj : pred : one
              pred : for
xcomp : subj : spec : det : pred : the
               headmod : 1 : num : sg
                             pers : 3
                             pred : investment
               num : sg
               pers : 3
               pred : community
        xcomp : subj : spec : det : pred : the
                       headmod : 1 : num : sg
                                     pers : 3
                                     pred : investment
                       num : sg
                       pers : 3
                       pred : community
                obj : spec : det : pred : a
                      adjunct : 2 : pred : speedy
                      pred : resolution
                      num : sg
                      pers : 3
                participiple : pres
                pred : anticipate
        pred : be
        tense : past
pred : have
tense : pres
```

F-structures are associated with strings and their parse trees (c-structures) in terms of annotating nodes in parse trees (and hence the corresponding CFG rules) with f-structure annotations (in simple cases, attribute value structure equations; more generally, expressions in a full equality logic including disjunction, negation etc. [Johnson, 1988]). The f-structure in our example would be induced by the following annotated CFG rules

is due to a reentrancy (↑ SUBJ = ↑ XCOMP SUBJ) annotation in the VP rule.

S	→	NP-SBJ ↑subj=↓	PP ↓∈↑adjn	VP ↑=↓
NP-SBJ	→	DT ↑spec=↓	NN ↓∈↑headmod	NN ↑=↓
PP	→	IN ↑=↓	NP ↑obj=↓	
NP	→	CD ↑=↓		
VP	→	VBZ ↑=↓	VP ↑xcomp=↓ ↑subj=↓subj	
VP	→	VBN ↑=↓	VP ↑xcomp=↓ ↑subj=↓subj	
VP	→	VBG ↑=↓	NP ↑obj=↓	
NP	→	DT ↑spec=↓	JJ ↓∈↑adjn	NN ↑=↓

together with equations resulting from lexical entries.

The question is how can we construct such an LFG grammar? Traditionally, LFG (and HPSG [Pollard and Sag, 1994]) grammatical resources have been constructed manually. For large coverage and rich grammars that scale to the Penn-II treebank data (about 1 million words of WSJ text) [Butt et al, 1999; Riezler et al, 2002], this can easily accumulate to a person decade.

Is there any way of (at least partially) automating the construction of large coverage, rich, unification-based grammatical resources? From the large literature on probabilistic parsing it is clear that given a treebank we can easily extract a probabilistic CFG (following the method of [Charniak, 1993; Charniak, 1996]. Indeed, such grammars are at the heart of many probabilistic parsing approaches such as [Charniak, 1993; Charniak, 1996; Collins, 1999; Hockenmaier and Steedman, 2002], to mention but a few. However, to the best of our knowledge, to date, there have been no attempts at semi-automatically deriving large coverage, rich unification-(constraint-) based grammar resources.

In order to obtain a unificationgrammar, we need the functional annotations (the f-structure constraints) in addition to the CFG rules. Of course, theoretically, we

could annotate the CFG rules extracted from a treebank manually. For the Penn-II treebank, however, we would be required to manually annotate more than 19,000 extracted rule types resulting in at least 2 person years of annotation work (on a very conservative estimate with 10 minutes annotation time for each rule type and no time budgeted for testing, comparison, improvement and verification cycles).

2.1 Previous Work

A number of researchers have addressed automatic functional structure identification or annotation in CFG trees or, alternatively, direct transformation of trees into functional structures.

To date, we can distinguish three different types of automatic f-structure annotation architectures:[2]

- annotation algorithms,
- regular expression based annotation,
- flat, set-based tree description rewriting.

All approaches are based on exploiting categorial and configurational information encoded in trees. Some also exploit the Penn-II functional annotations.[3]

Annotation algorithms come in one of two forms. They may

- *directly* (recursively) transduce a treebank tree into an f-structure – such an algorithm would more appropriately be referred to as a tree to f-structure transduction algorithm;
- *indirectly* (recursively) annotate CFG treebank trees with f-structure annotations from which an f-structure can be computed by a constraint solver.

As was recently pointed out to us by Shalom Lappin (p.c.), the earliest approach to automatically identify SUBJ, OBJ etc. nodes in CFG trees structures is probably [Lappin et al, 1989]. Their algorithm *identifies* nodes in CFG trees (output of the PET parser) corresponding to grammatical functions to facilitate the statement of

[2] These have all been developed within an LFG framework and although we refer to them as automatic f-structure annotation architectures, they could equally well be used to annotate treebanks with e.g. HPSG typed feature-structures [Pollard and Sag, 1994] or Quasi-Logical Form (QLF) [Liakata and Pulman, 2002] annotations.

[3] Note that apart from SBJ and LGS, functional annotations or tags in the Penn-II treebank do not provide LFG type predicate-argument style annotations but semantically classify e.g. modifying PP constituents as TMP (temporal), LOC (locative) etc. modifiers.

transfer rules in a machine translation project. It does not construct f-structures (or other attribute-value structures) although it could easily form the basis of such an algorithm for trees generated by the PET parser.

The first *direct* automatic f-structure annotation algorithm we are aware of is unpublished work by Ron Kaplan (p.c.) from 1996. Kaplan worked on automatically generating f-structures from the ATIS corpus to generate data for LFG-DOP applications. The approach implements a direct tree to f-structure transduction algorithm, which walks through the tree looking for different configurations (e.g. NP under S, 2nd NP under VP, etc.) and "folds" or "bends" the tree into the corresponding f-structure.

A regular expression-based, *indirect* automatic f-structure annotation methodology is described in [Sadler et al, 2000]. The idea is simple: first, the CFG rule set is extracted from the treebank (fragment); second, regular-expression based annotation principles are defined; third, the principles are automatically applied to the rule set to generate an annotated rule set; fourth, the annotated rules are automatically matched against the original treebank trees and thereby f-structures are generated for these trees. Since the annotation principles factor out linguistic generalisations, their number is much smaller than the number of CFG treebank rules. In fact, the regular expression-based f-structure annotation principles constitute a principle-based LFG c-structure/f-structure interface.

In a companion paper, [Frank, 2000] develops an automatic annotation method that in many ways is a generalisation of the regular expression-based annotation method. The idea is again simple: trees are translated into a flat set representation format in a tree description language and annotation principles are defined in terms of rules employing a rewriting system originally developed for transfer-based machine translation architectures. In contrast to [Sadler et al, 2000] which applies only to "local" CFG rule contexts, [Frank, 2000] can consider arbitrary tree fragments. Secondly, it can be used to define both order-dependent cascaded and order-independent annotation systems. [Liakata and Pulman, 2002] have recently developed a similar approach to map Penn-II trees to QLFs.

The approaches detailed in [Sadler et al, 2000; Frank, 2000] and compared in [Frank et al, 2002] are proof-of concept and operate on small subsets of the AP and Susanne corpora.[4]

2.2 A New Automatic Annotation Algorithm

In our more recent research [Cahill et al, 2002a; Cahill et al, 2002b], we have developed an algorithmic indirect annotation method for the > 49.000 parse annotated

[4] This is not to say that these approaches cannot be scaled to a complete treebank!

strings in the WSJ section of the Penn-II treebank.

The algorithm is implemented as a recursive procedure (in Java) which annotates Penn-II treebank tree nodes with f-structure information. The annotations describe what we call "proto-f-structures". Proto-f-structures

- encode basic predicate-argument-modifier structures;
- interpret constituents locally (i.e. do not resolve long-distance dependencies or "movement" phenomena encoded as traces in the Penn-II trees);
- may be partial or unconnected (the method is robust: in case of missing annotations a sentence may be associated with two or more unconnected f-structure fragments rather than a single complete f-structure).

Even though the method is encoded in the form of an annotation algorithm, we did not want to completely hardwire the linguistic basis for the annotation into the procedure. In order to support maintainability and reusability of the annotation algorithm and the linguistic information encoded within, the algorithm is designed in terms of three main components thatare applied in sequence:

L/R Context APs $\Rightarrow$ Coordination APs $\Rightarrow$ Catch-All APs

L/R Context Annotation Principles are based on a tripartition of the daughters of each local tree (of depth one, i.e. of CFG rules) into a prefix, head and suffix sequence. We automatically transform the Penn-II trees into head-lexicalised trees by adapting the rules of [Magerman, 1994; Collins, 1999]. For each LHS in the Penn-II CFG rule types we construct an annotation matrix. The matrix encodes information on how to annotate CFG node types in the left (prefix)and right (suffix) context. Table 1 gives a simplified matrix for NP rules.

NP	left context	head	right context
subcat	DT,CD: $\uparrow$=$\downarrow$	NN,NNS,NP: $\uparrow$=$\downarrow$	…
non-sub	ADJP: $\downarrow \in \uparrow$adjn NN,NNS,NP: $\downarrow \in \uparrow$headmod …		SBAR,VP: $\uparrow$relmod=$\downarrow$ PP: $\downarrow \in \uparrow$adjn NN,NNS,NP: $\uparrow$app=$\downarrow$

Table 1: Simplified,partial annotation matrix for NP rules

For each LHS category, the annotation matrices are populated by analysing the most frequent rules types such that the token occurrence of the rule types in the corpus covers at least 85%. To give an example, this means that instead of looking

at $>$ 6,000 NP rule types in the Penn-II corpus, we only look at the 102 most frequent ones to populate the NP annotation matrix.

To keep L/R context annotation principles simple and perspicuous, they only apply if the local tree does not contain coordination. Like and unlike coordinate structures are treated by the second component of our annotation algorithm, Finally, the algorithm has a catch-all and clean-up component. Lexical information is supplied via macros associated with each pre-terminal tag type.

The automatic annotation algorithm generates the following annotations. The annotations are collected, sent to a constraint solver and the f-structure shown before is generated.

```
(S
   (NP-SBJ[up-subj=down]
      (DT[up-spec:det=down] The[up-pred='the'])
      (NN[down-elem=up:headmod]
         investment[up-pred='investment',up-num=sg,up-
pers=3])
      (NN[up=down]
         community[up-pred='community',up-num=sg,up-
pers=3]))
   (, ,)
   (PP[down-elem=up:adjunct]
      (IN[up=down] for[up-pred='for'])
      (NP[up-obj=down]
         (CD[up=down] one[up-pred='one'])))
   (, ,)
   (VP[up=down]
      (VBZ[up=down]
          has[up-pred='have',up-tense=pres])
      (VP[up-xcomp=down,up-subj=down:subj]
         (VBN[up=down] been[up-pred='been',up-tense=past])
         (VP[up-xcomp=down,up-subj=down:subj]
            (VBG[up=down]
                anticipating[up-pred='anticipate',up-
participle=pres])
            (NP[up-obj=down]
               (DT[up-spec:det=down] a[up-pred='a'])
               (JJ[down-elem=up:adjunct] speedy[up-
pred='speedy'])
               (NN[up=down]
                  resolution[up-pred='resolution',up-
num=sg,up-pers=3])))))
   (. .))
```

Annotation coverage is measured in terms of f-structure fragmentation. The method is robust and in case of missing annotations may deliver unconnected f-structure fragments for a tree. Annotation accuracy is measured against a manually constructed gold-standard set f-structures for 105 trees randomly selected form Section 23 of the Penn-II treebank.[5]

# f-str.	[Cahill et al, 2002a]		[Cahill et al, 2002b]		current	
frags	# sent	percent	# sent	percent	# sent	percent
0	2701	5.576	166	0.343	120	0.25
1	38188	78.836	46802	96.648	48304	99.75
2	4954	10.227	387	0.799	0	0
3	1616	3.336	503	1.039	0	0
4	616	1.271	465	0.960	0	0
5	197	0.407	70	0.145	0	0
6	111	0.229	17	0.035	0	0
7	34	0.070	8	0.017	0	0
8	12	0.024	6	0.012	0	0
9	6	0.012	0	0	0	0
10	4	0.008	0	0	0	0
11	1	0.002	0	0	0	0

Table 2: Automatic proto-f-structure annotation fragmentation results

Table 2 shows the progress we have made over the last 6 months. Initially, 78.836% of the trees in the Penn-II treebank were associated with a single complete proto-f-structure with quite a number of trees having more than one proto-f-structure fragments and 2701 trees failing to get an f-structure because of inconsistent annotations. Currently our automatic annotation algorithm associates 99.75% of the trees with a complete (unfragmented) proto-f-structure while 120 trees do not receive any proto-f-structure.

Table 3 reports the quality of the f-structures generated in terms of precision and recall against our manually encoded gold-standard f-structures. We currently achieve P&R results of 0.94 and 0.87 for preds-only proto-f-structures.[6]

[5]Our gold-standard f-structures are available for inspection at `http://www.computing.dcu.ie/~away/Treebank/treebank.html`.

[6]Preds-only f-structures only show paths ending in a PRED feature.

	[Cahill et al, 2002b]		current	
	All annotations	Preds-only	All annotations	Preds-only
Precision	0.95	0.94	0.93	0.94
Recall	0.94	0.89	0.90	0.87

Table 3: Precision and Recall on descriptions of proto-f-structures

3 Two Parsing Architectures

Once we have the annotation algorithm and the annotated version of the Penn-II treebank, we can parse new text into trees and f-structures in two ways:

In our *pipeline* architecture we first extract a PCFG (probabilistic context free grammar) from the unannotated version of the Penn-II treebank and use this to parse new text. We then take the most probable tree associated with a string and send it to our automatic annotation algorithm. The algorithm annotates the tree with f-structure equations. We collect the equations and send them to a constraint solver to generate an f-structure:

Treebank → PCFG → text → Trees → f-str ann. → F-Str

In our *integrated* architecture we first annotate the Penn-II treebank trees with f-structure information using our automatic annotation algorithm and then extract an annotated PCFG (A-PCFG) from the annotated treebank. We treat strings consisting of CFG categories followed by one or more f-structure annotations, e.g. `NP[up-subj=down]`, as monadic categories. The effect of this is that the rules extracted in the A-PCFG will be different and will be associated with different probabilities compared to the simple PCFG rules extracted for the pipeline model. For instance, the A-PCFG distinguishes between subject and object NPs whereas the PCFG does not. We then parse text with these annotated rules and pick again the tree with the highest probability. We then collect the f-structure annotations from this tree and send them to a constraint solver to generate an f-structure:

Treebank → f-str ann. → A-PCFG → text → A-Trees → F-Str

We employ the following pre-processing steps to extract both the PCFG and the A-PCFG from the treebank: every tree is associated with a `Root` category node; we eliminate empty productions; we remove unary branches percolating daughter category information up in the tree and finally we annotate auxiliary verbs with an `Aux` category but only in the case where there is a sister `VP` node somewhere to the right of the `Aux` node.

Our grammars are extracted from sections 01 – 21 in the WSJ part of the Penn-II treebank and we measure our results against the held out section 23. All results are for sentences with length less than or equal to 40. We assume correct tagging of the roughly 2400 test sentences in section 23, i.e. we parse the tag sequences for those sentences given by the Penn-II treebank.

Our parsing experiments are based on a Java implementation of a CYK parser. For parsing, the grammar has to be transformed into Chomsky Normal Form (with binary branching productions). After parsing, the output trees are retransformed into the possibly n-ary branching treebank trees without loss of information. The parser is effi cient($O(n^3)$) and returns the single most probable tree.

We deliberately decided to use simple PCFG-based parsing technology in our experiments. The reason is that we want to be able to construct a system suitable for on-line parsing (this is not always possible with log-linear models as they may require unpacking of ambiguities). Mathematically speaking, however, this means that we do not use proper probability models. The reason is simple: PCFG technology is based on independence assumptions. The probability of a tree is the product of the probabilities of the productions in the tree (as each CFG production is assumed to be independent). What can happen in our model is that the parser returns the most probable tree but the f-structure equations generated for this tree (in the pipeline or the integrated architecture) are inconsistent and the constraint resolver cannot generate an f-structure. In such a case the probability mass associated with this tree is lost. Based on our experiments this case is extremely rare (less than 0.1%) so that the advantages of our engineering approach (speed and the ability to construct on-line applications) outweigh the disadvantages.

In order to evaluate our parsing results, we measure precision and recall using `evalb` on the 2400 trees in section 23. In order to evaluate the f-structures we measure f-structure fragmentation and precision and recall against manually encoded f-structures for 105 randomly extracted sentences from section 23.

We conducted a number of experiments. We extracted a simple PCFG for the pipeline model. We extracted an annotated A-PCFG for the integrated model. The integrated model with its f-structure annotations on CFG categories allows us to distinguish between subject and object `NP`s, for example, and associates different probabilities with rules expanding such `NP`s. A similar effect can be achieved in terms of a simple parent transformation on treebank trees: every daughter node receives an additional annotation involving its mother CFG category. An `NP-S`, e.g., is a an `NP` under `S` (i.e. a subject `NP` in English) while an `NP-VP` is a an `NP` under `VP` (i.e. an object `NP`). This approach (attributed to Charniak) has been studied extensively in [Johnson, 1999]. In our experiments we compare the parent transformation PCFG+P with our f-structure annotation pipeline and integrated approach. Finally, we conducted a number of experiments with simple thresholding grammar

compaction techniques [Krotov et al, 1998]. Our results are summarised in the following tables.

		Pipeline		Integrated	
		PCFG	PCFG+P	A-PCFG	A-PCFG+P
Full Grammar	# Rules	19439	30026	29216	35815
	F-Score Labelled	77.37	80.49	81.26	81.18
	F-Score Unlabelled	79.89	82.56	83.16	83.08
	F-Str Fragm.	95.80	95.64	94.69	94.93
	F-Score Gold Stnd.	51.98	56.28	60.66	60.21
	# Parses	2240	2227	2223	2207

All fractions are percentages. F-scores are calculated as $2\times$precision$\times$recall/(precision+recall). The first row shows the number of rules extracted for each grammar. PCFG+P is the standard PCFG with the parent transformation and A-PCFG+P is the annotated PCFG with the parent transformation (here the functional annotations of the mother nodes do not carry over to daughter nodes). Interestingly, PCFG+P and the automatically annotated A-PCFG are of roughly equal size. The next two rows show labelled and unlabelled f-score (composite precision and recall) on trees against the section 23 reference trees as computed by `evalb`. The parent transfom shows a 3% advantage over the PCFG while our A-PCFG performs slightly better with an almost 4% advantage for labelled f-score. Interestingly, combining the parent transform with our automatic f-structure annotation does not seem to improve results. The next row measures f-structure fragmentation. Our current results seem to indicate that (surprisingly perhaps) the pipeline parsing architecture outperforms the integrated model with respect to f-structure fragmentation. The next row reports f-score results against the hand-coded, gold-standard reference f-structures for the 105 randomly selected trees in section 23. The results show clearly that currently the f-structures generated by the integrated model are of higher quality than the ones generated by the pipeline model.

		Pipeline		Integrated	
		PCFG	PCFG+P	A-PCFG	A-PCFG+P
Threshold 1	# Rules	7906	12117	11545	13970
	F-Score Labelled	76.81	79.79	80.08	80.39
	F-Score Unlabelled	79.41	81.92	82.09	82.30
	F-Str Fragm.	96.14	93.92	94.79	95.41
	F-Score Gold Stnd.	51.99	56.07	59.15	60.42
	# Parses	2227	2173	2189	2115

This table reports on a simple thresholding grammar compaction experiment with a threshold set to 1. This means that every rule that occurs only once in the

Penn-II training set is discarded from the grammar. The first row in the table shows that this corresponds to a reduction in size of 60–70% against the original grammars with a very small loss in coverage as indicated in the last row in the table. The reduction in size results in a corresponding increase in parsing speed. According to our experiments, a threshold of 1 results in an increase in speed by a factor of 6–12. To give just one example, parsing section 23 (about 2400 sentences of length average around 20 words) takes 48.71 CPU hours with the full PCFG but "only" 7.91 CPU hours with the compacted PCFG under threshold 1. The table shows that labelled and unlabelled recall on trees as well as f-structure fragmentation and the quality of the f-structures measured as f-score against the gold standard f-structures suffer surprisingly little from this level of thresholding.

		Pipeline		Integrated	
		PCFG	PCFG+P	A-PCFG	A-PCFG+P
Threshold 2	# Rules	5433	8400	7924	9538
	F-Score Labelled	76.18	79.64	79.96	79.90
	F-Score Unlabelled	78.75	81.74	82.01	81.92
	F-Str Fragm.	96.38	96.34	95.10	95.75
	F-Score Gold Stnd.	50.95	55.65	59.07	57.46
	# Parses	2210	2104	2143	2025

Setting the threshold to 2 (i.e. discarding rules that are used less than or equal to 2), again shows remarkably little overall effect apart from a slight decrease in coverage and slightly worse overall f-score results. F-structure fragmentation, by contrast, even decreases in some cases. Parsing speed increases by a factor of 15–20 compared to the full grammar. To give an example, parsing section 23 takes 3.08 CPU hours with the compacted PCFG under a threshold of 2.

		Pipeline		Integrated	
		PCFG	PCFG+P	A-PCFG	A-PCFG+P
Threshold 5	# Rules	3246	4899	4520	5447
	F-Score Labelled	75.07	79.16	78.66	79.24
	F-Score Unlabelled	77.75	81.23	80.77	81.21
	F-Str Fragm.	96.19	96.36	95.77	96.30
	F-Score Gold Stnd.	48.99	54.63	56.26	58.80
	# Parses	2125	1922	1963	1784

A more severe threshold of 5 shows more marked results. On the one hand, all grammars now parse the 2400 sentences in section 23 in less than 1 CPU hour in our Java implementation. However, there is a marked decrease in both coverage and quality. In terms of coverage, the table shows that the more fine-grained grammars PCFG+P, A-PCFG and A-PCFG+P suffer more under severe compaction than the simple PCFG.

It is difficult to compare our approach with that of [Riezler et al, 2002]. The proto-f-structures generated in our approach are much more coarse-grained than the detailed proper f-structures delivered by their carefully handcrafted and optimised LFG grammar. Riezler et. al use an off-line exponential discriminative disambiguation method and achieve f-structure f-scores close to 80% (as against about 60% in our semi-automatic grammar development approach). They use partial bracketing derived from the Penn-II treebank in section 23 to guide their parses whereas we report our results for free (unguided) parses of tagged strings in section 23.

4 Current Work

In our work to date, we have shown how the development of large coverage, rich unification-based grammar resources can be partially automated. However, our research is only a first step in that direction. The reason is that the attribute-value structures we parse into are proto-f-structures. Proto-f-structures interpret linguistic material locally where it occurs in the tree and not where it should be interpreted semantically. Examples of such "non-local" phenomena are extraposition, topicalisation, wh-dependencies, distribution of subjects into VP-coordinate structures to mention but a few. Penn-II employs a rich arsenal of traces and empty productions (nodes that do not realise any lexical material) to coindex "displaced material" (and partly to indicate passive constructions) with positions where the material "originated" (or, to put it in more neutral terms, positions where it should be interpreted semantically). The proto-f-structure annotation algorithm ignores all such traces and empty productions. In our current work we have extended our automatic annotation algorithm to exploit this information. We do this in terms of a new fourth component (Traces) to our annotation algorithm:

L/R Context APs $\Rightarrow$ Coord APs $\Rightarrow$ Catch-All APs $\Rightarrow$ Trace APs

So far we have incorporated traces for A and A′ movement (movement to argument and non-argument positions) including traces for wh-questions, relative clauses, fronted elements and subjects of participle clauses, gerunds and infinitival clauses (including both controlled and arbitrary PRO) as reentracies in our f-structures. Null constituents are now treated as full nodes in the annotation (except passive empty object NP) and traces are recorded in terms of INDEX = n f-structure annotations. Traces without indices are translated into arbitrary PRO. The encoding of passive is important as LFG "surface-syntactic" grammatical functions such as SUBJ and OBJ differ from "logical" grammatical functions: surface-syntactic grammatical functions are identified in terms of e.g. agreement phenomena while logical

grammatical functions are more akin to thematic roles. The surface-syntactic subject of a passive sentence is usually a logical object, while a surface grammatical object of an optional by-prepositional phrase is usually the logical subject.

In order to evaluate the new "proper-" (rather than "proto-"), f-structures we have updated our manually encoded 105 gold-standard reference f-structures from section 23 with traces encoding reentrancies reflecting locations where linguistic material is encountered and where it should be interpreted. Our current results for proper f-structures (fragmentation and precision and recall) are summarised in the following tables:

# frags.	# sent.
0	507
1	47916
2	1

preds only	
Precision	0.93
Recall	0.87

Proper f-structure annotation precision and recall results suffer very little compared to the best (in terms of coverage) proto-f-structure annotation (preds only: precision 0.93 against 0.94; recall 0.87 against 0.87), indicating that the extended annotation algorithm can reliably determine traces for wh-questions, relative clauses, fronted elements and subjects of participle clauses, gerunds and infinitival clauses as well as passives, and reflect them accurately in terms of indices in the f-structure representations. At this stage, however, fragmentation goes up considerably. We are confident of being able to reduce the number of sentences that do not receive a proper f-structure significantly in further work.

5 Future Work

Penn-II trees annotated with proper f-structures reflecting non-local dependencies are an important ingredient in automatically deriving large coverage unification grammar resources. Recall, however, that the f-structure reentrancies were induced from traces and empty productions in full Penn-II trees relating linguistic material to where it should be interpreted semantically. Full Penn-II style trees with detailed coindexation traces and empty productions are not the standard fare in probabilistic parsing. Indeed, empty productions are usually eliminated in PCFGs and similar approaches to parsing.

In view of this, where do we now get our CFG backbone to parse into proper f-structures? Fortunately, LFG comes to the rescue: standardly, LFG assumes a very surface-oriented approach to its CFG backbone. Non-local phenomena are dealt with in terms of functional uncertainty equations on the level of f-structure representations. Given our annotated Penn-II treebank resource we can automatically

compute shortest paths through proper f-structure representations relating material coindexed in f-structure. For each of the long-distance phenomena we collect the paths and compact them into a regular expression to obtain e.g. the functional uncertainty equation for sentential TOPIC etc. During parsing (or annotation), our automatic annotation algorithm then associates every sentential TOPIC function with this equation and the reentrancy is resolved by the constraint solver. In order to make this work properly, we will need to enforce completeness and coherence constraints on our f-structures. Completeness and coherence constraints rely on semantic form values of PRED features. These semantic forms provide subcategorisation information in the form of the syntactic functions required by the predicate governing that level of f-structure. Semantic forms are supplied lexically. Our approach to date, however, is mostly non-lexical. Annotation is driven by categorial and configurationalinformation in Penn-II trees, templates for pre-terminal tags and occasionally by functional annotations (-TMP, -LOC, -CLR etc.) in the Penn-II treebank. How are we going to obtain the semantic forms? One possibility is to use a lexical resource such as COMLEX. Given our annotated version of the Penn-II treebank another option is available to us: if the quality of the automatically generated f-structures is good we can automatically read off semantic forms from these f-structures. Following [van Genabith et al, 1999], for each level of embedding in an f-structure we determine the value of the PRED function and collect all subcategorisable grammatical functions present at that level of f-structure. From these we construct a semantic form. We will explore this in the next stage of our research.

6 Conclusions

In this paper we have developed the first steps and presented initial results of a new methodology to partially automate the development of large coverage, rich unification-based grammar resources. The method is based on an automatic f-structure annotation algorithm that annotates trees in the Penn-II treebank with proto-f-structure information. We have presented two parsing architectures based on this resource: a pipeline model and an integrated model. Currently our best parsers achieve more than 81% f-score on the 2400 trees from section 23 and more than 60% f-score on gold-standard f-structures for 105 randomly selected trees from the same section. We have compared our results with the parent transform approach investigated in [Johnson, 1999] and have conducted a number of thresholding grammar compaction experiments [Krotov et al, 1998]. We have briefly compared our approach with that of [Riezler et al, 2002]. We have presented results of current work on automatic annotation of "proper-" (as opposed to "proto-") f-structures and outlined future research involving functional uncertainty equations and semantic

forms to semi-automatically develop grammatical resources that parse new text into full f-structures.

Acknowledgements

Ron Kaplan, Stefan Riezler, Mary Dalrymple, John Maxwell, Tracy King and Hiroshi Masuichi have provided comments and support. The research reported in the present paper is supported by Basic Research Grant SC/2001/186 from Enterprise Ireland.

References

[Bresnan, 2001] J. Bresnan 2001. *Lexical-Functional Syntax.* Blackwell, Oxford.

[Butt et al, 1999] Miriam Butt, T.H. King and F. Second, 1999. *A grammar writer's cookbook.* Stanford, Calif. : CSLI Publications

[Cahill et al, 2002a] Cahill, A., M. McCarthy, J. van Genabith and A. Way (2002). Automatic Annotation of the Penn Treebank with LFG F-Structure Information. in *Proceedings of the LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, Las Palmas, Canary Islands, Spain, pp.8–15.

[Cahill et al, 2002b] Cahill, A., M. McCarthy, J. van Genabith and A. Way (2002). Evaluating Automatic F-Structure Annotation for the Penn-II Treebank in *Proceedings of the Treebanks and Linguistic Theories (TLT'02) Workshop*, Sozopol. Bulgaria, Sept.19th-20th, 2002 to appear (2002)

[Charniak, 1993] Eugene Charniak, 1996. *Statistical language learning.* Cambridge, Mass : MIT Press, 1993.

[Charniak, 1996] Eugene Charniak, 1993. *Tree-bank Grammars*. in *Proceedings of the Thirteenth National Conference on Artificial Intelligence,* AAAI Press/MIT Press, Menlo Park, pp.1031–1036

[Collins, 1999] M. Collins 1999. *Head-driven Statistical Models for Natural Language Parsing.* Ph.D. thesis, University of Pennsylvania, Philadelphia.

[Dalrymple, 2001] Mary Dalrymple 2001. *Lexical-Functional Grammar*. San Diego, Calif.; London : Academic Press

[Frank, 2000] A. Frank. 2000. Automatic F-Structure Annotation of Treebank Trees. In: (eds.) M. Butt and T. H. King, *The fi fthInternational Conference on Lexical-Functional Grammar*, The University of California at Berkeley, 19 July - 20 July 2000, CSLI Publications, Stanford, CA.

[Frank et al, 2002] A. Frank, L. Sadler, J. van Genabith and A. Way 2002. From Treebank Resources to LFG F-Structures. In: (ed.) Anne Abeille, *Treebanks: Building and Using Syntactically Annotated Corpora*, Kluwer Academic Publishers, Dordrecht/Boston/London, to appear (2002)

[Hockenmaier and Steedman, 2002] Hockenmaier, Julia and Mark Steedman, 2002. Generative Models for Statistical Parsing with Combinatory Categorial Grammar Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02) Philadelphia, PA.

[Johnson, 1999] Mark Johnson, 1999. PCFG models of linguistic tree representations Computational Linguistics

[Johnson, 1988] Mark Johnson, 1988. Attribute Value Logic and Theory of Grammar. CSLI Lecture Notes Series, Chicago University Press.

[Kaplan and Bresnan, 1982] R. Kaplan and J. Bresnan 1982. Lexical-functional grammar: a formal system for grammatical representation. In Bresnan, J., editor 1982, *The Mental Representation of Grammatical Relations.* MIT Press, Cambridge Mass. 173–281.

[Krotov et al, 1998] Krotov, A., M. Hepple, R.Gaizauskas and Y. Wilks 1998. Compacting the Penn Treebank Grammar, in *Proceedings of COLING/ACL'98* pp.699–703

[Lappin et al, 1989] S. Lappin, I. Golan and M. Rimon 1989. *Computing Grammatical Functions from Confi gurationalParse Trees* Technical Report 88.268, IBM Israel, Haifa, Israel

[Liakata and Pulman, 2002] M. Liakata and S. Pulman. 2002. *From trees to predicate-argument structures*. COLING'02, Proceedings of the Conference, Taipei, 24 August – 1 September, 2002

[Magerman, 1994] Magerman, D. 1994. *Natural Language Parsing as Statistical Pattern Recognition* PhD Thesis, Stanford University, CA.

[Marcus et al, 1994] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, M. Ferguson, K. Katz and B. Schasberger 1994. The Penn Treebank: Annotating

Predicate Argument Structure. In: *Proceedings of the ARPA Human Language Technology Workshop.*

[Pollard and Sag, 1994] Pollard, C. and I. Sag. 1994. Head-Driven Phrase Structure Grammar, Chicago: University of Chicago Press, and Stanford: CSLI Publications.

[Riezler et al, 2002] Stefan Riezler, 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 Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02) Philadelphia, PA.

[Sadler et al, 2000] L. Sadler, J. van Genabith and A. Way. 2000. Automatic F-Structure Annotation from the AP Treebank. In: (eds) M. Butt and T. H. King, *The fi fthInternational Conference on Lexical-Functional Grammar*, The University of California at Berkeley, 19 July - 20 July 2000, CSLI Publications, Stanford, CA.

[van Genabith et al, 1999] J. van Genabith, L. Sadler and A. Way. 1999. Data-Driven Compilation of LFG Semantic Forms EACL 99, Workshop on Linguistically Interpreted Corpora (LINC-99), Bergen, Norway, June 12th, 1999, pp.69-76

Partitive Noun Phrases in Hungarian

ERIKA Z. CHISARIK

Department of Linguistics
University of Manchester
Oxford Road, Manchester, M13 9PL
UK England

Erika.Z.Chisarik@stud.man.ac.uk

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

PARTITIVE NOUN PHRASES IN HUNGARIAN*

Abstract

Although a variety of nominal constructions have been examined in Hungarian linguistics, partitive noun phrases have barely received any attention. In this paper, it is argued that there are four basic types of partitive construction in Hungarian: (i) *genitive,* (ii) *dative*, (iii) *"közül"*, and (iv) *elative.* In distinction to *genitive partitive*s, *dative*, *"közül"* and *elative partitives* allow splitting of head and dependent. What is more, split partitive noun phrases show diagnostics of long-distance dependencies: they are subject to the *Adjunct Island Constraint* and to the constraint imposed by *non-bridge verbs*. It is argued that the syntactic behaviour of split partitives can be accounted for in purely functional terms.

1. Introduction: the data

There are four types of partitive construction in Hungarian: two types of *possessive partitives* and two types of *oblique partitives*. These subsequently divide into *genitive* and *dative partitives*, and *"közül"* and *elative partitives*, as illustrated in (1) – (4):[1]

Possessive Partitives

(1) [[a könyvek]$_{NP}$ [bármelyik-**e**]$_{N}$]$_{NP}$ *genitive partitive*
the book.3PL.**GEN** any-**3SG.POSS**
'any of the books'

(2) [[a cukor-**nak**]$_{NP}$ a [fel-**e**]$_{N}$]$_{NP}$ *dative partitive*
the sugar.3SG-**DAT** the half-**3SG.POSS**
'half of the sugar'

Oblique Partitives

(3) [[bármelyik]$_{N}$ [a könyvek **közül**]$_{PP}$]$_{NP}$ *"közül" partitive*
any the book.3PL **from among**
'any of the books'

(4) [[bármennyi]$_{N}$ [a cukor-**ból**]$_{NP}$]$_{NP}$ *elative partitive*
any the sugar.3SG-**ELAT**
'any of the sugar'

* Very special thanks go to John Payne, Tracy Holloway King, Mary Dalrymple, Wim van der Wurff and Tibor Laczkó for invaluable comments and inspiring discussions on various versions of this paper. I am also grateful to the audience of the LFG02 Conference for insightful remarks. I am thankful to the Graduate School in the Arts at the University of Manchester for providing me with financial support to attend the LFG02 Conference.

[1] The suffixes *–nak/nek* and *–ból/ből* attach to the stem in accordance with vowel harmony rules.

What is common in all four constructions is that they consist of two main parts: an N and an embedded NP/PP. Morphological case-marking associated with the arguments of given verbs occurs on the N rather than on the embedded NP/PP. This case-marking pattern argues in favour of treating N as the head of the partitive NP, and the embedded NP/PP as its dependent.

Possessive partitives are morphologically analogous to possessor constructions in Hungarian: hence the name "possessive" partitive. Compare examples (5) and (6) with (7) and (8):

Possessive noun phrases

(5) [a diák-ok]$_{NP}$ [könyv-e]$_{N}$ ⇒ könyv-ük
the student-3PL.GEN book-3SG.POSS book-3PL.POSS
'the students' book' 'their book'

(6) [a diák-ok-nak]$_{NP}$ a [könyv-e]$_{N}$ ⇒ könyv-ük
the student-3PL-DAT the book-3SG.POSS book-3PL.POSS
'the students' book' 'their book'

Partitive noun phrases

(7) [a diák-ok]$_{NP}$ [bármelyik-e]$_{N}$ ⇒ bármelyik-ük
the student-3PL.GEN any-3SG.POSS any-3PL.POSS
'any of the students' 'any of them'

(8) [a diák-ok-nak]$_{NP}$ a [tíz százalék-a]$_{N}$ ⇒ tíz százalék-uk
the student-3PL -DAT the ten percent-3SG.POSS ten percent-3PL.POSS
'ten percent of the students' 'any of them'

As is illustrated in (5) and (6), in possessor constructions the possessive relation is marked on the possessum N by the suffix *–(j)a/(j)e* in accordance with vowel harmony rules; the possessor NP is either genitive or dative marked.[2] The same morphological marking occurs in *possessive partitives*: in (7) and (8), the N heads bear the 3rd singular possessive inflection *–e/a*, and the dependent NPs the genitive and the dative case (*–nak/nek*) respectively (hence the names *genitive* and *dative partitives*).

In Hungarian, overt 3rd singular possessums (NP) can co-occur with plural possessors (N), but when the possessor is omitted, its person/number marking appears on the possessum: hence the form *könyv-ük* [book-3PL.POSS] 'their book' in (5) and (6). The same pattern can be observed in

[2] According to standard analyses, the overtly unmarked possessor is assumed to bear nominative case, since nominative is the overtly empty case-marker in Hungarian. The possessor can also be dative-marked. Consequently, these two possessor constructions are referred to as the nominative possessor construction and the dative possessor construction (e.g. Szabolcsi 1994, Laczkó 1995, É.Kiss 2000). Contrary to these standard analyses, in Payne and Chisarik (2001) we have argued that the case of the "nominative" possessor should be analysed as genitive. This new genitive case developed through the reanalysis of the definite article *az* 'the' as a case prefix, and hence a marker of genitive case rather than nominative. Therefore, the two types of possessor noun phrases are referred to as genitive and dative, rather than nominative and dative.

possessive partitives in (7) and (8) resulting in the forms *bármelyik-ük* [any-3PL.POSS] 'any of them' and *tiz százalék-uk* [ten percent-3PL.POSS] 'ten percent of them'.[3]

Since *possessive partitives* are morphologically analogous to possessor constructions in Hungarian, it is reasonable to assume the same syntactic analysis for these NPs. In Chisarik and Payne (2001) we argue that possessor constructions require the postulation of two unrestricted argument functions: NCOMP (nominal complement) for genitive possessors and SUBJ (subject) for dative ones. By allowing the *partitive* semantic relation to be included in the range of semantic relations that can be encompassed by these unrestricted functions, the analysis of Chisarik and Payne (2001) can be straightforwardly extended to *possessive partitives*. The dependent in *genitive partitive*s then associates with the grammatical function NCOMP, while in the *dative partitive* it associates with SUBJ, as illustrated in (9) and (10):

(9) [[a diák-ok] **NCOMP** [bármelyik-e]$_N$]$_{NP}$ *genitive partitive*
the student-3PL.GEN any-3SG.POSS
'any of the students'

(10) [[a diák-ok-nak]**SUBJ** a [tíz százalék-a]$_N$]$_{NP}$ *dative partitive*
the student-3PL -DAT the ten percent-3SG.POSS
'ten percent of the students'

There is a difference between genitive possessor NPs and *genitive partitives*: the former show the definiteness effect while the latter do not. Therefore, in genitive possessor constructions, the embedded NP which is in complementary distribution with the definite article, in the absence of any specific indication of the indefiniteness of the noun phrase, is assumed to function as a definite determiner. In distinction to this, since the (in)definiteness of *genitive partitives* is determined by the head N, the embedded NP is treated as a pre-head complement, rather than a determiner. Consequently, the grammatical function NCOMP is allowed to associate with two distinct structural positions in the NP, that of a determiner and a complement. In *dative possessor* and *dative partitive* constructions, the embedded NP is structurally a pre-determiner preceding the definite article.

Partitive relations can also be expressed by *oblique partitive noun phrases* which subdivide to the *"közül" partitive* (or 'from among/between' partitive) and to the *elative partitive* (or 'from/out of' partitive). The *"közül"* and *elative partitives* consist of an N head and a PP/NP post-head complement.[4] As illustrated by the examples in (3) and (4), in the former the NP complement is marked with the case-like postposition *közül* 'from among/between' (hence the name *"közül" partitive*), while in the latter it is marked with the elative case-marker *-ból/ből* (hence the name *elative partitive*).[5] In the *"közül" partitive* the complement is associated with the grammatical function $OBL_{közül}$ and in the *elative partitive* with OBL_{ELAT}, as illustrated in (11) and (12):

[3] I assume that such forms are nouns incorporating the pronominal inflection.
[4] Post-head complements also occur in derived nominals, e.g. *János megérkez-és-e* ***Budapestre*** [John arrive-NOM-3SG Budapest-SUBL] 'John's arrival to Budapest'.
[5] For the classification of postpositions in Hungarian see Payne and Chisarik (2000).

(11) [[bármelyik]$_N$ [a könyvek közül]$_{\textbf{OBL közül}}$] $_{NP}$ *"közül" partitive*
any the book.3PL from among
'any of the books'

(12) [[bármennyi]$_N$ [a cukor-ból] $_{\textbf{OBL ELAT}}$] $_{NP}$ *elative partitive*
any the sugar-ELAT any
'any of the sugar'

The data suggest that there is no uniform syntactic expression of partitive relations in Hungarian. On the one hand, partitive relations can be expressed by possessor constructions. Koptevskaja-Tamm (1998) shows that this is not unusual cross-linguistically. On the other hand, similarly to Turkish, where partitives are formed with an N head taking an ablative-marked dependent, (e.g. *sut-ten biraz* [milk-ABL a little] 'a little of the milk'), Hungarian partitives are expressed with the help of oblique case-markers such as the postpositional *közül* 'from among/between' and the elative *-ból/ből* 'from/out of'.[6] To sum up, in Hungarian partitives are parasitic on existing constructions and therefore on existing grammatical functions.

2. Split and non-split partitives

*Genitive partitive*s cannot be split, whereas *dative, "közül"* and *elative partitives* allow splitting of head and dependent. Compare (13) with (14), (15) and (16):

(13) ***[A regények]$_{NP}$** elolvasta Péter **[__ [egyik-é-t]$_N$]$_{NP}$.**
the novel.3PL.GEN read.3SG.PAST Peter one-3SG.POSS-ACC
'Of the novels, Peter read one.'

(14) **[A tej-nek]$_{NP}$** Anna megitta **[__ a [fel-é-t]$_N$]$_{NP}$.**
the milk.3SG-DAT Anna drank.3SG.PAST the half-3SG.POSS-ACC
'Of the milk, Anna drank half.'

(15) **[A regények közül]$_{PP}$** Péter elolvasott **[[négy-et]$_N$ __]$_{NP}$.**
the novel.3PL from among Peter read.3SG.PAST four-ACC
'Of the novels, Peter read four.'

(16) **[A cukor-ból]$_{NP}$** Anna tett a kávéjába **[[valamennyi-t]$_N$ __]$_{NP}$.**
the sugar3SG-ELAT Anna put.3SG.PAST the coffee.3SG.POSS some-ACC
'Of the sugar, Anna put in her coffee some. '

The example in (13) illustrates that intervening elements between head and complement in *genitive partitive*s lead to ungrammaticality. In distinction to this, the examples in (14), (15) and (16) show that in the *dative*, *"közül"* and *elative partitives* complement and head do not need to be adjacent: the NP/PP dependents are displaced to a sentence-initial topic position, while the N heads remain in a post-verbal position. These examples illustrate extraction through a short path.

[6] For the analysis of Turkish partitives refer to Kornfilt (1996).

Extraction of constituents of *dative*, *"közül"* and *elative partitives* through a long path is also grammatical. Compare the examples (17), (18) and (19) with those in (14), (15) and (16):

(17) **[A tej-nek]$_{NP}$** úgy emlékszem, hogy Anna
the milk.3SG-DAT so think.1SG.PRES that Anna

megitta **[__ a [fel-é-t]$_{N}$]$_{NP}$.**
drink.3SG.PAST the half-3SGPOSS-ACC
'Of the milk, as far as I remember, Anna drank half.'

(18) **[A regények közül]$_{PP}$** úgy emlékszem, hogy Péter
the novel.3PL from among so remember.1SG.PRES that Peter

elolvasott **[[négy-et]$_{N}$ __]$_{NP}$.**
read.3SG.PAST four-ACC
'Of the novels, as far as I remember, Peter read four.'

(19) **[A cukor-ból]$_{NP}$** úgy emlékszem, hogy Anna
the sugar3SG-ELAT so remember.1SG.PRES that Anna

tett a kávéjá-ba **[[valamennyi-t]$_{N}$ __]$_{NP}$.**
put the coffee.3SG.POSS-into some-ACC
'Of the sugar, as far as I remember, Anna put some in her coffee.'

In (17), (18), and (19) the partitive noun phrases are embedded in clausal complements, i.e. the sentential objects of the matrix verb. The PP/NP dependents are extracted to the sentence-initial topic position across the clausal complements. Therefore, these examples are instances of long-distance topicalization.

Hungarian split partitive noun phrases are reminiscent of German split NP topicalization (Kuhn 1998) and of discontinuous NP constituents in Walpiri (Bresnan 2001, Simpson 1991) and in Wambaya (Nordlinger 1998), but their distinctive property is that the split constituents do not agree either in case or number.

3. Constraints on long-distance dependencies of split partitives

Extraction is universally subject to constraints. Various constraints have been proposed on long-distance extraction, such as the *Subject Island Constraint*, the *Complex NP Constraint*, the *Adjunct Island Constraint*, the constraint imposed by *non-bridge verbs*, etc.[7] The *Subject Island Constraint* does not hold for Hungarian, but extraction can be blocked by complex noun phrases, as well as by sentential adjuncts and non-bridge verbs (É.Kiss 2002). As examples of constraints

[7] Cf. Chomsky (1986) and Ross (1967) for transformational analyses, and Kaplan and Zaenen (1989), Bresnan (2001), Dalrymple (2001), Falk (2001), and Kuhn (1998) for feature-based accounts.

on long-distance extraction from partitive noun phrases, sentential adjuncts and complement clauses of non-bridge verbs are examined in this section.

3.1 Sentential adjuncts

Cross-linguistically, there is no common agreement on how to group constraints on long-distance dependencies involving modifying adjuncts.[8] In Hungarian, tensed and non-tensed adjunct clauses block long-distance movement (Kenesei *et al*. 1998, Komlósy 1994, É.Kiss 2002). This constraint can be extended to split partitive noun phrases, as illustrated by the examples in (20), (21), and (22):

(20) ***[A barátai-nak]** $_{NP}$ Péter nevetett amikor
the friends.3SG.POSS-DAT Peter laugh.3SG.PAST when

[__ a [fel-é-t]$_N$] $_{NP}$ beválasztották a csapatba.
the half-3SG.POSS-ACC select.3PL.PAST the team
'Of his friends, Peter laughed when half were selected for the team.'

(21) ***[A barátai közül]** $_{PP}$ Péter nevetett amikor
the friends.3SG.POSS from among Peter laughed when

[[kettő-t]$_N$ __] $_{NP}$ beválasztottak a csapat-ba.
two-ACC select.3PL.PAST the team-into
'Of his friends, Peter laughed when two were selected for the team.'

(22) ***[A cukor-ból]**$_{NP}$ Péter nevetni szokott, amikor Emese
the sugar-ELAT Peter laugh-INF used to when Emese

[[két kanál-lal]$_{NP}$ __] $_{NP}$ tesz a kávé-já-ba.
two spoon-INS put.3SG.PRES the coffee-3SG.POSS-into
'Of the sugar, Peter used to laugh when Emese puts two spoons into her coffee.'

In (20) and (21), the *dative* and the *"közül" partitives* are embedded in a tensed sentential adjunct, while in (22) the *elative partitive* is embedded in a non-tensed adjunct clause. Since the constituents of partitive noun phrases can be freely topicalized across complement clauses, the ungrammaticality of these examples is due to the modifying adjunct clauses.

Although sentential adjuncts block topicalization, phrasal adjuncts do not disallow it. Consider the split *"közül" partitive* illustrated in (23):

(23) **[A két leghíresebb egyetem közül]$_{PP}$** Péter tanított
the two most famous university from among Peter teach.3SG.PAST

[8] Refer to Williams (1992), Chinque (1990), Hornstein and Weinberg (1995), and Dalrymple (2001) for different analyses of long-distance dependencies involving modifying adjuncts.

biológiát **[[az egyik-en]$_{NP}$ __]$_{NP}$** .
biology the one-SUP
'Of the two most famous universities, Peter taught biology at one.'

In (23) the *"közül" partitive* functions as an adjunct phrase. Topicalizing the PP complement from within this partitive NP does not lead to ungrammaticality. Therefore, the general notion can be maintained that finite or non-finite adjunct clauses constrain long-distance movement in Hungarian, whereas adjunct phrases do not.

3.2 Non-bridge verbs

Although long-distance extraction is expected to be perfectly acceptable from sentential complements, it is not always the case that such extractions are grammatical. Similarly to English, in Hungarian it is possible to extract various constituents from a clausal complement if it is governed by so-called *bridge verbs* (verbs allowing extraction), but not from clausal complements governed by *non-bridge verbs* (verbs disallowing extraction).[9] This constraint also holds true for partitive noun phrases. Consider (24), (25), and (26):

(24) **[A vendégek-nek]**$_{NP}$ azt hiszem hogy Dávid
the guest.3PL-DAT that believe.1SG.PRES that David

ismeri **[__ a [fel-é-t]$_{N}$]**$_{NP}$
know.3SG.PRES the half-3SG.POSS-ACC
'Of the guests, I believe that David knows half.'

(25) **[A vendégek közül]**$_{PP}$ azt hiszem, hogy Dávid
the guest.3PL from among that believe.1SG.PRES that David

ismer **[[kettő-t]$_{N}$ __]**$_{NP}$
know.3SG.PRES two-ACC
'Of the guests, I believe that David knows two.'

(26) **[A tortá-ból]** $_{NP}$ azt hiszem hogy Dávid
the cake.3SG-ELAT that believe.1PL.PRES that David

elfogyasztott **[[két szelet-et]$_{NP}$ __]** $_{NP.}$
eat up.3SG.PAST two pieces-ACC
'Of the cake, I believe that David ate up two pieces."

The examples in (24), (25), and (26) illustrate that the verb *hisz* 'believe' allows long-distance extraction of the dependents of partitive NPs to a topic position. The verb *súg* 'whisper', however, blocks such extraction, as shown in (27), (28), and (29):

[9] For a discussion of extraction across bridge/non-bridge verbs refer to É.Kiss (2002).

(27) *[A **vendégek-nek**]$_{NP}$ azt súgták, hogy Dávid
the guest.3PL-DAT that whisper.3PL.PAST that David

ismeri **[__ a [fel-é-t]$_{N}$]$_{NP}$**
know.3SG.PRES the half-3SG.POSS-ACC
'Of the guests, they whispered that David knows half.'

(28) *[A **vendégek közül]$_{PP}$** azt súgták, hogy Dávid
the guest.3PL from among that whisper.3PL.PAST that David

ismer **[[kettő-t]$_{N}$ __]$_{NP}$**
know.3SG.PRES two-ACC
'Of the guests, they whispered that David knows two.'

(29) ***[A tortá-ból]** $_{NP}$ azt súgták, hogy Dávid **[[két szelet-et]$_{NP}$ __]** $_{NP}$
the cake.3SG-ELAT that whisper.3PL.PAST that David two pieces-ACC

fogyasztott el.
eat.3SG.PAST up
'Of the cake, they whispered that David ate up two pieces."

Bridge verbs in Hungarian are similar to those found in other languages. The group of bridge-verbs includes (i) modal predicates taking a subject or object clause *akar* 'want', *szeretne* 'would like', *kell* 'need', *szabad* 'may', *lehet* 'is possible', *nyilvánvaló* 'is obvious', *valószínű* 'is likely', etc.; (ii) verbs of saying and verbs denoting mental activities, such as *mond* 'say', *ígér* 'promise', *állít* 'claim', *gondol* 'think', *hisz* 'believe', etc. (É.Kiss 2002: 253).

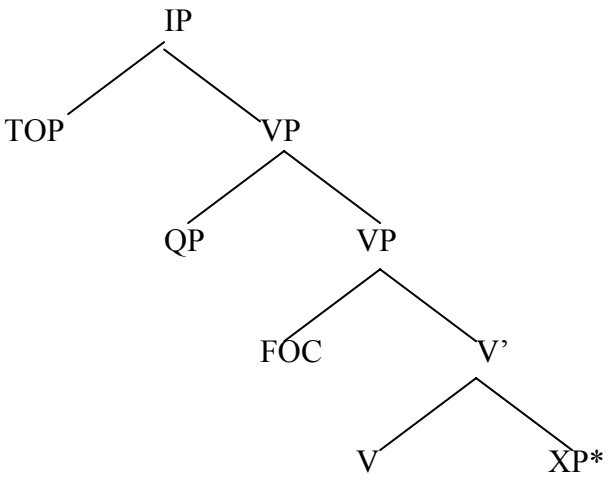
The Hungarian data has illustrated that long-distance dependencies of split partitive noun phrases are subject to at least two basic constraints, the *Sentential Adjunct Constraint* and the constraint imposed by *non-bridge verbs*: tensed and non-tensed sentential adjuncts and clausal complements which are not governed by bridge verbs clearly block extraction of constituents of partitive noun phrases.

4. The structure of the Hungarian pre-verbal periphery

Long-distance dependencies involve displacing a constituent from within a governable or non-governable grammatical function to a sentence-initial position associated with a particular discourse function. In Hungarian, different discourse functions are associated with different pre-verbal phrase structure positions. Constituents of split partitive noun phrases have to obey these syntactic restrictions.

Unlike post-verbal constituent order, the order of pre-verbal constituents is fixed. The immediate verb-adjacent position is reserved for FOCUSED constituents only. [10] When other constituents occur pre-verbally, they are placed to the left of this verb-adjacent position. TOPICS occur initially and they are considered to be outside the predicate, since they cannot receive stronger stress than the first element of the predicate.[11] A set of items receives stronger stress than the predicate, and therefore, must be considered internal to the predicate. In transformational theory such items are placed in QP, which occurs between a postulated topic and focus phrase.[12] The tree in (30) illustrates the basic structure of the Hungarian clause: IP is used as the category label for the sentence and VP for the predicate; the verb V, is followed by one or more arguments or modifying adjuncts, and preceded by a number of hierarchically ordered discourse-marked constituents:

(30) GB analysis (É.Kiss 1994)



In LFG the c-structure of the Hungarian sentence can avoid any confusion between category and function: it can contain solely phrasal categories, such as NP, PP, etc. The grammatical function of these phrasal constituents is expressed in the f-structure. The discourse functions TOPIC and FOCUS are standard LFG discourse functions. QP is, however, a category rather than a function.

---

[10] Focus can be either lexical (inherent) or structural in Hungarian. The focus position can be filled by the following inherent focus items:

(i) interrogative phrases (*ki* 'who', *mi* 'what');

(ii) positive and negative focus phrases (e.g. inherent focus phrases such as *kevés* 'few' or *sok* 'much', phrases modified by the adverb *csak* 'only' or *nemcsak* 'not only');

(iii) negative phrases (for instance, inherently negative quantifiers such as *kevés* 'few, little', inherenty negative adverbs such as *ritkán* 'seldom' and *rosszul* 'badly', negated universal quantifiers *nem mindenki* 'not everybody', or negative concord items *senki* 'no one/nobody' and *semmi* 'nothing/anything'.

[11] Any arguments of the verb can be topicalized, as well as modifiers, adverbials, and predicates.

[12] The predicate-initial position between topic and focus is obligatorily reserved for the following items:

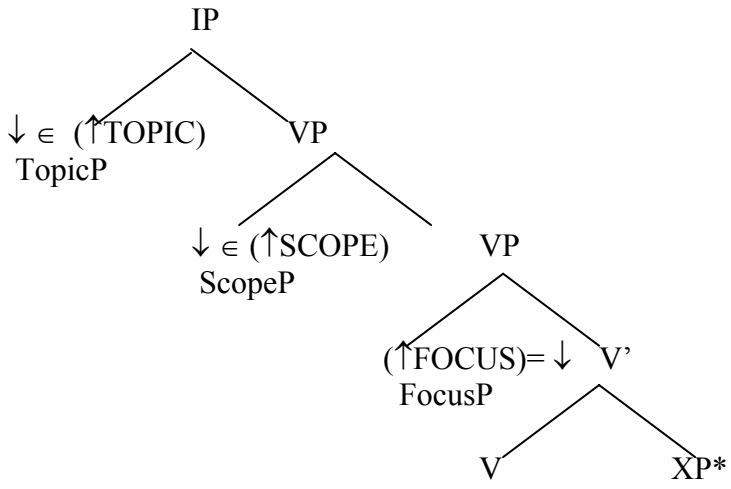
(i) positive universal quantifiers *mindenki* 'everyone', *minden* 'everything', *mindig* 'always', phrases containing determiners *minden* 'every', *mindegyik* 'each', *mindkét* 'both', *összes* 'all', *legtöbb* 'most'

(ii) phrases containing additive particles such as *is* 'also', *még .... is* 'even/also' or certain adverbials of frequency, degree and manner, such as *állandóan* 'constantly', *rettentően* 'terribly', *gyorsan* 'quickly'.

(iii)

Therefore, a new function, which I name SCOPE, is needed in Hungarian for constituents that occur between TOPIC and FOCUS. The LFG representation of the pre-verbal periphery is given in (31):

(31) LFG analysis



In the tree in (31) the constituent structure meta-category abbreviations such as TopicP, ScopeP and FocusP are used for the general description of the phrase structure categories (e.g. NP, PP, etc.) of the sentence-initial discourse functions.[13] The functional descriptions ↓∈(↑TOPIC), ↓∈(↑SCOPE), and (↑FOCUS)= ↓ capture the discourse functions of the given categories.

Items that occur in SCOPE are typically universal quantifiers and a small group of adverbs (cf. footnote 13). These items can occur post-verbally or pre-verbally. If two of these items co-occur post-verbally, or one of them co-occurs with another scopal element, their order does not effect their scopal relations: regardless of their precedence, they have scope over each other. If one of the universal quantifiers and adverbs or another scopal element occurs in pre-verbal position, a post-verbal universal quantifier or adverb has to move to the pre-verbal position preceding that other scopal element, if it has scope over it. In other words, universal quantifiers and adverbs are placed to the verb-initial position for scope reasons. Therefore, the label SCOPE is adopted for the function that associates with the syntactic position hosting them. It is worth noticing that since *scope* is a semantic notion, the function SCOPE cannot be considered strictly as a discourse function; rather, it is a semantic function.

5. Long-distance dependencies of split partitives in LFG

Kaplan and Zaenen (1989) and Bresnan (2001) have argued that long-distance dependencies obey functional rather than phrase structure constraints. Both analyses show that long-distance dependencies can be straightforwardly accounted for in LFG by employing functional

[13] For the role of meta-categories in syntactic description see Dalrymple (2001), Chapter 5.

uncertainty. Adopting this idea, I follow Kaplan and Zaenen (1989) in using outside-in functional uncertainty to model long-distance relations between the split constituents of partitive noun phrases in Hungarian.

5.1 Phrase structure rules

In LFG, constituent structure categories occur in the c-structure instead of functional categories. The permitted constituent structure categories for topic, scope and focus phrases are given in (32):

(32) TopicP ≡ { NP, PP, AP, ADVP, CP, VP}
ScopeP ≡ { NP, ADVP}
FocusP ≡ { NP, PP, AP, ADVP, CP, VP}

As was mentioned in section 4, sentence-initial discourse-marked constituents are ordered in Hungarian. Therefore, the following phrase structure rules are needed to describe the hierarchical left-periphery of the Hungarian sentence:[14]

(33) ***Phrase structure rules***

(a) IP → (TopicP* VP
↓∈ (↑TOPIC) ↑=↓
(↑TOPIC) = (↑BODY BOTTOM))

(b) VP → (ScopeP*) (FocusP V'
↓∈ (↑SCOPE) (↑FOCUS) =↓ ↑=↓
(↑SCOPE) = (↑BODY BOTTOM)) (↑FOCUS) = (↑BODY BOTTOM))

Besides capturing the order of the discourse-marked phrases, the rules in (33a) and (33b) state that topic, scope and focus phrases (TopicP, ScopeP, and FocusP) are linked to the appropriate discourse functions, namely TOPIC, SCOPE, and FOCUS by the functional equations ↓∈(↑TOPIC), ↓∈(↑SCOPE), and (↑FOCUS) =↓. That the discourse functions are identical to some grammatical function is indicated by the outside-in functional uncertainty equations (↑TOPIC)=(↑BODY BOTTOM), (↑SCOPE)=(↑BODY BOTTOM), (↑FOCUS)=(↑BODY BOTTOM) (Kaplan and Zaenen 1989).

[14] The *Kleene star operator* indicates that more than one TOPIC or SCOPE phrase can occur sentence-initially. The parenthesis mark the optional occurrence of the sentence-initial discourse-marked phrases.

5.2 Grammatical functions of the discourse-marked phrase

In long-distance dependencies, sentence-initial discourse functions have to be bound to certain within-clause positions. This is ensured by the *Extended Coherence Condition* (Zaenen 1980). For Hungarian the following modified version of this condition is relevant:

(34) ***Extended Coherence Condition:***
TOPIC, SCOPE and FOCUS must be linked to the semantic argument structure of the sentence in which they occur, either by functionally or by anaphorically binding an argument.

In the case of long-distance dependencies of split partitive noun phrases, TOPIC, SCOPE or FOCUS are related to their within-clause grammatical function functionally, rather than anaphorically.

Kaplan and Zaenen (1989) argue that the grammatical functions of the within-clause phrase are constrained: some grammatical functions can be related to discourse functions, whereas others cannot. This idea is also applicable to Hungarian: for instance, since arguments and modifiers can be freely topicalised, among others, the grammatical functions SUBJ, OBJ, COMP and ADJ can be related to the discourse function TOPIC. Also, *dative, "közül"* and *elative partitives* allow long-distance extraction of their complements to a sentence-initial topic position (refer back to examples (17), (18) and (19)). In the *dative partitive*, the complement is associated with the grammatical function SUBJ, whereas in the *"közül"* and *elative partitives* the complements are linked to the grammatical functions $OBL_{közül}$, and $OBL_{ELAT,}$ which are variants of OBL_{θ}. Thus, OBL_{θ} can associate with TOPIC as well. *Genitive partitive*s, which cannot be split, disallow any kind of extraction of their complement, NCOMP: in such NPs head and complement must be adjacent to each other (cf. example (13)). By disallowing the grammatical function NCOMP to associate with the grammatical function TOPIC, the ungrammaticality of (13) can be straightforwardly accounted for. That is, the inseparability of *genitive partitive*s then follows from a constraint on the within-clause grammatical functions (i.e. on the BOTTOM of the dependency path), formulated in (35):

(35) $(\uparrow \text{TOPIC}) = (\uparrow \text{GF-NCOMP})$

In (35) the annotation ($\uparrow$GF-NCOMP) states that any grammatical function except NCOMP can be associated with TOPIC. GF stands for all possible grammatical functions, and formally can be represented as a disjunction of such categories: GF $\equiv$ {SUBJ | OBJ | OBL_{θ} | COMP | NCOMP | ADJ }. Since *genitive partitive*s and genitive possessor noun phrases are syntactically identical, they behave in the same way with regard to extraction. The constraint in (35) simultaneously

accounts for genitive possessor noun phrases which are also inseparable.[15]

5.3 Constraining the grammatical functions on the path

In section 3, it has been demonstrated that long-distance dependencies involving a position inside a tensed or a non-tensed sentential adjunct are ruled out in Hungarian (refer back to long-distance topicalization of complements of partitive NPs in (20), (21), and (22)). In LFG the island constraints are accounted for by constraining the grammatical functions permitted on the path (i.e. constraining the BODY of the dependency). For sentential adjunct clauses, the constraint can be stated in the following way: the path to the within-clause function of the discourse-marked constituents may not include the grammatical function ADJ. However, phrasal adjuncts need to be excluded from this constraint (cf. example in (23)). The grammatical function of both sentential and phrasal adjuncts is the same; therefore, in the f-structure the same attribute is used, namely ADJ. What is constrained then is a particular value that clausal adjuncts have, but phrasal adjuncts lack. The formal description involves a general constraint supplemented by an off-path constraint, which can be illustrated as follows:

(36) $(\uparrow \text{TOPIC}) = (\uparrow \text{ADJ} \quad \text{GF-NCOMP})$
$\neg(\rightarrow \text{TENSE})$

The inside-out functional uncertainty expression (↑ TOPIC) = (↑ ADJ GF-NCOMP) states that TOPIC can be connected to any grammatical function besides NCOMP within an adjunct. The off-path constraint ¬(→TENSE) ensures that the grammatical function ADJ does not contain the attribute TENSE, since tense is a property of clauses rather than phrases. This off-path constraint is unavoidable, since the statement that any kinds of adjuncts allow long-distance dependencies is

[15] Similarly to genitive partitives, genitive possessors cannot be split; dative partitives and dative possessor NPs allow splitting of head and complement. Genitive and dative possessor NPs are illustrated in (i) and (ii) respectively:

Genitive Possessor NP

(i) ***[A kisfiú]$_{\text{NCOMP}}$** valószínű,
the little boy.3SG.GEN is likely

hogy megtalálták **[__ [bicikli-jé -t]$_{\text{N}}$]$_{\text{NP}}$**
that found.3PL.PAST bike-3SG.POSS-ACC
'The boy's, it is likely that they found bike.'

Dative Possessor NP

(ii) **[A kisfiú-nak]$_{\text{SUBJ}}$** valószínű,
the little boy.3SG-DAT is likely

hogy megtalálták **[__ a [bicikli-jé -t]$_{\text{N}}$]$_{\text{NP}}$**
that found.3PL.PAST the bike-3SG.POSS-ACC
'The boy's, it is likely that they found the bike.'

false. Although clausal and phrasal adjuncts differ c-structurally, it is more lucrative to formulate the adjunct island constraint f-structurally, because phrasal adjuncts can have a large variety of syntactic forms, all of which would need to be included as exceptions to the constraint. The vital characteristic of any phrase is the lack of tense (unless the phrase is verbal). Excluding the attribute TENSE from the grammatical function ADJ on the path can straightforwardly and economically capture the adjunct effects on long-distance dependencies in Hungarian.

5.4 Off-path constraints

Besides sentential adjuncts, non-bridge verbs also block long-distance dependencies in Hungarian. Recall (27), (28), (29) illustrating topicalization of NP/PP complements across the non-bridge verb *súg* 'whisper'. The distinction between bridge verbs and non-bridge verbs does not affect the within-clause grammatical function of the displaced constituent. In other words, the path to the inside of the clause remains invariable. As pointed out by Dalrymple (2001), there is no reason to assume that the grammatical function of the complements of these verbs differ, if syntactically they are the same. To account for the behaviour of non-bridge verbs therefore, following Dalrymple (2001), I assume an f-structure attribute LDD with the value –, which is lexically specified by a non-bridge verb as appearing in its sentential complement (COMP). F-structures containing such attributes cannot participate in long distance dependencies. To put it differently, a sentential complement (COMP) of a non-bridge verb has the feature LDD the value of which is minus. Bridge verbs lack this feature. The path in a long-distance dependency may not pass through an f-structure containing this feature. In formal terms, this requirement is stated as an off-path constraint, as illustrated in (37):

(37) $(\uparrow \text{TOPIC}) = (\uparrow \text{COMP} \quad \text{GF-NCOMP})$
$(\rightarrow \text{LDD} \neq -)$

The expression in (37) captures the notion that the TOPIC is connected to a grammatical function other than NCOMP which is embedded in a sentential complement. The off-path constraint $(\rightarrow \text{LDD} \neq -)$ ensures that the f-structure of COMP does not contain the attribute LDD with the value minus.

5.5 Hungarian TOPIC/SCOPE/FOCUS path

Besides TOPIC, heads/complements of split partitives as well as other sentence-initial constituents can be related to SCOPE and FOCUS. The path between SCOPE/FOCUS and their within-clause functions is subject to the same constraints as the topic path for the following reasons: (i) scope and focus phrases also involve movement to a pre-verbal discourse-marked position, (ii) the displaced constituent passes through the same path as topicalization. Therefore, taking into consideration the set of functional constraints outlined in sections 5.2, 5.3, and 5.4, the general TOPIC/SCOPE/FOCUS path for Hungarian can be formally characterised in the following way:

(38) ***Hungarian TOPIC/SCOPE/FOCUS path***

{	COMP	OBJ	SUBJ	}*	{ (ADJ ∈)	}	GF-NCOMP
	(→LDD ≠ -)				¬(→TENSE)		

This description allows the within-clause grammatical function other than NCOMP to be arbitrarily deeply embedded inside an infinite number of properly constrained COMP, OBJ or SUBJ functions, and optionally to appear as a member of a set of phrasal ADJ set of such a function. The *Kleene star operator* allows any number of COMP, OBJ or SUBJ attributes on the path. The off-path constraint (→LDD ≠ -) ensures that the within-clause function of the TOPIC does not involve a non-bridge verb, while the off-path constraint ¬(→TENSE) ensures that the ADJ does not contain any tense (that it is phrasal, rather than clausal).

5.6 An Example

Let us apply the general rules to a specific long-distance dependency between the constituents of a partitive noun phrase. Consider the following topicalization of the PP complement from the *"közül" partitive*:

(39)

[A	**betörő-k**	**közül]**$_{PP}$	mondták,	hogy
the	burglar-3PL	from among	say.3PL.PAST	that

Péter felismert **[[kettő-t]**$_{N}$ __ **]**$_{NP}$.
Peter recognize.3SG.PAST two-ACC
'Of the burglars, they said that Peter recognized two.'

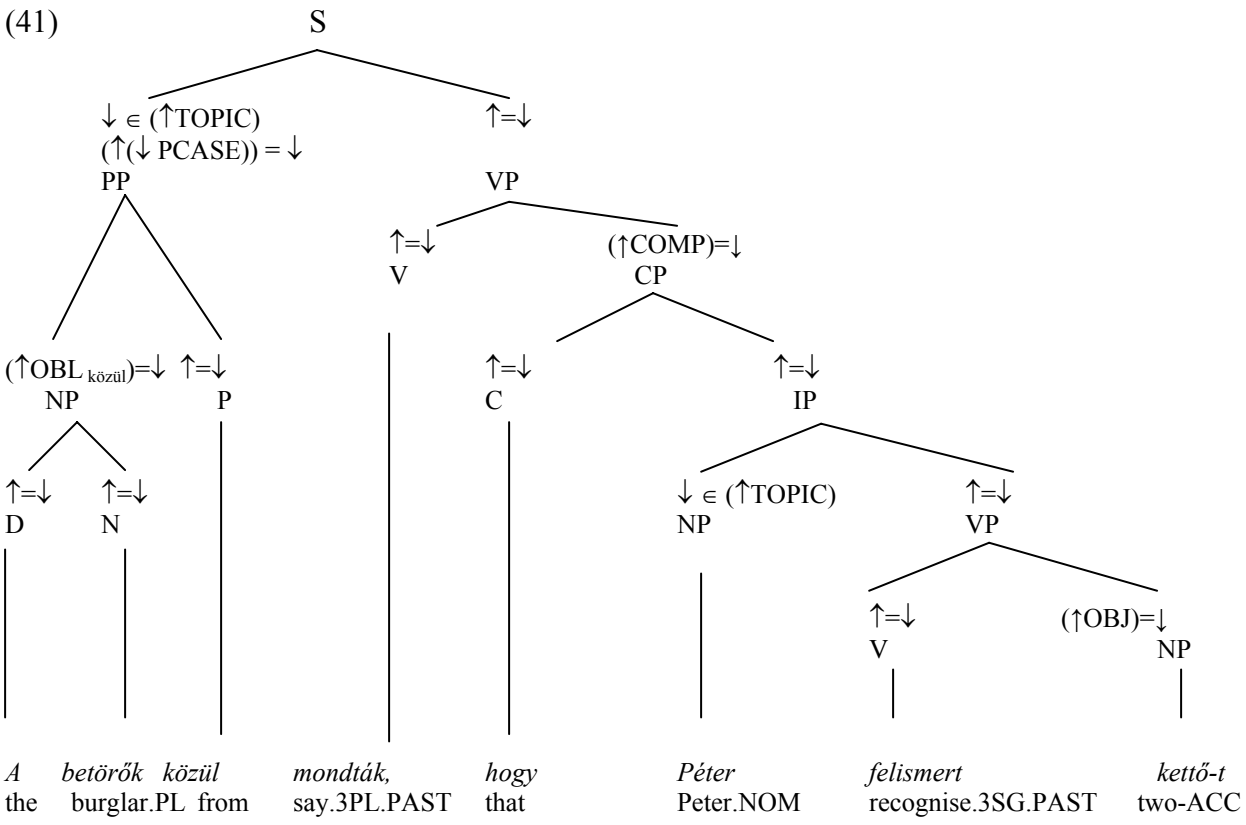
Taking into consideration the phrase structure rules, the constraint on syntactic categories, and the functional constraints on the path, we can formally account for the example in (39) as follows:

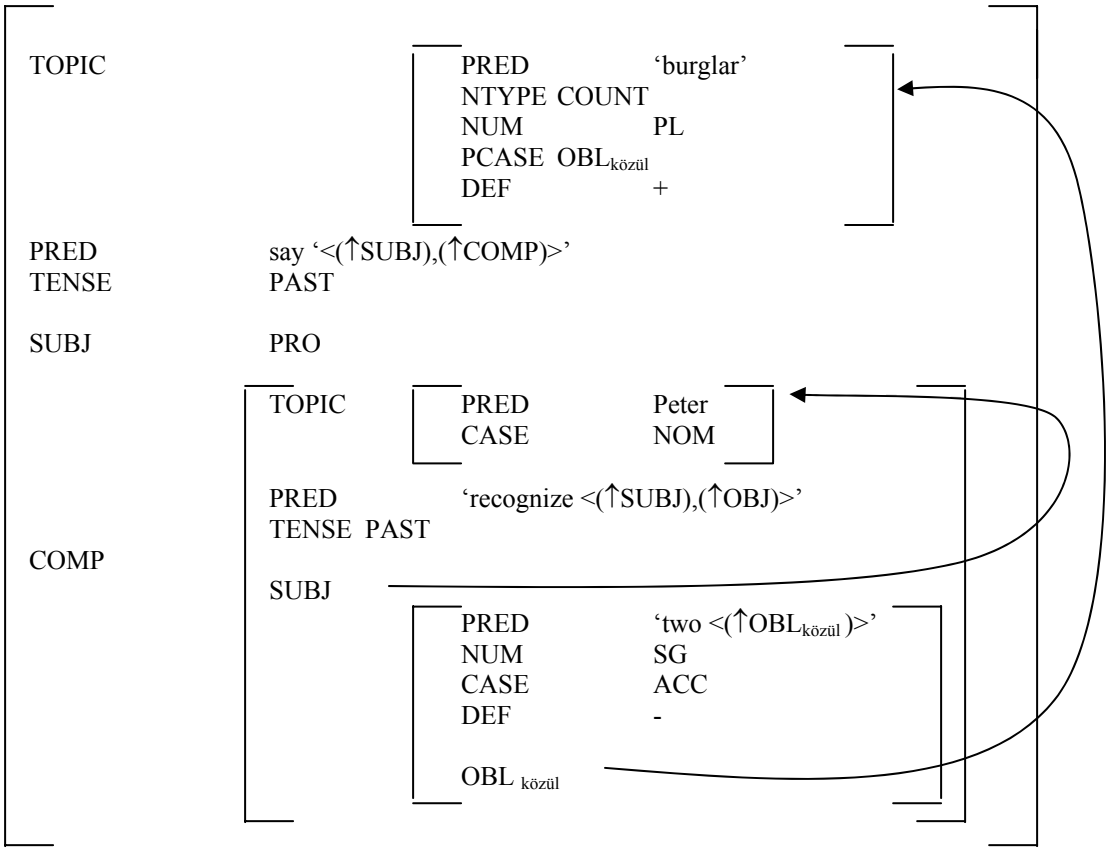
(40)

IP →	PP		VP
	↓ ∈ (↑TOPIC)		↑=↓
	(↑TOPIC) = (↑ COMP	OBL $_{közül}$)	
	(→LDD ≠ -)		

The c-structure and f-structure of the split partitive of (39) is given in (41):

(41)


S
↓ ∈ (↑TOPIC)
(↑(↓ PCASE)) = ↓
PP
↑=↓
VP
↑=↓
V
(↑COMP)=↓
CP
(↑OBL közül)=↓
NP
↑=↓
P
↑=↓
C
↑=↓
IP
↑=↓
D
↑=↓
N
↓ ∈ (↑TOPIC)
NP
↑=↓
VP
↑=↓
V
(↑OBJ)=↓
NP
A betörők közül
the burglar.PL from
mondták,
say.3PL.PAST
hogy
that
Péter
Peter.NOM
felismert
recognise.3SG.PAST
kettő-t
two-ACC


TOPIC
PRED ‘burglar’
NTYPE COUNT
NUM PL
PCASE OBL közül
DEF +
PRED say ‘<(↑SUBJ),(↑COMP)>’
TENSE PAST
SUBJ PRO
COMP
TOPIC
PRED Peter
CASE NOM
PRED ‘recognize <(↑SUBJ),(↑OBJ)>’
TENSE PAST
SUBJ
PRED ‘two <(↑OBL közül)>’
NUM SG
CASE ACC
DEF -
OBL közül

The rule in (41) shows that the topicalized phrase has the syntactic category PP and it is the sister of VP. The phrase structure rule correctly illustrates that the topicalized PP is outside the predicate. The annotation ↓ ∈ (↑TOPIC) captures the claim that PP belongs to a set of possible topicalizable phrases. The functional uncertainty equation (↑TOPIC) = (↑ COMP $\text{OBL}_{\text{közül}}$) represents the topic path, namely that the TOPIC is associated with the grammatical function $\text{OBL}_{\text{közül}}$ which is embedded into COMP. In other words, the $\text{OBL}_{\text{közül}}$ occurs deep within the embedded subordinate clause that functions as the complement of the matrix verb. The off-path constraint under COMP ensures that COMP does not contain the attribute-value pair <LDD ->. The f-structure correctly reflects the grammatical functions: it is complete and coherent and does not violate any constraint. The f-structure of the TOPIC is associated with the grammatical function $\text{OBL}_{\text{közül}}$, which is embedded in the sentential complement of the matrix verb. The f-structure of COMP does not contain the attribute-value pair <LDD ->. Therefore, the long-distance dependency between the split $\text{OBL}_{\text{közül}}$ complement and its head in this case is grammatical.[16]

6. Concluding remarks

In this paper, it has been shown that there are four basic types of partitive construction in Hungarian: *genitive*, *dative*, *"közül"*, and *elative.* In distinction to *genitive partitive*s, *dative*, *"közül"* and *elative partitives* allow splitting of head and dependent. Split partitive noun phrases show diagnostics of long-distance dependencies: they are subject to the *Adjunct Island Constraint* and to the constraint imposed by *non-bridge verbs*. I have proposed an analysis of split partitive noun phrases in purely functional terms. By excluding NCOMP from the within-clause grammatical functions that can be associated with the various discourse functions, the inability of *genitive partitive*s to split is accounted for straightforwardly. What is more, the explanation for the inseparability of genitive possessors from their head also follows from this constraint. The *Adjunct Island Constraint* has been captured by formalizing a functional constraint that excludes the attribute TENSE from the f-structure of adjuncts. This ensures that extraction is only allowed from adjunct phrases. The behaviour of non-bridge verbs was accounted for by a functional constraint ensuring that the f-structure of their complements does not contain the attribute-value pair <LDD = ->. Finally, I have defined a general long-distance path for Hungarian. Further research is required to look into the behaviour of other long-distance dependencies in Hungarian, for instance WH-question and relative clauses, in order to reveal whether a unified functional account of them can be achieved.

---

[16] In (41) another instance of a topicalized constituent also occurs, namely that of the subject of the embedded clause. The specially modified version of the general topicalization rule is given in (i):

(i)	IP →	NP	VP
		↓ ∈ (↑TOPIC)	↑=↓
		(↑TOPIC) = (↑SUBJ)	

The NP occurs as the topic of the embedded clause and it is outside the VP predicate. The topicalized NP constituent is related to the grammatical function SUBJ: the external argument of the verb.

References

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

Chisarik, Erika and John Payne (2001). Modelling Possessor Constructions in LFG: English and Hungarian. In Butt, M. and T. Holloway King (eds.) Online Proceedings of the LFG01 Conference. CSLI Publications: http://csli-publications.stanford.edu.lfg01.html.

Chomsky, Noam. (1986). *Barriers.* MIT Press: Cambridge, MA.

Cinque, G. (1990). *Types of A'-Dependencies.* Linguitics Inquiry Monographs. MIT Press: Cambridge, MA.

Dalrymple, Mary. (2001). *Syntax and Semantics 34: Lexical-Functional Grammar*. San Diego: Academic Press.

É.Kiss, Katalin. (1994). Sentence Structure and Word Order. In Kiefer, Ferenc and Katalin É.Kiss, *Syntax and Semantics: The Syntactic Structure of Hungarian 27*, Academic Press: San Diego, pp. 1-91.

É.Kiss, Katalin. (2000). The Hungarian noun phrase is like the English noun phrase, In Alberti, Gábor and István Kensei (eds), *Approaches to Hungarian* 7, Szeged: JATE, pp.119-150.

É.Kiss, Katalin. (2002). *The Syntax of Hungarian*, Cambridge University Press.

Falk, Jehuda. (2001). *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax.* CSLI Publications: Stanford University.

Hornstein, N. and A. Weinberg. (1995). The empty category principle. In *Government and Binding Theory and the Minimalist Program.* Blackwell Publishers: Oxford.

Kaplan, Ronald and Annie Zaenen (1989). Long-distance dependencies, constituent structure, and functional uncertainty. In Baltin, Mark and Anthony Kroch (eds). *Alternative Conceptions of Phrase Structure.* Chicago: Chicago University Press, pp.17-42.

Kenesei, Istvan, Vago, Robert, and Anna Fenyvesi. (1998). *Hungarian: Descriptive Grammars*. Routledge.

Komlósy, András. (1994). Complements and Adjuncts. In Kiefer, Ferenc and Katalin É.Kiss, *Syntax and Semantics: The Syntactic Structure of Hungarian.* Volume 27, Academic Press: San Diego, pp. 91-178.

Koptevskaja-Tamm, Maria. (1998). Genitives and Possessive NPs in the Languages of Europe. In F. Plank (ed), *Noun Phrase Structure in the Languages of Europe,* Mouton de Gruyter: Berlin.

Kornfilt, Jaklin. (1996). Naked Partitive Phrases in Turkish, In Hoeksema, Jakob. (ed). *Partitives: Studies on the Syntax and Semantics of Partitive and Related Constructions.* Berlin: Mouton de Gruyter, pp. 107-142.

Kuhn, Jonas. (1998). Resource sensitivity in the Syntax-Semantics Interface and the German Split NP Construction. In Kiss T. and D. Meurers (eds), *Proceedings of the ESSLLI X Workshop "Current topics in constraint based theories of Germanic syntax".* Saarbrücken.

Laczkó, Tibor. (1995). *The Syntax of Hungarian Noun Phrases.* Meta Linguistica 2, Frankfurt am Main: Peter Lang

Nordlinger, Rachel. (1998). *Constructive Case: Evidence from Australian Languages*. Stanford, California: CSLI Publications.

Payne, John and Erika Chisarik. (2000). Demonstrative constructions in Hungarian. In Alberti, Gábor and István Kenesei (eds), *Approaches to Hungarian* 7, Szeged: JATE, pp.179-198.

Payne, John and Erika Chisarik. (2001). The so-called "Nominative" Possessor Construction: A New Genitive? Paper presented at the 5th International Conference on the Structure of Hungarian, Budapest, 24-26 May 2001.

Ross, J. R. (1967). *Constraints on variables in Syntax.* Doctoral Dissertation, MIT.

Simpson, Jane. (1991). *Walpiri Morpho-Syntax: a Lexicalist Approach.* Dordrecht: Kluwer Academic Press.

Szabolcsi, Anna. (1994). The Noun Phrase. In Kiefer, Ferenc and Katalin É.Kiss, *Syntax and Semantics: The Syntactic Structure of Hungarian.* Volume 27, Academic Press: San Diego, pp. 179-274.

Williams, E. (1992). Adjunct control. In *Control and Grammar*, Larson *et al.* (eds), Kluwer Academic Press: Dordrecht, pp.297-322.

Zaenen, Annie. (1980). *Extraction Rules in Icelandic.* Doctoral dissertation, Harvard University. Reprinted by Garland Press, New York, 1985.

An LFG-type Grammar for German Based on the Topological Model

Lionel Clément Kim Gerdes Sylvain Kahane
Lionel.Clément@inria.fr kim@linguist.jussieu.fr sk@ccr.jussieu.fr
INRIA Rocquencourt Lattice
Domaine de Voluceau – BP 105 Université Paris 7 - case 7003
78153 Le Chesnay – France 75251 Paris Cedex 05 – France

Proceedings of the LFG02 Conference
National Technical University of Athens, Athens
Miriam Butt and Tracy Holloway King (Editors)
2002
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

This paper proposes a description of German word order in a LFG-type grammar. Contrary to earlier Lexical Functional Grammar description of German, our grammar uses as c-structure the topological model. This gives us a simpler grammar, which covers (partial) VP fronting, intraposition, extraposition, auxiliary flip and all other order possibilities for a verbal dependent. Our grammar is implemented in an LFG parser.

Introduction

The aim of this article is to propose a new description of German word order in the LFG formalism. Contrary to previous attempts, we explicitly use the topological model of German as the basis of our analysis.

The topological model subdivides the sentence into a hierarchy of topological domains that are themselves composed of fields (*Vorfeld*, *Mittelfeld*, *right bracket*, ...) to which specific rules are associated (Drach 1937) (Bech 1955). It has been successfully used in HPSG (Reape 1994) (Kathol 1995) and in dependency grammars (Duchier

and Debusmann 2001) (Gerdes and Kahane 2001). The present analysis is based on the description by (Gerdes and Kahane 2001), who use the hierarchy of topological domains as the only phrase structure. Translating this to LFG, we give the c-structure a topological interpretation. Therefore, contrary to earlier Dutch/German LFGs (Bresnan et al. 1982) (Zaenen and Kaplan 1995), we do not use X-bar syntax notations for our c-structure. Indeed, the topological notation corresponds better to the idea underlying LFG that the c-structure does not represent syntactic information such as subcategorization contrary to the X-bar phrase structure.

In Section 1, we propose some enrichments of the f-structure which avoid a multiplication of the phrase structure rules. The Topological model is presented in Section 2, and we present and comment the phrase structure rules of our German LFG in Section 3.

1 f-structure and dependency

We argue that word order in German (as in any natural language) depends exclusively on the syntactic links between words (the dependency tree) and the information structure (the communicative grouping of words). It has long been noted that the f-structure is closely related to classic syntactic dependency trees (Tesnière 1959) (Mel'čuk 1988). Nevertheless, there are some differences, starting with the fact that the f-structure is not in general a tree. It contains information of various types, in particular syntactic and semantic relationships. As an example consider the analysis of sentence (1) and its usual f-structure given in Figure 1.

(1) *The man seems to have slept.*

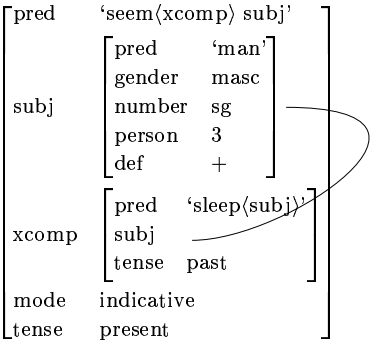


Figure 1: The usual f-structure for (1)

Traditionally, LFG's pred(icate) feature distinguishes between semantic arguments (carrying θ-roles) and purely syntactic arguments, by placing them inside and outside of the predicate's brackets. For instance, the subject of *seems* is not a semantic argument and is noted outside of the brackets. Yet, an argument that is only semantically and not syntactically motivated does not receive a special notation in the classical f-structure. The subject position of the infinitive, for example, is not syntactic; it encodes a more profound relation. In this work, we attempt to encode this difference explicitly by marking exclusively semantic arguments with a dot before the grammatical function. Figure 2 shows the f-structure we propose for the sentences in (1): The subject of *slept*/'sleep', which is only a semantic argument, is noted •subj.

Another problematic point concerns the status of the auxiliary: Today's customary analysis in LFG does not assign an independent predicator to the auxiliary.[1] It only verifies agreement features with the subject while adding grammatical morphemes to the main verb's feature structure. On a semantic level, we agree to treat complex verbal forms (*is sleeping*, *have slept*) in the same way as simple verbal forms (*sleeps*, *slept*); yet syntactically, they behave just like full verbs with verbal subcategorization. In German, for example, the complex verbal form *geschlafen haben* 'to have slept' has the same word order realizations as the control construction *schlafen wollen* 'to want to sleep'. Thus, in order to capture this common behavior, we introduce predicators for auxiliaries just as for control and raising verbs. We obtain a purely syntactic 'predicate' that has no semantic sub-categorization (and therefore no semantic brackets). Nevertheless, we want to indicate that the auxiliary and the participle will form one node in the semantic representation. Thus, the *xcomp* relation of the auxiliary receives a special mark, a circle, indicating that the two nodes it connects form a single semantic unit (Figure 2a,b,d). These new notations in the predicate function are completely compatible with the formalism itself, and moreover, they subsume the traditional analysis. However, our description allows distinguishing explicitly syntactic from semantic dependencies.

In our transcription of the predicate functions in Figure 2b, we denote with In Figure 2b, we indicate the predicate argument structures of the f-structure, depicting the semantic links with a dotted line and the syntactic links with a bold line; an arrow denotes the grammatical function that will be incorporated with another predicate when constructing the semantic graph 2c. Note that the subject link between 'have' and 'man' is neither syntactic nor semantic and simply ensures

[1] In (Bresnan 1982), auxiliaries still held a predicate function. In this sense, the f-structure has developed a more semantic ambition.

the percolation of the value of the syntactic subject of 'seem' to the semantic subject of 'sleep'.

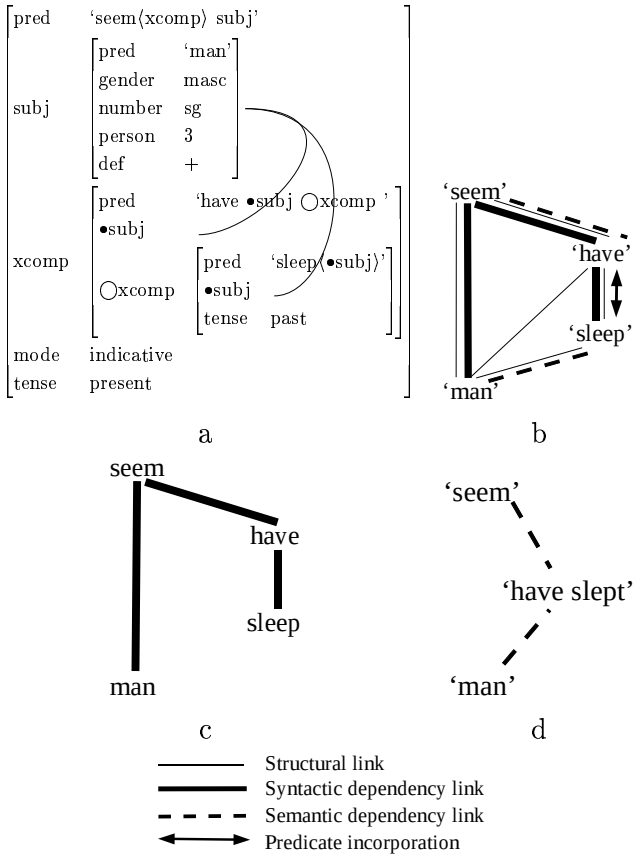


Figure 2: (a) Our f-structure for (1), (b) the underlying predicate-argument structure, and (c) the corresponding syntactic and (d) semantic structure.

2 German word order and the Topological model

Word order in German is much freer than in English. The f-structure of Figure 3, which will be our reference example, has a few dozen linearizations:

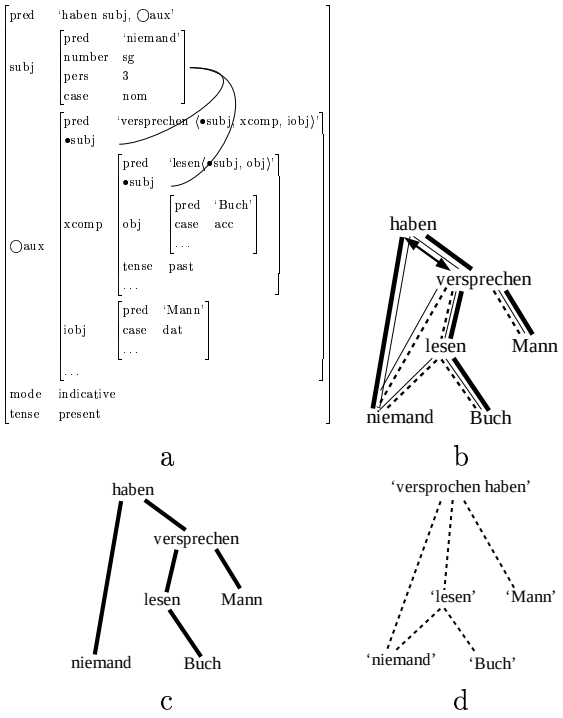


Figure 3: (a) An f-structure, (b) the graphical representation of its predicate functions, (c) the corresponding syntactic dependency tree, and (d) the semantic graph for the sentences in (2).

(2) a. *Niemand hat diesem Mann das Buch zu lesen versprochen*
nobody$_{nom}$ has this man$_{dat}$ the book$_{acc}$ to read promised
'Nobody promised this man to read the book.'

b. *Diesem Mann hat das Buch niemand zu lesen versprochen*

c. *Das Buch zu lesen hat diesem Mann niemand versprochen*

d. *Diesem Mann hat niemand versprochen, das Buch zu lesen*

e. *Diesem Mann hat, das Buch zu lesen, niemand versprochen*

f. *Zu lesen hat diesem Mann das Buch niemand versprochen*

g. *Das Buch hat niemand diesem Mann versprochen zu lesen*

2.1 The Topological model

To get all possible orders, we follow closely the classical topological model. This model supposes that a German sentence consists of a *main domain* composed of a sequence of five *fields*: *Vorfeld, left bracket, Mittelfeld, right bracket,* and *Nachfeld.* A domain is a constituent whose ordered compartments, called fields, can themselves accommodate new constituents.

Main domain

Vorfeld	left bracket	Mittelfeld	right bracket	Nachfeld

Figure 4: Fields of the main domain.

We start our description of word order with the root of the syntactic tree, whatever its semantic type (main verb, modal, or auxiliary); in our example in Figure 3, it is *hat* 'haben'. This word always takes the left bracket, which is the second position. A non-verbal syntactic dependent of the main verb, like *niemand* in our example, goes into one of the major fields, that is Vorfeld, Mittelfeld and Nachfeld. The case of a verbal syntactic dependent is more complicated: A non-finite verbal dependent of the main verb (here the past participle *versprochen* 'promised') can open two kinds of constituents:

- The first possibility is to open an *embedded domain*, consisting of three fields – Mittelfeld, right bracket, and Nachfeld. This embedded domain can accommodate all of the embedded verb's syntactic dependents.
- The second possibility is to construct a restricted phrase, called the *verb cluster*, with only one position for a verbal dependent at its left.

These two kinds of phrases must be placed in very different topological positions: an embedded domain goes to one of the major fields (just as a non verbal dependent)[2], while the verb cluster takes the right bracket. The verb in the right bracket can again have a dependent element. This dependent behaves exactly as a dependent of the main verb (in the left bracket): If it is a non verbal dependent (*diesem Mann*), it goes in one of the major fields;[3] if it is a verbal dependent (*zu lesen*),

[2]This is true for infinitives with *zu* 'to'. Bare infinitives and past participles can only create a new domain in the Vorfeld.

[3]This can be a major field of the domain of its governor, or a major field of a higher domain (2f,g). The subject, which is only a semantic dependent of the past participle, cannot join the embedded domain (except for ergative verbs and passive constructions).

either it creates an embedded domain in one of the major fields – in the Vorfeld for (2c,f), in the Mittelfeld for (2e) and in the Nachfeld for (2d,g) – or it joins its governor in the right bracket (2a,b). In this latter case, it takes the position at the left of its syntactic governor and creates a new restricted constituent, we call a *verb box*. This verb box proposes again only one position at its left reserved for a possible verbal dependent.

We show in the remaining part of this section how two other difficulties of German word order, separable verbal prefixes and auxiliary flip, can easily be integrated in the topological model.

2.2 Separable prefixes

Some verbal constructions such as *anfangen* consist of two parts: a verb, *fangen*, and a so-called separable prefix, *an*.

(3) a. *Die Schule hat um 9 Uhr angefangen*
The school$_{nom}$ has at 9 o'clock on_caught
'The school starts at 9AM.'

b. *Er fängt gleich zu schreien an*
He catches right away to shout on
'He begins to shout right away.'

The prefix *an* behaves just like a verbal dependent of *fangen*, i.e. it goes into the right bracket of the main domain.[4] When *fangen* is in the right bracket, the prefix *an* goes to its left, as a verbal dependent would do (3a). If *anfangen* has a verbal dependent, this dependent behaves as a syntactic dependent of *an* and, in particular, it can join the right bracket taken by *an* (3b).

For these reasons, we treat the verb and its prefix as two syntactic units although the writing conventions of German require joining them graphically when they are next to each other, and semantically, they clearly form an entity. This last point is taken care of in our LFG grammar by marking their syntactic link as semantic incorporation.

2.3 Auxiliary flip

Another difficulty of German verb placement is known as auxiliary flip (or *Oberfeldumstellung*).

(4) *Er wird das Buch haben lesen können*
He will the book have read can
'He will have been able to read the book.'

[4]The separable prefix cannot open an embedded domain. In case of contrast with another prefix, it can go in the Vorfeld.

Contrary to the usual order of verbal sub-categorization in the verb cluster (V_1, V_2V_1, $V_3V_2V_1$, etc.), the Oberfeldumstellung allows the auxiliaries *werden* and *haben* to place their verbal dependent to their right (V_1V_2). We handle this possibility by allowing auxiliaries to open a field for their dependents not only to their left (Oberfeld) but also to their right (Unterfeld).[5] The subsequent verbal dependent can join its governor's Oberfeld ($V_1V_3V_2$ - see example 4) or even the auxiliary's Oberfeld ($V_3V_1V_2$, *Zwischenstellung*).

We would like to stress that all the above word order rules are exclusively based on the syntactic dependencies and semantic dependencies do not intervene.

We will now present our formalization of this topological analysis in LFG.

3 Formalization

The topological model distinguishes constituents and fields. The purpose of fields is to provide a sentence position for different types of constituents, that otherwise would have to be encoded in a large number of different rewriting rules. For instance the Vorfeld can hold very different kind of constituents like nominal, verbal, adjunct, and complementizer phrases, and the same kind of constituents can go in the Mittelfeld and the Nachfeld.

We can see the constituents as boxes, one put into the other; the fields are the compartments of these boxes. In our grammar, this dichotomy is reflected in two types of labels, one for constituents, behaving in the usual way, the other for fields, passing up automatically all of the functional information. Moreover, fields do not verify the constraint that force the f-structure associated to a constituent to have a *pred* feature. In our notation, field names are preceded by underscores (following the notation of the XLFG parser).

In 3, we present the simplified phrase structure rules for our German LFG. This grammar covers only a fragment of German, leaving aside, among others, the internal structure of NPs, relative phrases, and complementizers.

[5] The Unterfeld can only be taken by an infinitive, and the past participle has to surface as an *ersatz-infinitive* if it goes into the Unterfeld. Furthermore, the governed verbs V_2 taking the Unterfeld form a closed class including modal and perception verbs (and some others like *helfen*, 'help', the causative/permissive *lassen* 'make/let' ... – *haben* 'have' itself also allows this right-placement, which suffices to explain the cases of 'double flip' giving $V_1V_2V_3$, $V_1V_2V_4V_3$).

Phrase structure rules for German:[6]

MD $\rightarrow$ _VF _LB (_MF) (_RB) (_NF)
ED $\rightarrow$ (_MF) _RB (_NF)
_VF $\rightarrow$ _XF
_VF $\rightarrow$ ED
($\uparrow$ xcomp) = $\downarrow$
_XF $\rightarrow$ ADVP
($\uparrow$ *xcomp** *adjunct*) $\ni$ $\downarrow$
_LB $\rightarrow$ V
$\uparrow$=$\downarrow$
_MF $\rightarrow$ _XF*
_RB $\rightarrow$ VC
$\uparrow$=$\downarrow$
_RB $\rightarrow$ VC
($\uparrow$ *xcomp*) =$\downarrow$
_NF $\rightarrow$ _XF*
_XF $\rightarrow$ ED
($\uparrow$ *xcomp*$^+$) =$\downarrow$
_XF $\rightarrow$ NP
($\uparrow$ *subj*) =$\downarrow$
_XF $\rightarrow$ NP
($\uparrow$ *xcomp** $\{obj, iobj\}$) =$\downarrow$
VC $\rightarrow$ (_O) _H (_U)
VB $\rightarrow$ (_O) _H
_H $\rightarrow$ V
$\uparrow$=$\downarrow$
_O $\rightarrow$ VB
($\uparrow$ *xcomp*$^+$) =$\downarrow$[7]
_U $\rightarrow$ VC
($\uparrow$ *xcomp*) =$\downarrow$
($\uparrow$ *type*) = *aux*
($\downarrow$ *modal*) = +
($\downarrow$ *ersatz*) = +
($\downarrow$ *tense*) $\neq$ zu-inf

Our initial symbol is the main domain (MD). In our grammar, the embedded domain (ED) takes the place of verbal phrases

[6]We suppose that cases (nom, acc, dat) are imposed in subcategorization rules in the lexicon. Nevertheless, it could be possible to consider them as grammatical rules. In such case, we would add cases in order rules (nom for subj, acc for obj, and dat for iobj). It must be remarked that some verbs have a gen subject or two accusative complements.

[7]If we left out the + sign, we could exclude the Zwischenstellung, not accepted by all German natives.

for infinitives and past participles. Other projections of verbs, which cannot take noun phrases, are the verb cluster (VC) and the verbal box (VB).

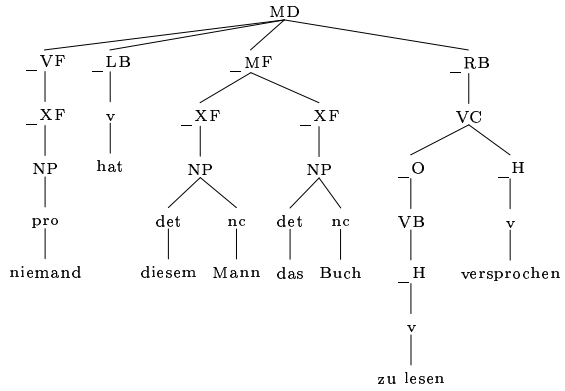
We note fields that can remain empty as optional. In order to generalize the similar behavior of the three major fields – Vorfeld (_VF), Mittelfeld (_MF) and Nachfeld (_NF) – in the LFG formalism, we have to introduce the additional label _XF, although it does not really have the status of a field.

As usual, long-distance dependencies are taken care of by functional uncertainty. We easily obtain topicalization (placement in the Vorfeld, including VP fronting) and scrambling (mixed order of complements of different verbs in the Mittelfeld).

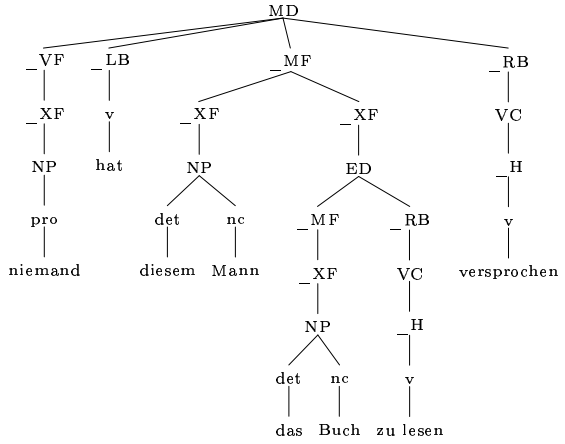
The placement of a verbal subordinative chain (a *xcomp* chain) into a topological domain obeys a simple principle: We place the verbs in descending order. The left bracket (_LB) is taken at first, then the right bracket (_RB), then the Oberfeld (_O) or the Unterfeld (_U) inside this right bracket, and so on. In the case of an embedded domain, which does not have a left bracket, the same principle applies, starting with the placement in the right bracket. Any phrase structure based formalism as LFG and HPSG does not allow expressing this generality, and our grammar thus contains two rules for the right bracket; one for the right bracket of the main domain, where the verb's f-structure gets the *xcomp* position of the main f-structure, and a second rule for the right bracket of the embedded domain, where the verb is the head of the domain.

Our analysis of the verb cluster allows all the correct orderings, in particular Oberfeldumstellung and Zwischenstellung. It is easy to exclude this last possibility, only accepted by parts of the native German speakers, by suppressing the + in the functional equation of the verb box (VB) in the rule of the Oberfeld (_O).

(a)



(b)



(c)

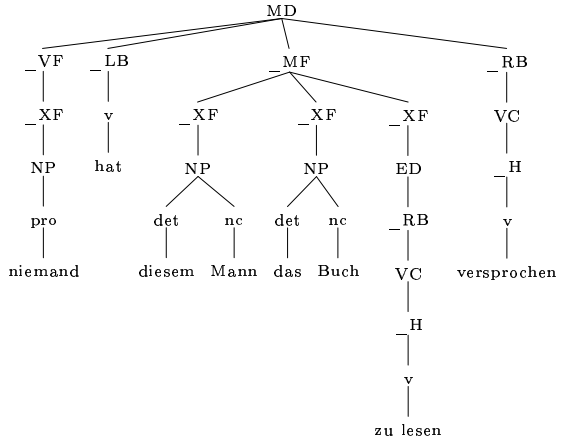


Figure 5: Three c-structures of (2a) corresponding to different topological (and communicative) structures.

Our grammar produces the desired f-structure (Figure 3a) for sentence (2a) using three different c-structures (Figure 5a,b,c). In the first analysis, *zu lesen versprochen* forms a verb cluster. This corresponds to the possibility that the two verbs form a unique prosodic pattern, with stress only on the first syllable of the first verb's radical. The second analysis reflect the fact, that embedded domains can be placed in the Mittelfeld just as in example (2e), where the verb cluster analysis is impossible. Here, the embedded domain forms an independent prosodic group. The last c-structure corresponds to the additional possibility that the embedded domain does not contain all of its verbs complements, as in (2f,g), where this is the only possible analysis. The existence of these different prosodic patterns for the same word order has been demonstrated in (Gerdes and Kahane 2002). They argue that these groupings correspond to different information structures.

Conclusion

We have shown that it is possible to create a Lexical Functional Grammar for German based on the classical topological model. We obtain a grammar of remarkable simplicity that can handle an important number of grammatical phenomena, usually considered as complex. The grammar is implemented in XLFG (Clément and Kinyon 2001), version 3.4.3, for the moment only with a toy lexicon.

In this study we are not concerned with coordination and with the order in the Mittelfeld. We believe these restrictions to be of a different nature: (Lenerz 1977) and (Choi 1999) among others have shown that the German ordering constraints for the Mittelfeld depend mostly on the information structure of the sentence. However in Dutch, a closely related language with a very similar topological structure, the word order in the Mittelfeld is constraint primarily by the syntactic position. This requires an enrichment of the LFG formalism (the *f-precedence*) proposed by (Zaenen and Kaplan 1995).

The transfer of the topological model into the LFG formalism gave us the opportunity to reexamine the theoretical status of the two principal structures of LFG:

- The f-structure can differentiate syntactic and semantic dependencies.
- The c-structure can encode word order and (prosodic and informational) groupings of words. We obtain a c-structure that is completely liberated from its functional burden inherited from X-bar syntax.

It comes out that the clear distinction of a topological level, a syntactic level, and a semantic level is just as useful for an adequate linguistic description as for an economic formalization.

We hope that this study can contribute to a convergence of various formalisms that can handle a topological description of German word order, like LFG, HPSG and dependency grammars.

References

Bech, G. 1955. *Studien über das deutsche Verbum infinitum.* Tübingen: Niemeyer.

Bresnan, J. 1982. Control and Complementation. *Linguistic Inquiry* 13(3). Also in Joan Bresnan, ed. 1982. The Mental Representation of Grammatical Relations, Chapter 5, pp. 282–390. Cambridge, MA: The MIT Press.

Bresnan, J., R. M. Kaplan, S. Peters, and A. Zaenen. 1982. Cross-serial dependencies in Dutch. *Linguistic Inquiry* 13(4):613–635. Reprinted in W. Savitch et al. (eds) The Formal Complexity of Natural Language, 286–319. Dordrecht: D. Reidel.

Choi, H.-W. 1999. *Optimizing Structure in Context - Scrambling and Information Structure.* Stanford: CSLI Publications.

Clément, L., and A. Kinyon. 2001. XLFG-an LFG parsing scheme for French. In *LFG 2001*, Hong Kong. http://www.lionel-clement.net/xlfg/.

Drach, E. 1937. *Grundgedanken der deutschen Satzlehre.* Frankfurt: Diesterweg.

Duchier, D., and R. Debusmann. 2001. Topological Dependency Trees: A Constraint-Based Account of Linear Precedence. In *ACL 2001*, Toulouse.

Gerdes, K., and S. Kahane. 2001. Word Order in German: A Formal Dependency Grammar Using a Topological Hierarchy. In *ACL 2001*, Toulouse.

Gerdes, K., and S. Kahane. 2002. Phrasing it differently. In L. Wanner (Ed.), *Works in Meaning-Text Theory in honour of Igor Mel'čuk*, Philadelphia. Benjamins.

Kathol, A. 1995. *Linearization-based German Syntax*. PhD thesis, Ohio State University.

Lenerz, J. 1977. Zur Abfolge nominaler Satzglieder im Deutschen. In *TBL Verlag Günter Narr*, Tübingen.

Mel'čuk, I. 1988. *Dependency Syntax: Theory and Practice*. New-York: SUNY Press.

Reape, M. 1994. Domain Union and Word Order Variation in German. In J. N. et al. (Ed.), *German in Head-Driven Phrase Structure Grammar*. Stanford: CSLI Lecture Notes.

Tesnière, L. 1959. *Éléments de syntaxe structurale*. Paris: Kliencksieck.

Zaenen, A., and R. M. Kaplan. 1995. Formal devices for linguistic generalizations: West Germanic word order in LFG. In J. S. Cole, G. M. Green, and J. L. Morgan (Eds.), *Linguistics and Computation*, 3–27. Stanford, CA: CSLI Publications.

GETARUN PARSER
A parser equipped with Quantifier Raising and Anaphoric Binding based on LFG

Rodolfo Delmonte
Ca' Garzoni-Moro, San Marco 3417
Università "Ca Foscari"
30124 - VENEZIA
Tel. 39-41-2578464/52/19 - Fax. 39-41-5287683
E-mail: delmont@unive.it - website: project.cgm.unive.it

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

GETARUN PARSER

Rodolfo Delmonte

Abstract

GETARUN, the system for text understanding developed at the University of Venice, is equipped with three main modules: a lower module for parsing where sentence strategies are implemented; a middle module for semantic interpretation and discourse model construction which is cast into Situation Semantics; and a higher module where reasoning and generation takes place. We assume that from a psycholinguistic point of view, parsing requires setting up a number of disambiguating strategies, basically to tell arguments apart from adjuncts and reduce the effects of backtracking. The system is based on LFG theoretical framework and has a highly interconnected modular structure. It is a top-down depth-first DCG-based parser written in Prolog which uses a strong deterministic policy by means of a lookahead mechanism with a WFST to help recovery when failure is unavoidable due to strong attachment ambiguity. It is divided up into a pipeline of sequential but independent modules which realize the subdivision of a parsing scheme as proposed in LFG theory where a c-structure is built before the f-structure can be projected by unification into a DAG. As to multilinguality, the basic tenet of the parser is based on a UG-like perspective, i.e. the fact that all languages share a common core grammar and may vary at the periphery: internal differences are taken care of by parameterized rules. The DCG grammar allows the specification of linguistic rules in a highly declarative mode: it works topdown and by making a heavy use of linguistic knowledge may achieve an almost complete deterministic policy. Parameterized rules are scattered throughout the grammar so that they can be activated as soon as a given rule is entered by the parser.

1. Introduction

GETARUN, the system for text understanding developed at the University of Venice, is equipped with three main modules: a lower module for parsing where sentence strategies are implemented (Delmonte, 1990; Delmonte and Bianchi and Pianta, 1992); a middle module for semantic interpretation and discourse model construction which is cast into Situation Semantics; and a higher module where reasoning and generation takes place (Delmonte, 2000a; Delmonte and Bianch, 2002).

We assume that from a psycholinguistic point of view, parsing requires setting up a number of disambiguating strategies, basically to tell arguments apart from adjuncts and reduce the effects of backtracking.

The system is based on LFG theoretical framework (see Bresnan, J. 2001) and has a highly interconnected modular structure. It is a top-down depth-first DCG-based parser written in Prolog which uses a strong deterministic policy by means of a lookahead mechanism with a WFST to help recovery when failure is unavoidable due to strong attachment ambiguity.

It is divided up into a pipeline of sequential but independent modules which realize the subdivision of a parsing scheme as proposed in LFG theory where a c-structure is built before the f-structure can be projected by unification into a DAG. In this sense we try to apply in a given sequence phrase-structure rules as they are ordered in the grammar: whenever a syntactic constituent is successfully built, it is checked for semantic consistency, both internally for head-spec agreement, and externally, in case of a non-substantial head like a preposition dominates the lower NP constituent; other important local semantic consistency checks are performed with modifiers like attributive and predicative adjuncts. In case the governing predicate expects obligatory arguments to be lexically realized they will be searched and checked for uniqueness and coherence as LFG grammaticality principles require.

Whenever a given predicate has expectancies for a given argument to be realized either optionally or obligatorily this information will be passed below to the recursive portion of the parsing: this operation allows us to implement parsing strategies like Minimal Attachment, Functional Preference and other ones (see Delmonte and Dolci 1989; Delmonte and Dolci 1997).

As to multilinguality, the basic tenet of the parser is based on a UG-like perspective, i.e. the fact that all languages share a common core grammar and may vary at the periphery: internal differences are predicted by parameters. The DCG grammar allows the specification of linguistic rules in a highly declarative mode: it works topdown and by making a heavy use of linguistic knowledge may achieve an

almost complete deterministic policy. Parameterized rules are scattered throughout the grammar so that they can be made operative as soon as a given rule is entered by the parser.

In particular, a rule may belong either to a set of languages, e.g. Romance or Germanic, or to a subset thereof, like English or Italian, thus becoming a peripheral rule. Rules are activated at startup and whenever a switch is being operated by the user, by means of logical flags appropriately inserted in the right hand side of the rule. No flags are required for rules belonging to the common core grammar.

Some such rules include the following ones: for languages like Italian and Spanish, a Subject NP may be an empty category, either a referential little pro or an expletive pronoun; Subject NPs may be freely inverted in postverbal position, i.e. preverbal NP is an empty category in these cases. For languages like Italian and French, PP or adverbial adjuncts may intervene between Verb and Object NP; adjectival modifiers may be taken to the right of their head Noun. For languages like English and German, tense and mood may be computed in CP internal position, when taking the auxiliary or the modal verb. English allows an empty Complementizer for finite complement and relative clauses, and negation requires do-support. Italian only allows for a highly genre marked (literary style) untensed auxiliary in Comp position.

Syntactic and semantic information is accessed and used as soon as possible: in particular, both categorial and subcategorization information attached to predicates in the lexicon is extracted as soon as the main predicate is processed, be it adjective, noun or verb, and is used to subsequently restrict the number of possible structures to be built. Adjuncts are computed by semantic compatibility tests on the basis of selectional restrictions of main predicates and adjuncts heads.

Syntactic rules are built using CP-IP functional maximal projections. Thus, we build and process syntactic phenomena like wh- movement before building f-structure representations, where quantifier raising and anaphoric binding for pronominals takes place. In particular, all levels of Control mechanisms which allow coindexing at different levels of parsing give us a powerful insight into the way in which the parser should be organized.

Yet the grammar formalism implemented in our system is not fully compliant with the one suggested by LFG theory, in the sense that we do not use a specific Feature-Based Unification algorithm but a DCG-based parsing scheme. In order to follow LFG theory more closely, unification should have been implemented. On the other hand, DCGs being based on Prolog language, give full control of a declarative rule-based system, where information is clearly spelled out and passed on and out to higher/lower levels of computation. In addition, we find that topdown parsing policies are better suited to implement parsing strategies that are essential in order to cope with attachment ambiguities (but see below). We use XGs (extraposition grammars) introduced by Pereira(1981;1983). Prolog provides naturally for backtracking when allowed, i.e. no cut is present to prevent it. Furthermore, the instantiation of variables is a simple way for implementing the mechanism for feature percolation and/or for the creation of chains by means of index inheritance between a controller and a controllee, and in more complex cases, for instance in case of constituent ellipsis or deletion. Apart from that, the grammar implemented is a surface grammar of the chosen languages. Also functional Control mechanisms – both structural and lexical - have been implemented as close as possible to the original formulation, i.e. by binding an empty operator in the subject position of a propositional like open complement/predicative function, whose predicate is constituted by the lexical head.

Being a DCG, the parser is strictly a top-down, depth-first, one-stage parser with backtracking: differently from most principle-based parsers presented in Berwick et al.(1991), which are two-stage parsers, our parser computes its representations in one pass. This makes it psychologically more realistic. The final output of the parsing process is an f-structure which serves as input to the binding module and logical form: in other words, it constitutes the input to the semantic component to compute logical relations. In turn the binding module may add information as to pronominal elements present in the structure by assigning a controller/binder in case it is available, or else the pronominal expression will be available for discourse level anaphora resolution. As to the most important features of DCGs, we shall quote from Pereira and Warren(1980) conclusions, in a comparison with ATNs:

"Considered as practical tools for implementing language analysers, DCGs are in a real sense more powerful than ATNs, since, in a DCG, the structure returned from the analysis of a phrase may depend on items which have not yet been encountered in the course of parsing a sentence. ... Also on the practical side, the greater clarity and modularity of DCGs is a

vital aid in the actual development of systems of the size and complexity necessary for real natural language analysis. Because the DCG consists of small independent rules with a declarative reading, it is much easier to extend the system with new linguistic constructions, or to modify the kind of structures which are built. ... Finally, on the philosophical side, DCGs are significant because they potentially provide a common formalism for theoretical work and for writing efficient natural language systems."(ibid,278).

1.1 Implementing LFG theory

As said above, there are a number of marked differences in the treatment of specific issues, concerning Romance languages, which are not sufficiently documented in the linguistic literature (however see Bresnan, 2001). In particular,
- we introduced an empty subject pronominal - little pro - for tensed propositions, which has different referential properties from big PRO; this has an adverse effect on the way in which c-structure should be organized. We soon realized that it was much more efficient and effective to have a single declarative utterance-clause level where the subject constituent could be either morphologically expressed or Morphologically Unexpressed (MUS). In turn MUS or little pros could be computed as variables in case the subject was realized in postverbal position. At the beginning, LFG posited the existence of a rule for sentence structure which could be rewritten as VP both in case there was no subject, and in case the subject was expressed in postverbal position, an approach that we did not implement;
- as to functional constituents: CP typically contains Aux-to-Comp and other preposed constituents, adjuncts and others; IP contains negation, clitics, and tensed verbal forms, simple and complex, and expands VPs as complements and postverbal adjuncts;
- each constituent is semantically checked for consistency before continuing parsing; we also check for Uniqueness automatically by variable instantiation. But sometimes, in particular for subject-verb agreement we have to suspend this process to check for the presence of a postverbal NP constituent which might be the subject in place of the one already parsed in preverbal position(but see below);
- syntactic constituency is replicated by functional constituency: subject and object are computed as constituents of the annotated c-structure, which rewrite NP - the same for ncomp - this is essential for the assignment of the appropriate annotated grammatical function; this does not apply to VP, a typical LFG functional non-substantial constituent;
- our lexical forms diverge from the ones used in the theoretical framework: we introduced aspectual categories, semantic categories and selectional restrictions in the main lexical entry itself, which are used to compute tense/aspect structural representations;
- we also have semantic roles already specified in the lexical form and visible at the level of syntactic-semantic parsing;
- rather than generating a c-structure representation to be mapped onto the f-structure, we generate a fully annotated c-structure representation which is then checked for Grammatical Principles Consistency at the level of number/type of arguments and of Adequacy for adjuncts, on the basis of lexical form of each predicate and semantic consistency crossed checks for adjuncts.

1.2 Disambiguating constituency with functional mapping

As shown in Fig.2 below, the parser is made up of separate modules:
1. The Grammar, based on DCGs, incorporates Extraposition to process Long Distance Dependencies, which works on annotated c-structures: these constitute the output to the Interpretation Module;
2. The Interpretation Module checks whether f-structures may be associated to the input partially annotated c-structure by computing Functional Uniqueness, Coherence and Completeness. Semantic roles are associated to the input grammatical function labels at this level, after semantic selectional restrictions are checked for membership;
3. The Mapping scheme, to translate trees into graphs, i.e. to map c-strutures onto f-structures. The parser builds annotated c-structure, where the words of the input sentence are assigned syntactic constituency and functional annotations. This is then mapped onto f-structure, i.e. constituent information is dropped and DAGs are built in order to produce f-structure configuration.
Mapping into f-structure is a one-to-many operation: each major constituents may be associated with different functional values:
a. NP --> SUBJect, both in preverbal and postverbal position - VP internally, VP adjoined and IP adjoined (see Delmonte, 1987) - with any kind of verbal category; OBJect, usually in VP internal position, but also in preverbal position at Spec CP in case of

reversed transitive structures; NCOMP predicative function - if not proper noun - occuring with copulative, and ECM verbs like "consider, believe"; closed ADJunct with [temporal] value, as the corresponding English example "this morning", which however in Italian can be freely inserted in sentence structure;
b. AP --> Modifier of an NP head, occurring as attribute in prenominal and as predication in postnominal position; ACOMP predicative function occurring with copulative, and ECM verbs; open XADJunct occurring freely at sentence level. Other examples of open adjuncts are: floating quantifiers, which however may only occur VP internally; doubling emphatic pronoun "lui" which also occurs VP internally and is computed as open adjunct;
c. AdvP --> Open or closed Adjuncts according to its selectional properties, occurring anywhere in the sentence according to their semantic nature;
d. PP --> OBLiques, when selected by a given predicate; PCOMP predicative function, when selected by a given predicate - both these two types of argument usually occur VP internally but may be fronted; open XADJunct or closed ADJunct according to semantic compatibility checks;
e. VP' --> VCOMP infinitivals, when selected by a given predicate; SUBJect propositional clauses; closed ADJuncts with semantic markers like "for"; VP' gerundive and participial, which are always computed respectively as closed ADJuncts the former and as open ADJuncts the latter;
f. S' -->or CP as main clauses, or subordinate clauses, as well as sentential complements and SUBJect propositional clauses;
g. Clitics and Pronominal elements are also computed as Nps or PPs, because they are assigned grammatical functions when not associated to NP dislocation in preverbal position: in that case, the clitic is simply erased and TOPic function is associated with the binder NP.

1.3 Quantifier Raising

Since we know that quantifiers and quantified NPs usually take scope at propositional and NP level, we assume f-structure to be an adequate level of representation in which quantifer scope can be computed. In this we partially follow Halvorsen's proposal (see Halvorsen; Halvorsen & Kaplan), which however requires a further mapping from f-structures to s-structures in order to do that. We proceed as follows: after assigning Q-Markers to quantifiers and quantified NPs and adding this information as attribute-value pair at f-structure, we perform Quantifier Raising by traversing f-structure until we reach a propositional node. At that level we deposit a Quantifier-Operator(Q-Op), in an attribute that has a list as its value. Once Q-Ops have been produced, we are in a position to assign quantifier scope. In case more than one Q-Op is present in the list, the algorithm simply reorders the operators according to their quantifying force, and/or to grammatical function. Otherwise, a search downward is performed in the f-structure for other q-ops. When some q-marker is found another attribute-value pair is added at pred level indicating a quantified interpretation. We could duplicate the procedure at NP level by taking into account all NP modifiers in case they are quantified (but see Delmonte & Bianchi, 1992; Delmonte 1997; Dibattista et al. 1999).

1.4 The Binding Module

The output of grammatical modules is fed then onto the Binding Module(BM) which activates an algorithm for anaphoric binding in LFG terms using f-structures as domains and grammatical functions as entry points into the structure. Pronominals are internally decomposed into a feature matrix which is made visible to the Binding Algorithm(BA) and allows for the activation of different search strategies into f-structure domains. Antecedents for pronouns are ranked according to grammatical function, semantic role, inherent features and their position at f-structure. Special devices are required for empty pronouns contained in a subordinate clause which have an ambiguous context, i.e. there are two possible antecedents available in the main clause. Also split antecedents trigger special search strategies in order to evaluate the possible set of antecedents in the appropriate f-structure domain. Special care is paid to pronominals bound by quantifiers or quantified NPs in order to detect crossover violations. The output of the BA is then passed on to an Interpretation Module which operates locally in order to spot the presence of conditions for Specific or Arbitrary Reading for pronominal expressions (see Delmonte & Bianchi, 1991). Eventually, this information is added into the original f-structure graph and then passed on to the Discourse Module(DM) (see Delmonte & Bianchi, 1999; Delmonte 2002).
In Fig.1 We show the interrelations existing between the parser architecture as it has been described so far and the LFG theoretical background. The different

levels of representations required by LFG are coupled with modules and processes of the parser.

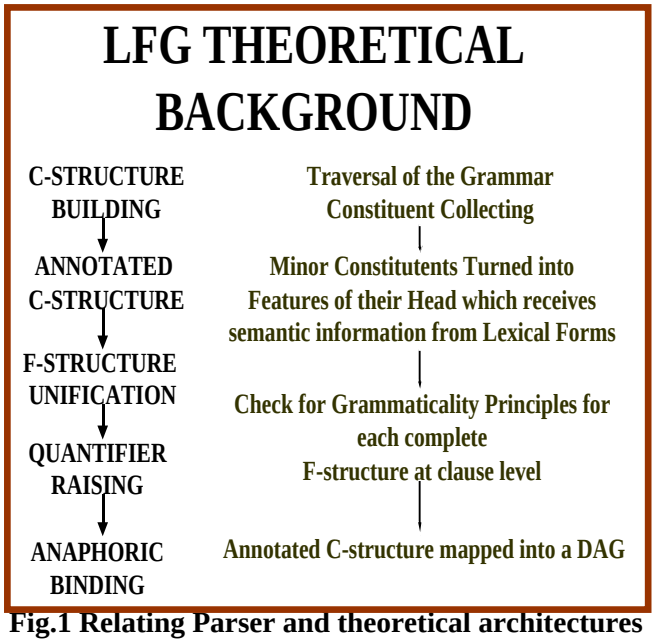


Fig.1 Relating Parser and theoretical architectures

2. Tracing c-structure rules

The parser looks for syntactic constituents adjoined at CP level: in case of failure, it calls for IP level constituents, including the SUBJect which may either be a clause or an NP. This is repeated until it reaches the Verbal Phrase: from that moment onward, the syntactic category associated to the main verb - transitive, unergative, unaccusative, impersonal, atmospheric, raising, psych, copulative - and the lexical form of the predicate, are both used as topdown guidelines for the surface realization of its arguments. Italian is a language which allows for empty or morphologically unexpressed Subjects, so that no restriction may be projected from the lexicon onto c-structure: in case it is empty, a little pro is built in subject position, and features are left as empty variables until the tensed verb is processed.

The grammar is equipped with a lexicon containing a list of fully specified inflected word forms where each entry is followed by its lemma and a list of morphological features, organized in the form of attribute-value pairs. However, morphological analyzers for Italian and English are also available with big root dictionaries (90,000 for Italian, 25,000 for English) which only provide for syntactic subcategorization, though. The fully specified lexicon has been developed for Italian, English and German and contains approximately 5,000 entries for each language. In addition to that there are all lexical form provided by a fully revised version of COMLEX, and in order to take into account phrasal and adverbial verbal compound forms, we also use lexical entries made available by UPenn and TAG encoding. Their grammatical verbal syntactic codes have then been adapted to our formalism and is used to generate an approximate subcategorization scheme with an approximate aspectual and semantic class associated to it. Semantic inherent features for Out of Vocabulary Words , be they nouns, verbs, adjectives or adverbs, are provided by a fully revised version of WordNet in which we used labels similar to those provided by CoreLex.

Once the word has been recognized, lemmata are recovered by the parser in order to make available the lexical form associated to each predicate. Predicates are provided for all lexical categories, noun, verb, adjective and adverb and their description is a lexical form in the sense of LFG. It is composed both of functional and semantic specifications for each argument of the predicate: semantic selection is operated by means both of thematic role and inherent semantic features or selectional restrictions. Moreover, in order to select adjuncts appropriately at each level of constituency, semantic classes are added to more traditional syntactic ones like transitive, unaccusative, reflexive and so on. Semantic classes are of two kinds: the first class is related to extensionality vs intensionality, and is used to build discourse relations mainly; the second class is meant to capture aspectual restrictions which decide the appropriateness and adequacy of adjuncts, so that inappropriate ones are attached at a higher level.

Grammatical functions are used to build f-structures and the processing of pronominals. They are crucial in defining lexical control: as in Bresnan (1982), all predicative or open functions are assigned a controller, lexically or structurally. Lexical control is directly encoded in each predicate-argument structure, but see below.

Structural information is essential for the assignment of functions such as TOPic and FOCus. Questions and relatives, (Clitic) Left Dislocation and Topicalization are computed with the Left Extraposition formalism presented by Pereira(1981;1983). In particular, Extraposition Grammars allows for an adequate implementation of Long Distance Dependencies: restrictions on which path a certain fronted element may traverse in order to bind its empty variable are very easily described by allowing the Prolog variable associated to the element in question - a wh- word or a

relative pronoun - to be instantiated in a certain c-structure configuration.
Eventually, structural information is translated into functional schemata which are a mapping of annotated c-structures: syntactic constituency is now erased and only functional attribute-value pairs appear. Also lexical terminal categories are erased in favour of referential features for NP's determiners, as well as temporal and modal features. Some lexical element disappears, as happens with complementizers which are done away with and substituted by the functional attribute SCOMP or COMP i.e., complement clause - in Italian FCOMP.
As said above, we think it highly important to organize c-structure rules for sentence level representation by means of the introduction of functional major constituents at the following basic levels:

```
CP -->  Spec, C'
C'  -->  C, IP
IP  -->  Spec=NP(subject),  I'
I'  -->  Inflected Tensed Verb Form, VP.
```

According to this configuration, adjuncts and constituents like wh- words for questions and topicalized NPs, adjoined at sentence level, will be computed at first in a CP constituent and then passed down to the lower level of analysis. This organization of constituency allows for complementizers, i.e. the head of CP, to be kept separate in C' level so that a nice interaction may be possible, if needed.
However, rule ordering may clash with the need to produce the most adedquate structure as soon as possible without incurring in an inefficient andc psychologically unmotivated backtracking.
When IP is reached, the NP subject or sentential subject should be computed: at this point there are at least two possible parsing strategies to be followed, both theoretically plausible. The former is in line with LFG traditional view that no empty category should be produced unless it is strictly required by language typology. The latter is in line with Chomsky's assumption of the necessity to pose a basic structural or deep structure configuration which is equal for all languages. In the former case no empty subject NP should arise in case the structure to be analysed is an inverted construction: this is justified by the fact that the Subject NP is actually to be found in inverted VP internal, or VP adjoined position. Since no NP movement is postulated in LFG, there would be no possibility to adequately bind the empty category previously generated in preverbal position. Thus, the sentential structure of inverted, presentational constructions corresponds directly to a VP. In the latter case, the subject position is filled by an empty category and it should be erased when parsing the actual lexical subject NP in postverbal position. In case we choose the first strategy, this is how the reasoning proceeds with parsing: since Italian freely allows the subject to be left lexically empty, and since we do not want to produce an empty little pro in case the lexical subject is present in postverbal position, the rule for marked presentational IP must be accessed first. In case the sentence has a canonical structure, failure would have to take place in order to start the second rule for canonical IP. The reason to let the presentational structure come first is due to the fact that in case the sentence starts with a lexical NP before the VP (computed at first as subject), a fail is performed very soon. Here we should note exceptions like bare NPs with a head noun homograph with a verb - which is a common case in English - less so in Italian. In case no lexical NP is present, there are still two possibilities: we either have a canonical structure with an empty little pro as subject, or we have a fully inverted structure – i.e. preposed Object NP followed by postverbal Subject NP.
At first we must assume that no subject is available and try to compute an inverted Subject: clearly this might fail, in case the NP computed in the VP is not interpretable as Subject but as Object of the main predicate. However, we take the marked option to be more frequent and less extendible than the other way round: not every verb class may undergo subject inversion, which is not completely free (see Delmonte, 91). And even if it does, there is quite a number of restrictions that may be made to apply to the inverted subject, as to its referential features (definiteness, etc.), which do not apply to the canonical object NP.
As can be easily gathered, the number of drawbacks from the point of view of parsing strategies is quite high: failure requires backtracking to be performed and this might be very heavy, depending mainly on what has been previously computed as inverted Subject. Not to mention the fact that VP rules should be duplicated in part.
As to the second choice, there will be only one general procedure for parsing grammatical sentence structure, which would postulate the existence of a subject position to be filled either by lexical material or by an empty constituent. In other words, in case the

sentence starts with a verb we let typologically determined parameters decide whether it is possible to build an empty subject NP or not: in case we are parsing Italian texts this parameter would be active, but in case we are parsing a text belonging to Germanic languages, it would be disactivated. When we generate an empty category in subject position it remains to be decided what to do with it in case a lexical NP in postverbal position is computed, and this is interpreted as the actual Subject function of the sentence, the trace should be discarded.

C-structure building in our parser corresponds to a partial interpretation of each constituent: in fact, when a parse is completed, we assign a structurally determined grammatical function label which could match semantic checking procedures performed when annotated c-structure is built, or it might be rejected as semantically inappropriate, due to selectional restrictions associated to that constituent. Grammatical function assignment at a c-structure level is required in all cases in which a presentational construction has been parsed: it is just on the basis of the structural position of a given constituent, the postverbal NP, that we know what is the pragmatic import of the entire utterance. And this will be registered only in the grammatical function assigned to one of the arguments of the predicate, which is computed either as Subj_Foc, or Subj_Top according to whether it is an indefinite or definite NP respectively. The empty NP subject is not bound to the actual lexical NP found in inverted position, and it is simply discarded from the final representation. In this way, the annotated c-structure outputted by the parser is CP rewritten as VP, but the postverbal subject is computed with an adequate grammatical function. Backtracking is thus totally eliminated, and there is only one single procedure which applies to all sentential structures.

At the highest level we want to differentiate between direct speech and other utterances, which are all called by the rule standard_utterance. Only simplex utterances are accepted here and not complex utterances. A simple utterance can either be started by the SPEC of CP containing a ±wh element, i.e. it can be a question, a topicalization or a left dislocation, or a yes-no question. These are fairly general rules applying to all languages: there is a call to adjuncts at CP level, and a call to aux-to-comp elements which however is typologically restricted. It applies to Germanic languages in particular, where auxiliaries may be computed in Comp position, as will be discussed below in more detail. In case the call to canonical structures fails, we try topicalized and dislocated constructions.

The first of these calls, is a call to impersonal SI: in case of reverse constructions, these are usually associated to passive voice. Then we have reverse constructions with transitive verbs which may have the object in sentence initial position: this NP cannot be used to trigger Agreement with the Verb, and must be taken at Top level. Two possibilities exist now: in the first case, we have a typical left dislocation construction, which has the following essential structure: NP Object, NP Subject, resumptive clitic, VP structure, and may be exemplified by the sentence,

1."Il libro Gino lo ha comprato"/The book John it has bought.

In the second case, left dislocation is accompanied by subject inversion, i.e. the essential structure, NP Object, resumptive clitic, tensed verb, NP subject, as in the following example,

2."Il libro lo ha comprato Gino"/The book it has bought.

Thus, when a clitic is present and the Subject is in inverted postverbal position, this is captured by the rule where the topicalized Object NP is linearly followed by a clitic which has accusative case, and no intervening lexical NP can be computed.

From this structural level, either a VP could be straightforwardly computed, or else, an empty NP Subject be postulated and then discarded. We prefer the first option since from structural representation we can already tell that the subject must be empty, owing to the presence of an object clitic. In the former case, the clitic is present but the SUBJect is in preverbal position. Or else, which is the option available in all languages, as in

3."Ski John loves",

we have a Topicalization or focalization, i.e. the OBJect is in Top CP, and the SUBJect in preverbal position. No clitic appears. This is achieved partly by constituent check when building annotated c-structure, and partly by Interpretation at sentence level, when all constituents have been recovered and constructed. The presence of a bound clitic for clitic left dislocation, or else the absence of a clitic and the type of constituent can now be adequately dealt with respectively, as a case of left clitic dislocation with subject focalization in the first case, left clitic dislocation in the second and topicalization in the third case. In the former case, the inverted subject will be interpreted as Foc; in the latter case the preposed object will be interpreted as Top; and in the third case the preposed object as Foc. Notice also that no lexical subject might be present,

thus resulting in a simple clitic left dislocated structure with an empty NP subject.
It is interesting to note that all this will follow independently by letting the adequate structure building and constituent check at VP level. After CP has been correctly built, we activate the call to IP where subject NP and negation may be parsed; then a call to i_one_bar, will activate calls to Clitics and Infl, for all inflected verbal forms. The call to Clitics, is allowed both for German and Italian; it also applies exceptionally to English "there", provided no NP subject has been analyzed. Infl is a call which is specialized for different languages and the subsequent typologically marked constructions of Italian.
Parsing the appropriate VP structure requires the instantiation of the appropriate syntactic verb class of the main predicate: in this case, it may either belong to the class of psychic or copulative verbs. Theoretically speaking, c-structure is now represented with a verbal phrase which contains no verb, which has been raised to infl, in case it is a tensed finite verb. In order to let the analysis enter the call for inchoativized verb_phrase, aspectual class is needed; in addition, Subject NP should be an empty pro, in Italian.
All subject inverted constructions at VP level, are constrained by a check on the subject NP: it must be an empty category. This check also applies to impersonal-si constructions and to dislocated constructions. In this way, no backtracking will be allowed. In addition, syntactic category of the main verb should always be checked accordingly. In particolar, inchoative constructions and impersonal-si constructions are also typologically marked, since they are only allowed in Romance languages; also fully inverted transitive constructions and intransitive reflexive structures are only present in Romance languages. The call to intransitive verbal phrases is subsequently further split into the four syntactic classes: {atmospheric, unaccusative, inergative, impersonal}. Transitive structures are differentiated according to the complement type: i.e. adverbial objects require a separate treatment owing to differences in the interpretation of its NP, see
4."John spent three days in Venice"
5."Mary weighs 45 kilos"
and so on. Transitive verbs with open complements are also special in that their object is nonthematic and is interpreted in the open complement, see verbs like
6."believe John stupid"
7."see Mary in the shower",
"consider" and so on. The presence of syntactic classes in verbal entries listed in the lexicon is used as a filter in the construction of VP might be regarded as redundant information, but from a computational point of view it turns out to be a very powerful tool. In some cases, however, interpretation will folow from the actual structural representation rather than from lexical information alone: this is the case of all non subcategorized cases of secondary predication, like resultative structures and other cases of attributive open complements. This is especially so, seen that Italian verbs select auxiliaries according to syntactic class! In particular, unaccusatives require "essere/be" and unergatives "avere/have".
The rule for copulative VPs starts by checking whether a "lo" clitic has been found, in that case this will constitute the open complement, as in
8."Gino lo è" = John it is (happy),
where "lo" is the resumptive invariable clitic for open complements in Italian. In case another clitic has been computed, this can only be treated as a complement or adjunct of the open complement, and is consequently included as first element in the list of constituents passed onto the open complement call. The XCOMP call can be instantiated with any of the allowable lexical heads X=P,A,N,V,Adv, and its associated main constituents. Finally, there is a check on the specifier and referentiality of the preverbal NP computed: in case it is a deictic pronoun, or the Xcomp is a proper noun, this structure will be locally computed as inverted structure as appears in sentences like:

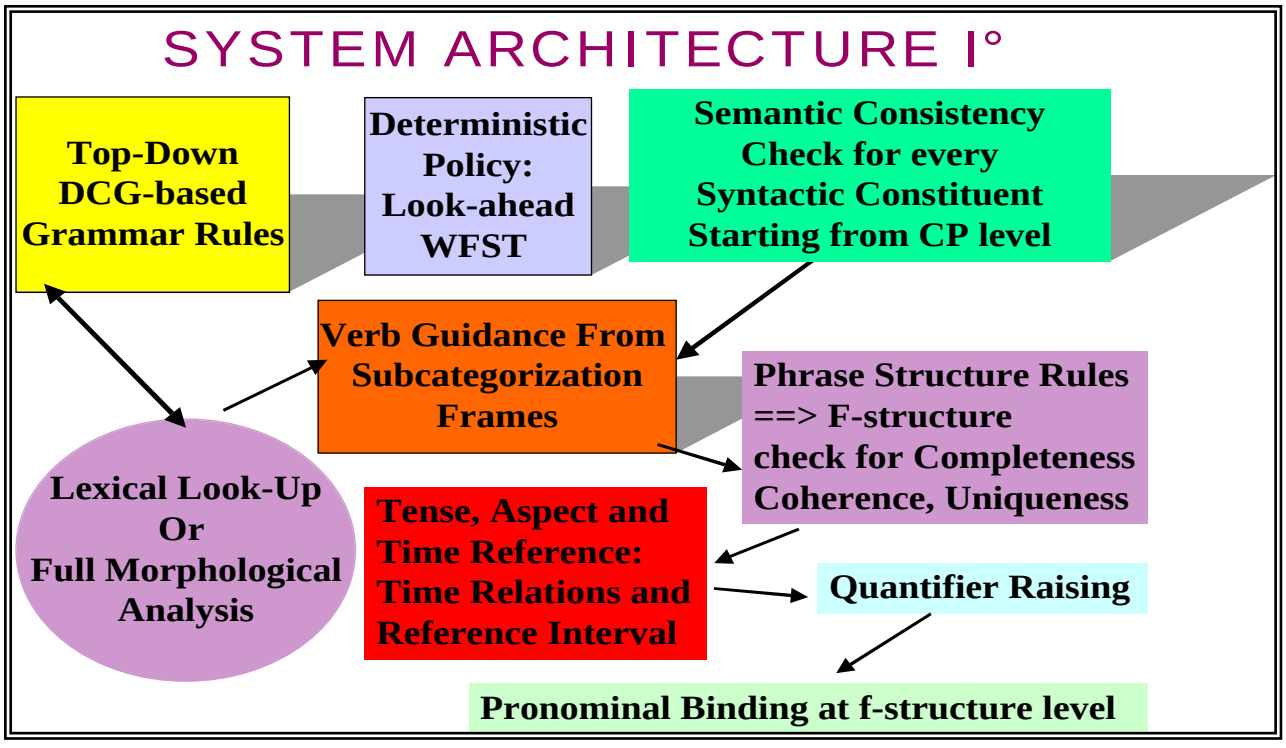
9.The murdered is John,
10.This is a spy story.
Here below we list some of the higher rules of the grammar with one of the interpretation rules for copulative constructions:

```
utterance          --> assertion_direct
utterance          --> standard_utterance
standard_utterance--> wh_question
standard_utterance--> yes_no_question
standard_utterance--> assert_cp

assert_cp-->    aux_to_comp
                adjunct_cp
                i_double_bar
assert_cp-->    object
                adjunct_cp
                pro=SI
                verb_phrase_impersonal
assert_cp-->    object
```

```
                adjunct_cp
                negat
                pro=CLI, {Case=acc}
                verb_phrase_focalized

assert_cp-->    object
                adjunct_cp
                i_double_bar

i_double_bar-->         subject
                negat
                adjs_preverbal
                parenthetical
                i_one_bar

i_one_bar--> verb_phrase_pass_canonic
i_one_bar--> clitics,
             { germanic_aux,
               clitics,
               adjs_post_aux,
               germanic_vp  ;
                                all_languages_vp }

verb_phrase_copulative-->  adv_phrase
                           check_clitic_object
                           xcomp
                           prepositional_phrases
interpret_copulative:-
        lexical-form& predicate-argument_structure
          interpret_subject
          interpret_xcomp
          assign_control_xcomp
          interpret_adjuncts
```

Notice that i_one_bar rewrites as passive VP and in case of failure as active VP: again this is required by the need to activate the appropriate interpretation rule for transitive verb which in most languages is morphologically determined by the presence of the appropriate auxiliary/ies and the past participle of the main verb.

Fig.2 GETARUN Parser Architecture



In this way also the Inflection rule is kept separate from that used for active verbs, which is complicated by the presence of germanic languages: in case an auxiliary has already been taken at CP level, it will have to be copied down in the following VP structure to build the adequate verbal compound.

3. Mapping Functional Structures to an XML output

Lately, the output of the parser, which is computed as a DAG, i.e. a list of interconnected arcs, nodes and

leaves, has been ported to an XML format in order to be viewed in any browser. We decided to treat f-structures and subsidiary f-structures as arcs which contain nodes consisting of features which are attribute-value pairs. A distinctive node has been preserved for predicates, which are called Semantic Words <sw>. We also assigned a special structure to tense/aspect structural nodes which are then treated at the same level of higher f-structures and assigned to an arc. We show here below a simple example based on the sentence "Did Mary like John?"

```
f([did, mary, like, john, ?], es34).
rete(es34).
leaf(n2, f1, es34).
leaf(n3, yes/no_question, es34).
leaf(n4, like, es34).
leaf(n5, [sn/sogg/esperiente/[umano, animato], sn/ogg/tema_emot/
[stato, umano, animato, oggetto]], es34).
leaf(n6, active, es34).
leaf(n7, ind, es34).
leaf(n8, past, es34).
leaf(n9, emotivo, es34).
leaf(n11, sn1, es34).
leaf(n12, [umano], es34).
leaf(n13, mary, es34).
leaf(n15, fem, es34).
leaf(n16, sing, es34).
leaf(n17, 3, es34).
leaf(n19, 0, es34).
leaf(n21, sn2, es34).
leaf(n22, [umano], es34).
leaf(n23, john, es34).
leaf(n25, mas, es34).
leaf(n26, sing, es34).
leaf(n27, 3, es34).
leaf(n29, 0, es34).
leaf(n14, [+ref, -pro, -ana, -class], es34).
leaf(n24, [+ref, -pro, -ana, -class], es34).
leaf(n30, stato, es34).
leaf(n31, [tr(f1_es34)<td(f1_es34)], es34).
leaf(n32, [tr(f1_es34)=tes(f1_es34)], es34).
leaf(n33, -, es34).
leaf(n34, [tr(f1_es34)], es34).
arc(n1, n2, indice, es34).
arc(n1, n3, perf, es34).
arc(n1, n4, pred, es34).
arc(n1, n5, lex_form, es34).
arc(n1, n6, voice, es34).
arc(n1, n7, modo, es34).
arc(n1, n8, tempo, es34).
arc(n1, n9, cat, es34).
arc(n1, n10, sogg/esperiente, es34).
arc(n10, n11, indice, es34).
arc(n10, n12, cat, es34).
arc(n10, n13, pred, es34).
arc(n10, n15, gen, es34).
arc(n10, n16, num, es34).
arc(n10, n17, pers, es34).
arc(n10, n18, spec, es34).
arc(n18, n19, def, es34).
arc(n1, n20, ogg/tema_emot, es34).
arc(n20, n21, indice, es34).
arc(n20, n22, cat, es34).
arc(n20, n23, pred, es34).
arc(n20, n25, gen, es34).
arc(n20, n26, num, es34).
arc(n20, n27, pers, es34).
arc(n20, n28, spec, es34).
arc(n28, n29, def, es34).
arc(n10, n14, tab_ref, es34).
arc(n20, n24, tab_ref, es34).
arc(n1, n30, aspetto, es34).
arc(n1, n31, rel1, es34).
arc(n1, n32, rel2, es34).
arc(n1, n33, definitezza, es34).
arc(n1, n34, ref_int, es34).
```

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<text words="did mary like john ?">
<sent init="es34">
<node type= "index" ><leaf  val="f1"> </leaf></node>
<node type= "perf" ><leaf val="yes/no_question">
</leaf></node>
<sw id="sw_167" type="pred" ><leaf  val="like"> </leaf></sw>
<node type= "lex_form" ><leaf  val="[sn/sogg/esperiente/[umano,
animato], sn/ogg/tema_emot/[stato, umano, animato, oggetto]]">
</leaf></node>
<node type= "voice" ><leaf  val="active"> </leaf></node>
<node type= "mood" ><leaf  val="ind"> </leaf></node>
<node type= "tense" ><leaf  val="past"> </leaf></node>
<node type= "cat" ><leaf  val="emotivo"> </leaf></node>
<arc type= "subj/experiencer" ><node type= "index" ><leaf
val="sn6"> </leaf></node>
        <node type= "cat" ><leaf  val="[umano]"> </leaf></node>
        <sw id="sw_168" type="pred" ><leaf  val="mary">
</leaf></sw>
          <node type= "gen" ><leaf  val="fem"> </leaf></node>
          <node type= "num" ><leaf  val="sing"> </leaf></node>
          <node type= "pers" ><leaf  val="3"> </leaf></node>
          <node type= "spec" ><node type= "def" ><leaf
val="'0'"> </leaf></node>
</node>
          <node type= "tab_ref" ><leaf  val="[+ref, -pro, -ana, -
class]"> </leaf></node>
</arc>
<arc type= "obj/emot_theme" ><node type= "index" ><leaf
val="sn16"> </leaf></node>
        <node type= "cat" ><leaf  val="[umano]"> </leaf></node>
        <sw id="sw_169" type="pred" ><leaf  val="john">
</leaf></sw>
          <node type= "gen" ><leaf  val="mas"> </leaf></node>
          <node type= "num" ><leaf  val="sing"> </leaf></node>
          <node type= "pers" ><leaf  val="3"> </leaf></node>
        <node type= "spec" ><node type= "def" ><leaf  val="'0'">
</leaf></node>
</node>
    <node type= "tab_ref" ><leaf  val="[+ref, -pro, -ana, -class]">
</leaf></node>
</arc>
<arc type= "'tense/aspect'" ><leaf  val="state"> </leaf>
<node type= "rel1" ><leaf val="[after(tr(f1_es34),
td(f1_es34))]"> </leaf></node>
```

```
    <node type= "rel2" ><leaf  val="[tr(f1_es34)=tes(f1_es34)]">
</leaf></node>
      <node type= "definiteness" ><leaf  val="-"> </leaf></node>
      <node type= "ref_int" ><leaf  val="[tr(f1_es34)]">
</leaf></node>
</arc>
</sent>
</text>
```

4. Parsing Strategies and Preferences

The parser has been built to simulate the cognitive processes underlying the grammar of a language in use by a speaker, taking into account the psychological nuances related to the wellknown problem of ambiguity, which is a pervading problem in real text/communicative situation, and it is regarded an inseparable benchmark of any serious parser of any language to cope with.

In order for a parser to achieve psychological reality it should satisfy three different types of requirements: psycholinguistic plausibility, computational efficiency in implementation, coverage of grammatical principles and constraints. Principles underlying the parser architecture should not conform exclusively to one or the other area, disregarding issues which might explain the behaviour of the human processor. In accordance with this criterion, we assume that the implementation should closely mimick phenomena such as Garden Path effects, or an increase in computational time in presence of semantically vs. syntactically biased ambiguous structures. We also assume that a failure should ensue from strong Garden Path effects and that this should be justified at a psycholinguistic interpretation level (see Pitchett.

Differently from what is asserted by global or full paths approaches (see Schubert, 1984; Hobbs et al., 1992), we believe that decisions on structural ambiguity should be reached as soon as possible rather than deferred to a later level of representation. In particular, Schubert assumes "...a full paths approach in which not only complete phrases but also all incomplete phrases are fully integrated into (overlaid) parse trees dominating all of the text seen so far. Thus features and partial logical translations can be propagated and checked for consistency as early as possible, and alternatives chosen or discarded on the basis of all of the available information(ibid., 249)." And further on in the same paper, he proposes a system of numerical 'potentials' as a way of implementing preference trade-offs. "These potentials (or levels of activation) are assigned to nodes as a function of their syntactic/semantic/pragmatic structure and the preferred structures are those which lead to a globally high potential. Other important approaches are represented by Hindle et al., 1993, who attempt to solve the problem of attachment ambiguity in statistical terms (see Delmonte 2000b,2000c,2000d).

Among contemporary syntactic parsing theories, the garden-path theory of sentence comprehension proposed by Frazier(1987a, b), Clifton & Ferreira (1989) among others, is the one that most closely represents our point of view. It works on the basis of a serial syntactic analyser, which is top-down, depth-first - i.e. it works on a single analysis hypothesis, as opposed to other theories which take all possible syntactic analysis in parallel and feed them to the semantic processor. From our perspective, it would seem that parsing strategies should be differentiated according to whether there are argument requirements or simply semantic compatibily evaluation for adjuncts. As soon as the main predicate or head is parsed, it makes available all lexical information in order to predict if possible the complement structure, or to guide the following analysis accordingly. As an additional remark, note that not all possible syntactic structure can lead to ambiguous interpretations: in other words, we need to consider only cases which are factually relevant also from the point of view of language dependent ambiguities.

Parsing theories are of two kinds: the first garden-path theory (hence GPT) is syntactically biased, and the second incremental-interactive theory (hence IIT) is semantically biased. There is a crucial difference between the two theories: whereas GPT claims that a single analysis of a syntactic structure ambiguity is initially constructed and passed to the semantic module, IIT claims that the syntactic module offers all grammatical alternatives to the semantic one, in parallel, to be evaluated. There is evidence that is favourable to both sides on this issue (see G.Altman, 1989).

The basic claims of GPT are that the sentence processing mechanism (the parser) uses a portion of its grammatical knowledge, isolated from world knowledge and other information, in initially identifying the relationships among the phrases of a sentence. IIT permits a far more intimate and elaborate interaction between the syntax and the semantic/referential modules. Altmann(1989) offers a functional argument against a system in which choices are initially made by a syntactic processor, and later

corrected by appeal to meaning and context. He says that if referential or discourse information is available, only a strange processor would make decisions without appealing to it. It is also our opinion that all lower level constraints should work concurrently with higher level ones: in our parser all strategies are nested one inside another, where Minimal Attachment (hence MA) occupies the most deeply nested level.

The list of examples here below includes sentences used in the literature to support one or the other of the two parsing theories, GPT and IIT (see Altman (ed), 1989).

11. Mary put the book on the table
12. Mary put the book on the table in her bag
13. Mary saw the cop with the binoculars
14. Mary saw the cop with the revolver
15. The thieves stole the painting in the museum
16. The thieves stole the painting in the night
17. John saw Mary in the kitchen
18. John saw Mary from the bathroom

Steedman & Altmann(1989) note that several of the ambiguities present in these sentences and resolved by Minimal Attachment involve contrasts between NP modification and other structures. However, examples 12, 13 and 15 require some more knowledge, linguistic one.

In example 11 we assume that the PP should be computed as an argument of the main predicate "put", thus following a MA strategy when parsing the NP "the book". However, the same strategy would lead to a complete failure in example 12, where the PP should be taken as np modifier. If we look at subcategorization requirements of the verb, example 1 constitutes a clear case of Verb Guidance and of semantic role satisfaction requirements: the main predicate requires an argument which is a locative, so at every decision point in which a PP might be taken, argument requirements should be accessed and a MA strategy imposed locally by FP.

Example 13 contains an instrumental adjunct: when the head preposition "with" is met, the parser will not close the NP "the cop" and will continue building an internal PP modifier since preposition "with" heads a compatible NP modifier, a comitative. In our dictionary the verb "see" has one single lexical entry but a list containing two different lexical forms. The first form in the list has a higher number of arguments,

I. see <SUB/perceiv, OBJ/theme, PCOMP/locat>

where the Pcomp predicates a location of the Object; and a second form, where the Pcomp is absent. In order for a Location PP to be accepted, the head preposition should be adequate, and "with" does not count as such. In examples 13 and 14, the first decision must be taken when computing NP structure. In fact, a PP headed by preposition "with" is a semantically compatible np modifier - a comitative - and the analysis should be allowed to continue until the PP is fully analysed. In other words the parser should verify PP attachment consistency inside the NP constituent.

In the following examples (15, 16), argument structure plays no role whatsoever: instrumentals, comitatives, locatives with predicate "steal" are all cases of sentential adjuncts, and only semantic criteria can apply. As a matter of fact, "in the museum" might be freely attached lower at NP level as well as higher at Sentence level. This is due to the fact that head preposition "in" constitutes a viable local NP modifier and there are no argument requirements from the main verb predicate. However, "in the night" is not a possible NP modifier and example 16 is a clear case of minimal attachment sentence. On the contrary, in example 15 PP attachment is ambiguous between a np internal modifier and VP level attachment.

In example 17 we understand that the location at which Mary was when the seeing event took place is the kitchen: we also understand that John might have been in the same location or in a different one, already provided by the previous context, and this can be achieved by FP which activates MA and makes the locative PP available at VP level.

In example 18, on the contrary, we understand that the location from the which the seeing event took place is the bathroom and that John was certainly there; however we are given no information whatsoever about Mary's location. This case is treated as the previous one, except that the PP is computed as sentence adjunct rather than as VP complement.

4.1 Two mechanisms at work

We implemented two simple enough mechanisms in order to cope with the problem of nondeterminism and backtracking. At bootstrapping we have a preparsing phase where we do lexical lookup and we look for morphological information: at this level of analysis of all input tokenized words, we create the lookahead stack, which is a stack of pairs input wordform - set of preterminal categories, where preterminal categories are a proper subset of all lexical categories which are actually contained in our lexicon. The idea is simply to prevent attempting the construction of a major constituent unless the first entry symbol is well

qualified. The following list of preterminal 14 symbols is used:

19. PRETERMINAL SYMBOLS
1. v=verb-auxiliary-modal-clitic-cliticized verb
2. n=noun – common, proper;
3. c=complementizer
4. s=subordinator;
5. e=conjunction
6. p=preposition-particle
7. a=adjective;
8. q=participle/gerund
9. i=interjection
10. g=negation
11. d=article-quantifier-number-intensifier-focalizer
12. r=pronoun
13. b=adverb
14. x=punctuation

As has been reported in the literature (see Tapanainen, 1994; Brants, 1995), English is a language with a high level of homography: readings per word are around 2 (i.e. each word can be assigned in average two different tags). Lookahead in our system copes with most cases of ambiguity: however, we also had to introduce some disambiguating tool before the input string could be safely passed to the parser. Disambiguation is applied to the lookahead stack and is operated by means of Finite State Automata. The reason why we use FSA is simply due to the fact that for some important categories, English has unambiguous tags which can be used as anchoring in the input string, to reduce ambiguity. I'am now referring to the class of determiners which is used to tell apart words belonging to the ambiguity class [verb,noun], the most frequent in occurrence in English.

In order to cope with the problem of recoverability of already built parses we built a more subtle mechanism that relies on Kay's basic ideas when conceiving his Chart(see Kay, 1980; Stock, 1989). Differently from Kay, however, we are only interested in a highly restricted topdown depthfirst parser which is optimized so as to incorporate all linguistically motivated predictable moves. Any already parsed NP/PP is deposited in a table lookup accessible from higher levels of analysis and consumed if needed. To implement this mechanism in our DCG parser, we assert the contents of the structure in a table lookup storage which is then accessed whenever there is an attempt on the part of the parser to build up a similar constituent. In order to match the input string with the content of the stored phrase, we implemented a WellFormed Substring Table(WFST) as suggested by Woods(1973).

Now consider the way in which a WFST copes with the problem of parsing ambiguous structure. It builds up a table of well-formed substrings or terms which are partial constituents indexed by a locus, a number corresponding to their starting position in the sentence and a length, which corresponds to the number of terminal symbols represented in the term. For our purposes, two terms are equivalent in case they have the same locus and the same length.

In this way, the parser would consume each word in the input string against the stored term, rather than against a newly built constituent. In fact, this would fit and suit completely the requirement of the parsing process which rather than looking for lexical information associated to each word in the input string, only needs to consume the input words against a preparsed well-formed syntactic constituent.

Lookahead is used in a number of different ways: it may impose a wait-and-see policy on the topdown strategy or it may prevent following a certain rule path in case the stack does not support the first or even second match:

a. to prevent expanding a certain rule

b. to prevent backtracking from taking place by delaying retracting symbols from input stack until there is a high degree of confidence in the analysis of the current input string.

It can be used to gather positive or negative evidence about the presence of a certain symbol ahead: symbols to be tested against the input string may be more than one, and also the input word may be ambiguous among a number of symbols. Since in some cases we extend the lookahead mechanism to include two symbols and in one case even three symbols, possibilities become quite numerous.

Consider now failure and backtracking which ensues from it. Technically speaking, by means of lookahead we prevent local failures in that we do not allow the parser to access the lexicon where the input symbol would be matched against. It is also important to say that almost all our rules satisfy the efficiency requirement to have a preterminal in first position in their right-hand side. This is usually related to the property belonging to the class of Regular Languages. There are in fact some wellknown exceptions: simple declarative sentence rule, yes-no questions in Italian. Noun phrase main constituents have a multiple symbols lookahead, adjectival phrase has a double symbol lookahead, adverbial phrase has some special

cases which require the match with a certain word/words like "time/times" for instance. Prepositional phrase requires a single symbol lookahead; relative clauses, interrogative clauses, complement clauses are all started by one or more symbols. Cases like complementizerless sentential complements are allowed to be analysed whenever a certain switch is activated.

Suppose we may now delimit failure to the general case that may be described as follows:

- a constituent has been fully built and interpreted but it is not appropriate for that level of attachment:

failure would thus be caused only by semantic compatibility tests required for modifiers and adjuncts or lack of satisfaction of argument requirements for a given predicate.

Technically speaking we have two main possibilities:

A. the constituent built is displaced on a higher level after closing the one in which it was momentarily embedded.

This is the case represented by the adjunct PP "in the night" in example 16 that we repeat here below:

16. The thieves stole the painting in the night.

The PP is at first analysed while building the NP "the painting in the night" which however is rejected after the PP semantic features are matched against the features of the governing head "painting". The PP is subsequently stored on the constituent storage (the WFST) and recovered at the VP level where it is taken as an adjunct.

B. the constituent built is needed on a lower level and there is no information on the attachment site.

In this case a lot of input string has already been consumed before failure takes place and the parser needs to backtrack a lot before constituents may be safely built and interpreted.

To give a simple example, suppose we have taken the PP "in the night" within the NP headed by the noun "painting". At this point, the lookahead stack would be set to the position in the input string that follows the last word "night". As a side-effect of failure in semantic compatibility evaluation within the NP, the PP "in the night" would be deposited in the backtrack WFST storage. The input string would be restored to the word "in", and analysis would be restarted at the VP level. In case no PP rule is met, the parser would continue with the input string trying to terminate its process successfully. However, as soon as a PP constituent is tried, the storage is accessed first, and in case of non emptiness its content recovered. No structure building would take place, and semantic compatibility would take place later on at sentence level. The parser would only execute the following actions:

- match the first input word with the (preposition) head of the stored term;
- accept new input words as long as the length of the stored term allows it by matching its length with the one computed on the basis of the input words.

As said above, the lookahead procedure is used both in presence and in absence of certain local requirements for preterminals, but always to confirm the current choice and prevent backtracking from taking place. As a general rule, one symbol is sufficient to take the right decision; however in some cases, more than one symbol is needed. In particular when building a NP, the head noun is taken at first by nominal premodifiers, which might precede the actual head noun of the NP. The procedure checks for the presence of a sequence of at least two nouns before consuming the current input token. In other cases the number of preterminals to be checked is three, and there is no way to apply a wait-and-see policy.

Reanalysis of a clause results in a Garden Path(GP) in our parser because nothing is available to recover a failure that encompasses clause level reconstruction: we assume that GP obliges the human processor to dummify all naturally available parsing mechanisms, like for instance lookahead, and to proceed by a process of trial-and-error to reconstruct the previously built structure in order not to fall into the same mistake. The same applies to our case which involves interaction between two separate modules of the grammar.

As an example, consider processing time 5.6 secs with strategies and all mechanisms described above activated, as compared to the same parse when the same are disactivated – 17.3 secs, in relation to the following highly ambiguous example taken from Legal Texts in the Appendix:

Producer means the manufacturer of a finished product, the producer of any raw material or the manufacturer of a component part and any person who by putting his name, trade mark or other distinguishing feature on the product presents himself as its producer.

Computation time is calculated on a Macintosh G4.

In more detail, suppose we have to use the information that "put" is a verb which requires an oblique PP be present lexically in the structure, as results from a check in its lexical form. We take the verb in I position and then open the VP complement structure, which at first builds a NP in coincidence with "the book". However, while still in the NP

structure rules, after the head has been taken, a PP is an option freely available as adjunct.
We have implemented two lookahead based mechanisms which are used in the PP building rule and are always triggered, be it from a position where we have a noun as head and we already built part of the corresponding constituent structure; be it from a position where we have a verb as head and we want to decide whether our PP will be adequate as argument rather than as adjunct - in the latter case it will become part of the Adjunct Set.
The first one is called,
- Cross Compatibility Check (CCC)
This mechanism requires the head semantic features or inherent features to be checked against the preposition, which in turn activates a number of possible semantic roles for which it constitutes an adequate semantic marker. For instance, the preposition "on" is an adequate semantic marker for "locative" semantic role, this will cause the compatibility check to require the presence in the governing heading of inherent or semantic features that allow for location. A predicate like "dress" is computed as an object which can be assigned a spatial location, on the contrary a predicate like "want" is computed as a subjective intensional predicate which does not require a spatial location. However, in order to take the right decision, the CCC must be equipped with the second mechanism we implemented;
The second one is called,
- Argument Precedenc (AP)
The second mechanism we implemented allows the parser to satisfy the subcategorization requirements in any NP constituent it finds itself at a given moment if the parsing process. Suppose that after taking "put" as the main verb, this mechanism is activated, by simply copying the requirements on PP oblique locative present in the lexical form associated with the predicate "put" in the lexicon, in the AP. As soon as the NP "the book" is opened, after taking "book" as N at the head position, the parser will meet the word "on", which allows for a PP adjunct. While in the P head position, the parser will fire the CCC mechanism first to see whether the preposition is semantically compatible, and in case it is, the second AP mechanism will be fired. This will cause the system to do the following steps:
i. check whether the requirements are empty or not;
ii. and in case it is instantiated, to control the semantic role associated with it;
iii. to verify whether the P head is a possible semantic marker for that semantic role: in our case, "on" is a possible semantic marker for "locative" semantic role;
iv. finally to cause the parser to fail on P as head of a PP adjunct of the head noun;
v. produce a closure of NP which obeys Minimal Attachment principle.

4.2 Some examples

In the texts reported in the Appendix there is a great number of examples which can be used as empirical evidence for the need to use lexical information in order to reduce parsing loads resulting from backtracking procedures. We use examples taken from the Legal text included in the Appendix at the end of the paper: we mark decision points with a bar,

20. Council directive | **of** july 1985 | **on** the approximation | **of** the laws, | regulations and | administrative provisions | **of** the Member States | concerning liability | **for** defective products.

At the first boundary we have "of" which is non semantically marked and no prediction is available, so that the default decision is to apply Late Closure, which turns out to be the correct one. When the second preposition is found we are in the NP of the PP headed by "of", and we have taken the date "1985": this will cause the CCC to prevent the acceptance of the preposition "on" as a semantically compatible marker thus preventing the construction of the NP headed by "approximation".
Notice, that in case that would be allowed, the NP would encompass all the following PPs thus building a very heavy NP: "the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products". In case the parser had a structure monitoring strategy all this work would have to be undone and backtracking would have to be performed. Remember that the system does not possibly know where and how to end backtracking unless by trying all possible available combination along the path. In our case, the presence of a coordinate structure would render the overall process of structure recoverability absolutely untenable.
Another important decision has be taken at the boundary constituted by the participial head "concerning": in this case the CCC will take the inherent features of the head "States" and check them with the selection restrictions associated in the lexical

form for the verb "concern". Failure in this match will cause the NP "the Member States" to be closed and will allow the adjunct to be attached higher up with the coordinated head "laws, regulations and administrative provisions". In this case, all the inherent features are collected in a set that subsumes them all and can be used to fire CCC.

Notice that the preposition "for" is lexically restricted in our representation for the noun "liability", and the corresponding PP that "for" heads interpreted as a complement rather than as an adjunct. We include here below the relevant portion of each utterance in which the two mechanisms we proposed can be usefully seen at work. We marked with a slash the place in the input text in which, usually when the current constituent is a NP a decision must be taken as to whether causing the parser to close (MA) or to accept more text (LC) is actually dependent upon the presence of some local trigger. This trigger is mostly a preposition; however, there are cases in which, see e., f., h., i., the trigger is a conjunction or a participle introducing a reduced relative clause. Coordinate NPs are a big source of indecision and are very hard to be detected if based solely on syntactic, lexical and semantic information. For instance, e. can be thus disambiguated, but h. requires a matching of prepositions; In the case represented by i. we put a boundary just before a comma: in case the following NP "the Member State" is computed as a coordination - which is both semantically, syntactically and lexically possible, the following sentence will be deprived of its lexical SUBJect NP - in this case, the grammar activates a monitoring procedure independently so that backtracking will ensue, the coordinate NP destroyed and the comma computed as part of the embedded parenthetical (which is in turn an hypothetical within a subordinate clause!!). Notice also that a decision must be taken in relation to the absolutives headed by a past participle which can be intended as an active or a passive past participle: in the second case the head noun would have to be computed as an OBJect and not as a SUBJect

b. a differing degree of protection of the consumer | **against** damage caused by a defective product | **to** his health or property

c. in all member states | **by** adequate special rules, it has been possible to exclude damage of this type | **from** the scope of this directive

d. to claim full compensation for the damage | **from** any one of them

e. the manufacturer of a finished product, the producer of any raw material or the manufacturer of a component part | **and** any person

f. The liability of the producer | **arising** from this directive

g. any person who imports into the community a product | **for** sale, hire or any form of distribution | **in** the course of his business

h.both by a defect in the product | **and** by the fault of the injured person

i.However, if... the commission does not advise the Member State | **concerned** that it intends submitting such a proposal | **to** the council | , the Member State

4.3 Principles of Sound Parsing

o Principle One: Do not perform any unnecessary action that may overload the parsing process: follow the Strategy of Minimal Attachment;

o Principle Two: Consume input string in accordance with look-ahead suggestions and analyse incoming material obeying the Strategy Argument Preference;

o Principle Three: Before constructing a new constituent, check the storage of WellFormed Substring Table(WFST). Store constituents as soon as they are parsed on a stack organized as a WFST;

o Principle Four: Interpret each main constituent satisfying closer ties first - predicate-argument relations - and looser ties next - open/closed adjuncts as soon as possible, according to the Strategy of Functional Preference;

o Principle Five: Erase short-memory stack as soon as possible, i.e. whenever clausal constituents receive Full Interpretation.

oStrategy Functional Preference: whenever possible try to satisfy requirements posed by predicate-argument structure of the main governing predicate as embodied in the above Principles; then perform semantic compatibility checks for adjunct acceptability.

oStrategy Minimal Attachment: whenever Functional Preference allows it apply a Minimal Attachment Strategy.

The results derived from the application of Principle Four are obviously strictly linked to the grammatical theory we adopt, but they are also the most natural ones: it appears very reasonable to assume that arguments must be interpreted before adjuncts can be and that in order to interpret major constituents as arguments of some predicate we need to have completed clause level structure. In turn adjuncts need to be interpreted in relation both to clause level properties like negation, tense, aspect, mood, possible

subordinators, and to arguments of the governing predicate in case they are to be interpreted as open adjuncts.
As a straightforward consequence, owing to Principle Five we have that reanalysis of a clause results in a Garden Path(GP) simply because nothing is available to recover a failure that encompasses clause level reconstruction: we take that GP obliges the human processor to dummify all naturally available parsing mechanisms, like for instance look-ahead, and to proceed by a process of trial-and-error to reconstruct the previously built structure in order not to fall into the same mistake.

As a peculiar case of parser architecture interaction with theoretically driven strategies, consider the following couple of examples of the Extraposed Relative Clause containing a Short Anaphor taken from the Appendix in the Made-Up sentences reported here below,
21a. The doctor called in the son of the pretty nurse who hurt herself.
21b. The doctor called in the son of the pretty nurse who hurt himself.

In the second example we have the extraposition of the relative clause (hence RC), a phenomenon very common in English but also in Italian and other languages. The related structures theoretically produced, could be the following ones:

21a.
s[np[The doctor],
 ibar[called in],
 vp[np[the son, pp[of, np[the pretty nurse,
 cp[who, s[pro, ibar[hurt],
 vp[sn[herself]]]]]]]]]

21b.
s[np[The doctor],
 ibar[called in],
 vp[np[the son, pp[of, np[the pretty nurse]],
 cp[who, s[pro, ibar[hurt],
 vp[sn[himself]]]]]]]

If this is the correct input to the Binding Module, it is not the case that 10a. will be generated by a parser of English without special provisions. The structure produced in both cases will be 1a. seen that it is perfectly grammatical, at least before the binding module is applied to the structure and agreement takes place locally, as required by the nature of the short anaphor. It is only at that moment that a failure in the Binding Module warns the parser that something wrong has happened in the previous structure building process. However, as the respective f-structures show, the only output available is the one represented by 10b, which wrongly attaches the RC to the closest NP adjacent linearly to the relative pronoun:

10b.
s[np[The doctor],
 ibar[called in],
 vp[np[the son,
 pp[of, np[the pretty nurse,
 cp[who, s[pro, ibar[hurt],
 vp[sn[himself]]]]]]]]]
The reason why the structure is passed to the Binding Module with the wrong attachment is now clear: there is no grammatical constraint that prevents the attachment to take place. The arguments of the governing predicate HURT are correctly expressed and are both coherent and consistent with the information carried out by the lexical form. At the same time the Syntactic Binding has taken place again correctly by allowing the empty "pro" in SUBJect position of the relative adjunct to be "syntactically controlled" by the relative pronoun, which is the TOPic binder, in turn syntactically controlled by the governing head noun, the NURSE. There is no violation of agreement, nor of lexical information, nor any other constraint that can be made to apply at this level of analysis in order to tell the parser that a new structure has to be produced.
The question would be to prevent Failure since we do not want Constituent Structure Building to be dependent upon the Binding of the Short Anaphor. The only way out of this predicament is that of anticipating in Sentence Grammar some of the Agreement Checking Operations as proposed above. So the Parser would be able to backtrack while in the Grammar and to produce the attachment of the Relative Clause at the right place, in the higher NP headed by the masculine N, “the son”. The important result would be that of maintaining the integrity of Syntax as a separate Module which is responsible in “toto” of the processing of constituent structures. The remaining Modules of the Grammar would be fully consistent and would use the information made available in a feeding relation, so that interpretation will follow swiftly.
The reason why the structure is passed to the Binding Module with the wrong attachment is now clear: there is no grammatical constraint that prevents the attachment to take place. The arguments of the

governing predicate HURT are correctly expressed and are both coherent and consistent with the information carried out by the lexical form. At the same time the Syntactic Binding has taken place again correctly by allowing the empty "pro" in SUBJect position of the relative adjunct to be "syntactically controlled" by the relative pronoun, which is the TOPic binder, in turn syntactically controlled by the governing head noun, the NURSE. There is no violation of agreement, nor of lexical information, nor any other constraint that can be made to apply at this level of analysis in order to tell the parser that a new structure has to be produced (see .

To integrate this suggestion coming from Implementation problems, into the theoretical Framework of LFG or other similar theories we need to integrate GRAMMATICALITY PRINCIPLES as they have been stipulated so far, to be consisting of:

- UNIQUENESS
- COHERENCE
- COMPLETENESS

with the additional restriction:

- BOUND ANAPHORA AGREEMENT

i.e. short anaphors should be checked before leaving sentence grammar, for agreement with their antecedents iff available in their Minimal Nucleus.

Bibliography

Delmonte R. 1990. Semantic Parsing with an LFG-based Lexicon and Conceptual Representations, Computers & the Humanities, 5-6, 461-488.

Delmonte R. 2000a. Generating and Parsing Clitics with GETARUN, Proc. CLIN'99, Utrech, pp.13-27.

Delmonte R., D. Bianchi. 2002. From Deep to Partial Understanding with GETARUNS, Proc.ROMAND2002, Università Roma2, Roma, pp.57-71.

Delmonte R., D.Bianchi, E.Pianta. 1992. GETA_RUN - A General Text Analyzer with Reference Understanding, in Proc. 3rd Conference on Applied NLP, Systems Demonstrations, Trento, ACL, 9-10

Bresnan, Joan. 2001. Lexical-Functional Syntax. Blackwells.

Delmonte R., R.Dolci. 1989. Parsing Italian with a Context-Free Recognizer, Annali di Ca' Foscari XXVIII, 1-2,123-161.

Delmonte R. R.Dolci. 1997. Sound Parsing and Linguistic Strategies, Atti Apprendimento Automatico e Linguaggio Naturale, Torino, pp.1-4.

Pereira F. 1981. Extraposition Grammars, American Journal of Computational Linguistics 7, 4, 243-256.

Pereira F. 1983. Logic for Natural Language Analysis, Technical Note 275, Artificial Intelligence Center, SRI International.

Berwick, R.C., S. Abney, and C. Tenny. 1991. Principle-Based Parsing: Computation and Psycholinguistics. New York: Kluwer Academic Publishers.

Pereira F. and Warren D. 1980. Definite Clause Grammars for Language Analysis. A Survey of the Formalism and a Comparison with Augmented Transition Networks. Artificial Intelligence, 13, 231-278.

Delmonte R., D.Bianchi. 1992. Quantifiers in Discourse, in Proc. ALLC/ACH'92, Oxford(UK), OUP, 107-114.

Delmonte R. 1997. Lexical Representations, Event Structure and Quantification, Quaderni Patavini di Linguistica, 15, 39-93.

Dibattista D., E.Pianta, R.Delmonte(1999), Parsing and Interpreting Quantifiers with GETARUN, Proc. VEXTAL, Unipress, 215-225.

Delmonte R., D.Bianchi. 1991. Binding Pronominals with an LFG Parser, Proceeding of the Second International Workshop on Parsing Technologies, Cancun(Messico), ACL 1991, pp. 59-72.

Delmonte R., D.Bianch. 1999. Determining Essential Properties of Linguistic Objects for Unrestricted Text Anaphora Resolution, Proc. Workshop on Procedures in Discourse, Pisa, pp.10-24.

Delmonte R. 2002a. From Deep to Shallow Anaphora Resolution: What Do We Lose, What Do We Gain, in Proc. International Symposium RRNLP, Alicante, pp.25-34.

Delmonte R. 1987. Grammatica e Ambiguità in Italiano, Annali di Ca' Foscari, XXVI, 1-2, 257-333.

Halvorsen P.K. and R.Kaplan. 1988. Projections and semantic description, Proceedings of the International Conference on Fifth Generation Computer System, Tokyo, 1116-1122.

Halvorsen P.K. 1983. Semantics for Lexical Functional Grammar,Linguistic Inquiry 4,567-616.

Bresnan J.(ed.). 1982. The Mental Representation of Grammatical Relations, MIT Press, Cambridge Mass.

Delmonte R. 1991. Empty Categories and Functional Features in LFG, Annali di Ca'Foscari, XXX, 1-2,79-140.

Pritchett B.L. 1992. Grammatical Competence and Parsing Performance, The University of Chicago Press, Chicago.

Schubert L.K. 1984. On Parsing Preferences, Proc. of COLING, 247-250.

Hobbs J.R., D.E.Appelt, J.Bear, M.Tyson. 1992. Robust Processing of Real-World Natural-Language Texts, Proc.2nd Conference on NLP, 186-192.
D.Hindle & M.Roth. 1993. Structural Ambiguity and Lexical Relations, Computational Linguistics 19, 1, 103-120.
Frazier L. 1987a. Sentence processing, in M.Coltheart(ed), Attention and Performance XII, Hillsdale, N.J., Lawrence Elbaum.
Delmonte R. 2000b. Shallow Parsing And Functional Structure In Italian Corpora, LREC, Atene, pp.113-119.
Delmonte R. 2000c. Parsing Preferences and Linguistic Strategies, in LDV-Forum - Zeitschrift fuer Computerlinguistik und Sprachtechnologie - "Communicating Agents", Band 17, 1,2, (ISSN 0175-1336), pp. 56-73.
Delmonte R. 2000d. Parsing with GETARUN, Proc.TALN2000, 7° confèrence annuel sur le TALN,Lausanne, pp.133-146.
Frazier L. 1987b. Theories of sentence processing, in J.Garfield(ed), Modularity in knowledge representation and natural language understanding, Cambridge Mass., MIT Press, 291-308.
Clifton C., & F. Ferreira. 1989. Ambiguity in Context, in G.Altman(ed), Language and Cognitive Processes, op.cit., 77-104.
Altman G.T.M.(ed.). 1989. Language and Cognitive Processes 4, 3/4, Special Issue - Parsing and Interpretation.
Steedman M.J. & Altmann G.T.M. 1989. Ambiguity in Context: a Reply, in Altman(ed), op.cit., 105-122.
Tapanainen P. and Voutilainen A. 1994. Tagging accurately - don't guess if you know, Proc. of ANLP '94, pp.47-52, Stuttgart, Germany.
Brants T. & C.Samuelsson. 1995. Tagging the Teleman Corpus, in Proc.10th Nordic Conference of Computational Linguistics, Helsinki, 1-12.
Kay Martin. 1980. Algorithm Schemata and Data Structures in Syntactic Processing, CSL-80-12, Xerox Corporation, Palo Alto Research Center.
Stock O. 1989. Parsing with flexibility, dynamic strategies and idioms in mind, in Computational Linguistics, 15, 1, 1-18.
Woods W.A. 1973. An Experimental Parsing System for Transition Network Grammars, in Rustin(ed.) 1973. Natural Language Processing, Algorithmic Press, New York, pp.111-154.
Delmonte R. 2002b. Relative Clause Attachment And Anaphora: A Case For Short Binding, Proc.TAG+6, Venice, 84-89.
Delmonte R.(1992), Linguistic and Inferential Processing in Text Analysis by Computer, Padova, Unipress.

APPENDIX

We report in the appendix the texts that have been used to test the functioning of the parser. We only include English and Italian texts. There is no space here to show the output of the parser, which is however freely available from the author (but see also Delmonte, 1992).

SECTION 1. ENGLISH TEXTS

LEGAL TEXT – European Council Directive

Council directive of july 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerns liability for defective products.
The Council of the European Communities has adopted this directive.
Having regard to the proposal from the commission.
Whereas approximation of the laws of the Member States concerning the liability of the producer for damage caused by the defectiveness of his products is necessary because the existing divergences may entail a differing degree of protection of the consumer against damage caused by a defective product to his health or property.
Whereas liability without fault should apply only to movables which have been industrially produced.
Whereas protection of the consumer requires that all the producers involved in the production process should be made liable in so far as their finished product, component part or any raw material supplied by them was defective.
Whereas, to the extent that liability for nuclear injury or damage is already covered in all Member States by adequate special rules, it has been possible to exclude damage of this type from the scope of this directive.
Whereas to protect the physical well-being and property of the consumer the defectiveness of the product should be determined by reference not to its fitness for use but to the lack of the safety which the people at large is entitled to expect.
Producer means the manufacturer of a finished product, the producer of any raw material or the manufacturer of a component part and any person who by putting his name, trade mark or other distinguishing feature on the product presents himself as its producer.
The producer shall be liable for damage caused by a defect in his product.
The injured person shall be required to prove the damage, the defect and the causal relationship between defect and damage.
Where, as a result of the provisions of this directive, two or more persons are liable for the same damage.
They shall be liable jointly and severally, without prejudice to the provisions of national law concerning the rights of contribution or recourse.
A product is defective when it does not provide the safety which a person is entitled to expect, taking all circumstances into account.

The liability of the producer arising from this directive may not, in relation to the injured person be limited or excluded by a provision limiting his liability or exempting him from liability.
This directive shall not apply to injury or damage arising from nuclear accidents and covered by international conventions ratified by the Member States.
Without prejudice to the liability of the producer any person who imports into the community a product for sale, hire or any form of distribution in the course of his business shall be deemed to be a producer within the meaning of this directive.
However, if within three months of receiving the said information the commission does not advise the Member State concerned that it intends submitting such a proposal to the council, the Member State may take the proposed measure immediately.
The liability of the producer may be reduced or disallowed when, having regard to all the circumstances, the damage is caused both by a defect in the product and by the fault of the injured person or any person for whom the injured person is responsible.
Any member state may provide that a producer's total liability for damage resulting from a death or personal injury and caused by identical items with the same defect shall be limited to an amount which may not be less than 70 million ecu.

LITERARY TEXTS – excerpts taken from Virginia Woolf's novels

John gave Mary a rose.
She took it and put it in her hair.
She knew that she had been given a present, something precious.
When Steve faced them saying : "are you enjoying yourselves ?".
"It was horrible ! It was shocking !".
Not for herself.
She felt only hostility and his determination to ruin that wonderful moment.
John smiled and went away embarassed.
The three friends went all outdoors.
As they were walking in the garden, John said to himself "Sara will marry that man", without any resentment.
Richard would marry Sara.
He felt strongly about that.
She was the right person for a man like Richard.
For himself he was absurd.
His demands upon Sara were absurd.
She would have accepted him still if he had been less absurd.
Richard began to sing.
Mary picked up the phone and called Jason.
Her husband, she thought, would have considered such a move as untruthful and utterly base.
Perhaps there was something bad in herself that could not help but do the wrong thing at the wrong time.
Jason answered immediately.
John went into a restaurant.
There was a table in the corner.
The waiter took the order.
The air was nice and clean.
The atmosphere was warm and friendly.
He began to read his book.

LINGUISTICALLY BIASED MADE-UP TEXTS

John beats the donkey that he saw.
John beats every donkey that he saw on his house.
John loves every boy that saw his sister.
John spoke to his wife.
John spoke to his wife because she loves his sister.
Every man that has a donkey beats it.
John talked to Mary about herself on the bus.
John talked to Frank about himself.
John talked to Mary about himself.
Mary talked to Frank about herself.
Which boy saw his sister?
His wife likes John.
John gave him a book.
John wants Mary to read a book.
The farmer beats his donkey because it does not obey him.
Which boy said that he loves his sister?
The farmer does not beat his donkey because he prefers to feed it.
Which farmer beats every donkey that he owns?
Who beats his donkey?
John, who spoke about Mary, said that she loves him.
Who said that John loves Mary ?
Talking about himself pleases Mary.
John beats his donkey with the telescope.
John beats his donkey from the bus.
John does not beat his wife.
His brother.
Who does beat John?
Who does not beat Mary?
Is Mary wise?
Did Mary love John?
While he spoke about his brother_in_law, Frank said to Mary , who insulted him, to let him go.
John will not marry Mary because she is rich.
John told her that Mary is nice.
What did John give to Mary?
What did John say?
What is it about?
Who do his friends like?
Who likes his friends?
His friends like John.
About his friends.
Mary put the book on the table.
Mary saw the cop with the binoculars.
Mary saw the cop with the revolver.
The cop killed the man with the revolver.
Mary promised his brother that she would come.
Mary knew the answer very well.
The doctor called in the son of the pretty nurse who hurt herself.
The doctor called in the son of the pretty nurse who hurt himself.
Mary will say that it rained yesterday.
Mary said that it will rain yesterday.
Mary will say that it rained tomorrow.
Mary said that it will rain tomorrow.
The authorities refused permission to the demonstrators because they feared violence.
The authorities refused permission to the demonstrators because they supported the revolution.
The thieves stole the painting in the night.
The thieves stole the painting in the museum.

SECTION II: ITALIAN TEXTS

o La storia dei tre porcellini / The story of the three little pigs
C'erano una volta tre fratelli porcellini che vivevano felici nella campagna. Nello stesso luogo però viveva anche un terribile lupo che si nutriva proprio di porcellini grassi e teneri. Questi allora, per proteggersi dal lupo, decisero di costruirsi ciascuno una casetta. Il maggiore, Jimmy che era saggio, lavorava di buona lena e costruì la sua casetta con solidi mattoni e cemento. Gli altri, Timmy e Tommy, pigri se la sbrigarono in fretta costruendo le loro casette con la paglia e con pezzetti di legno. I due porcellini pigri passavano le loro giornate suonando e cantando una canzone che diceva: chi ha paura del lupo cattivo. Ma ecco che improvvisamente il lupo apparve alle loro spalle. Aiuto, aiuto, gridarono i due porcellini e cominciarono a correre più veloci che potevano verso la loro casetta per sfuggire al terribile lupo. Questo intanto si leccava già i baffi pensando al suo prossimo pasto così invitante e saporito. Finalmente i porcellini riuscirono a raggiungere la loro casetta e vi si chiusero dentro sbarrando la porta. Dalla finestra cominciarono a deridere il lupo cantando la solita canzoncina: chi ha paura del lupo cattivo. Il lupo stava intanto pensando al modo di penetrare nella casa. Esso si mise ad osservare attentamente la casetta e notò che non era davvero molto solida. Soffiò con forza un paio di volte e la casetta si sfasciò completamente. Spaventatissimi i due porcellini corsero a perdifiato verso la casetta del fratello. "Presto, fratellino, aprici! Abbiamo il lupo alle calcagna". Fecero appena in tempo ad entrare e tirare il chiavistello. Il lupo stava già arrivando deciso a non rinunciare al suo pranzetto.Sicuro di abbattere anche la casetta di mattoni il lupo si riempì i polmoni di aria e cominciò a soffiare con forza alcune volte. Non c'era niente da fare. La casa non si mosse di un solo palmo. Alla fine esausto il lupo si accasciò a terra. I tre porcellini si sentivano al sicuro nella solida casetta di mattoni. Riconoscenti i due porcellini oziosi promisero al fratello che da quel giorno anche essi avrebbero lavorato sodo.

o La stroria di Avveduti /The story of Avveduti
Fino a tre anni fa Franco Avveduti non si era mai immischiato col mondo della pubblica amministrazione. Come burocrate, era un immigrato che veniva dal di fuori. Figlio di buona famiglia, a venti anni decise di iscriversi alla accademia militare di cavalleria. Era un buon allievo con ottime qualifiche. Più tardi fu un ufficiale di successo. Poi nel 1945 Avveduti si dimise dallo esercito. I militari lo avevano deluso. Avveduti, deposta la divisa, si iscrisse alla università. Nel 1947 era laureato e nel 1948 era procuratore legale. Intanto a Verona aveva conosciuto Paola, figlia di Antonio Alberti, potente senatore democristiano, e la aveva sposata. Il senatore poteva considerarsi il più influente uomo politico veronese. Gli elettori lo mandavano al parlamento coprendolo di voti preferenziali. Al seguito di Alberti, che era diventato vicepresidente del senato, Franco Avveduti nello immediato dopoguerra si trasferì a Roma. Tuttavia, da principio si tenne lontano dalla sfera di interessi del suocero.
Gli piaceva parlare del suocero come di una facile occasione mancata che chiunque altro avrebbe sfruttato ma che lui, Avveduti, preferiva lasciare perdere. Solo verso il 1950 decise di accettare un posto nella organizzazione della fiera di Verona. Lo nominarono delegato cioè una specie di funzionario viaggiante con incarichi diplomatici di tenere i rapporti con le delegazioni commerciali, curare i produttori stranieri, le grandi ditte, la stampa. Questo era un compito che corrispondeva bene alla sua vocazione e nel quale Avveduti sapeva giostrare con notevole agilità. Quando il suocero morì, egli non perse il posto. A Verona il collegio di Alberti lo aveva ereditato Trabucchi e col collegio aveva ereditato la presidenza della fiera. Trabucchi continuò a valersi della collaborazione di Avveduti. L'ex-ufficiale del Novara-Cavalleria gli era simpatico. La sua distinzione lo impressionava. Lo confermò nell'incarico alla fiera. Avveduti funzionava benissimo come segretario particolare. Sapeva mobilitare prefetti e questori. Tutti gli invidiavano il suo segretario particolare.

o La storia dei tre porcellini rivisitata / The story of the little pigs revisited
Questa è la storia di tre porcellini che andarono per il mondo a cercare fortuna. I loro nomi erano Timmy, suonatore di flauto, Tommy, violinista e Jimmy, grande lavoratore. Giunti in un bel bosco, decisero di costruire ognuno una comoda casetta. A Timmy non piaceva per niente lavorare così pensò di costruirsi rapidamente una capanna di paglia. In breve la casetta fu pronta e Timmy decise allora di andare a vedere che cosa stavano facendo i suoi fratellini. Incontrò dapprima Tommy il violinista. Anche lui non aveva molta voglia di faticare così costruiva con dei pezzi di legno una semplice casetta. Ben presto anche la casa di legno fu pronta. Come quella di paglia, non era certo molto resistente. Ma i due porcellini scansafatiche se la erano sbrigata in poco tempo ed ora potevano tranquillamente divertirsi. Mentre Timmy suonava il flauto, Tommy lo accompagnava con il suo violino e insieme se la spassavano allegramente. Poi stanchi di fare baldoria, decisero di andare a vedere che cosa stava facendo il loro fratellino. Si misero in cammino e ben presto raggiunsero Jimmy. Il bravo porcellino stava costruendo anche lui la sua casetta. Ma poiché Jimmy era previdente e non aveva paura di lavorare sodo, la costruiva con mattoni e cemento. Jimmy voleva una casa robusta perché sapeva che il lupo cattivo viveva nel bosco vicino. Quando i due pigri porcellini videro Jimmy impegnato nel suo duro lavoro, si misero a ridere a crepapelle. Ma quei due sciocchi porcellini non pensavano al pericolo. Così continuarono a prendere in giro il saggio Jimmy, canticchiando e suonando con il flauto e il violino. I due porcellini, sempre suonando e ballando, tornarono ciascuno alla propria fragile casetta. Ma appena Timmy aprì la porta, sbucò fuori dal bosco il lupo cattivo. Il porcellino lo vide e tremante di paura si chiuse immediatamente in casa. Il lupo cattivo cominciò a chiamarlo. "Apri la porta e fammi entrare nella tua casetta di paglia."

LINGUISTICALLY BIASED MADE-UP TEXTS

Jimmi è in casa.
C è jimmi in casa.
C è jimmi.
Jimmi ha un libro in casa.
Jimmi ha paura del lupo.
Jimmi c'ha un libro in casa.
Jimmi c'ha paura del lupo.
Jimmi è un porcellino.
Jimmi è saggio.
Si costruirono molte case.
Sembra essersi lavorato bene.
Sembra essersi sbrigato in fretta.
Avendo jimmy lavorato bene timmy era felice.
Costruì tutto.
Ne costruì tutto.
Costruì tutti.
Li costruì tutti.
Ne costruì tutti.

Ne costruì molti.
Costruì molto.
Costruì molti.
Avvisò molti.
Apparvero molti.
Apparvero tutti.
Ne apparvero tutti.
Ne apparvero molti.
Avvisò tutti.
Jimmi costruì molte delle sue case con il cemento.
Jimmi ne costruì molte con il cemento.
I porcellini si sono costruiti una casa.
I porcellini gli hanno costruito una casa.
Parlare del suocero era importante.
Parlare del suocero era importante per avveduti.
Il lupo costruì loro una casa.
Il lupo costruì loro una casa ciascuno.
Il lupo gli leccava i baffi.
Il lupo gli riempì i polmoni di aria.
Jimmi gli ha rotto un mattone sulla spalla.
Gli si è rotto un mattone sulla spalla.
Le case sono state ereditate.
Sono state ereditate due case.
Le case furono costruite da jimmi.
Jimmi si è rotto un mattone sulla spalla.
Mario non sapeva a chi parlava.
I suoi lavori alla propria casa.
Il suo attaccamento alla propria famiglia.
La liberazione dei prigionieri da parte del ministro.
La uscita della statale 592.
La uscita dalla statale 592.
Il libro di jimmi di maria.
La interruzione della statale 592 di canelli per lavori che dureranno fino a due settimane.
Questa è la storia di tre porcellini.
Jimmi costruì molte delle sue case coi mattoni.
Però ne costruì alcune con la paglia.
Furono costruite molte case.
Non è arrivato nessuno.
Nessuno è arrivato.
Marco non è arrivato.
Non è arrivato marco.
Il padrone insultò gli studenti.
Il padrone insultò studenti.
Mario corre.
Mario è corso.
Mario correrà da maria.
Mario corre da maria.
Mario è corso per ore.
Mario è corso per tre ore.
Mario domani corre.
Mario domani corre da maria.
Mario ieri corre.
Mario ieri è corso da maria.
Mario giovedì corre da maria.
Mario ogni giorno corre da maria.
Mario ogni tre giorni corre da maria.
Mario giovedì è corso da maria.
Mario è corso tre volte da maria.
Mario sta correndo alle tre.
Mario domani sarà corso da maria.
Mario giovedì era malato.
Mario fa mangiare la mela a maria.
Mario fa avere mangiato la mela a maria.
Mario dice a maria che luigi mangia le mele.
Mario dice a maria che luigi mangia mele.
Mario era malato quando luigi è arrivato.
Mario era malato quando luigi giovedì è arrivato.
Mario partiva quando luigi è arrivato.
Mario partiva quando luigi arrivava.
Maria è stata malata mentre era a verona.
Maria è con luigi.
Ogni uomo che ha un asino lo picchia.
L'asino è arrabbiato.
È un animale disobbediente.
Sono animali disobbedienti.
Uno studente ha letto tutti i libri.
È mio fratello.
Uno studente sta leggendo un libro.
È molto bello.
Uno studente legge un libro.
È necessario per la sua formazione.
Tre ragazzi hanno mangiato un pesce.
Erano salmoni.
Tre ragazzi hanno mangiato un pesce ciascuno.
Gino vuole leggere un libro.
È un libro di storia.
Gino vuole conoscere uno studente che studi marx.
È un amico di maria.
Gino ama molto marx.
Gino vuole conoscere uno studente che studia marx.
Ognuno ama un libro che ha letto.
Era un libro di trabucchi.
Ognuno ama un libro che ha letto.
Erano libri di trabucchi.
Tutte le donne costruirono una casa.
La casa era solida.
Era solida.
Erano solide.
Una donna dice che ogni uomo la ammira.
Vuole anche essere felice.
Una donna vuole che ogni uomo la ammiri.
Gino la conosce.
Ogni donna ha mangiato un pesce.
Era un salmone.
Una donna ha detto che ogni uomo la ammira.
Sono vanitose.
Ogni uomo che ha un asino lo batte.
Parlare di se stesso piace.
La propria salute è necessaria.
La propria salute era necessaria.
Parlare di se stesso piaceva.
La propria salute preoccupa ognuno.
La sua salute preoccupa ognuno.
Ognuno ama i film che ha visto.
I film che ha visto piacciono a ognuno.
La madre di ogni ragazzo pensa che sia un genio.
Mario vede ogni trasmissione che parli della propria storia.
La salute della propria famiglia è importante.
Una donna desidera che ogni uomo la ammiri.
Una donna desidera che gino la ammiri.
Gli uomini che hanno un libro di trabucchi li leggono.
I porcellini hanno visto un lupo da ogni casetta.
I porcellini in ogni casa gridarono aiuto.
Il porcellino in ogni casa gridò aiuto.
Mario telefonò a luigi perché voleva delle informazioni.

Mario criticava luigi perché ha rovinato il file.
Mario critica luigi perché è ipercritico.
Mario criticava luigi perché voleva il file.
La guardia sparò al ladro perché stava scappando.
Le autorità rifiutarono il permesso ai dimostranti perché temevano la violenza.
Le autorità rifiutarono il permesso ai dimostranti perché sostenevano la rivoluzione.
I ladri rubarono i quadri nella notte.
I ladri rubarono i quadri nel museo.
Mario promise a suo fratello che sarebbe tornato.
Il dottore chiamò il figlio della infermiera che si era fatta male.
Il dottore chiamò il figlio della infermiera che si era fatto male.
Parlando di suo suocero, trabucchi ha ordinato a avveduti che lo aspettava di lasciare perdere.
Avveduti ritiene che maria ami la propria famiglia.
Avveduti ritiene che lui ami la propria famiglia.
Avveduti ritiene che la propria sorella ami gino.
Lui ritiene che la propria sorella ami gino.
L'insegnante promosse lo studente poiché era preparato.
La ragazza che la propria salute preoccupa è tua sorella.
La ragazza che sua madre ama è sua sorella.
Mario ha visto lo asino sopra la propria casa.
Mario ha visto lo asino dalla propria casa.
Mario batte un asino sulla propria casa.
Mario batte un asino dalla propria finestra.
Mario ha visto lo asino sopra la sua casa.
Mario ha visto lo asino sopra se stesso.
Mario ha visto lo asino sopra di sé.
Mario ha visto lo asino sopra di lui.
Quale ragazzo hai detto che mario ha visto ?
Chi dici che mario ha visto ?
Chi dici che ha visto se stesso ?
Mario ha visto un asino col cannocchiale.
A se stesso gino ritiene che mario non ci pensi mai.
Di se stesso gino ritiene che mario non parli mai.
Gino ritiene che mario non parli mai di se stesso.
Gino ritiene che mario ci pensi mai a se stesso.
Ognuno ama il suo asino.
Ogni uomo ha detto che batte il suo asino.
Trabucchi ha visto ogni uomo che batte il suo asino.
Mario ritiene che la propria libertà sia importante.
Ognuno ama i film che ha visto.
I film che ha visto piacciono a ognuno.
La madre di ogni ragazzo pensa che sia un genio.
Mario vede ogni trasmissione che parli della propria storia.
Mario ha visto un asino col cannocchiale sopra la propria casa.
Quale libro mario ha perso ?
Di quale libro mario parlava ?
Con quale uomo mario parlava ?
Chi batte lo asino ?
Di chi parlava mario ?
Con chi parlava mario ?
Quale oceano che confina con l'Africa è inquinato ?
Quale è l'oceano che confina con i paesi della Africa ?
Quale padrone che ha un asino lo batte ?
Chi batte l'asino che lui ha visto ?
Quale è l'asino che il padrone batte ?
Di quale ragazzo sua madre parlava ?
Di quale ragazzo la propria madre parlava ?
Quale ragazzo la propria salute preoccupa ?
Chi sai che mario ha visto ?
Avveduti ha incontrato trabucchi che lo accusava di avere perso il suo libro.
Avveduti ha incontrato trabucchi che lo accusava di avere rubato il libro a lui.
Avveduti ha incontrato trabucchi che si accusava di avere preso il libro a lui.
Partire in quel modo provocò a trabucchi dispiacere.
Partire in quel modo gli provocò male di testa.
Io ho incontrato trabucchi che mi ha detto di dire a avveduti che lo aspetta.
Non abbiamo votato per nessun candidato dal momento che il presidente lo aveva raccomandato.
I suoi sostenitori non hanno votato per nessun candidato.
L'uomo che tu dovresti avvertire ogni qualvolta tu lo incontri.
Noi non siamo stati in grado di leggere nessuna confessione prima che lo autore decidesse di renderla pubblica.
Lo uomo che tu dovresti avvertire ogni qualvolta tu incontri.
Noi non siamo stati in grado di leggere nessuna confessione prima che il suo autore decidesse di renderla pubblica.
A loro interessano se stessi.
Avveduti sperava che i giornali parlassero di sé.
Avveduti sperava che i giornali parlassero di lui.
Maria riteneva ognuno innamorato di sé.
Maria riteneva ognuno innamorato di lei.
Mario vide il toro sopra di lui.
Se stessi interessano loro.
Avveduti ritiene che quella casa appartiene alla propria famiglia.
Lui ritiene che gino sia amato dalla propria sorella.
Ad ognuno sta a cuore la propria salute.
Avveduti mi insultò pesantemente quando lo interrogai.

Resumptive Pronouns in LFG*

Yehuda N. Falk

The Hebrew University of Jerusalem

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

* I would like to express my thanks to Ash Asudeh, Aaron Broadwell and Mary Dalrymple for comments on this paper. They should not be held responsible for anything I say here.

1. Introduction

Since the beginnings of generative syntax, filler-gap constructions have attracted a great deal of attention. What has attracted significantly less attention is the other type of long-distance dependency: the resumptive pronoun construction. In this paper, we will outline an analysis of resumptive pronouns in LFG, based primarily on Hebrew but with consideration of other languages.

The major questions that need to be addressed by a theory of resumptive pronouns are the following:

- In what ways are filler-resumptive constructions similar to filler-gap constructions and in what ways are they different? An adequate analysis must account for both the similarities and the differences.

- Why pronouns? In other words, how is it that pronouns come to be used as a way of marking the lower end of a long-distance dependency. The importance of this question is reinforced by the fact that even in languages that do not "have" resumptive pronouns, like English, there is a limited marginal use of resumptive pronouns as a way of circumventing island constraints.

2. Resumptive pronouns vs. gaps

The relationship between filler-gap and filler-resumptive constructions has been discussed in much of the literature on resumptive pronouns. It has led Vaillette (2001) to analyze resumptive pronoun constructions as essentially the same as gap constructions, and Sharvit (1999) to analyze them as being different.

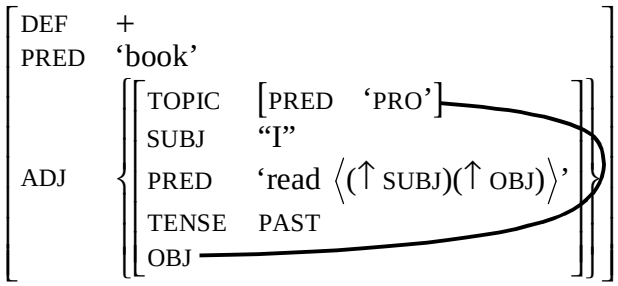
The main similarity between gaps and resumptive pronouns is that both are linked to a discourse function or operator.

(1)	a.	ha-	sefer	še	kara-	ti	oto
		the-	book	that	read.PST-	1SG	it
	b.	ha-	sefer	še	kara-	ti	
		the-	book	that	read.PST-	1SG	

'the book that I read'

This invites an analysis in which the two constructions are essentially the same, with a single f-structure element having two distinct grammatical functions.

(2)



Such an analysis has the advantage of being consistent with the strongest version of the Extended Coherence Condition:

(3) **Discourse Function Clause of Extended Coherence Condition (strong version)**
Discourse functions must be identified with argument or adjunct functions.

While many statements of the Extended Coherence Condition have allowed an anaphoric link, it is not clear that this is required independently of resumptive pronouns, and it is too weak for non-resumptive pronoun languages like English, which require identity. This version of Extended Coherence is clearly too strong for topic-oriented languages like Chinese, in which the sentential topic may be only loosely related to the arguments of a clause, but Hebrew is not such a language.

A strong argument for a long-distance dependency analysis of resumptive pronoun constructions comes from the interaction between that construction and reflexive anaphora in Swedish, as reported by Zaenen, Engdahl, and Maling (1981). The possessive reflexive *sina* is a nuclear anaphor, bound in the minimal complete nucleus. As expected, a reflexive in a fronted phrase has the same anaphoric possibilities as it would in the clause-internal position.

(4) Vilken av sina$_i$ flickvänner tror du att Kalle$_i$ inte längre träffar?
which of self's girlfriends think you that Kalle no longer sees
'Which of his girlfriends do you think that Kalle no longer sees?'

This follows because the f-structure corresponding to 'which of self's girlfriends' has the function of OBJ of 'see'. Crucially, the same thing happens in a resumptive pronoun construction.

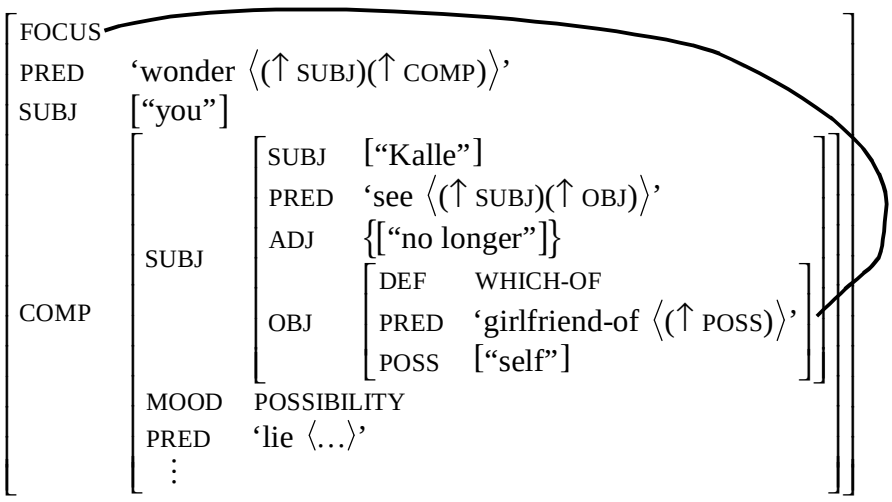
(5) [Vilken av sina$_i$ flickvänner]$_j$, undrade du om det att Kalle$_i$ inte längre
which of self's girlfriends wonder you if it that Kalle no longer
fick träffa henne$_j$ kunde ligga bakom hans dåliga humör.
sees her could lie behind his bad mood
'Which of his girlfriends, do you wonder if the fact that Kalle no longer sees her could lie behind his bad mood.'

This only makes sense under a long-distance dependency analysis of resumptive pronouns (6a) or, equivalently (6b).

(6) a.

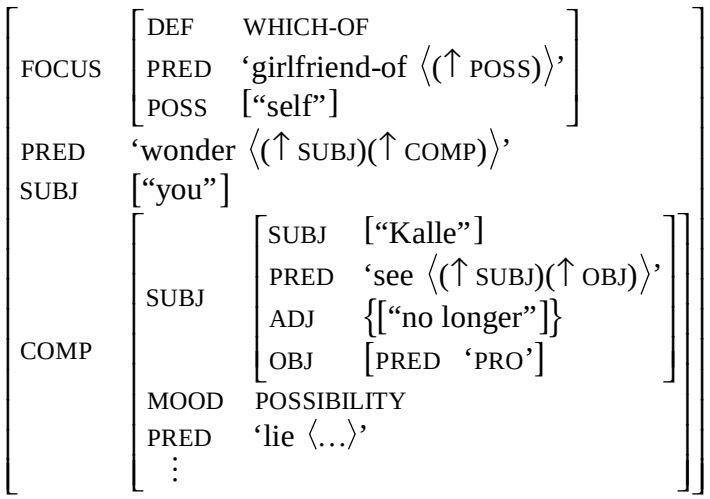
$$\begin{bmatrix} \text{FOCUS} & \begin{bmatrix} \text{DEF} & \text{WHICH-OF} \\ \text{PRED} & \text{'girlfriend-of } \langle(\uparrow \text{POSS})\rangle\text{'} \\ \text{POSS} & [\text{"self"}] \end{bmatrix} \\ \text{PRED} & \text{'wonder } \langle(\uparrow \text{SUBJ})(\uparrow \text{COMP})\rangle\text{'} \\ \text{SUBJ} & [\text{"you"}] \\ \text{COMP} & \begin{bmatrix} \text{SUBJ} & [\text{"Kalle"}] \\ \text{PRED} & \text{'see } \langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle\text{'} \\ \text{ADJ} & \{[\text{"no longer"}]\} \\ \text{OBJ} & \end{bmatrix} \\ & \begin{bmatrix} \text{MOOD} & \text{POSSIBILITY} \\ \text{PRED} & \text{'lie } \langle\ldots\rangle\text{'} \\ \vdots & \end{bmatrix} \end{bmatrix}$$

b.



Under an anaphoric analysis of the resumptive pronoun, the reflexive would not be bound in its minimal nucleus.

(7)



Another relevant property is the licensing of parasitic gaps (Sells 1984, Shlonsky 1992, Vaillette 2001). Resumptive pronouns do not license parasitic gaps in adjuncts, but do license them in subjects.

(8) a. *Elu hasfarim še Dan tiyek otam bli likro *pg*.
these the.books that Dan filed them without to.read
‘These are the books that Dan filed without reading.’

b. ?Zo- hi habaxura še haanašim še tearu *pg* lo hikiru ota hetev.
this- is the.girl that the.people that described not knew her well
‘This is the girl that the people who described didn’t know very well.’

A full understanding of this would require a theory of parasitic gaps in general, and the contrast between parasitic gaps in subjects and those in adjuncts in particular. However, on the standard assumption that parasitic gaps are licensed by long-distance dependencies, the ability of resumptive pronouns to license any kind of parasitic gap indicates that resumptive pronoun constructions are long-distance dependency constructions.

Another similarity, which we will not review here, is susceptibility to crossover effects

(Sells 1984, Shlonsky 1992, Vaillette 2001). Since crossover effects are based on the operator-gap relation (Bresnan 1995), this also argues for a long-distance dependency analysis.

Another piece of evidence that has been cited is the fact that across-the-board extraction is satisfied in structures where one conjunct has a gap and the other has a resumptive pronoun.

(9) ha- sefer še kaniti ve še divaxti al- av
the- book that I.bought and that I.reported on- it
'the book that I bought and reported on'

Under the analysis of the across-the-board phenomenon proposed by Falk (2000), this is not direct evidence for an LDD analysis. Under that analysis, the discourse function is distributed between the conjuncts, so there is a separate dependency in each conjunct.

Other properties of resumptive pronouns point to differences between them and gaps. For example, in most languages, including Hebrew, resumptive pronoun constructions are not subject to island constraints.

(10) a. **Coordinate Structure**
ha- sefer še karati oto / *∅ ve nirdamti
the- book that I.read it and fell.asleep
'the book that I read it and fell asleep'

b. **"Complex NP"**
ha- sefer še riayanti et ha- iša še katva oto / *∅
the- book that I.interviewed ACC the- woman that wrote it
'the book that I interviewed the woman who wrote (it)'

c. **Object of Preposition**
ha- sefer še šamati al- av / *∅
the- book that I.heard about it
'the book that I heard about'

It is this fact, combined with the approximate complementary distribution of gaps vs. resumptives that led Shlonsky (1992) to propose that resumptive pronouns are a last resort device, used to circumvent island constraints.

On the other hand, it is not universally true that no island constraints apply to resumptive pronouns. In Igbo, as reported by Goldsmith (1981), both gaps and resumptive pronouns obey what Goldsmith identifies as the Complex NP Constraint.

(11) a. *Nke-a bụ uno m maalu nwoke lulu (ya).
this is house I know man built (it)
'This is the house that I know the man who built it.'
b. *Nke-a bụ uno m maalu onye lulu (ya).
this is house I know who built (it)
'This is the house that I know who built it.'

Similarly, in Palauan (Georgopoulos 1990) extraction from an adjunct is ungrammatical, even with a resumptive pronoun.

(12) *ng- oingerang a mlarngii a betok el ’ad el mle
CLFT- when REAL.PST.be many LNK man COMP AUX
songerenger (se er ngii)
starving when P it
‘When were there many people who starved (then)?’

Sells (1984) reports that in Swedish resumptive pronouns are subject to most of the island constraints to which gaps are subject. However, this situation seems to be relatively unusual.

Another difference that argues against too close a relationship between gaps and resumptive pronouns has to do with special morphological marking on the long-distance dependency path. In some languages, as discussed by Zaenen (1983), there is special marking on either the verb or the complementizer of every clause between the filler and gap. Irish is one such language, and it is also a language with resumptive pronouns (McCloskey 1979). In the resumptive pronoun construction, the special marking is only on the main clause of the construction (the one with the operator), but not on lower clauses.

(13) a. an t-úrscéal aL mheas mé aL thuig mé ___
the novel COMP.WH thought I COMP.WH understood I
b. an t-úrscéal arL mheas mé gurL thuig mé é
the novel COMP.RESUMP thought I COMP understood I it
‘the novel that I thought I understood’

On the other hand, Vaillette (2001) points out that in Palauan resumptive constructions behave the same as gap constructions as regards marking of the path. It is interesting to note, though, that Palauan is one of the few languages in which resumptive constructions are subject to island constraints.

It is significant that this evidence that resumptive pronoun constructions differ from gap constructions relates to the path between the filler and gap or resumptive. In the LFG theory of long-distance dependencies (Kaplan and Zaenen 1989, Falk 2001), properties of the path relate not to the dependency itself but rather the nature of the licensing of the dependency. Islands are the result of an illicit grammatical function on or adjacent to the path, as defined by the language-specific functional uncertainty expression defining a well-formed extraction path. Special marking along the path, such as the complementizer *aL* in Irish, is analyzed in Dalrymple’s (2001) reworking of Zaenen’s original analysis as an off-path constraint in the functional uncertainty expression. Thus, this evidence does not contradict our earlier conclusion that resumptive pronoun constructions are long-distance dependencies; it simply requires the dependencies to be licensed differently from gap dependencies.

3. Pronouns

Lying at the heart of the phenomenon of resumptive pronouns is the question of why pronouns can be used as the lower end of a long-distance dependency. As observed above, this includes languages like English which do not have a grammatical phenomenon of resumptive pronouns but nevertheless seem to marginally allow pronouns in place of gaps (what Sells 1984 refers to as “intrusive pronouns”) in islands.

The fact that the use of pronouns in this construction is not accidental is emphasized by observations that have been made from time to time concerning the referential possibilities for resumptive pronouns. The essential observation is that the reference of the resumptive pronoun is what one would expect from an ordinary pronoun. For example, Sharvit (1999) discusses the inability of resumptive pronouns to be interpreted as being in the scope of a quantifier in the same clause, unlike gaps. Note the contrast between the gap, which is ambiguous, and the

resumptive pronoun, which only has the referential reading.[1]

(14) a. ha- iša še kol gever hizmin ___ hodeta lo.
the- woman that every man invited thanked him
(i) 'The [one] woman every man invited thanked him [=one particular man].'
(ii) 'For every man *x*, the woman that *x* invited thanked *x*.'

b. ha- iša še kol gever hizmin ota hodeta lo.
the- woman that every man invited her thanked him
'The [one] woman every man invited thanked him [=one particular man].'

Unlike the gap, the resumptive pronoun must be interpreted as referential. Another case that Sharvit discusses is the distinction between the *de re* (i) and *de dicto* (ii) readings in the following example.

(15) a. Dan lo yimca et haiša še hu mexapes ___.
Dan not will.find ACC the.woman that he looks.for
(i) 'Dan will not find the [specific, existing] woman he is looking for.'
(ii) 'Dan will not find the woman he is looking for [who may not exist].'

b. Dan lo yimca et haiša še hu mexapes ota.
Dan not will.find ACC the.woman that he looks.for her
'Dan will find the [specific, existing] woman he is looking for.'

The gap allows both the de dicto reading, in which the object of 'look-for' is not referential, and the de re reading, in which it is referential. The resumptive pronoun only allows the referential reading. That this is true of pronouns in general is shown by the following.

(16) a. Dan mexapes iša.
Dan looks.for woman
'Dan is looking for a woman.' (ambiguous)

b. Dan mexapes iša. Gam Ram mexapes ota.
Dan looks.for woman also Ram looks.for her
'Dan is looking for a [specific, existing] woman. Ram is also looking for her.'

This example clearly shows that the referential properties of resumptive pronouns are related to the referential properties of ordinary pronouns. Similar effects have been noted in other languages, with the same conclusion. Thus, in discussing resumptive pronouns in Spanish, Suñer (1998: 358) observes that "resumptive pronouns in restrictive relatives act like regular pronouns with respect to the antecedent."

Another point that emerges from the literature on resumptive pronouns is that the antecedent of the resumptive pronoun has some kind of discourse-related prominence, characterized by Erteschik-Shir (1992) as "restrictive focus" (identification as part of a set defined by the context), and by Sharvit (1999) as "D(iscourse)-linking." Erteschik-Shir contrasts the following two sentences.

[1]Sharvit also observes that the facts are different in specificational sentences, and explains this in terms of a theory of pronoun interpretation.

(17) a. Hine ha- simla še kaniti.
here.is the- dress that I.bought
b. Hine ha- simla še kaniti ota.
here.is the- dress that I.bought it
'Here is the dress that I bought.'

As described by Erteschik-Shir, (17b) would be used if the hearer knew not only that the speaker went to town to buy a dress, but also that she had a few specific dresses in mind. That is to say, there is a contextually defined set, and this example identifies a particular dress as a member of the set. Another piece of evidence is provided by Sharvit, who notes that while it is usually stated that in Hebrew resumptive pronouns are only used in relative clauses, and it is true that questions generally disallow them, some varieties of Hebrew allow them in 'which' questions.

(8) a. im mi nifgšta?
with who you.met
'Who did you meet with?'

b. *mi nifgašta ito?
who you.met with.him
'Who did you meet with?'

c. eyze student nifgašta ito? [grammatical for some speakers of Hebrew]
which student you.met with.him
'Which student did you meet with?'

This fits with Erteschik-Shir's description of the situation: in 'which' questions there is an assumed set, presumably defined by the context, and the purpose of the question is to choose a member of the set.

We will not discuss all the intricacies of pronoun interpretation, nor will we formalize our observations in terms of glue-language semantics. (For a glue-based account of pronouns, see Dalrymple 2001.) However, we will need some informal rudimentary assumptions. Following such work as Reinhart (1983) and Bresnan (2001), we distinguish between the referential use of pronouns and the bound-variable use. As argued by Reinhart, bound-variable pronouns are syntactically constrained while referential pronouns are not. Since syntactic constraints on binding are based on notions of rank at various syntactic levels, including the functional level, and the discourse functions are not part of the relational hierarchy of grammatical functions, we assume that a bound-variable account of the relation between the operator and the resumptive pronoun is not available. We also assume that, since the reference of referential pronouns is essentially governed by pragmatics, that "D-linking" can be included in a full account of the referential properties of resumptive pronouns.

The essence of (referential) pronouns is referentiality. A pronoun is an element which refers, but has no inherent reference of its own. Therefore, it must pick up its reference from something else in the discourse, usually something relatively prominent in the discourse. We take it to be uncontroversial in LFG that referentiality is represented at some non-syntactic level of representation. For concreteness, we will assume a ρ projection from f-structure, represented as a list of elements which have entered into the discourse. This referential structure should probably take the form of a DRT-like representation, but we will use a simplified representation

here.[2] It is possible that this referential structure corresponds to what Dalrymple (2001) calls the "context list."

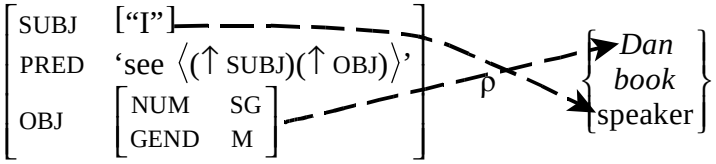
Given the ρ projection, the basic referentiality of a pronoun can be represented lexically as:

(18) $\uparrow_{\rho}$

This is a statement that the pronoun has a reference, without providing it with a reference. The pronoun is thus free to pick up a reference from the discourse. A pronoun will also typically have number and gender features specified lexically.

Consider the context in (19a) and the sentence in (19b). Assuming that there is no other context, the f-structure and its ρ projection will be (19c).

(19) a. Dan is reading a book.
b. I see him.
c.



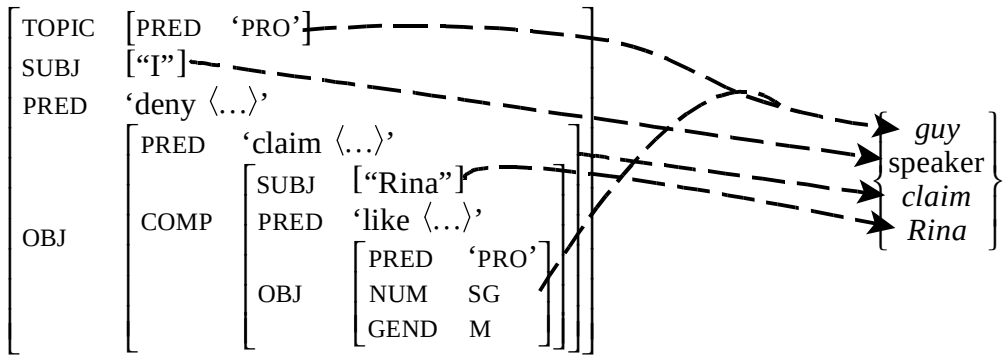
This results in the interpretation where *him* is coreferential with *Dan*. However, the f-structure is ill-formed: specifically, it is incoherent, since the OBJ lacks a PRED feature. The usual device to circumvent this problem is the dummy PRED value 'PRO'. But under the approach being taken here, [PRED 'PRO'] is not an essential property of pronouns, merely a formal f-structure device to allow pronouns to satisfy the Coherence Condition.

Next, consider the relative clause in (20).

(20) ??(the guy) that I denied the claim that Rina likes him

The f-structure and ρ projection are as follows:

(21)



This is an example of an "intrusive" pronoun: a resumptive pronoun in a language that doesn't

[2]Actually, much of this characterizes bound-variable pronouns as well. They too are characterized by being identified with something else. Within the framework of the projection architecture of LFG, it is possible that the coreference of referential pronouns is determined at what I am calling the ρ projection, while the antecedence of bound-variable pronouns is determined at the (semantic) σ projection.

have resumptive pronouns. Under the version of the Extended Coherence Condition we are assuming, this is ungrammatical. On the other hand, the existence of an anaphoric link makes this interpretable, even if it violates a technical requirement of the syntax. This seems to conform to the intuitive "feel" of a sentence like this. It is odd, but usable since there is no other way to say this.[3]

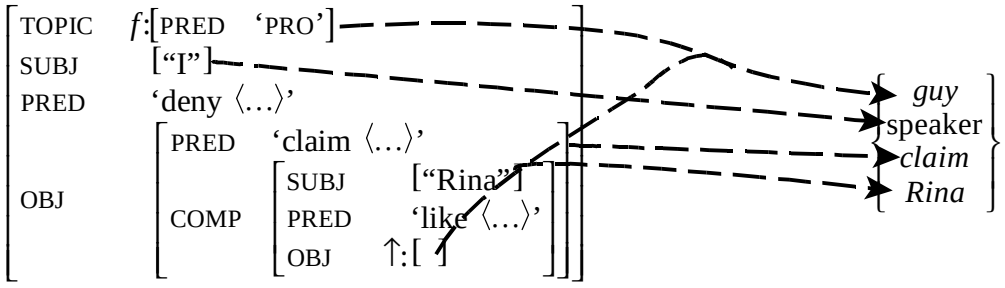
The difference between resumptive pronouns in English and resumptive pronouns in Hebrew is that in Hebrew there is an actual long-distance dependency. This can be accounted for by letting Hebrew pronouns have the following specification as an alternative to the [PRED 'PRO'] feature.

(22) $f \in \rho^{-1}(\uparrow_\rho) \wedge (\mathrm{DF}\ f) \Rightarrow \uparrow = f$

Like the ordinary [PRED 'PRO'] specification, (22) is a realization of the referentiality which we claim is the essential property of pronouns. The fact that the same pronouns are typically used for resumption as for ordinary pronominal uses is captured here by taking $\uparrow_\rho$ to be the core, with universal grammar allowing different realizations for it.[4]

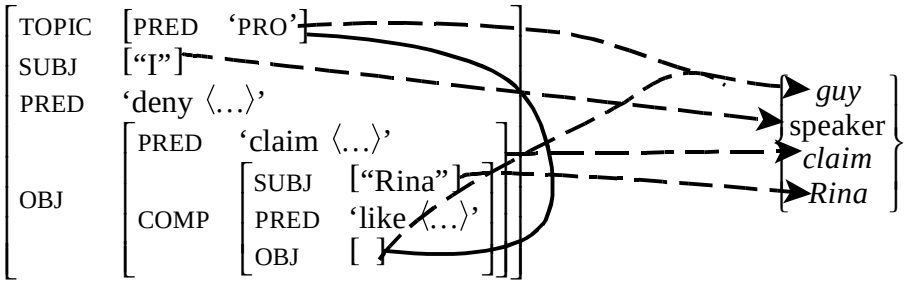
In the Hebrew equivalent of (20), the f-structure and ρ projection are the following if we ignore the specification in (22) and also do not assign the pronoun the [PRED 'PRO'] feature.

(23)



In this f-structure, the '↑' and '*f*' of (22) are labeled. (22) licenses establishing identity between these two f-structure elements, resulting in an ordinary long-distance dependency which is licensed not by a functional uncertainty equation but by (22).

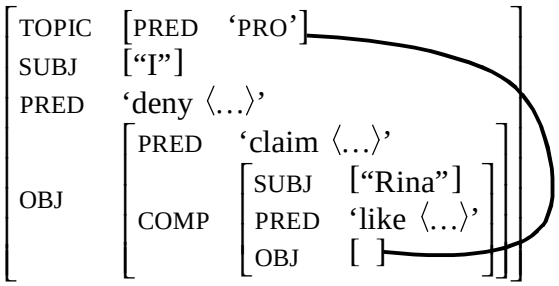
(24) a. f-structure + ρ projection



[3]My non-linguist wife refers to this as "talking yourself into a corner."

[4]On the other hand, nothing in this account requires all languages to use the same forms for resumption and for pronouns. It allows for a marked situation in which a language might have the lexical specification (22) on a form which is not otherwise a pronoun. It has been pointed out to me by George Aaron Broadwell (personal communication) that such a situation obtains in the Mayan language Kaqchikel. Kaqchikel has a resumptive pro-PP *wi*, which is optional, but appears to be more natural with a greater distance between filler and gap, a common situation for resumptive pronouns. However, *wi* is not used as an ordinary anaphor. So it appears to be a resumptive which is not a pronoun.

b. just f-structure



Note that there is no syntactic restriction on the path between the filler and the resumptive pronoun in Hebrew; in fact, the path is not even mentioned. This accords with the observation that filler-resumptive relations are not governed by island constraints. In languages in which they are so governed, an extra conjunct will be added to the premise of the conditional specifying the relation between the two f-structure elements in the form of a conventional inside-out functional uncertainty equation.

Some of the differences between real resumptive pronouns and “intrusive” pronouns may be related to the syntactic link. For example, Sells (1984) claims that in a relative clause embedded in a quantified nominal phrase, the quantifier can bind resumptive pronouns but not intrusive pronouns.

(25) every linguist that Mary couldn’t remember if she had seen ___/*him before

(26) a. Which of the linguists do you think that if Mary marries ___ then everyone will be happy. [___ could be a list]
b. Which of the linguists do you think that if Mary marries him then everyone will be happy. [*him* is a single linguist]

(27) kol gever še Dina xoševet še hu ohev et Rina
every man that Dina thinks that he loves ACC Rina
‘every man that Dina thinks loves Rina’

We will not pursue this here.

4. Distribution of Resumptive Pronouns and Gaps

We have not yet accounted for the relative distribution of resumptive pronouns and gaps. In Hebrew, subjects in the main clause of the relative clause must be gaps, objects and embedded subjects can be either gaps or resumptive pronouns, and oblique objects must be resumptive pronouns. This is nearly complementary distribution, but not entirely complementary. (Since an oblique object is unextractable in Hebrew, we hypothesize that OBL_θ is not allowed on the extraction path, and thus that obliques are islands in Hebrew.)

In LFG, it has been proposed that c-structure is constrained by the Economy of Expression principle, which disallows syntactic nodes which are not necessary for either f-structure well-formedness or semantic expressivity (Bresnan 2001). Interpreted strictly, Economy of Expression should allow resumptive pronouns in islands, because there is no other

way to license the same f-structure, but not in non-island contexts.[5] That is to say, it suggests complete complementarity.[6] This complementarity is approximately what we find, but not exactly. We note in passing that Economy of Expression can be invoked here only because we have analyzed resumptive pronouns as involving long-distance dependencies; if it were simply an anaphoric dependency the f-structures would be different and the two types of relatives would not be in competition with each other.

So the question is why we do not find absolute complementarity.[7] We propose that Economy of Expression is only part of the story. While Economy of Expression can account for certain interesting patterns (such as the distribution of relative pronouns and complementizers in English relative clauses, as discussed in Falk 2001), there are other constructions which blatantly violate Economy of Expression. One particularly striking case is the complementizer *that* (or, more precisely, the CP which it heads) in complements to verbs in English. The following sentences produce identical f-structures:

(28) a. I believe [$_{IP}$ the world is flat].
b. I believe [$_{CP}$ that [$_{IP}$ the world is flat]].

The question is why (28b) is grammatical, given Economy of Expression. Intuitively, the complementizer is useful for the hearer: it marks the beginning of the clause, thus making the sentence easier to parse. We propose that there is another principle (or perhaps family of principles) in competition with Economy of Expression. We will call it Sufficiency of Expression, and state it informally as follows.

(29) **Sufficiency of Expression**
Syntactic elements which provide clues to parsing are exceptions to Economy of Expression. Such elements include markers of clause boundaries and extraction sites.

This will allow resumptive pronouns where they compensate for parsing difficulty.

There are several reasons to think that the presence of resumptive pronouns in positions where they are not necessary is conditioned by parsing. For example, Erteschik-Shir (1992) notes that in many languages distance from filler improves the grammaticality of the resumptive pronoun, as in the following examples from English and Hebrew.

(30) a. This is the girl that John likes ___/*her.
b. This is the girl that Peter said that John likes ___/??her.

---

[5]An anonymous reviewer has remarked that my use of Economy of Expression differs from that of Bresnan (2001), in that Bresnan's version of the principle assumes fixed lexical choice. However, I do not believe that this is an accurate reading of Bresnan. On pp.147–8 she discusses pronominal clitics in Spanish, which have (at least in dialects that allow clitic doubling) lost their [PRED 'PRO'] feature and become merely agreement markers. She suggests, assuming the clitic is adjoined to the verb, that the higher V node is subject to Economy of Expression, and thus that *lo vio a Juan* (him s/he.saw ACC Juan) should be less economical than the version without the clitic: *vio a Juan*., and then goes on to suggest ways to circumvent this conclusion for clitics. It is clear that Bresnan views the two versions of the sentence as being in competition, even though the lexical choice is different (including the clitic *lo* in one case but not the other).

[6]This is true of principles that have been proposed in other theoretical frameworks as well, such as the "Avoid Pronoun" Principle in GB.

[7]This question is raised in other theories as well. Shlonsky (1992) is forced to hypothesize two separate complementizers *še* in relative clauses to account for the lack of complementarity.

c. This is the girl that Peter said that John thinks that Bob likes ___/?her.
d. This is the girl that Peter said that John thinks that yesterday his mother had given some cakes to ?___/her.

(31) a. ?Šošana hi ha- iša še nili ohevet ota.
Shoshana is the- woman that Nili loves her
'Shoshana is the woman that Nili loves.'

b. Šošana hi ha- iša še dani siper še moše rixel še nili ohevet ota.
Shoshana is the- woman that Dani said that Moshe gossiped that Nili loves her
'Shoshana is the woman that Dani said that Moshe gossiped that Nili loves her.'

Sells (1984) observes that in Swedish resumptive pronouns are used for multiple crossing dependencies, and also when there are two clauses between the filler and the extraction site. Both distance and multiple crossing dependencies introduce potential parsing complexity; it is plausible that the resumptive pronouns are used to overcome this complexity. Glinert (1989) explicitly notes that while resumptive pronouns are not usually used for objects, they are used in long, complex relative clauses.

The general resistance of resumptive pronouns to appear as SUBJ in the matrix of the long-distance dependency can also be explained by an appeal to ease of parsing. The SUBJ function[8] is an "overlay" or discourse-like function (Bresnan 2001, Falk 2000). and thus has a natural affinity to other discourse functions. The matrix SUBJ is thus the most natural extraction site, and therefore the easiest to parse. Sufficiency of Expression is inapplicable, and Economy of Expression rules out the resumptive pronoun.

There is an exception to the generalization that the matrix SUBJ of the long-distance dependency cannot be a resumptive pronoun in Hebrew. As observed by Borer (1984) and Shlonsky (1992), it can be a resumptive pronoun if there is a topicalized phrase.

(32) a. ha- iš še rak al kesef hu / ??___ xošev
the- man that only on money he thinks
'the man who only thinks about money'

b. ha- iš še al politika hu / ??___ lo ohev ledaber
the- man that on politics he NEG likes to.speak
'the man who doesn't like to talk about politics'

Following the argumentation of this section, this should be explained on the grounds of additional complexity due to the topicalization. It is plausible that, in a subject-initial language like English or Hebrew, pre-subject material in the clause would make parsing more difficult. The attribution of the resumptive pronoun to processing considerations is also supported by Shlonsky's observation that Hebrew speakers disagree on the acceptability of the version with no resumptive pronoun. There is independent evidence that such complexity is introduced by topicalization. Culicover (1993) observes that the *that*-trace effect is suspended in English if there is a topicalized adverbial intervening between the complementizer and the clause.

[8]In the framework of Falk (2000), this is the PIV(ot) function. In "syntactically ergative" languages with resumptive pronouns, it would be the OBJ in a transitive clause that can't be resumptive. This is confirmed by Mosel and Hovdhaugen (1992) for Samoan and Chung (1978) for Tongan.

(33) a. Robin met the man that/who Leslie said **that** [for all intents and purposes] was the mayor of the city.
b. Leslie is the person who I said **that** [under no circumstances] would run for president.

In these sentences, the bolded complementizer would be ungrammatical if the bracketed phrase were not topicalized. This kind of effect is unexpected under almost any theory of the *that*-trace effect.[9] If there is some condition which disallows the complementizer *that* from coexisting with subject extraction in the same clause, the presence of a topicalized phrase should be irrelevant. However, on the assumption that a topicalized phrase introduces additional computational complexity, the Sufficiency of Expression principle becomes relevant. By marking the beginning of the clause, the complementizer aids the language hearer in parsing the sentence.

Languages also may differ in exactly what constitutes parsing complexity. For example, Suñer (1998) states that while top-level SUBJ resumptive pronouns are marked in Spanish, they are not as dispreferred as in Hebrew:

(34) Conozco a un tipo que él me aconseja a mí.
I.know ACC a guy that he me.DAT advises to me
'I know a guy that (he) advises me.'

The contrast may have to do with the greater flexibility of subject expression in Spanish than in Hebrew.

An interesting case of resumptive pronouns where Economy and Sufficiency can explain an otherwise puzzling distribution is discussed (from a Minimalist perspective) in Aoun, Choueri, and Hornstein (2001). The language in question is Lebanese Arabic. Subject pronouns are independent words, while other pronouns are incorporated into the head of which they are arguments. Economy of Expression, which constrains syntactic nodes, is therefore relevant for subject pronouns but not for non-subject pronouns. Pronouns and epithets can serve as resumptive pronouns. Resumption is used fairly freely.

(35) ha- l- muttahame ʕrəfto ʔənno hiyye nħabasit.
this- the- suspect.F know.2PL that she imprisoned.3FSG
'This suspect, you know was imprisoned.'

However, if the fronted element is quantified, a full resumptive pronoun (or epithet) is possible only if the extraction path crosses an island boundary and an incorporated pronoun is possible even in a non-island context.

(36) a. *kəll muttahame ʕrəfto ʔənno ha- l- mažduube nħabasit.
each suspect.F know.2PL that this- the- idiot.F imprisoned.3FS
'Each suspect, you know that this idiot was imprisoned.'

b. kəll muttahame saʔalto ʔəza ha- l- mažduube nħabasit.
each suspect.F asked.2PL whether this- the- idiot.F imprisoned.3FS
'Each suspect, you asked whether this idiot was imprisoned.'

[9]For more on the *that*-trace effect in LFG, see Appendix A to this paper.

c. kəll məžrim fakkarto ʔənno l- bolisiyye laʔaṭu- u.
each criminal.M thought.2PL that the- police caught.3P- him
'Each criminal, you thought that the police had caught him.'

We do not expect pronouns to be able to resume quantified expressions, since they are not referential and do not add discourse referents to the context (Dalrymple 2001). This use of pronouns looks like the bound-variable interpretation, which is generally not available with discourse-function antecedents. We hypothesize that Lebanese Arabic exceptionally allows discourse-function quantifiers to bind pronouns, and that a long-distance dependency can be licensed by it. However, perhaps because it is a marked kind of resumptive pronoun, a bound-variable-type resumptive pronoun seems not to trigger Sufficiency. With this assumption, Economy and Sufficiency derive the correct distribution of forms.

(37) a. Extraction of nonquantified subject without crossing island
Gap: ✓
Resumptive: not ruled out by Economy because the pronoun is referential, so it satisfies Sufficiency

b. Extraction of quantified subject without crossing island
Gap: ✓
*Resumptive : has to be a bound-variable pronoun, not a referential pronoun, so Sufficiency is irrelevant. Economy is violated.

c. Extraction of nonquantified nonsubject without crossing island
Gap: ✓
Resumptive: incorporated pronoun, so not subject to Economy

d. Extraction of quantified nonsubject without crossing island
Gap: ✓
Resumptive: incorporated pronoun, so not subject to Economy

e. Extraction of nonquantified element across island
*Gap: not generable (because of island)
Resumptive: ✓

f. Extraction of quantified element across island
*Gap: not generable (because of island)
Resumptive: ✓

The resumptive pronoun facts concerning quantifiers and islands are thus derived.

What is not yet clear is the exact nature of Sufficiency of Expression, and its interaction with Economy of Expression. Unlike Economy of Expression, Sufficiency of Expression seems not to be an entirely competence-based principle; rather, it is tied to linguistic performance. It is not clear whether it is possible to define the relevant notion of computational complexity formally, and interspeaker variation suggests that it might not be. A stochastic Optimality Theory approach may be possible, but we will not pursue one here.

5. Conclusion

This paper has argued for an analysis of resumptive pronouns in LFG under which they participate in long-distance dependency constructions. These long-distance dependencies are not

licensed in the normal way by functional uncertainty equations, but rather by establishing a referential (anaphoric) identity between the two positions. This analysis is able to account for both the similarities and differences between gaps and resumptive pronouns. It also crucially depends on the parallel projection-based architecture of LFG, and the analysis of long-distance dependencies as a static identification of two functions rather than a derivational process of movement.

Appendix A. *That*-Trace Effect

One very common use of resumptive pronouns is to circumvent the "*that*-trace" effect, as in the following examples from Sells (1984) from Hebrew and Swedish, respectively.

(38) a. Eize xešbon kol maškia lo zoxer im hu noten ribit tova?
which account every investor NEG remembers if he gives interest good
'Which account does every investor not remember if it gives good interest?'

b. Det finns mycket som man önskar att det skulle vara annorlunda.
there is much that one wishes that it should be difficult
'There is much that one wishes should be difficult.'

This poses a problem for the analysis of the *that*-trace effect proposed by Falk (2000). This appendix offers a solution to the problem.

The analysis of the *that*-trace effect in Falk (2000) is based on the idea that complementizers which mark functionally more-independent subordinate clauses formalize this greater independence by disallowing their SUBJ (actually PIV(ot), but we will use SUBJ here for simplicity) from being identified with an element in a higher clause. Formally, complementizers like English *that* and Hebrew *im* have the following lexical specification:

(39) $(\uparrow \text{SUBJ}) \neq ((\text{GF}^{+} \uparrow) \text{GF})$

Consider the resumptive pronoun-less version of (38a).

(40) *Eize xešbon kol maškia lo zoxer im noten ribit tova?
which account every investor NEG remembers if gives interest good
'Which account does every investor not remember if it gives good interest?'

The f-structure is:

(41)
$$\begin{bmatrix} \text{FOCUS} & [\text{"which account"}] \\ \text{SUBJ} & [\text{"every investor"}] \\ \text{POL} & \text{NEG} \\ \text{PRED} & \text{'remember } \langle(\uparrow \text{SUBJ})(\uparrow \text{COMP})\rangle\text{'} \\ \text{COMP} & f\text{:} \begin{bmatrix} \text{TYPE} & \text{Q} \\ \text{SUBJ} & \\ \text{PRED} & \text{'give } \langle(\uparrow \text{SUBJ})\ (\uparrow \text{OBJ})\ (\uparrow \text{OBL}_{\text{Goal}}\ \text{OBJ})/\varnothing\rangle\text{'} \\ \text{OBJ} & [\text{"good interest"}] \end{bmatrix} \end{bmatrix}$$

The f-structure labeled *f* is headed by the complementizer *im*, and thus is associated with the constraint (39). Since (*f* SUBJ) = ((COMP *f*) FOCUS), the constraint is violated and the sentence is

ungrammatical. The problem is that the grammatical (38a) has the same f-structure.

The difference between the grammatical sentence with the resumptive pronoun and the ungrammatical sentence without is that in the grammatical sentence the SUBJ is partially represented in c-structure in *im*'s clause, whereas in the ungrammatical sentence it is completely outside of *im*'s clause. (Note that this only goes through under an analysis in which there is no c-structure trace in the position of the extracted subject. This is consistent with either a completely traceless analysis, as in Kaplan and Zaenen 1989 and Dalrymple, Kaplan, and King 2001, or a mixed analysis in which there is a trace for everything except subject extraction, as in Falk 2000, 2001.) If we consider c-structure, then, there is a way in which the (semi-)independence of the *im* clause's SUBJ is still the issue. The mistake in Falk (2000) was doing it entirely at f-structure.

Semi-formally, we want to replace (39) with something like the following:

(42) If $\uparrow$ is represented in c-structure (i.e. if $\phi^{-1}(\uparrow)$ exists)[10], one of the nodes in $\phi^{-1}(\uparrow)$ must immediately dominate one of the nodes in $\phi^{-1}(\uparrow \text{ SUBJ})$

More formally, we can define an f-structure-aware notion of immediate dominance, similar to such concepts as f-precedence. We will call this the f-ID relation.[11]

(43) For any f-structures f_1 and f_2, f_1 f-IDs f_2 ($f_1 \rightarrow_f f_2$) iff there exists a node n_1 in $\phi^{-1}(f_1)$ and a node n_2 in $\phi^{-1}(f_2)$ such that n_1 immediately dominates n_2.

We can now restate the lexical constraint on *that*-trace complementizers:

(44) $\phi^{-1}(\uparrow) \Rightarrow \uparrow \rightarrow_f (\uparrow \text{ SUBJ})$

We now have an account of the *that*-trace effect which retains the original insight of Falk (2000) and also explains the use of resumptive pronouns to circumvent the effect.

Appendix B. Pronoun Fronting in Hebrew

Although it is only marginally related to the question of resumptive pronouns, no discussion of Hebrew relativization would be complete without mentioning pronoun fronting. In addition to (45a), (45b,c,d) are also grammatical.

(45) a. ha- sefer še ani xošev še karata oto
the- book that I think that you.read it
b. hasefer še ani xošev še oto karata
c. hasefer še oto ani xošev še karata
d. hasefer oto ani xošev še karata
'the book that I think you read'

That is to say, the pronoun can be fronted, either partially or completely, and if it is fronted completely the complementizer can be omitted.

The description in the previous sentence has often been taken to be an accurate

[10]This natural condition prevents the effect from applying in the case of "empty operator" LDD constructions, such as *that* relatives in English.

[11]Thank you to Ron Kaplan (p.c.) for help with the formalization.

description of the situation. Borer (1984: 223) takes the fronting of the pronoun to "demonstrate clearly that a major strategy of relative clause formation in Hebrew involves movement of some sort." Glinert (1989) notes the variable positioning of the pronoun, and also notes that preposing the pronoun can substitute for having a complementizer.

However, as argued by Vaillette (2001), there are serious problems with an analysis which sees the movement of the pronoun as part of the process of relativization in Hebrew, or sees the complementizer-less version as just another minor variation. Vaillette observes that Hebrew allows free partial or complete topicalization independently of relative clauses. The simplest analysis, then, is to see the fronting of the resumptive pronoun as a case of topicalization. In fact, pronouns other than the resumptive pronoun can be fronted in relative clauses.

(46) ha- rofe še otam šalaxti elav
the- doctor that them I.sent to.him
'the doctor that I sent them to'

There is thus no reason to see fronting as part of relativization in Hebrew.

The form without the complementizer is different, though. In the first place, the complementizer *še* is generally obligatory, unlike the English *that*. This renders Borer's (1984) free deletion analysis of the absence of *še* somewhat dubious. Secondly, as observed by Vaillette, the fronting of the pronoun in the complementizerless version behaves differently from topicalization: the fronting must be all the way to the matrix of the relative clause, and other elements cannot front instead.

(47) *ha- rofe otam šalaxti elav
the- doctor them I.sent to.him
'the doctor that I sent them to'

This looks more like the fronting of a relative pronoun in a language like English. Finally, as observed by Sharvit (1999), the fronted pronoun in the complementizerless relative does not have the referential properties of resumptive pronouns.

(48) a. Ha- iša ota kol gever hizmin higia ito.
the- woman her every man invited arrived with.him
(i) 'The [one] woman every man invited arrived with him [=one particular man].'
(ii) 'For every man *x*, the woman that *x* invited arrived with *x*.'

b. Ha- iša še ota kol gever hizmin higia ito.
the- woman that her every man invited arrived with.him
'The [one] woman every man invited arrived with him [=one particular man].'

This observation of Sharvit's confirms Vaillette's analysis, under which the two fronted-pronoun variants are very different constructions: the one with a complementizer involving a resumptive pronoun that happens to be fronted, and the one without a complementizer involving a homophonous relative pronoun.

References

Aoun, Joseph, Lina Choueiri, and Norbert Hornstein (2001) "Resumption, Movement, and Derivational Economy." *Linguistic Inquiry* 32: 371–403.

Borer, Hagit (1984) “Restrictive Relatives in Modern Hebrew.” *Natural Language And Linguistic Theory* 2: 219–260.

Bresnan, Joan (1995) “Linear Order, Syntactic Rank, and Empty Categories: On Weak Crossover.” in Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, eds., *Formal Issues in Lexical-Functional Grammar*. Stanford, Calif.: CSLI Publications. 241–274

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

Chung, Sandra (1978) *Case Marking and Grammatical Relations in Polynesian.* Austin: University of Texas Press.

Culicover, Peter W. (1993) “Evidence Against ECP Accounts of the That-t Effect.” *Linguistic Inquiry* 24: 557–561.

Dalrymple, Mary (2001) *Syntax and Semantics 34: Lexical-Functional Grammar.* New York: Academic Press.

Dalrymple, Mary, Ron Kaplan, and Tracy Holloway King (2001) “Weak Crossover and the Absence of Traces.” in Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG01 Conference, University of Hong Kong* On-line: CSLI Publications. 66–82. **http://cslipublications.stanford.edu/LFG/6/lfg01.html**.

Erteschik-Shir, Nomi (1992) “Resumptive Pronouns in Islands.” in H. Goodluck and M. Rochemont, eds., *Island Constraints.* Dordrecht: Kluwer. 89–108.

Falk, Yehuda N. (2000) “Pivots and the Theory of Grammatical Functions.” in Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG00 Conference, University of California, Berkeley* On-line: CSLI Publications. 122–138. **http://cslipublications.stanford.edu/LFG/5/lfg00.html**.

Falk, Yehuda N. (2001) *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax.* Stanford, Calif.: CSLI Publications.

Georgopolous, Carol (1990) *Syntactic Variables: Resumptive Pronouns and A′ Binding In Palauan.* Dordrecht: Kluwer.

Glinert, Lewis (1989) *The Grammar of Modern Hebrew.* Cambridge: Cambridge University Press.

Goldsmith, John (1981) “The Structure of Wh-Questions in Igbo.” *Linguistic Analysis* 7: 367–393.

Kaplan, Ronald M., and Annie Zaenen (1989) “Long-Distance Dependencies, Constituent Structure, and Functional Uncertainty.” in Mark R. Baltin and Anthony S. Kroch, eds., *Alternative Conceptions of Phrase Structure.* Chicago: University of Chicago Press. 17–42.

McCloskey, James (1979) *Transformational Syntax and Model Theoretic Semantics: A Case Study in Modern Irish.* Dordrecht: D. Reidel.

Mosel, Ulrike, and Even Hovdhaugen (1992) *Samoan Reference Grammar.* Oslo: Scandinavian University Press.

Reinhart, Tanya (1983) *Anaphora and Semantic Interpretation.* London: Croon Helm.

Sells, Peter (1984) *Syntax and Semantics of Resumptive Pronouns.* Doctoral dissertation, University of Massachusetts, Amherst

Sharvit, Yael (1999) “Resumptive Pronouns in Relative Clauses.” *Natural Language and Linguistic Theory* 17: 587–612.

Shlonsky, Ur (1992) “Resumptive Prononous as a Last Resort.” *Linguistic Inquiry* 23: 443–468.

Suñer, Margarita (1998) Resumptive Restrictive Relatives: A Crosslinguistic Perspective.” *Language* 74: 335–364.

Vaillette, Nathan (2001) “Hebrew Relative Clauses in HPSG.” in Dan Flickinger and Andreas Kathol, eds., *Proceedings of the 7th International HPSG Conference, UC Berkeley (22–23 July 2000).* On-line: CSLI Publications. 305–324.

http://cslipublications.stanford.edu/HPSG/1/hpsg00.html.
Zaenen, Annie (1983) “On Syntactic Binding.” *Linguistic Inquiry* 14: 469–504.
Zaenen, Annie, Elisabet Engdahl, and Joan M. Maling (1981) “Resumptive Pronouns Can Be Syntactically Bound.” *Linguistic Inquiry* 12: 679–682.

A (DISCOURSE) FUNCTIONAL ANALYSIS OF ASYMMETRIC COORDINATION

Anette Frank

Language Technology Lab
German Research Center for Artificial Intelligence
DFKI GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken
frank@dfki.de

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

A (Discourse) Functional Analysis of Asymmetric Coordination

Abstract

A long-standing puzzle in the analysis of coordination is the so-called SGF coordination (Subject Gap in Finite/Fronted constructions) in German, first discussed by Höhle (1983a). The syntactic analysis of SGF constructions is challenging for any type of syntactic framework, as they seem to violate basic assumptions of accessibility or distribution in coordination constructions.

SGF constructions have been analysed in terms of asymmetrically embedded constituents (Wunderlich 1988; Höhle 1990; Heycock and Kroch 1993; Büring and Hartmann 1998) or symmetric conjuncts (Steedman 1990; Kathol 1995, 1999). Asymmetric embedding is problematic as it involves extraction asymmetries, or an analysis of coordination as adjunction. Symmetric analyses need to assume special licensing conditions that are not independently motivated. In particular, we argue that the symmetric analysis of Kathol (1999) is lacking independent syntactic motivation, and fails to account for related asymmetric coordinations of verb-last and verb-fronted (VL/VF) sentences.

We present a multi-factorial LFG analysis of asymmetric coordination, building on independently motivated principles of correspondence between c–structure, f–structure, and i(nformation)–structure. SGF coordination is analysed as symmetric coordination in c–structure. Binding of the (prima facie) inaccessible subject of the first conjunct is enabled, at the level of f–structure, by asymmetric projection of a "grammaticalised discourse function (GDF)", a TOPIC, FOCUS or SUBJ function (Bresnan 2001). Asymmetric GDF projection is motivated by relating the semantic and discourse-functional properties of asymmetric coordination to well-known discourse subordination effects of modal subordination (Frank 1997; Frank and Kamp 1997). In conjunction with word order constraints in the optimality model of Choi (2001), our analysis explains the mysterious word order constraints of asymmetric coordination, and some puzzling scoping properties.

1 Introduction

Coordination for efficient and economic linguistic realisation

Coordination is a perfect syntactic means to support efficient and economic linguistic realisation. The contrasts in (1) and (2) exemplify that redundancy in overt linguistic expression is successfully avoided by use of an appropriate coordination construction.[1]

(1) a. The hunter went into the forest and *the hunter* caught a rabbit.

b. The hunter went into the forest and caught a rabbit.

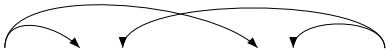
(2) a. Fred knows Rome and *Fred* loves *Rome*.

b. Fred knows and loves Rome.

Coordinations (1) and (2) are instances of standard constituent coordination – VP and V coordination, respectively. As illustrated in (3), the subcategorisation requirements of the coordinated heads are not fulfilled within the coordinated constituents proper. Instead, the *unique* arguments realised outside the coordinate structure need somehow to be *distributed* over the conjuncts, in order to satisfy the subcategorisation requirements of the individual coordinated heads.

[1] Note that the (a.) and (b.) examples are truth-conditionally equivalent only with coreferent interpretation of the redundant phrases.

(3) a. 
The hunter [[$_{VP}$ went into the forest] and [$_{VP}$ caught a rabbit]].

b. 
Fred [[$_{V}$ knows] and [$_{V}$ loves]] Rome.

Thus, redundancies that are avoided in coordination constructions lead – prima facie – to violations of basic syntactic principles, most prominently, agreement and subcategorisation requirements. Theories of formal syntactic frameworks provide specific mechanisms to apply in coordination constructions that account for their special "reductionist" properties, while excluding ungrammatical constructs. Phenomena of "regular" constituent coordination are in this sense well understood, and successfully handled by all major syntactic formalisms.[2]

The challenge of asymmetric coordination

In this paper, we are concerned with a special case of *asymmetric coordination*, the so-called SGF coordination (Subject Gap in Finite/Fronted constructions) in German, first discussed by Höhle (1983a).[3] This construction, illustrated in (4), is very frequent,[4] and not restricted to specific registers or style. The syntactic properties of SGF coordination represent a challenge for modern syntactic theories, as they seem to violate the basic assumptions of accessibility (or distribution) as established for cases of regular constituent coordination: the subject of the left conjunct is realised in a middle field position, and is thus – under standard analyses of constituent coordination – not accessible from within the second conjunct, which is missing a subject (hence "subject gap").

(4) a. In den Wald ging der Jäger und fing einen Hasen.
Into the forest went the hunter and caught a rabbit
'The hunter went into the forest and caught a rabbit'

b. Nimmt man den Deckel ab und rührt die Füllung um , steigen Dämpfe auf.
Takes one the lid off and stirrs the contents round , rise fumes
'If one takes the lid off and stirrs the contents, fumes will rise'

SGF constructions have been analysed in terms of asymmetrically embedded constituents (Wunderlich 1988; Höhle 1990; Heycock and Kroch 1993; Büring and Hartmann 1998) or symmetric conjuncts (Steedman 1990; Kathol 1995, 1999). Asymmetric embedding is problematic as it involves extraction asymmetries, or an analysis of coordination as adjunction. Symmetric analyses need to assume special licensing conditions that are not independently motivated. Especially the word order conditions of Kathol (1999) are lacking independent syntactic motivation, and fail to account for related asymmetric coordinations of verb-last and verb-fronted (VL/VF) sentences (5).

(5) Wenn Du in ein Kaufhaus kommst und (Du) hast kein Geld, kannst Du nichts kaufen.
if you into a shop come and you have no money, can you nothing buy
'If you enter a shop and (you) don't have any money, you can't buy anything'

[2]This does not hold for the wide variety of so-called non-constituent coordinations: gapping, left or right conjunction reduction, ellipsis, etc. See Kehler (2002) for a recent overview and account of gapping and VP-ellipsis.

[3]We will most of the time stick to "classical" examples from previous work in Höhle (1983a), Wunderlich (1988), Büring and Hartmann (1998), Kathol (1999), and avoid repeated glossing.

[4]In a corpus study based on the NEGRA corpus, we determined 13.8% SGF coordinations, compared to 20.7% subject-initial verb-fronted sentence coordinations in an evaluation corpus consisting of 406 sentences involving sentential coordination (see Frank 2001).

A multi-factorial LFG analysis of asymmetric coordinations

In this paper we develop a multi-factorial LFG analysis of asymmetric coordination constructions (4) and (5), building on independently motivated principles of correspondence between c–structure, f–structure, and i(nformation)–structure (cf. Choi 2001). SGF coordination is analysed as symmetric coordination in c–structure. Binding of the (prima facie) inaccessible subject of the first conjunct is enabled, at the level of f–structure, by asymmetric projection of a "grammaticalised discourse function (GDF)", a TOPIC, FOCUS or SUBJ function (Bresnan, 2001). Asymmetric GDF projection is motivated by relating the semantic and discourse-functional properties of asymmetric coordination to well-known discourse subordination effects of modal subordination (Frank 1997; Frank and Kamp 1997). In conjunction with word order constraints in the optimality model of Choi (1999, 2001), our analysis explains the mysterious word order constraints of asymmetric coordination, as well as some puzzling scoping properties.

Overview

The paper is structured as follows. In Section 2 we give a brief introduction to the analysis of constituent coordination in Lexical-Functional Grammar. Section 3 characterises the challenge of asymmetric coordination constructions within the LFG treatment of coordination. We give an overview of the characteristic syntactic (and semantic) properties of SGF and VL/VF coordinations, to be accounted for by any successful analysis of asymmetric coordinations. Section 4 briefly reviews previous approaches to SGF coordination in German, focusing on the symmetric analysis of Kathol (1999). In Section 5 we develop our own symmetric analysis of asymmetric coordination. Section 6 concludes.

2 Coordination in LFG

Multi-level syntactic representation

Lexical-Functional Grammar provides two main syntactic representation levels: c–structure and f–structure. C–structure is a tree representation that encodes constituency and word order, while f–structure is an attribute-value representation that encodes functional-syntactic properties, in particular grammatical functions and morpho-syntactic information.

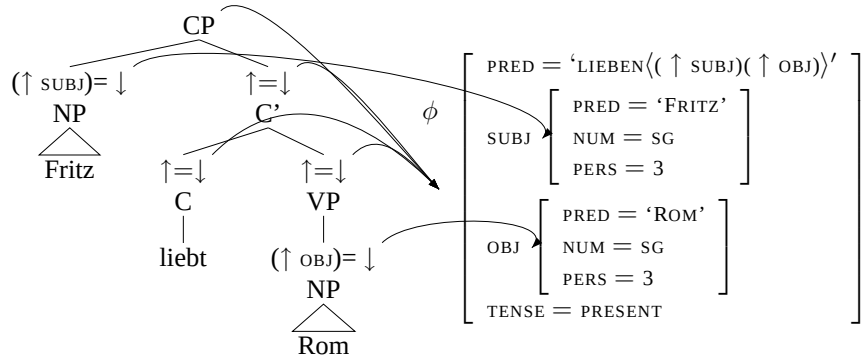
C– and f–structure are set into correspondence by functional annotation of c–structure nodes. These define the correspondences between c–structure nodes and their associated functional representation in the f–structure, in terms of a functional mapping, the so-called the ϕ–correspondence. The familiar abbreviations $\uparrow$ and $\downarrow$ are defined as in (6).

(6) $\phi(n) =_{def} \downarrow$ $\downarrow$ refers to the f–structure corresponding to the local c–structure node n.

$\phi(M(n)) =_{def} \uparrow$ $\uparrow$ refers to the f–structure corresponding to the mother $M(n)$ of the local c–structure node n.

Thus, in (7), the annotation ($\uparrow$ OBJ)= $\downarrow$ on the VP-internal NP node defines that the f–structure projected by the NP node – containing PRED = 'ROM' – plays the role of OBJ within the f-structure corresponding to the VP.

(7)



C–/f–structure correspondence for *Fritz liebt Rom. – Fritz loves Rome.*

Both representation levels are subject to principles of wellformedness: c–structure obeys principles of X-bar theory for lexical and functional categories. Grammatical functions in f–structure are classified as argument vs. non-argument functions. Argument functions need to be subcategorised by their local predicator (PRED) (Coherence Principle), and vice versa, all argument functions subcategorised by a predicator need to be realised (Completeness Principle). Finally, the Principle of Economy of Expression states that of all valid c–/f–structure representations only those are considered "optimal", and thus grammatical, that are maximally economic. In Bresnan (2001) Economy of Expression is measured in terms of the number of syntactic c–structure nodes.[5]

Coordination: Set-valued f–structures and distribution

In LFG, a special c–structure rule schema defines coordinated phrases of like constituents (8). In the associated f–structure, the coordinated phrase is represented as a set-valued f-structure. Each of the conjuncts is represented as an element within the set, by the functional annotations $\downarrow \in \uparrow$.

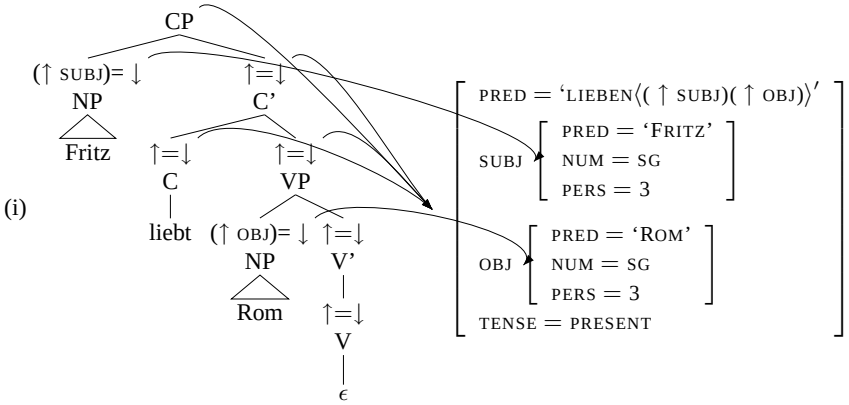
(8)

XP →	XP	Conj	XP
	$\downarrow \in \uparrow$	$\uparrow = \downarrow$	$\downarrow \in \uparrow$

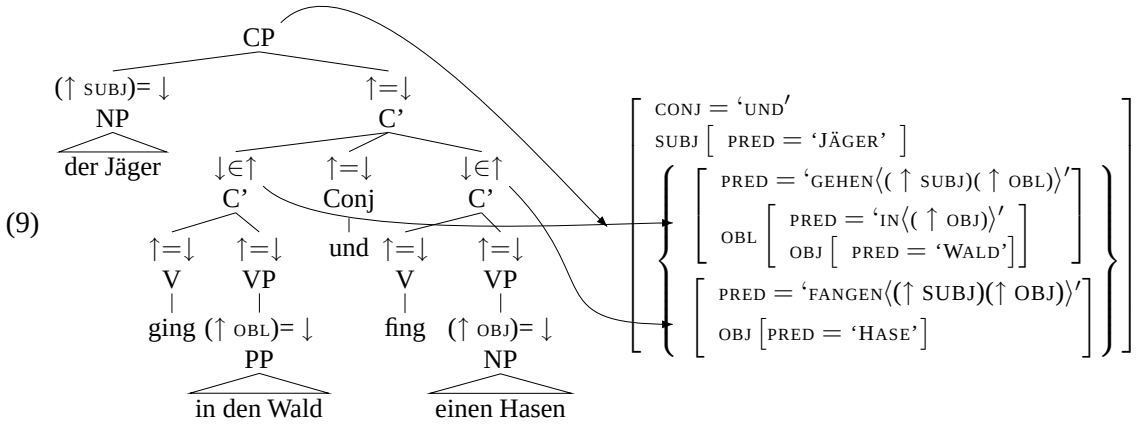
(9) displays the resulting c–/f–structure pair for a coordination of C' constituents with a shared SUBJ outside the coordinated phrase. Without further assumptions, the f–structure is incomplete regarding the elements of the set, which are both missing a SUBJ.

[5]The Principle of Economy of Expression qualifies an analysis of verb-second (V2) involving an empty verbal head as in (i) as unoptimal – as opposed to (7) above, which does not assume an empty verb head, while projecting an identical f–structure. See Section 5.2.2 for more detail.

(i)



(9)

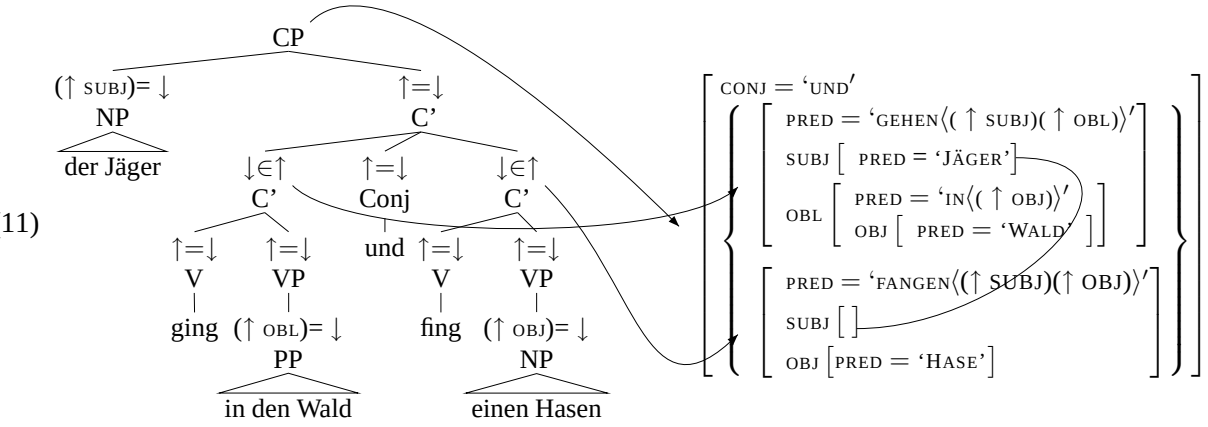


To account for shared arguments in coordinate structures (cf. (3)), the operation of *distribution* (10) is automatically applied to all features that are declared *distributive*. In particular, all grammatical functions are distributive features.

(10) **Distribution of features into set elements**
If a is a distributive feature and s is a set of f–structures, then $(s\ a) = v$ holds if and only if $(f\ a) = v$ for all f–structures f that are members of the set s. (Dalrymple 2001, p.158)

As a result, we obtain a wellformed f–structure in (11). The distributed SUBJ f–structure satisfies the completeness condition in both conjuncts.

(11)



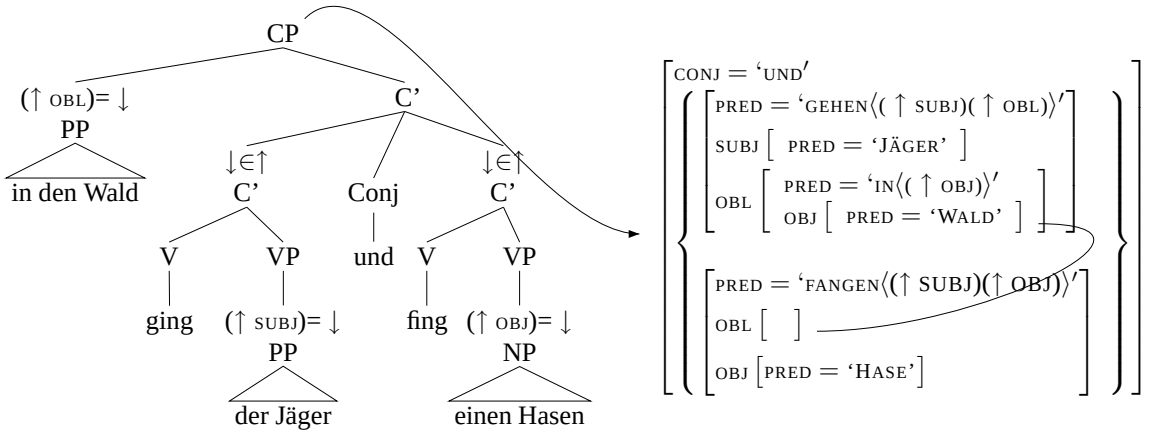
3 Asymmetric Coordination

3.1 Problems of Standard Coordination Analysis

Let us now consider the problem of SGF coordination in view of the standard coordination analysis.

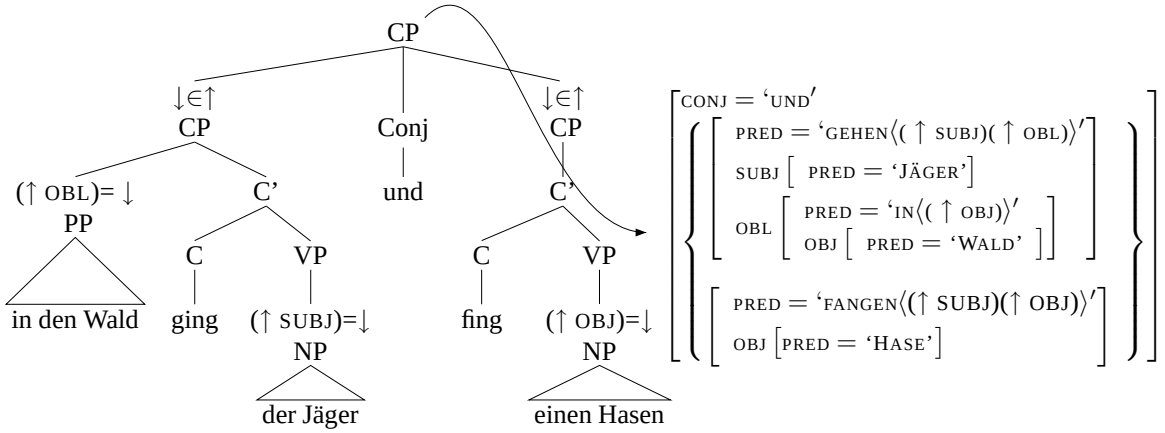
If we analyse (12) as a coordination of C' constituents, distribution applies to the topicalised OBLique PP *in den Wald*. While this yields a wellformed f–structure for the first conjunct, distribution into the second conjunct violates coherence: *fangen* (catch) does not subcategorise for an OBLique argument. Moreover, since the subject is realised within the first conjunct, it is not distributed to the second conjunct. That is, although we understand the second conjunct as a predication over the same subject as the first conjunct, it is missing a SUBJect function, violating Completeness.

(12)



If we analyse SGF coordination as involving symmetric CP coordination as in (13), we avoid illicit distribution of the topicalised phrase into the second conjunct, but still encounter the problem of a conjunct internal subject that cannot distributed – the notorious "subject gap" problem.

(13)



3.2 Syntactic Properties of Asymmetric Coordination

Having illustrated the problems we encounter when applying established principles of regular constituent coordination to SGF coordination constructions, we now review the major syntactic (and semantic) characteristics of SGF coordinations (see Kathol 1999, for concise overview). We can distinguish three types of basic syntactic (and semantic) properties that need to be accounted for by any successful analysis of SGF coordination.

Number and Type of Gaps Example (14) illustrates that SGF coordination does not license additional gaps in the right conjunct(s), besides the characteristic subject gap.

(14) * Einen Wagen$_j$ kaufte Hans$_i$ und meldete e$_i$ e$_j$ an.
A car$_j$ bought Hans$_i$ and registered e$_i$ e$_j$
'A car bought Hans and registered'

Only subjects can be "gapped" in asymmetric coordination constructions. Equivalent examples with a non-subject (here: object) gap are ungrammatical.

(15) * Gestern kaufte Hans den Wagen$_i$ und meldete sein Sohn e$_i$ an.
Yesterday bought Hans the car$_i$ and registered his son e$_i$
'Yesterday Hans bought the car and his son registered'

Word Order Properties SGF coordination shows a peculiar word order restriction, preventing the structural specifier position of CP in the right conjunct to be overtly realised: whereas (16.a) with a topicalised object in SpecCP is a perfectly grammatical sentence in German, the specifier position cannot be occupied in (16.b). Only the serialisation in (16.c) is acceptable.[6]

(16) a. Einen Hasen fing der Jäger.
A rabbit caught the hunter
'A rabbit, the hunter caught'

b. * In den Wald ging der Jäger und einen Hasen fing.
Into the forest went the hunter and a rabbit caught
'Into the forest went the hunter and a rabbit caught'

c. In den Wald ging der Jäger und fing einen Hasen.
Into the forest went the hunter and caught a rabbit
'Into the forest went the hunter and caught a rabbit'

Quantifier Scope As observed by Büring and Hartmann (1998) and Kathol (1999), the same interpretation is obtained for (17.a) and (17.b), irrespective of the position of the quantified subject: in both examples the quantified subject takes scope over both conjuncts: the interpretation is that for almost no one it is the case that he or she both buys a car and takes the bus. This is surprising for the SGF construction (17.b), as the quantified subject *die wenigsten Leute* occupies the middlefield position *within* the first conjunct, from where it does not structurally outscope the second conjunct.

(17.c), on the other hand, is problematic for analyses that assume an empty PRO subject in SGF constructions: (17.c) with an overt (repeated) quantified subject only allows for a narrow scope reading, where almost no one buys a car *and* almost no one takes the bus – that is, almost no one seems to need transportation.

(17) a. Die wenigsten Leute kaufen ein Auto und fahren mit dem Bus.
Almost no one buys a car and takes the bus
Almost no one buys a car and takes the bus.

b. Daher kaufen die wenigsten Leute ein Auto und fahren mit dem Bus.
Therefore buys almost no one a car and takes the bus
Therefore almost no one buys a car and takes the bus.

c. Daher kaufen die wenigsten Leute ein Auto und fahren die wenigsten Leute mit dem Bus.
Therefore buys almost no one a car and takes almost no one the bus
Therefore almost no one buys a car and almost no one takes the bus.

[6] Note that (16.b) is intended as a verb-fronted structure, not a verb-final construction. This is more evident in examples involving separable verb prefixes:

(i) * Gestern kaufte Fritz ein Auto und eine Ampel fuhr um.
Yesterday bought Fritz a car and a red light ran down
'Yesterday, Fritz bought a car and a red light ran down'

(ii) Gestern kaufte Fritz ein Auto und fuhr eine Ampel um.
Yesterday bought Fritz a car and ran a red light down
'Yesterday, Fritz bought a car and ran down a red light'

Asymmetric verb-last/verb-first (VL/VF) coordination constructions

Coordinations of verb-last/verb-first sentences were first brought to attention by Wunderlich (1988, p.312). Coordination of these clause types is only supported in the order order VL/VF (cf. (18.b)).

(18) a. [$_{\mathrm{CP}}$ Wenn Du in ein Kaufhaus kommst] und [$_{\mathrm{CP}}$ (Du) hast kein Geld],
if you into a shop come and you have no money,
'If you enter a shop and (you) don't have any money,

b. * [$_{\mathrm{CP}}$ In ein Kaufhaus kommst Du und [$_{\mathrm{CP}}$ wenn (Du) kein Geld hast],
into a shop come you and if you no money have,
'A shop you enter and if (you) don't have any money,

kannst Du nichts kaufen.
can you nothing buy
you can't buy anything'

Asymmetric VL/VF coordinations are closely related to SGF constructions: the subject of the right conjunct can be omitted, in which case we find a similar accessibility paradox, since the subject within the first conjunct cannot be distributed to the second conjunct (19.a). A gap in the second conjunct is only licensed for subjects (19.b), and as for SGF coordination, we cannot have multiple gaps (19.c). Finally, similar to SGF coordinations, the second conjunct's SpecCP position cannot be filled by a non-subject constituent (19.d).

(19) a. [$_{\mathrm{CP}}$ Wenn <u>Du</u> in ein Kaufhaus kommst] und [$_{\mathrm{CP}}$ (Du) hast kein Geld], ...
if you into a shop come and you have no money, ...
'If you enter a shop and (you) don't have any money, ...

b. * Wenn Du einen Kunden$_j$ hast und Du beleidigst e$_j$ / e$_j$ beleidigst Du , ...
if you a customer have and you offend / offend you,
'If you have a customer and you offend ,...'

c. * Wenn Du$_i$ ein Stück$_j$ übst und (Du$_i$) führst e$_j$ auf, ...
if you a play practice and (you) perform ,
'If you practice a play and (you) perform ,...'

d. * Wenn Du in ein Kaufhaus kommst und kein Geld hast (Du), ...
if you into a shop come and no money have (you),
'If you enter a shop and no money (you) have,...'

In Section 5, we will discuss special semantic and discourse-functional properties of asymmetric coordinations of both types: SGF and VL/VF coordination. This will lead us to a unified account of the functional and word order properties observed for both types of asymmetric coordination.

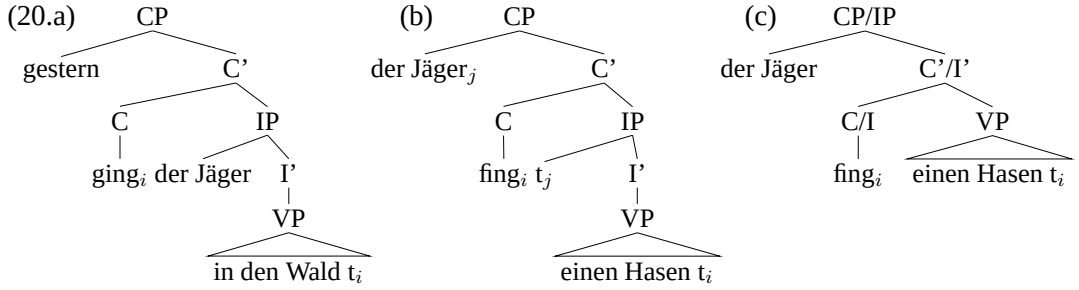
4 Previous Approaches

Before developing our own analysis of asymmetric coordination, we briefly review the two types of approaches that have been explored in previous work: analysis by asymmetrically embedded constituents, and coordination of symmetric conjuncts.

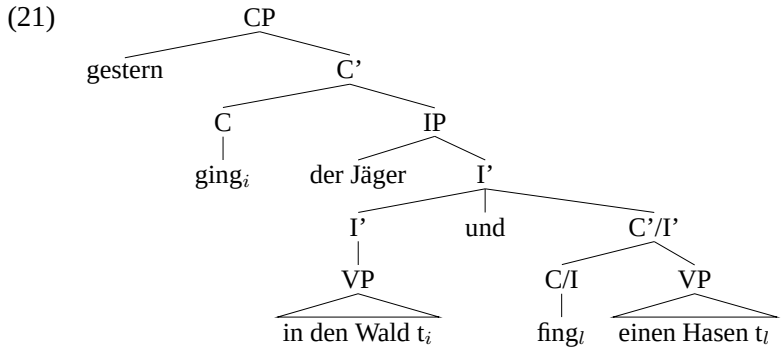
4.1 Asymmetric Analyses

Heycock and Kroch (1993) proposed an analysis in the P&P model that is similar – at a conceptual level – to the early analyses of Wunderlich (1988) and Höhle (1990). It will be discussed here as representative of the class of analyses that admit coordination of unlike constituents to account for the observed asymmetry of SGF coordinations.

The analysis builds on independent assumptions about the phrase structure of verb second (V2) languages like German. V2 is analysed as I-to-C movement. The specifier of CP can be filled by a non-subject phrase, as in (20.a). In subject initial V2 sentences, however, the subject must move from the SpecIP position to SpecCP, leaving behind an empty I projection (20.b). Similar to Haider (1988), the empty I projection and the structurally isomorphic C projection are "folded" into a *matching projection* of a complex category C/I in (20.c).



Heycock and Kroch's analysis of SGF coordinations naturally emerges from this *matching projection* analysis of subject-initial V2 sentences: An SGF coordination – omitting redundant subject phrases *der Jäger* – can be constructed from (20.a) and (20.c) by coordination of I' and C'/I' constituents, which are unlike, but share the categorial features of I. The resulting SGF coordination structure is displayed in (21).



Due to low coordination at the level of I', the shared subject governs both conjuncts, accounting for the main syntactic properties of SGF constructions: the restriction to subject gaps and wide scope of quantified subjects. However, the analysis necessarily involves extraction asymmetries that are otherwise ungrammatical.

It is well-known that extraction from coordinated phrases is only possible "across-the-board". The structure assigned in (21), by contrast, involves head movement out of the first conjunct only. The analysis of (22), with a topicalised argument, clearly violates the ATB extraction constraint and results in a fully evacuated first conjunct. Finally, the analysis needs to explain why a topicalised adjunct does not necessarily take scope over the second conjunct (as discussed by Höhle).

(22) In den Wald$_j$ ging$_i$ der Jäger [[e$_i$ e$_j$] und [fing einen Hasen]].

Büring and Hartmann (1998) present an asymmetric analysis of SGF coordination that avoids extraction asymmetries by considering it as an instance of adjunction, rather than coordination. Their analysis accounts for new data on scope, but nevertheless suffers from two problems:[7] First, as opposed to classical adjunction constructions, SGF coordination does not admit topicalisation of the adjoined material (23) (cf. Kathol 1999, p.309).

(23) a. [Ohne sie anzuschauen]$_i$ hat Fritz Maria geküsst e$_i$.
Without her to.look.at has Fritz Maria kissed
'Fritz kissed Maria without looking at her'

b. [(Und) fing einen Hasen]$_i$ ging der Jäger in den Wald e$_i$.

More importantly, while Büring and Hartmann motivate their analysis by special binding and scoping phenomena to be found in SGF constructions, they must concede that the same type of data can be found in uncontroversial VP coordination structures. Our conclusion is therefore that instead of reanalysing classical VP coordinations as adjunction, we need to account for such special scoping and binding asymmetries in a different way.

4.2 Symmetric Analyses

Symmetric analyses of SGF coordination have been proposed by Steedman and Kathol.

Steedman (1990) accounts for SGF coordination within his general theory of gapping. He proposes special functional application rules for coordination in the CCG framework that operate in gapping and SGF coordination constructions alike. While the analysis is very general, it fails to explain important restrictions of the SGF construction, in particular the restriction to apply to a unique grammatical function, the subject.

Kathol (1995, 1999) developed a linearization-based model of German syntax that is extended to account for SGF coordination. In Kathol (1999) special licensing conditions are defined to account for word order constraints of SGF coordination. We discuss the analysis and the problems it encounters in more detail below.

Kathol: Symmetric constituents and asymmetric linearisation

Kathol starts from the observation that the two coordinations in (24) are merely linearisation variants of a unique underlying predicate coordination structure, with a shared subject.

(24) a. *Der Jäger* {*ging in den Wald*} *und* {*fing einen Hasen*}.

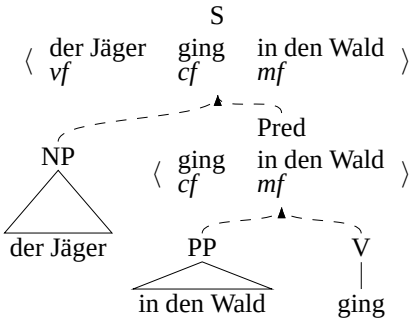
b. {*In den Wald ging*} *der Jäger und* {*fing einen Hasen*}.

This intuition, however, cannot be formalised in a phrase structure tree, which encodes constituency *and* word order at the same time. He therefore develops a "linearisation-based model of syntax" that provides a modular representation of constituency and (variable) linearisation.

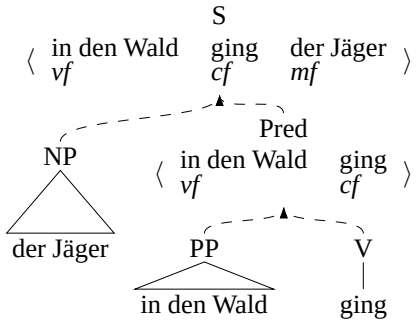
An illustration is given in (25.a,b), where the same constituent tree (represented by dotted arcs) is associated with different word orders (displayed in square brackets). Restrictions on possible linearisations are defined in terms of topological constraints (26) that need to be met by the assignment of topological labels *vf, cf, mf, vc, (nf)* in a sentential clause.[8] Both linearisations in (25) satisfy the topological linearisation constraints (26) that require, in particular, *vf* to precede *cf*, and *cf* to precede *mf*.

[7] The analysis of Büring and Hartmann (1998) is extremely interesting and thoroughly worked out, in particular for the new data on scoping facts it accommodates, but cannot be discussed here in detail, for reasons of space. We have to reserve detailed discussion to a later occasion.

(25) (a)



(b)



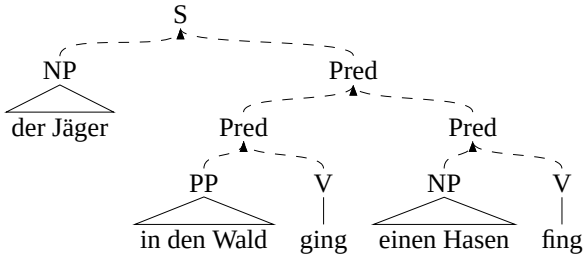
(26) $vf < cf < mf < vc$

Topological linearisation in coordination constructions

To account for coordination, the linearisation model needs to accommodate for the distributional behaviour of "shared" material outside the coordinated phrases.

Kathol introduces the notion of a *combinatorial factor* for phrases that are shared among coordinated phrases. Moreover, a combinatorial factor needs to be "linearised" to the second conjunct's tier. If this happens, it is called a *linear factor*. An example is given in (27). Here, the additional tabular representation represents the linearisation of the combinatorial factor (*der Jäger*) (in bold face) to the second conjunct's tier (indicated by brackets and underlining of the linearised phrase). Linearisation of the combinatorial factor preserves its topological label (here *vf*). The coordination structure is wellformed iff the Topological Construal Condition (28) is satisfied.

(27)

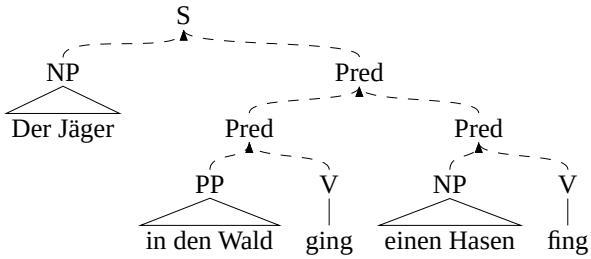


der Jäger	ging	in den Wald	(und)		
vf	*cf*	*mf*			
[der Jäger]			fing	einen Hasen	
vf			*cf*	*mf*	

(28) **Topological Construal Condition** (Kathol 1999, p.329)
A coordinated construction is well-formed if the linear factor's topological assignment yields a valid topological sequence on each conjunct tier.

However, if this model is applied to an SGF construction (here an interrogative V1 variant), linearisation of the combinatorial factor yields an *invalid* topological sequence (29).

(29)



Ging	**der Jäger**	in den Wald	(und)		
cf	*mf*	*mf*			
	[der Jäger]		fing	einen Hasen	
	* *mf*		*cf*	*mf*	

[8]The underlying topological field model of German syntax goes back to early descriptive grammarian work, and was introduced in formal syntactic theory by Höhle (1983b). The model gives a topological characterisation of German clausal syntax: Argument and adjunct phrases can occur in three phrasal fields: Vorfeld (*vf*), Mittelfeld (*mf*) or Nachfeld (*nf*). They are delimited by the complementizer field *cf* and the verbal complex *vc*, where *cf* can only host complementizers or the finite verb, while *vc* admits verbal and particle elements.

To account for the special type of *asymmetric* (SGF) coordination structures, Kathol introduces a Subject Functor Linearisation condition (clause A), which is later extended to clause B.

(30) **Subject Functor Linearization** (Kathol 1999, p.332,334)

A. The subject of a verb-initial conjoined predicate counts as a linear factor only if it occurs in the *Vorfeld*.

B. In the absence of any other linear factor, a constituent occurring in the *Vorfeld* counts as a linear factor (regardless of its status as combinatorial factor).

Clause A restricts linearisation of a *subject* combinatorial factor *in verb-initial coordination structures* to those subjects that occur in the vorfeld position. Since in verb-initial structures the subject is either in a vorfeld or a middle field position, clause A excludes linearisation of the combinatorial subject *exactly* in those – exceptional – cases that characterise the SGF coordination construction: if the subject is contained in the middle field of a verb-fronted coordination structure, but interpreted as the subject of both conjuncts.

Condition A adjusts the analysis of SGF coordination from (30) to (31): none of the combinatorial SGF subjects is linearised to the second tier. While this yields the correct results for (a) and (b) ($cf < mf$ is a valid topological sequence), it also admits the ungrammatical serialisation (c).

(31) a.

Ging	**der Jäger**	in den Wald	(und)		
cf	*mf*	*mf*			
			fing	einen Hasen	
			cf	*mf*	

b.

In den Wald	ging	**der Jäger**	(und)	
vf	*cf*	*mf*		
			fing	einen Hasen
			cf	*mf*

c.

*In den Wald	ging	**der Jäger**	(und)	
vf	*cf*	*mf*		
			einen Hasen	fing
			vf	*cf*

Here then, clause B comes into being, positing that "In the absence of any other linear factor, [any] constituent occurring in the *Vorfeld* counts as a linear factor", i.e. disregarding its status as combinatorial factor. This further amendment does, in the end, account for the facts (32), but at a high price: linearisation of phrases to the second tier could be motivated for *combinatorial factors*, but is lacking any justification for phrases that are not shared with the second conjunct. As a consequence, clause B weakens the otherwise crucial notion of a *combinatorial factor*.

(32) b.

<u>In den Wald</u>	ging	**der Jäger**	(und)	
vf	*cf*	*mf*		
[In den Wald]			fing	einen Hasen
vf			*cf*	*mf*

c.

*<u>In den Wald</u>	ging	**der Jäger**	(und)	
vf	*cf*	*mf*		
[In den Wald]			einen Hasen	fing
**vf*			*vf*	*cf*

In sum, the *Subject Functor Linearisation* conditions – designed to account for the special properties of SGF coordination – are introduced without motivation or supporting evidence. They are far from being motivated by independent grammatical notions or observations, and considerably weaken the notion of a *combinatorial factor*.

5 A Multi-Factorial LFG Analysis of Asymmetric Coordination

In the remainder of this paper we develop a multi-factorial LFG analysis of asymmetric coordination. It builds on well-established grammatical principles of the LFG theory, in particular principles of correspondence between c–structure, f–structure, and i–structure, and the notion of *grammaticalised discourse functions* (GDF).

Our analysis of asymmetric coordination introduces a new concept – *asymmetic* GDF *projection* – that is motivated by relating the semantic and discourse-functional properties of asymmetric coordination to the well-known discourse subordination effects of modal subordination. In conjunction with word order constraints in the optimality model of Choi (2001), our analysis explains the mysterious word order restrictions of asymmetric coordination.

5.1 Symmetric Analysis with Asymmetric GDF-Projection

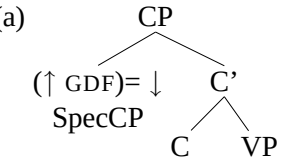
Grammatical functions can be classified according to properties of various dimensions, e.g., argument vs. non-argument functions, discourse functions vs. non-discourse functions (cf. Bresnan 2001, p.97f). (Bresnan 2001, p.98) further introduces the notion of a *grammaticalised discourse function* (GDF), covering the functions TOPIC, FOCUS, and SUBJ (33): "These functions are the most salient in discourse and often have c–structure properties that iconically express this prominence, such as preceding or c-commanding other constituents in the clause."

(33) *Grammaticalised Discourse Functions*

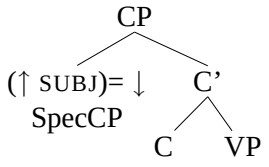
$$\overbrace{\underbrace{\text{TOPIC}\ \ \text{FOCUS}}_{\textit{Discourse Functions}}\ \ \underbrace{\text{SUBJ}\ \ \text{OBJ}\ \ \text{OBJ}_{\Theta}\ \ \text{OBL}_{\Theta}\ \ \text{COMP}}_{\textit{Argument Functions}}}\ \ \text{ADJ}$$

Within a verb second language like German, we can characterise the GDF functions as the class of functions that occupy the specifier position of CP. From the abstract functional annotation principle in (34.a) we can derive alternative GDF instantiations in (34.b) and (34.c).[9] This language-specific characterisation of GDF functions corresponds to Bresnan's general characterisation: functions that occupy the specifier position of CP qualify as most salient in discourse (cf. Choi 2001), and are obviously c–structurally prominent, in terms of both precedence and c–command.

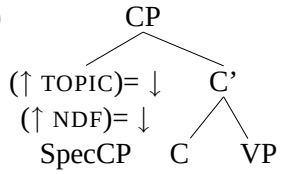
(34) (a)



(b)



(c)



Our formal analysis of asymmetric coordination can be summarised in the following (extended) definition of the CP coordination rule: (35) defines symmetric CP coordination in c–structure, with symmetric projection of the conjunct's f–structures in terms of the classical $\downarrow\in\uparrow$ annotations. As an extension to this classical *symmetric* coordination analysis we allow – at the level of f–structure – for *optional, asymmetric projection of a* GDF *function* of the left conjunct to the level of the coordination. As we shall see, this exten-

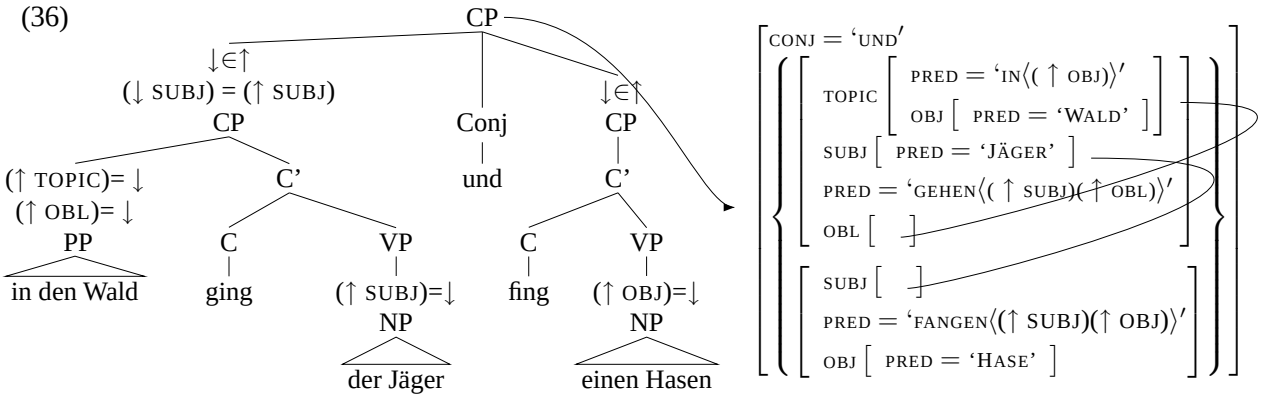
[9] Projection of a discourse function typically involves additional projection of a non-discourse function NDF (34.c).

sion accounts for the major syntactic properties of SGF coordination.

(35) CP ⟶ CP Conj CP
↓∈↑ ↑=↓ ↓∈↑
((↓ GDF) = (↑ GDF))

An example analysis is given in (36). Here, GDF is chosen to instantiate to SUBJ. The annotation (↓ SUBJ) = (↑ SUBJ) defines the first conjunct's SUBJ (*Jäger*) as the SUBJ of the coordination as a whole, i.e. the set-valued f–structure. Due to the distributional character of grammatical functions, the SUBJ defined for the set is distributed to *all* elements of the set. While it is already defined for the left conjunct, it is now introduced for the right conjunct, filling the notorious subject gap.

(36)



5.2 Syntactic Properties Revisited

We can now investigate the predictions of the analysis, reconsidering the syntactic and semantic properties of SGF coordinations discussed in Section 3.2.

5.2.1 Number and Type of Gaps

We had seen, in Section 3.2, that asymmetric SGF coordination is restricted to a *single* gap, and to *subject* gaps only. The examples are reproduced in (37) and (38), respectively. How does our analysis by asymmetric GDF-projection account for these restrictions?

(37) * Einen Wagen$_j$ kaufte Hans$_i$ und meldete e$_i$ e$_j$ an.
A car$_j$ bought Hans$_i$ and registered e$_i$ e$_j$
'A car bought Hans and registered'

(38) * Gestern kaufte Hans den Wagen$_i$ und meldete sein Sohn e$_i$ an.
Yesterday bought Hans the car$_i$ and registered his son e$_i$
'Yesterday Hans bought the car and his son registered'

We need to consider two cases: Instantiation of GDF to (i) SUBJ, or (ii) a discourse function DF.

(i) Instantiation of GDF **to** SUBJ**:** In (37) asymmetric projection of SUBJ enables distribution of the first conjunct's SUBJ (*Hans*) to the second conjunct, satisfying the completeness constraint of *anmelden* regarding its SUBJ. However, the obligatory OBJ function is not locally defined, and *cannot* be satisfied by alternative means: asymmetric GDF projection in (35) can only be instantiated once, and has been chosen to project the SUBJ. The sentence is ungrammatical due to the missing object.

As for (38), the ungrammaticality of non-subject gaps is explained as follows: Since the subjects of the

first and second conjunct are distinct, asymmetric projection of SUBJ (by instantiation of GDF to SUBJ) leads to an inconsistency in f–structure regarding the definition of the SUBJ in the second conjunct. Moreover, due to SUBJ projection, the object gap cannot, at the same time, be asymmetrically projected from the first to the second conjunct.

(ii) Instantiation of GDF **to** TOPIC/FOCUS**:**[10] Our account of (37) and (38) is of course only valid if we can prove that the examples are equally ruled out in the alternative case, instantiation of GDF to a discourse function, e.g. TOPIC.

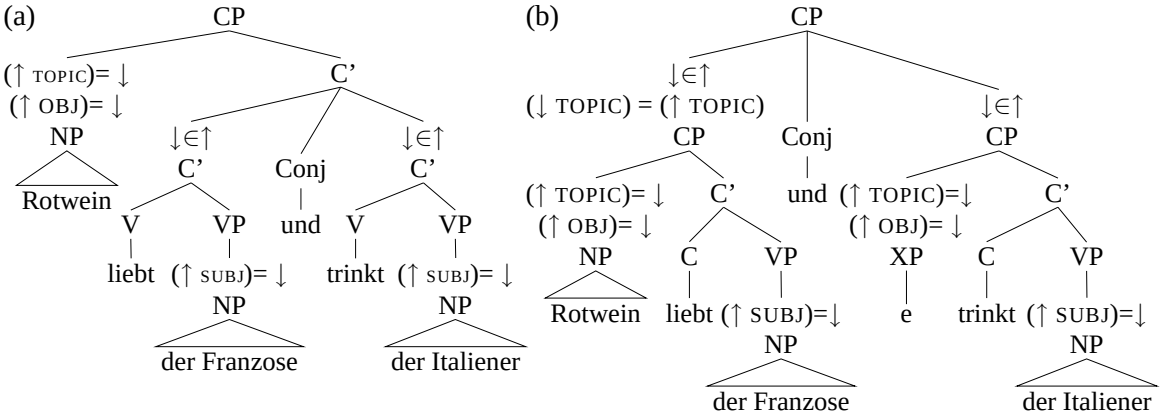
In (37) the TOPIC (*Wagen*) is identical to the first conjunct's OBJ (cf. (34.c)). Asymmetric projection of TOPIC leads to the distribution of the TOPIC to the second conjunct. The SUBJ function of the second conjunct, by contrast, remains unfilled; the structure is ruled out as ungrammatical.

In a similar way, (38) with a non-subject gap is ruled out if GDF is set to TOPIC. The structural TOPIC position is occupied by a non-OBJect function, here an adjunct. Its projection to the second conjunct does no harm, but leaves the crucial object gap unfilled, leading to ungrammaticality.

5.2.2 Principle of Economy of Expression

Asymmetric GDF projection as defined in (35) predicts the basic functional properties of SGF coordination. However, besides the cases discussed above, it predicts an asymmetric analysis of data such as (39), which are – however – cases of classical, symmetric ATB-extraction.

(39) Rotwein liebt der Franzose und trinkt auch der Italiener.
Red wine loves the Frenchman and drinks also the Italian.



(c)
$$\left[\left\{\left[\begin{array}{l}\textsc{topic}\left[\ \textsc{pred} = \text{`}\textsc{Rotwein}\text{'}\ \right]\\ \textsc{obj}\left[\ \right]\\ \textsc{pred} = \text{`}\textsc{lieben}\langle(\uparrow \textsc{subj})(\uparrow \textsc{obj})\rangle\text{'}\\ \textsc{subj}\left[\ \textsc{pred} = \text{`}\textsc{Franzose}\text{'}\ \right]\end{array}\right]\right.\right.$$
$$\left.\left.\left[\begin{array}{l}\textsc{topic}\left[\ \right]\\ \textsc{obj}\left[\ \right]\\ \textsc{pred} = \text{`}\textsc{trinken}\langle(\uparrow \textsc{subj})(\uparrow \textsc{obj})\rangle\text{'}\\ \textsc{subj}\left[\ \textsc{pred} = \text{`}\textsc{Italiener}\text{'}\ \right]\end{array}\right]\right\}\right]$$

The classical analysis of ATB extraction examples like (39) is given in (39.a). The topicalised OBJ is realised outside the C' coordination. The (coreferent) OBJ and TOPIC functions are distributed to both conjuncts, as displayed in (39.c).

[10] We restrict our discussion to the TOPIC function, the case of FOCUS being equivalent.

However, the same f-structure is now obtained by an alternative analysis, in terms of asymmetric GDF projection, as displayed in (39.b). In c–structure, the (shared) topic is now realised *within* the first CP conjunct. With GDF instantiated to TOPIC, the TOPIC is asymmetrically projected to the second conjunct. In addition, an empty SpecCP position is required within the second conjunct, to equate TOPIC and OBJ functions. The analysis projects the very same f–structure that we obtain for the regular ATB extraction analysis, namely (39.c).[11]

This unwarranted spurious ambiguity is, however, ruled out on the basis of the *Principle of Economy of Expression*. This principle basically requires the choice of the smallest c–structure that allows for the satisfaction of f–structure constraints and the expression of the intended meaning (cf. Dalrymple 2001, p.85).

(40) **Economy of expression** (Bresnan 2001, p.91)
All syntactic phrase structure nodes are optional and are not used unless required by independent principles (completeness, choherence, semantic expressivity).

The alternative analyses (39.a,b) yield in fact identical f-structure representations, on the basis of different c–structure representations. In particular, the structural complexity – measured in terms of the number of syntactic nodes employed, excluding lexical and preterminal nodes – is higher for the asymmetric coordination analysis (10 nonterminal nodes) as opposed to the regular ATB extraction analysis (9 syntactic nodes).

Following the Principle of Economy of Expression, then, the more "verbose" structural backbone, the asymmetric analysis in (39.b), is not admitted as an alternative grammatical analysis.

5.2.3 Quantifier Scope

Before discussing the more intricate word order properties, let us first review the scope phenomena discussed in Section 3.2. Example (17) – repeated below as (41) – shows the peculiar property of SGF coordination to allow wide scope of the quantified subject, from the middlefield position of the first conjunct. That is, the SGF coordination (41.b) is semantically equivalent to the regular VP coordination construction (41.a) (modulo the topicalised adverbial in (41.b)).

(41) a. Die wenigsten Leute [kaufen ein Auto] und [fahren mit dem Bus].
Almost no one buys a car and takes the bus
Almost no one buys a car and takes the bus.

b. [Daher kaufen die wenigsten Leute ein Auto] und [fahren mit dem Bus].
Therefore buys almost no one a car and takes the bus
Therefore, almost no one buys a car and takes the bus.

The key answer to this puzzling behaviour is already implied by our asymmetric GDF projection analysis, where the inherent asymmetry of the construction is captured in the c– to f–structure correspondence: by asymmetric projection of the SUBJ to the second conjunct we derive the very same f–structure representations for the symmetric and asymmetric coordination examples (again, modulo the causal adjunct in (41.b)).

Since in the LFG theory semantic interpretation, including quantificational scope, is computed on the basis of the f–structure representation, we predict the equivalent f–structures of symmetric and asymmetric coordinations in (41) to yield identical scopal interpretations.

[11] Equivalent examples can be constructed for symmetric coordination with a shared, topicalised SUBJ. These cases are similarly accounted for by consideration of the Principle of Economy.

In the *Glue Semantics* approach (see e.g. Dalrymple 1999), meaning is constructed compositionally, and in parallel to a linear logic derivation that assembles and consumes parts of the f–structure that contribute to the sentence meaning.

For coordination with shared arguments, such as the quantified subjects in (41), the semantics is built on exactly identical f–structure representations, schematically displayed in (42). Several proposals have been made for semantics construction for shared arguments in coordination (see Dalrymple 2001, p.376ff). An analysis attributed to Dick Crouch and Ash Asudeh is sketched in (42): the semantic contributions of the conjoined predicates (corresponding to $h_\sigma \multimap f1_\sigma$ and $h_\sigma \multimap f2_\sigma$ in the glue part) are consumed first, leading to an open, conjoined predicate in the corresponding meaning part: $\lambda X.[P(X) \wedge Q(X)]$. Quantifying in of the shared subject, referred to by h_σ in the glue part, then leads to a wide scope reading in case of a quantified subject subject.

The important steps of the derivation for example (41) are illustrated in (43).

(42) f [CONJ 'UND' { $f1$ [PRED = ... SUBJ h [ ]], $f2$ [PRED = ... SUBJ h [ ]] }]

$\lambda P.\lambda Q.\lambda X.[P(X) \wedge Q(X)] :$
$[h_\sigma \multimap f1_\sigma] \multimap [[h_\sigma \multimap f2_\sigma] \multimap [h_\sigma \multimap f_\sigma]]$

(Dalrymple 2001, p.379)

(43) [CONJ = 'UND' { $f1$ [SUBJ h [PRED = 'LEUTE'] PRED = 'KAUFEN⟨(↑ SUBJ)(↑ OBJ)⟩' OBJ [PRED = 'WAGEN']], $f2$ [SUBJ h [ ] PRED = 'FAHREN⟨(↑ SUBJ))⟩' ADJ { [PRED = 'MIT⟨(↑ OBJ)⟩' OBJ [PRED = 'BUS']] }] }]

$\lambda X.[\lambda x.kaufen(x, wagen)(X) \wedge \lambda x.fahren\_mit(x, bus)(X)] : h_\sigma \multimap f_\sigma$
$wenige(x, leute(x), kaufen(x, wagen) \wedge fahren\_mit(x, bus)) : f_\sigma$

5.2.4 The puzzle of word order asymmetry

We are left with the special word order restrictions observed for SGF coordination in Section 3.2. In particular, we need to explain why the specifier position of the right conjunct CP cannot be overtly realised. That is, why is (44.b) ungrammatical, as opposed to the general availability of topicalised non-subjects as in (44.c)?

These order restrictions are particularly challenging for a symmetric c–structure analysis, where the second conjunct offers a SpecCP position, and thus predicts (44.b) to be grammatical.

(44) a. In den Wald ging der Jäger und fing einen Hasen.

b. * In den Wald ging der Jäger und einen Hasen fing.

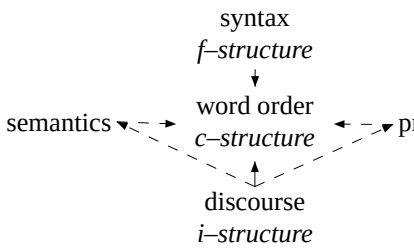
c. Einen Hasen fing der Jäger.

In discussion of Kathol's approach we have argued that his attempt to derive these word order restrictions from structural and/or functional constraints leads to rather ad-hoc conditions, lacking independent grammatical motivation.

In contrast, we will investigate these data within a broader perspective, by integration of well-established constraints on the interaction of *word order* and *discourse properties*. In particular, we build on the OT-based

LFG model of word order developed in Choi (1999, 2001). It derives word order properties observed in various, typologically distinct languages from a set of interacting constraints between different levels of grammatical description, in particular structural, functional-syntactic and discourse properties represented in c–structure, f–structure and i–structure (45).

(45)



(Choi 1999)

In the optimality-based model of Choi, word order is determined by interaction of – possibly conflicting – word order constraints that are imposed by the different levels of representation and their respective notions of "prominence". The optimality-based model is grounded on the principle of "harmonic alignment", i.e. the constraints are globally aimed at the most harmonic alignment of prominence hierarchies among the different levels of representation.

Choi (2001) assumes the following constraints to predict the word order patterns of various language (type)s: English, German, and Catalan. (46.a) predicts that word orders are most optimal if functional and word order prominence are harmonically aligned, e.g. if the most prominent grammatical function, the SUBJ is realised in the most prominent c–structure position, such as the SpecCP position in German. Concurrently, discourse properties need to be aligned with word order properties in such a way that discourse prominence is in harmonic alignment with structural (word order) prominence (46.b), where languages differ as to which direction (L/R) they choose to align the 'prominent' elements [+prom] or [+new]. In German, the constraints are set to constrain information that is already known in discourse, [-new], to precede [+new] information, and prominent elements [+prom] to precede non-prominent elements (see Choi 2001, for more detail).

(46) a. f–structure/c–structure constraints: (Choi 2001, p.29)
SBJ: The SUBJect aligns with most prominent c–structure position
CMPL: Complements align according to the 'grammatical prominence' hierarchy

b. i–structure/c–structure constraints: (Choi 2001, p.34)
PROM-L/R: [+prom] aligns left/right in the clause
NEW-L/R: [+new] aligns left/right in the clause

c. Optimality-based resolution of conflicts: e.g. [+prom]-LEFT >> SUBJ-LEFT

The model predicts, for a given i–structure representation, an optimal (most harmonic) word order. (47.a), e.g., is situated in a context where no element is discourse-prominent (e.g. focussed), and *Buch* is the only [+new] element in the discourse. The principles for German predict that prominent word order position of the subject yields the most harmonic, i.e. optimal serialisation.

(47) a. Context: *Was hast Du dem Kind geschenkt? – What did you give to the child?*
i–str: [ich]$_{[-prom,-new]}$ [dem Kind]$_{[-prom,-new]}$ [das Buch]$_{[-prom,+new]}$ [geschenkt]$_{[-prom,-new]}$
c–str: Ich$_{\mathrm{subj}}$ habe dem Kind das Buch geschenkt.

b. Context: *Was war mit dem Buch? Wem hast Du das Buch geschenkt?*
What happened to the book? To whom did you give the book?
i–str: [ich]$_{[-prom,-new]}$ [dem Kind]$_{[+prom,+new]}$ [das Buch]$_{[-prom,-new]}$ [geschenkt]$_{[-prom,-new]}$
c–str: Dem Kind$_{[+prom]}$ habe ich das Buch geschenkt.

In the optimality-based model, mismatches between the different word order constraints in (46.a,b) are resolved by language-specific constraint rankings. E.g., a language like German may define that precedence of discourse-prominent elements is more important (or more optimal) than precedence of a SUBJ function (46.c). In (47.b) this leads to an optimal serialisation where the prominent element *dem Kind* is left-aligned, while the competing subject takes a non-initial position.

However, if we apply this model to the word order properties of SGF coordination, it remains mysterious why the order in (48) should be ruled out as suboptimal. After all, we can imagine a discourse context where the object *Hase* (rabbit) is a discourse-prominent element, as rendered e.g. by emphatic stress. So, are we back to square one?

(48) Context: *Wohin ging der Jäger und was tat/fing er?*
Where did the hunter go and what did he do/catch?
i–str: [Jäger]$_{[-prom,-new]}$ [Wald]$_{[+prom,+new]}$ [Hase]$_{[+prom,+new]}$
c–str: * In den Wald$_{[+prom]}$ ging der Jäger und einen Hasen$_{[+prom]}$ fing.

5.3 A Discourse-Functional Analysis

What the previous section shows is that the general word order model of Choi (2001) fails to predict the special word order restrictions of SGF coordination. However, we argue that the analysis needs to accommodate special discourse-functional properties of asymmetric coordination. In what follows, we relate these properties to well-known discourse subordination effects of modal subordination. We establish general licensing conditions for this kind of discourse-functional subordination. In conjunction with the basic word order model of Choi (2001), these will explain the mysterious word order restrictions of both SGF and VL/VF coordination.

5.3.1 Discourse-functional properties of asymmetric coordination

The following set of examples gives pairwise contrasts between "regular" coordination or discourse sequences, as opposed to what we will call "discourse(-functional) subordination contexts" (see also Frank 1994).

For (49.a,a') we observe a striking contrast of interpretation between the symmetric (VL/VL) and the asymmetric (VL/VF) coordination:[12] the asymmetric variant only allows for a nonsensical interpretation where I like to go for walks if it is summer and winter at a time, whereas in the symmetrical (a) example I like to go for walks either way. (49.b,b') involving SGF coordination shows a related contrast: in the symmetrical case, the question focusses on possibly different points in times: the time when Peter calls the dog and the time he takes him for a walk. The SGF construction, though, can only be understood as a question about the time of a single, complex event or situation, when Peter calls the dog to take him for a walk.

(49) a. [[Wenn es Sommer ist] und [wenn es Winter ist]], gehe ich gerne spazieren.

a'. $\neq$ [[Wenn es Sommer ist] und [es ist Winter]], gehe ich gerne spazieren. VL/VF
'When it is summer and (# when) it is winter, I like to go for walks.'

b. [Wann ruft Peter den Hund] und [wann geht Peter mit ihm spazieren]?

b'. [Wann ruft Peter den Hund] und [geht mit ihm spazieren]? SGF
'When does Peter call the dog and (when does Peter) take him for a walk?'

[12] This example was brought up in discussion by Ellen Brandner about 10 years ago (see also Frank 1994).

c. Wenn Fritz ein Pferd hätte, würde er es lieben. # Er reitet *es* jeden Tag.

c'. Wenn Fritz ein Pferd hätte, würde er es lieben. Er würde *es* jeden Tag reiten. MS

If Fritz had a horse, he would love it. He (#rides | would ride) *it* every day.

The modal subordination examples in (49.c,c') show a related pattern: The first sentence of the sequence is – under standard analyses of the discourse semantics of conditionals – an island for the binding of anaphoric pronouns like *es*. However, in (49.c') the same syntactic configuration seems to allow for the extension of the conditional's scope, as indicated by the binding of *es* to *ein Pferd*.

While analyses of modal subordination differ in various aspects (cf. Frank 1997), an abstract characterisation of the crucial aspects involved can be stated as follows: modal subordination can occur in contexts of complex situations (or eventualities), by extension of the scope of a modal operator to otherwise inaccessible material. Domain extension is only licensed if the discourse-subordinated elements do not display *independent* domain marking. This condition is violated in (49.c), where indicative mood signals reference to the actual world; as opposed to (49.c'), where subjunctive mood accords with the context of hypothetical worlds set up by the subordinating modal operator.

We can generalise these conditions to a more abstract characterisation of generalised *discourse subordination*, involving (i) the subordinating *domain extension* of an *operator*, (ii) in a complex situation, (iii) lacking *independent* domain marking of the discourse-subordinated elements.

5.3.2 Licensing conditions for asymmetric GDF-projection

We consider asymmetric coordination as a syntactic instance of this general notion of *discourse subordination*. Unlike extension of a modal operator's scope, we encounter extension of a syntactic, discourse-functional domain, which is marked by a complementiser or a genuine discourse function, both typical elements of the clause's functional projection. This extension of the default discourse-functional domain is brought about and modeled by our notion of (asymmetric) projection of a grammaticalised discourse function GDF, and is subject to various constraints. In particular, extension of a discourse-functional domain is incompatible with *independent domain marking* of the subordinated elements, by complementisers or genuine discourse functions TOPIC or FOCUS.

The conditions summarised in (50) apply to the asymmetric examples (49.a',b'): the functional domain established by the first conjunct (by a complementiser or FOCUS phrase) is extended to the second conjunct, lacking independent domain marking by a complementiser or discourse function.

(50) Asymmetric Coordination as discourse-functional domain extension

- Complementisers (C) and genuine discourse functions TOPIC, FOCUS are syntactic markers of discourse functional domains.
- Extension of a discourse functional domain is modeled by (asymmetric) projection of a grammaticalised discourse function (GDF).
- It occurs in coordinated conjuncts, conceived or presented as a complex situation.
- Independent domain marking of functionally subordinated conjuncts by complementisers (C) or TOPIC/FOCUS marking is prohibited.

5.3.3 Word order properties explained

The assumptions summarised in (50) account for the word order properties of asymmetric coordination. In (51) we associate the different serialisations of both types of asymmetric coordinations with their respective discourse-functional domain markers: it is brought out that an introducing domain marked by a TOPIC or

complementiser (COMPL) may be extended (by asymmetric GDF projection), provided the subordinated conjunct is not independently domain-marked by another complementiser or a genuine discourse function. A SUBJ function in the second conjunct is in this respect a neutral element for functional domain marking.

(51) a. In den Wald ging der Jäger und fing einen Hasen. TOPIC-OBL & SUBJ
* In den Wald ging der Jäger und einen Hasen fing. * TOPIC-OBL & TOPIC-OBJ

b. Wenn Du in ein Kaufhaus kommst und hast kein Geld, ... COMPL & SUBJ
Wenn Du in ein Kaufhaus kommst und Du hast kein Geld, ... COMPL & SUBJ
* Wenn Du in ein Kaufhaus kommst und kein Geld hast Du, ... COMPL & TOPIC-OBJ

Final support for relating asymmetric coordination to a general notion of discourse subordination is suggested by the general forward direction of domain extension to the right (cf. the ungrammatical backwards serialisations in (52)). Restriction of forward-directed scope extension is also observed for modal subordination.

(52) a. * Ging in den Wald und gestern fing der Jäger einen Hasen.
Went into the forest and yesterday caught the hunter a rabbit.

b. * Kommst in ein Kaufhaus und wenn Du kein Geld hast, kannst Du nichts kaufen.
Enter a shop and if you no money have, can you nothing buy

6 Conclusion

Our analysis of asymmetric coordination is built on a minimal extension of the classical LFG analysis of constituent coordination. Due to the flexible correspondence architecture of LFG theory, the asymmetry is captured in the c– to f–structure mapping, by asymmetric projection of a grammaticalised discourse function. This analysis predicts the basic functional syntactic and semantic properties of asymmetric coordinations.

We motivated asymmetric GDF projection by taking into account the discourse properties of asymmetric coordination. We argued that asymmetric coordination is a special instance of a more general notion of *discourse subordination*, by relating it to modal subordination. From this notion of discourse subordination we derived special licensing conditions for functional-syntactic discourse subordination that account for the peculiar word order restrictions of asymmetric coordination.

We conclude that our LFG account of asymmetric coordination makes a case for the projection architecture of LFG, where independent levels of representation constrain each other.

There are many open question that we wish to pursue in future work. An obvious question to ask is why asymmetric coordination (and thus GDF projection) is restricted to Germanic languages. Moreover, we need to investigate whether instantiation of GDF to a discourse function licenses asymmetric coordination in other languages, where FOCUS phrases can be clause internal.

Acknowledgements

We thank the audiences of the LFG'02 conference, the Dublin Computational Linguistics Research Seminar, and the Linguistics Departments at the University of Konstanz and the University of Stuttgart for valuable comments and discussion. Many interesting observations could not be addressed here. Special thanks go to Stefanie Dipper for comments and discussions on the basis of an early draft.

References

Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell Publishers, Oxford.

Büring, D. and Hartmann, K. (1998). Asymmetrische Koordinationen. *Linguistische Berichte*, 174:172–201.

Choi, H.-W. (1999). *Optimizing Structure in Context*. CSLI Publications, Stanford.

Choi, H.-W. (2001). Phrase Structure, Information Structure, and Resolution of Mismatch. In Sells, P., editor, *Formal and Empirical Issues in Optimality Theoretic Syntax*, pages 17–62. CSLI Publications, Stanford.

Dalrymple, M., editor (1999). *Semantics and Syntax in Lexical Functional Grammar*. MIT Press.

Dalrymple, M. (2001). *Lexical-Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press.

Frank, A. (1994). V2 by underspecification or by lexical rule. Arbeitspapiere des SFB 340 Nr. 43, University of Stuttgart. 77 pages.

Frank, A. (1997). *Context Dependence in Modal Constructions*. PhD thesis, Stuttgart University. 411 pages, published in: Arbeitspapiere des Sonderforschungsbereichs 340, Sprachtheoretische Grundlagen für die Computerlinguistik, Nr. 91.

Frank, A. (2001). Treebank Conversion. Converting the NEGRA Treebank to an LTAG Grammar. In *Proceedings of the Workshop on Multi-layer Corpus-based Analysis*, pages 29–43, EUROLAN 2002 Summer Institute, Iasi, Romania.

Frank, A. and Kamp, H. (1997). On Context Dependence in Modal Constructions. In *Proceedings of SALT VII*, Stanford University. CLC Publications, Cornell University. 19 pages.

Haider, H. (1988). Matching projections. In Cardinaletti, A., Cinque, G., and Giusti, G., editors, *Constituent Structure: Papers from the 1987 GLOW Conference*, Venezia: Annali di Ca'Foscari XXXVII, pages 101–121.

Heycock, C. and Kroch, A. (1993). Verb movement and coordination in a dynamic theory of licensing. *The Linguistic Review*, 11:257–283.

Höhle, T. (1983a). Subjektlücken in Koordinationen. Unpubl. manuscript, University of Cologne.

Höhle, T. (1983b). Topologische felder. Unpublished manuscript, University of Cologne.

Höhle, T. (1990). Assumptions about asymmetric coordination in German. In Mascaró, J. and Nespor, M., editors, *Grammar in Progress*, pages 221–235. Foris, Dordrecht.

Kathol, A. (1995). *Linearization-based German Syntax*. PhD thesis, Ohio State University.

Kathol, A. (1999). Linearization vs. phrase structure in German coordination constructions. *Cognitive Linguistics*, 10(4):303–342.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. CSLI Publications, Stanford.

Steedman, M. (1990). Gapping as constituent coordination. *Linguistics and Philosophy*, 13:207–264.

Wunderlich, D. (1988). Some problems of coordination in german. In Reyle, U. and Rohrer, C., editors, *Natural Language Parsing and Linguistic Theories*, pages 289–316. Reidel, Dordrecht.

BULGARIAN WORD ORDER AND THE ROLE OF THE DIRECT OBJECT CLITIC IN LFG

T. Florian Jaeger
Veronica A. Gerassimova

Linguistics Department, Stanford University

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

ABSTRACT – This paper provides an LFG account of the Bulgarian direct object clitic's interaction with information structure (i.e. topic-focus structure) and word order. We show that the direct object clitic has at least two functions (it is both a topical object agreement marker and default pronoun) and then demonstrate how our account correctly predicts in which syntactic environment which of the two functions can be chosen. In order to achieve this we allow for two different ways to identify a 'topic' in LFG – a move, which reduces the necessary claims about the direct object clitic's behaviour to the most general principles of LFG (i.e. Uniqueness, Completeness, Extended Coherence). The proposed analysis is based on extensive evidence (our own online experiment, Leafgren 1997a,b, 1998, and Avgustinova 1997), and incorporates recent findings on the discourse-configurationality of the left periphery in Bulgarian clauses (cf. Rudin 1997, Arnaudova 2001, Dimitrova-Vulchanova & Hellan 1998). Although covering a much broader range of data from spoken Bulgarian than other formal accounts, our account makes the right predictions about possible word orders and the optional, or obligatory presence/absence of the direct object clitic. Unlike almost all other recent accounts, our analysis does not rely on the assumption of configurationality, which has been shown to be problematic for Bulgarian (cf. Gerassimova & Jaeger 2002).

I Introduction*

Contemporary, colloquial Bulgarian allows for clitic doubling of objects in certain contexts. The object clitics can occur as the only realization of the object, as in (1), double an NP or double a long form pronoun, as in (2). Although there is also an indirect object clitic whose distribution is for the most part parallel with the direct object clitic's, we restrict ourselves to the investigation of the direct object clitic (henceforth DOC).[1] All examples given in this paper only contain the direct object clitic. The DOC can also occur in an embedded sentence from which the direct object has been extracted. An example for extraction out of an adjunct clause, is given in (3). (4) is an example of object extraction out of a sentential subject. For ease of understanding, the DOC and the coreferential object (if present) are underlined.

(1) Decata ja običat.[2]
children$_{DEF.PL}$ DOC$_{3.SG.FEM}$ love$_3$
The children love her.

(2) Decata ja običat Marija/neja.
children$_{DEF.PL}$ DOC$_{3.SG.FEM}$ love$_3$ Maria/her$_{3.SG.FEM.ACC}$
The children love Maria/her.

(3) Radioto, koeto Todor otide na plaž [bez da (go)
radio$_{DEF}$ which Todor went$_3$ on beach without SBJ DOC$_{3.SG.MASC}$
izključi], e na Elena.
switch-off is of Elena
The radio which Todor went to the beach without switching off is Elena's.

* Our special thanks go out to Peter Sells, Joan Bresnan, Elizabeth Traugott, Chris Manning, Arnold Zwicky, and Tracy H. King for their advice and support all throughout the progress of our research. We want to thank Mary Dalrymple, Tracy H. King, and Jonas Kuhn for their help with some formal aspects of LFG. We also want to very much thank Ruth Kempson for making us aware of several interesting questions and providing good ideas how to approach them, as well as Iskra Iskrova for discussing the relevant data with us. Last but not least, we benefited from the questions and suggestions from the anonymous reviewers of the LFG02 abstracts, Shiao-Wei Tham, Judith Tonhauser, Andrew Koontz-Garboden, and especially Lev Blumenfeld. We are also grateful for the great feedback we got at the LFG02 and at an earlier presentation of our work for the Linguistics Department, Stanford. Thanks to Tracy H. King (again) and Lev Blumenfeld for feedback on the final draft of this paper. All remaining mistakes remain ours and must not be reproduced without our permission, ;-).

[1] We use the term direct object clitic (DOC) to refer to the set of linguistic *forms* of the direct object clitic, not their meaning. These are the following forms: SG – 1st *me*, 2nd *te*, 3rd masc./neut. *go*, fem. *ja*; PL – 1st *ni*, 2nd *vi*, 3rd *gi*.

[2] We use the following glosses: 1, 2, 3 – first, second, and third person; DEF – definite suffix, INDEF – indefinite specific article; FEM – feminine, MASC – masculine, NEUT – neuter; PL – plural, SG – singular; REFL – reflexive pronoun, SBJ – subjunctive marker. SMALL CAPS indicate emphatic accent.

(4) Todor e jasno, če Ivan *(go) e vidjal.
Todor is clear that Ivan $DOC_{3.SG.MASC}$ is seen
Todor, it is clear that Ivan has seen him.

In this paper, we discuss different functions of the Bulgarian direct object clitic and its interaction with syntax (especially word order) and information structure. Our research relates to research on D[iscourse] F[unction] and G[rammatical] F[unction]-configurationality, an issue which has been identified as the primary, so far unresolved issue in the literature on (South) Slavic Syntax (cf. Siewierska & Uhlířová 1998:143 in their review of the recent literature on the word order of Slavic languages).

We propose an analysis of the DOC, which - we argue - accounts for a whole range of data that so far have not been explained, including examples of free word order in a non-dependent marking language. We argue that the DOC has not one but several functions, one of which has not been recognized at all in the literature. This is at least partially the reason why the issue of the DOC's functions is still unresolved in the literature. First, in the clitic doubling construction (we explain what we mean by this in the next paragraph), the DOC is a non-anaphoric direct object TOPIC-agreement-marker. Second, the DOC is the default direct object pronoun. Third, the DOC is an intrusive direct object pronoun in extractions (cf. Sells 1984). Due to lack of space, we only discuss the first two functions here. We argue that some of the confusion about these functions in the literature is due to different notions of topic and suggest a way to resolve this issue within LFG. Furthermore, we account for the range of possible word orders given the presence or absence of the DOC.

The last point is especially important since – to the best of our knowledge – all existing accounts either hardly, if at all, capture the generalizations relating to possible word orders, or only account for a relatively small subset of them. To guarantee a broad coverage of data, we test and compare the predictions of our analysis with the data provided in Avgustinova (1997; elicited question-answer pairs) and Leafgren (1997a,b, 1998, 2001; corpus studies of written/spoken, informal/formal Bulgarian). Moreover, we use the case of island violations to show how the distribution of the DOC as default pronoun or topic marker is correctly predicted.

Before we provide an outline of the structure of this paper, we will briefly clarify our use of the term 'clitic doubling'. With clitic doubling (henceforth CD) we refer to the overt doubling of a constituent, usually an argument (here the direct object), by a phonologically weak, syntactically non-projecting[3] lexical element, i.e. a clitic (here the DOC). CD is a prominent topic in the literature on Slavic and Balkan linguistics (e.g. Franks & King 2000, Rudin 1990/1991, 1996, 1997, Dyer 1992, Guentcheva 1994, a.o.), the typology of pronouns, agreement (e.g. Bresnan & Mchombo 1987), configurationality (e.g. Baker's (1991) *pronominal object hypothesis*), and case assignment (e.g. Rudin 1997). The aspects of CD that are addressed here include the following. First, how is coreference between the clitic and the doubled NP established? In MP/GB this comes down to the question whether, for example, fronted objects are moved or anaphorically bound by the DOC. In LFG terms, this corresponds to the issue of functional control vs. anaphoric binding. Second, does the clitic mark a grammatical function (GF) or a discourse function (DF)[4] or both? Third, is the clitic and/or the lexical object NP the object argument? This is interesting since according to some theories (e.g. GB, MP) only one constituent can be assigned CASE. The LFG framework is less restrictive in this respect. As long as UNIQUENESS (cf. Bresnan 2001:47) is fulfilled, information belonging to the same GF can be distributed among several syntactic constituents. Nevertheless, translated into LFG, the above-mentioned question remains, namely whether the clitic provides information on OBJ PRED (i.e. the PRED value of the object).

In the remaining sections, we proceed as follows. In section II, we introduce some basic facts about Bulgarian, including some phrase structure rules describing the internal order of the predicate clitic cluster and capturing the fact that Bulgarian is not configurational. In section III, we briefly

[3] See Toivonen (2001:chapter 3) for a typology of non-projecting words.
[4] Note that, with 'discourse function', we do not refer to discourse function as defined in Schiffrin (1988) or Fraser (1988). We comply to the naming convention of LFG and use the term discourse function (DF) to refer to what more precisely could be called f-structure correlate of an information structural role.

describe some earlier analyses of the DOC's function. In section IV, we introduce recent findings on the discourse-configurationality in Bulgarian, incorporate them into our analysis, and formalize the direct object clitic's (DOC) properties in CD. In this context, we discuss our proposal in the light of the known data and show how the interaction of the proposed lexical entry of the DOC and the proposed phrase structure rules make the right predictions about grammaticality of certain word orders and their information structural correlate (we will elaborate on this below). We also use section IV to introduce our model of the information structure (henceforth IS) component and its interface with other components (e.g. f-structure). In section V, we discuss a second function of the DOC, which has so far been ignored in the literature, namely its use as the default pronoun of Bulgarian. In section VI, we briefly survey islands in Bulgarian to show how our account makes the right predictions about the distribution of the different types of DOCs. Last, we will summarize the conclusions and mention some open issues in section VII.

II An introduction to some aspects of Bulgarian

Bulgarian is a South-Slavic language spoken by approximately 9 million speakers[5] world wide. If not mentioned otherwise, we will use the term Bulgarian to refer to contemporary, colloquial, spoken Bulgarian. Bulgaria has a strong prescriptive tradition and the differences between written vs. spoken and formal vs. informal Bulgarian seem to be immense.[6] Clitic doubling (henceforth CD) is very rare in formal and written Bulgarian. Leafgren (2001:4) shows that the frequency of CD in formal written texts (0.5% of all object occurrences) contrasts sharply with the 10% frequency of CD in informal oral texts. Furthermore, we restrict ourselves to those dialects of Bulgarian which make productive use of the object clitics, i.e. mostly the Western dialects (cf. Leafgren 1997a:119).

Since Bulgarian is in many respects the most atypical Slavic language and has some typologically uncommon properties, we sketch those characteristics of Bulgarian that will turn out to be relevant for understanding the analysis presented in section IV.

Unlike all other Slavic languages (except for Macedonian) Bulgarian has lost its case marking system. Some scholars have argued that the definiteness suffix (singular: masc. *-a*, fem.: *-ta*, neut.: *-to*; plural: *-te*/*-ta*) identifies the subject. This is wrong since the definiteness suffix can also be attached to an object. The only dependent-marking device in Bulgarian is the preposition *na* which among other things identifies the indirect object. In certain environments even this last bit of dependent-marking can be dropped (cf. Vakareliyska 1994).

Despite the almost complete lack of dependent-marking, Bulgarian allows very free word order. With different requirements on the context, the intonation and morpho-syntactic marking, all theoretically possible word orders can actually be observed (cf. Siewierska & Uhlířová 1998:107-10 for ditransitives and implicitly Avgustinova 1997:112). While we provide more details on the effect of the DOC on word order in section IV, it is generally true that some word orders are not possible without the DOC. In other words, the DOC seems to 'license' certain word orders. Two examples for alternative word orders with the DOC are given below (based on Avgustinova 1997:112).

(5) Parite *(gi) vze Olga.
money$_{DEF}$ DOC$_{3.PL}$ took$_{3.SG}$ Olga

(6) vze *(gi) Olga parite.
took$_{3.SG}$ DOC$_{3.PL}$ Olga money$_{DEF}$
Olga took the money.

Note, however, that Bulgarian shows a clear preference for a SUBJ-V-DO-IO surface order, a tendency noted by several scholars (cf. Leafgren 2002:1, Dyer 1992:63, Avgustinova 1997:114, a.o.). Leafgren (2002:1) argues that averaged over all registers and genres about 80.5% of all

[5] Data gathered in 1995. For more information, refer to Ethnologue, Barbara F. Grimes, eds. 13th Edition.
[6] During an online experiment that we designed to get native speaker judgments on contemporary, colloquial, spoken Bulgarian we first ran into problems since our informants were so strongly influenced by the idea that they had to judge the prescriptive correctness instead of 'what they actually say'.

sentences are SVO. Dyer (1992) shows that SVO is not only statistically the most common constituent order but also *stylistically neutral*.

The lack of stringent word order and case marking is – at first – surprising. However, Bulgarian has other means to identify grammatical functions, namely intonation and head-marking. Here, we focus on head-marking, more precisely one kind of head-marking in Bulgarian, clitic doubling by the direct object clitics. Before we turn to the interaction of the direct object clitic and word order, we want to briefly mention other morphosyntactic means of Bulgarian. First, the sentence predicate agrees with the subject in person and number, and participles (i.e. subjunctives) agree also in gender. The predicate combines with the clausal clitics into the predicate clitic cluster. Because there is an extensive literature on the internal order of the predicate clitics cluster (e.g. Avgustinova 1997, Siewierska & Uhlířová 1998; see Franks & King 2000:234ff. for a summary), we do not discuss this issue here. To understand the examples given later in this paper it is sufficient to bear in mind the following, simplified schema for the internal order of the predicate clitic cluster, where IOC stands for 'indirect object clitic' and DOC for 'direct object clitic' (cf. Englund 1977:109-19). For our purpose, the annotated phrase structure rule in (8) captures the generalization in (7).

(7) aux (not 3.SG) > IOC > DOC > aux (3.SG)

(8) $V \rightarrow$ (V_{CL}) (N_{CL}) (N_{CL}) (V_{CL}) V'
(↑SUBJ PERS)≠3 (↑OBJ2)=↓ (↑OBJ)=↓ (↑SUBJ PERS)=3 ↑=↓
(↑SUBJ NUM)≠SG (↑SUBJ NUM)=SG

The clitic cluster as a whole is preverbal except for the cases where this would cause the clitics to be clause-initial. In those cases, the verb is preposed to the clitic cluster. In other words, the positioning of the Bulgarian clitic cluster is subject to the Tobler-Mussafia effect (cf. Tomić 1997, 1996, Rudin et al. 1998:566; for an OT account to typology of clitic positioning, see Billings 2000) and not to Wackernagel's Law (unlike the clausal clitics in almost all other Slavic languages). The object clitics belong to the clausal clitics. In the case of clitic doubling, the object clitic(s) agree in person, number and gender (only for 3.SG) with the reduplicated object. Unlike the object clitics, which can only occur in the clitic cluster, the second kind of pronouns in Bulgarian, namely the long form pronouns, have the same syntactic distribution as full lexical NPs. The long form pronouns, when occurring alone, mark contrastive or emphatic focus (cf. Avgustinova 1997:116 Vakareliyska 1994:125; see Leafgren 1997a:118 for a table of all clitic pronouns and long form pronouns), in which case they always receive stress (compare (9) and (10) below).

(9) Decata обичат НЕЯ.[7]
children$_{DEF.PL}$ love$_3$ her$_{3.SG.FEM.ACC}$
The children love HER.

(10) Decata ja обичат neja.
children$_{DEF.PL}$ DOC$_{3.SG.FEM}$ love$_3$ her$_{3.SG.FEM.ACC}$
The children love her.

To sum up what has been said so far, Bulgarian is a non-case marking, partially head-marking, free word order language with optional clitic doubling of objects. Another important aspect of Bulgarian that has been ignored in the literature so far is the lack of evidence for G[rammatical] F[unction]-configurationality. Although already Rudin (1985) mentions that there seems to be no such evidence, GF-configurationality plays a crucial role in most recent analyses of Bulgarian syntax (including those on CD). We have shown elsewhere (cf. Gerassimova & Jaeger 2002) that it is difficult if not impossible to find evidence *for* GF-configurationality. More precisely, some tests, such as weak crossover tests, variable binding tests, extraction tests, etc., clearly argue for non-configurationality of Bulgarian. Therefore we do not assume GF-configurationality here.

[7] In our examples throughout the paper, we mark emphatic accent/stress with SMALL CAPS. Although only a part of the word receives emphatic accent we will just mark the whole word as prosodically emphasized.

The annotated phrase structure rule in (11) captures this and describes a flat VP with unordered constituents (c.f. Kiss 1995:11 for Hungarian).[8]

(11) VP → (XP) , (PP), V'
(↑GF)=↓ (↑OBJ2)=↓ ↑=↓

In section IV, we show that the flat VP hypothesis is necessary for or at least highly compatible with the formal account of CD and its interaction with possible word orders presented here. Before we turn to our own analysis of the DOC in CD and of its use as default pronoun of Bulgarian, we briefly summarize previous analyses of the DOC.

III Previous analyses of the DOC

All of the accounts discussed here have exclusively dealt with C[litic] D[oubling] (sometimes also referred to as clitic replication in the literature) and ignored other uses/functions of the DOC. To the best of our knowledge, the function of the DOC as default pronoun (cf. section V) and its interaction with the use of the DOC in the CD construction have not been described by anyone yet. The existing accounts of the DOC can be distinguished according to their basic hypothesis. We will discuss each of them in the order they are listed below.

(H1) The object clitics mark non-canonical word orders.
(H2) The object clitics mark the case (of the doubled constituent).
(H3) The object clitics mark definite objects.
(H4) The object clitics mark specific objects.
(H5) The object clitics mark topical objects.

Both (H1) and (H2), i.e. the word order marker and the case marker hypotheses, suggested in AG (1983,3:187-188, 282-283), Popov (1963:166, 229-230), Cyxun (1968:110) and Georgieva (1974:75), have in common the claim that CD together with word order serves to disambiguate case roles. Leafgren (1997a:124) concludes that under this view sentences with CD should be unambiguous even if both subject and object have the same gender, number, etc. However, this is not the case. Sentences with CD can be ambiguous. For example, as shown below both VOS and VSO word orders are possible with the same stress assignment as long as the clitic is present.

(12) Parite gi VZE Olga.
money$_{DEF}$ DOC$_{3.PL}$ took$_{3.SG}$ Olga
(13) VZE gi parite Olga.
took$_{3.SG}$ DOC$_{3.PL}$ money$_{DEF}$ Olga
Olga took the money.

Furthermore, the word order marker hypothesis cannot explain why the DOC is optional and why it can occur in both the unmarked and the marked word order, and the case marker hypothesis fails to account for the optionality of the object clitics. The definiteness-marker hypothesis, (H3), as proposed in Cyxun (1962:289-290), Minčeva (1969:3), Ivančev (1957:139), Georgieva (1974:75),[9] has been shown to be wrong by Ivančev (1968:164) and Kazazis & Pentheradoukis (1976:399-400), since indefinite *specific* NPs can be doubled (cf. Leafgren 1997a:122), as shown in (14). *Edno* is an instance of the Bulgarian indefinite, specific article.[10]

[8] We use the XP annotated with (↑GF)=↓ to express that all kinds of core arguments can occur in this position (including e.g. COMPs).
[9] Also see Popov & Popova (1975:48) and Popov (1973:173), who, probably aiming at specificity, require the doubled NP to be 'articulated' (cf. Leafgren 1997a:121).
[10] The specific, indefinite article has the following paradigm: Singular: masc. *edin*, fem. *edna*, neut. *edno* 'a certain, a particular'; Plural: *edni* 'certain' (cf. Vakareliyska 1994:122). More precisely, this article requires an

(14) Edno dete go vidjax da pluva.
a-cerain child DOC$_{3.SG.NEUT}$ saw$_{1.SG}$ SBJ swim$_{3.SG}$
I saw a (certain) child swimming.

Avgustinova (1997:92-95) is a recent proponent of the specificity-marker hypothesis, (H4).[11] She distinguishes between [+/-limited] nominal material and further divides [+limited] nominal material into [+/-specific] and [-limited] nominal material into [+/ -generic]. In her terminology only [+limited, +specific] objects can be doubled. The specificity-marker hypothesis is motivated by the contrast between (14) and (15). In (15) the fronted, [-specific] object cannot be doubled although the corresponding sentence (16) with neutral word order and without CD is grammatical.

(15) *Njakoja po-nova kola iskam da si ja kupja.
some$_{-SPEC}$ newer car want$_{1.SG}$ SBJ REFL$_{1.SG}$ DOC$_{3.SG.FEM}$ buy
Intended: *I want to buy (for myself) some newer car.*

(16) Iskam da si kupja njakoja po-nova kola.
want$_{1.SG}$ SBJ REFL$_{1.SG}$ buy some$_{-SPEC}$ newer car
I want to buy (for myself) some newer car.

This point is further supported by the exceptions to the generalization that *edni* is [+specific] (cf. footnote 10 above). In (17) *edni po-iziskani drexi* is [-specific] (cf. Avgustinova 1997:95) and the fronted object cannot be doubled.

(17) *Edni po-iziskani drexi gi dadoxa na Ivan.
some$_{-spec}$ stylish clothes DOC$_{3.PL}$ gave$_{3.PL}$ to Ivan
Intended: *Some stylish clothes, they gave (them) to Ivan.*

However, (H4) has also proven to be insufficient since generics *can* and in some cases even *must* be doubled, as illustrated in (18). Independently of our observations, Alexandrova (1997) and Guentchéva (1994), too, point out that generics and interrogatives can be doubled (for the doubling of interrogatives, cf. also Jaeger 2002).

(18) Slonovete *(gi) obučavat xorata.[12]
elephants$_{DEF}$ DOC$_{3.PL}$ train$_{3.PL}$ people$_{DEF}$
The elephants, (the) people train.

So far we have shown that [-limited, +generic], e.g. (18), and [+limited, +specific] object NPs, e.g. (14), can be doubled while [+limited, -specific] object NPs cannot be doubled, as shown in (15) and (17). This raises the question if [-limited, -generic] object NPs can also be doubled. As for the examples above, we use the object fronting construction to test this.[13] The examples (19) and (20) are taken from Avgustinova (1997:92). The corresponding CD examples, (21) and (22), are ungrammatical.

(19) Tuk kupuvam knigi.
here buy$_{1.SG}$ books$_{-DEF}$
I buy books here.

NP not marked by the definiteness suffix. For a formal description of the semantics of *edin*, see Izvorski (1994) who, among other things, shows that, in her terminology, *edin* is not always [+specific].

[11] See also Kazazis & Pentheradoukis (1976) and Vakareliyska (1994:122).

[12] Actually, (18) is grammatical without the DOC if *slonovete* is realized with emphatic stress and thus receives the exclusive focus. This is what we would expect since this is a case of FOCUS-fronting (see section IV). In this paper, we are only interested in non-focus object fronting, i.e. object fronting without emphatic stress on the object. Therefore, whenever we star an example with a fronted object that is not given in small caps, we always mean that this example is ungrammatical for fronted non-focused objects.

[13] This will become clearer in section IV. In short, a fronted object without focus intonation must be doubled by the corresponding object clitic if this is possible at all. If doubling is not possible (like for e.g. [+limited, -specific] object NPs) the resulting clause is ungrammatical.

(20) Târsja prijateli.
look-for$_{1.SG}$ friends$_{-DEF}$
I am looking for friends.

(21) *Knigi tuk gi kupuvam.
books$_{-DEF}$ here DOC$_{3.PL}$ buy$_{1.SG}$
Intended: *Books, I buy here.*

(22) *Prijateli gi târsja.
friends-$_{DEF}$ DOC$_{3.PL}$ look-for$_{1.SG}$
Intended: *Friends, I am looking for.*

To sum up, we have shown that [-limited, +generic] and [+limited, +specific] objects can be doubled, whereas [-limited, -generic] and [+limited, -specific] objects cannot be doubled. In the following, we adopt a slightly different but equally common classificatory system where nominal material is [+/-generic], and the [-generic] NPs are further divided into [+/-specific]. Then, the generalization is captured as follows: [-generic, -specific] NPs can *not* be doubled.[14] Note that is is typologically common that [+specific] and [+generic] NPs pattern together (Shiao-Wei Tham, p.c.). Our observations, like those of Alexandrova (1997) and Guentchéva (1994), contrast with Avgustinova's (1997) claim that only [+limited, +specific] NPs can be doubled. Our data also rejects Rudin's (1997) analysis that the DOC only doubles (topical) [+specific] NPs.

Now consider the topic-marker hypothesis, (H5), as formulated in Leafgren (1997a,b, 1998), Avgustinova & Andreeva (1999), and to some extent Ivančev (1974), Georgieva (1974), Minčeva (1969), Popov (1963:167) and the AG (1983,3:188). According to this hypothesis, the above-mentioned restriction on the doubled object is an indirect effect of the requirement that the doubled object has to be topical. Leafgren (1997a:136ff.) further shows that topicality marking in Bulgarian cannot be reduced to agentivity or subjecthood, two scales that correlate with the scale of topicality in many languages (for a discussion of those hierarchies, cf. Givon 1976). However, Leafgren (1997a,b, 1998) does not show how his proposal (i.e. (H5) as stated above) accounts for the contrast between (14) and (15) or the ungrammaticality of (17), (21), and (22). In fact, (H5) turns out to be to drastic in its formulation. Consider examples (23) and (24). In our classificatory system, *njakolko* is a [-definite; -generic, +specific] quantifier, *malko* a [-definite; -generic, -specific] quantifier. *Njakolko*, unlike *malko*, is compatible with and sometimes even requires CD.[15] However, there is no apparent reason why *njakolko spisanija* in (23) should be a topic and *malko spisanija* in (24) not. Thus it seems hard to explain the difference between (23) and (24) by (H5).[16]

(23) Ima njakolko spisanija koito mnogo xora (gi) xaresvat.
have a-few$_{+SPEC}$journals$_{-DEF}$ which$_{3.PL}$ lots people DOC$_{3.PL}$ like$_{PL}$
There are a few (certain) journals that a lot of people like (them).

(24) Ima malko spisanija, koito mnogo xora (*gi) xaresvat.
have a-few$_{-SPEC}$ journals$_{-DEF}$ which$_{3.PL}$ that people DOC$_{3.PL}$ like$_{PL}$
There is a small number of journals that a lot of people like.

There are two ways out of this problem. One is to take typological evidence as, for example, sketched in Lambrecht (1994:155-56) who claims that topics have to be "referring expressions" to

[14] Thanks to Shiao-Wei Tham for discussing different classificatory systems for the semantics of nominal material with one of the authors (F.J.). Remaining mistakes are, of course, due to the authors.

[15] We are thankful to Ruth Kempson for pointing us to this data and helping us to gather it. We also are very grateful for the patience of Iskra Iskrova who explained and discussed (23) - (25) (and other material) with one of us (F.J.) in detail.

[16] Interestingly, one of our informants pointed out that for her (24) is only grammatical if either only *koito* 'which' or only the DOC *gi* 'them' is realized. This relates to the third use of the DOC as an intrusive pronoun in extractions (cf. Sells 1984), which we cannot discuss here due to lack of space. For V.G. and another informant, (24), as given above, is grammatical.

show that there are universal restrictions on the semantics of topics.[17] This approach will result in a notion of topic that will be qualitatively quite different from that of Leafgren (1997a:127) who defines the topic to be 'what the clause is about'. Second, one could claim that CD in Bulgarian has more than one constraint on the semantics and information structural role of the doubled object, namely a) doubled objects have to be topical, and b) doubled objects cannot be [-generic, -specific]. As the comparison between (23) and (25) shows, this is necessary anyway to explain why CD is *obligatory* in some cases and *optional* in others.

(25) Njakolko spisanija mnogo xora *(gi) xaresvat.
a-few$_{+SPEC}$ journals$_{-DEF}$ lots people DOC$_{3.PL}$ like$_{PL}$
A few (certain) journals, a lot of people like (them)

Here we are mainly interested in the differences between cases of obligatory and optional CD and therefore do not care to commit ourselves to either of the two ways. The account presented here (cf. section IV) is compatible with additional constraints on the semantics (e.g. specificity). Although we are aware that the inherently vague and widely varying definition of topic is problematic for (H5), we take this hypothesis as the starting point for a formalization of the properties of the DOC in the CD construction, which we introduce in the next section. In other words, we adopt an approach similar to that in Lambrecht (1994): topics cannot be [-generic, -specific]. We leave the details open to future research. Finally, note that none of the above-mentioned approaches captures the fact that the DOC can also be the default pronoun. It is exactly the interaction between this use and its use as a topical object agreement marker that provides interesting evidence for our analysis. We come back to this issue in section VI. Next, we present our analysis of the DOC in the CD construction and in its use as the default pronoun.

IV DF-configurationality and the DOC in clitic doubling

There is good evidence from the extensive literature on the left periphery of the Bulgarian clause that Bulgarian is DF-configurational (see Dimitrova-Vulchanova & Hellan 1998, Rudin 1994, 1990/1991, 1985, Arnaudova 2001, Lambova 2002, Dyer 1992, 1993, Leafgren 1997c, a.o. on Bulgarian; Kiss 1995, 2001 for DF-configurationality). Bulgarian allows hanging topics (cf. Cinque 1977), or EXTERNAL-TOPIC (cf. Aissen 1992, Kiss 1994:80; also King 1995 for Russian), for which we account by the following annotated phrase structure rule:[18]

(26) EP → ({NP, PP, AP, SubjP}) CP
(↑E-TOPIC)=↓ ↑=↓

Also, there is extensive evidence for fronted TOPICs[19] in a position preceding the complementizer (in principal an arbitrary number of TOPICs can be fronted; cf. Rudin 1994, 1990/1991, 1985:24-25). Consider example (27), which is accounted for by the proposed phrase structure rule (28).

(27) Toj kaza Marija če šte ja vidi.
He said Marija that will her see
He said that he will meet Maria.

(28) CP → {NP, PP, AP, SubjP}* C'
↓∈(↑TOPIC) ↑=↓

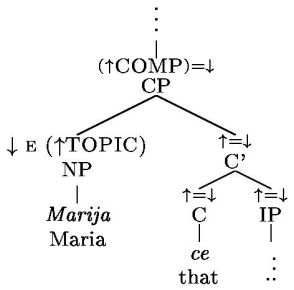
[17] See also Givon (1992:308-309) who claims that contrastive topics can be [+referring, +definite], or [–referring, –definite] but not [+referring, –definite].
[18] We use the abbreviation SubjP to refer to a subjunctive phrase.
[19] Throughout the paper, we use capital letters for DFs, which are part of the f-structure, and non-caps for IS-roles.

Finally, Bulgarian has a FOCUS-position following the TOPIC-position. In subordinate clauses the FOCUS – unlike the fronted topic(s) – follows the complementizer, as in example (29). We thus propose the two phrase structure rules presented in (30) and (31). We apply the annotated phrase structure rules (28), (30), and (31) to the examples in (27) and (29), and present the resulting partial c-structures under (31).

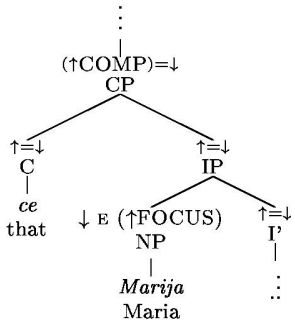
(29) Toj kaza če MARIJA šte vidi.
He said that Marija will see
He said that he will meet MARIJA.

(30) C' → C IP
↑=↓ ↑=↓

(31) IP → {NP, PP, AP, SubjP}* I'
↓∈(↑FOCUS) ↑=↓



c-structure 1



c-structure 2

The preliminary results of an online experiment[20] designed by us suggest that fronted, topical objects (i.e. not the hanging EXTERNAL TOPICs) are *always* doubled. Note that this still allows for *non-topical* fronted objects (i.e. FOCUS objects). Without going into further detail here, we assume that focused fronted objects can be distinguished from topical fronted objects by the different stress assigned to them. Our results are supported by the observations in Dimitrova-Vulchanova & Hellan (1998:xviii), and implicitly Avgustinova (1997:112). In order to capture this fact and Leafgren's (1997a,b, 1998) claim that CD always marks topicality of the doubled object, we propose that the syntactic topic position is assigned the following outside-in functional uncertainty equation. The rule in (32) is the updated rule from (28).[21]

(32) CP → {NP, PP, AP, SubjP}* C'
↓∈(↑TOPIC) ↑=↓
(↑XP* [GF])=↓

The DOC is identified as the direct object by its lexical semantics and the phrase structure rule for the predicate clitic cluster (see (8) above on p. 5). The agreement between the DOC and the doubled object guarantees that no spurious ambiguities are predicted, even in the case of multiple object fronting. Below we give a representative lexical entry for *ja*, the 3.SG.FEM form of the DOC.

[20] Human Subjects Application #0102-655, approved by the Human Subjects Panel, Stanford. The experiment can be found at http://symsys.stanford.edu/experiment/. In this experiment subject where asked to judge Bulgarian sentences after being primed for colloquial spoken language. All judgments were elicited using magnitude estimation, i.e. subjects were asked to assign a gradual value for the "goodness" of each sentence in respect to an always present reference sentence.

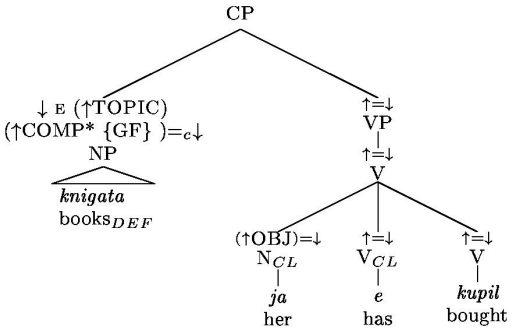
[21] The squared parentheses are a convention used to express that the bracketed part of the equation is not defining (Dalrymple, p.c.). Note that (↑XP* [GF])=↓ ≡ (↑[XP* GF])=↓. A similar rule seems to be necessary for FOCUS-fronting but in that case the whole equation is defining.

```
ja: N_CL - DOC
    (OBJ ↑)
    (↑PERS) = 3
    (↑NUM) = SG
    (↑GEN) = FEM
```

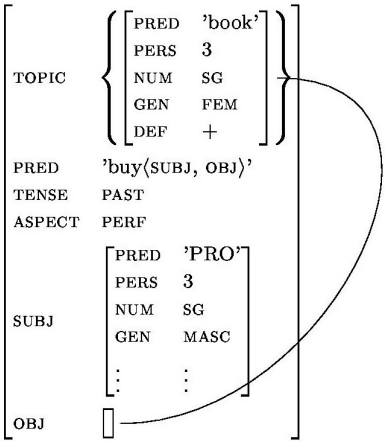
Figure 1 Simplified lexical entry for the DOC *ja* (preliminary version)

The proposal predicts that fronted objects must be doubled since the DOC is the only way to *define* the object function without violating UNIQUENESS (ignoring UNIQUENESS, one could wrongly generate a second object in the VP to satisfy COHERENCE and COMPLETENESS).[22] As an example, consider the sentence in (33) with subject pro-drop and a fronted topical object. The corresponding c- and f-structure are given below.[23] We leave it to the reader to convince herself that the f-structure is the only predicted one in our account.[24]

(33) Knigata ja e kupil.
book$_{PL.DEF}$ DOC$_{3.SG.FEM}$ AUX$_{3.SG}$ bought
The books, he has bought.



c-structure 3



f-structure 1

Crucially, our proposal captures the intuition that it is the absence or presence of a fronted topical object that causes obligatory CD. However, Leafgren argues that the following two generalization hold (the second point is also supported by Vakareliyska 1994:125):

(34) *All* doubled objects are topics.
(35) Object doubling is *always* optional.

In other words, CD is just one option of identifying an object as topical.[25] Unfortunately, Leafgren (1997a,b,c, 1998, 2001, 2002) does not formalize his working definition of 'topic' any

[22] If required we can rule out generation of the DOC in the normal object location by phonological rules like the Tobler-Mussafia effect (see section II).
[23] Due to formatting reasons, we use curly brackets in the tree where we use the standard notation, i.e. square parentheses, in the phrase structure (32) rule above.
[24] Note that the subject function is defined through the verbal subject agreement morphology (including an optional PRED PRO since subject drop is common in Bulgarian), so that subjects, too, can be in the fronted position.
[25] Leafgren (2001:4) shows that in 1200 object occurrences, 0% of the non-topical objects are doubled. This contrasts with 10.8% doubled topical objects in spoken Bulgarian. Alternative means of topical object marking depend on the register. In informal, spoken Bulgarian, speakers may also use marked word order (i.e. object-topic fronting) or intonation or just not mark the topicality of the object when the context unambiguously identifies the object to be the topic (cf. Leafgren 1997b:128). In more formal registers, passivization or impersonal reflexive constructions can be used to mark that the semantic object is topical (cf. Leafgren 2001).

more precisely than 'What a clause provides or requests information about' (cf. Leafgren 1997a:127, referring to Sgall 1975:303; see also Sgall 1993). Leafgren gives the following example to illustrate his topic definition:[26]

(36) Vanja$_i$ ne ja$_i$ vâlnuvat tezi nešta ...
Vanja NEG DOC$_{3.SG.FEM}$ worry$_{3.PL}$ these things
These things don't worry Vanja ...

The generalization in (35) conflicts with Dimitrova-Vulchanova & Hellan's (1998) and our own observations. Leafgren's work is based on a corpus study of more then 7,000 object occurrences in written texts (1997a,b; including ~200 cases of CD), more than 3,000 object occurrences in spoken texts (1998: including ~200 cases of CD), and a comparative study of 1,200 object occurrences each in informal oral, formal oral, and formal written texts (Leafgren 2001). In light of such extensive evidence, we should try to resolve the mismatch between Leafgren's and our observations. There are two main sources for this mismatch aside from the apparent problem with informal topic definitions. First, although Leafgren (2001) considers informal oral texts, (35) is based on Leafgren's (1997a,b) work on written corpus (consisting of 2 novels and 2 short stories). The online experiment done by us (cf. above) aims at judgments about informal contemporary spoken Bulgarian. Secondly, and more importantly, Leafgren does not control for fronted *focused* phrases. Actually, Leafgren (1997a:132) explicitly allows for topics to be "focused" (in his terminology). Although this admittedly has to be done at some point, it is not the purpose of this paper to determine the exact semantic and/or pragmatic function of what we have called 'topic' so far (for Bulgarian). Here the crucial point is that Bulgarian seems to have two sentence initial positions, here labeled TOPIC and FOCUS (see above) that can be distinguished in terms of the stress contours that go along with them. Thus we have an independent motivation for those two positions[27], which we label TOPIC and FOCUS. One of those two positions, namely TOPIC, requires CD if it is filled by an object. Thus the distinction of TOPIC and FOCUS allows us to capture a generalization, which Leafgren misses, without additional stipulation. For simplicity's sake, we will assume that the TOPIC and FOCUS position each encode at least their corresponding I[nformation] S[tructural] roles, namely topic and focus (again, here we are not concerned with the meaning of the two IS-roles). Somewhat more formally, this constraint can be stated as in (37), where DF is the set of f-structure features that encode discourse functions, and for a given input DF the function IS-role(DF) yields the corresponding IS-role (e.g. topic for TOPIC).

(37) $X \in \text{DF} \Rightarrow x \in$ IS-role(DF), where X is the f-structure correspondence of a linguistic form w, and x is the denotation of w.

Similarly to EXTENDED COHERENCE (cf. Bresnan 2000), we can formulate a constraint INFORMATION PACKAGING COHERENCE that guarantees that the generalization in (37) holds for all DFs of an f-structure.

INFORMATION PACKAGING COHERENCE (IPC) - preliminary version
(38) An F-structure FS fulfills IPC iff every discourse function DF in FS fulfills (37).

Similarly to other authors (e.g. Choi 1999), we assume an IS-component which has interfaces not only to f-structure but also to the Prosodic Structure (PS) and the Lexical Structure (LS), see Figure 2. Here we are not interested in the interface between PS and IS but in the interface between

[26] Note one important detail in Leafgren's definition. The topic is defined on the level of a clause, not a sentence. This allows for topics in, for example, subordinate clauses. Examples like (27) above clearly show that this is necessary.
[27] More precisely, we have a motivation for a formal distinction, which we choose to capture in terms of c-structure position.

LS and IS.[28] Note that the one-way implication of (37) works to our advantage. While we want every phrase that is fronted to TOPIC to be part of the IS-topic, we want to allow for non fronted constituents to bear the role of the IS-topic, too.

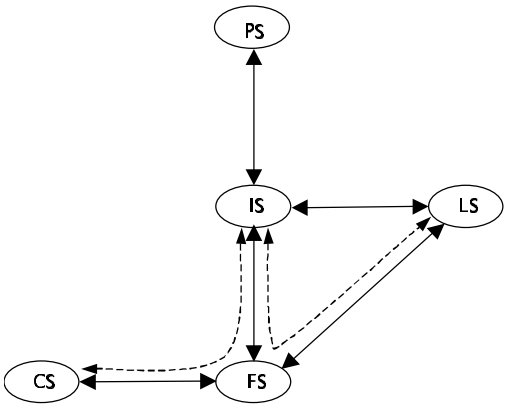


Figure 2 Relations between parts of the grammar that are relevant for IS.

Now with the proposed model of IS, LS, and FS interaction in mind, we can restate Leafgren's generalizations in (34) and (35) more precisely as (39) - (41).

(39) *All* doubled objects are *IS*-topics.
(40) TOPIC objects *must* be doubled.
(41) IS-topic objects *can* be doubled.[29]

In order to predict that CD implies that the doubled object is part of the IS-topic, i.e. to guarantee (39), we have to slightly modify the lexical entry of the DOC(s). Again, the 3.SG.FEM form *ja* is given as a representative example in Figure 3. The upwards-pointing arrow with the subscript 'IS' indicates that the referent identified by the DOC is mapped onto information structure (where it is identified as a part of the IS-topic).

ja: N_{CL} - DOC
($\uparrow_{IS} \in topic_{IS}$)
(OBJ↑)
(↑PERS) = 3
(↑NUM) = SG
(↑GEN) = FEM

Figure 3 Revised lexical entry for the topic-marking DOC *ja*.

The lexical entry in Figure 3 together with the revised annotated phrase structure rule for the TOPIC position in (32) captures all of the above-mentioned generalizations, (39) - (41), and therefore resolves the apparent conflict between Leafgren's work and e.g. Dimitrova-Vulchanova & Hellan's claims. Moreover, our account predicts optional CD for fronted FOCUS-objects as long as they are part of the IS-topic. If we adopt a two-dimensional IS-component[30], following Choi (1999), implicitly Leafgren (1997a,b), a.o., this is not surprising at all. Indeed, reduplication of fronted FOCUS objects can be observed in Bulgarian. First, CD of fronted object wh-phrases (cf. Jaeger

[28] In a model like the one presented here, encoding of IS through CS (and therefore within LFG through F-structure, FS) corresponds to what is commonly called discourse configurationality (henceforth, DF-configurationality).
[29] Here we do not address the pragmatic factors which determine in which contexts speakers tend to make use of this mechanism (CD to mark IS-topicality of the object). See Givon (1987) for a general discussion of this.
[30] By two dimensional, we mean that there is not only one dimension along which information structural roles differ e.g. topic-comment or link-tail-focus (cf. Vallduví 1993, 1992). Instead informational structural roles differ along two dimensions, e.g. they can be [+/- prominent] and [+/- given] (cf. Choi 1999).

2002; see also Dimitrova-Vulchanova & Hellan 1998:xxi-xxii), which are usually considered to be in FOCUS (see, for example, Rudin et al. 1998), is possible.

(42) Kogo kakvo go iznenada?
whom what DOC$_{3.SG.MASC}$ suprised$_{3.SG}$
Whom did what surprise?

Second, of all non-wh, focused constituents, only contrastive topics can be doubled. The proposal presented here therefore accounts, among other things, for the fact that contrastive topics can be doubled, a fact that Avgustinova's (1997) analysis of CD cannot straightforwardly account for since she employs the one dimensional IS-component proposed in Vallduví (1992, 1993).

As mentioned in the introduction, one aim of this paper is to provide a formal account for CD and its interaction with word order and IS. The current section has done exactly this. Second, we wanted to resolve the discrepancy between the different empirical approaches to Bulgarian CD and the theoretical literature. For one part, we have already done this by resolving the mismatch between our own empirical studies, Leafgren's work and the theoretical literature on CD and DF-configurationality in Bulgarian. We did this by distinguishing between two independently motivated phrase structural positions and their correspondences in the IS-component. The analysis resulting from this is able to capture both the generalization from the extensive empirical work and predicts the right restrictions resulting from certain word orders (i.e. obligatory CD of TOPIC objects). Next, we use the second source of data for spoken Bulgarian mentioned above, Avgustinova's (1997) elicited question-answer pairs, to briefly test if the presented proposal makes correct predictions about possible word orders beyond the fronted TOPIC construction.

It is beyond the scope of this paper to provide a detailed analysis for all of the patterns (i.e. word order-intonation-information structure mappings) described by Avgustinova (1997:112). Although this issue is open for further research, we suggest that Bulgarian has some kind of 'default ordering' within the flat VP (see above, phrase structure rule (11) in section II). Among other features, such as definiteness, person, referentiality, etc., topicality of a phrase seems to be one – maybe the major – determining factor for the constituent order with the VP.[31] Leafgren (1997c:5ff.) shows that topic-before-comment seems to be the more important ordering mechanism in Bulgarian than subject-before-object or agent-before-patient, both in terms of frequency[32] and in that all violations of the two other conditions serve to satisfy the topic-before-comment condition or another discourse or information structure constraint (e.g. CD and object fronting). The assumption of a default order similar to the one suggested by the Prague school (cf. Functional Sentence Perspective, henceforth FSP; Sgall 1993) but only applied to the flat VP instead of the whole clause explains why a certain default constituent order can be observed in Bulgarian while, at the same time, only a few strict rules (like the above-mentioned TOPIC object fronting) seem to hold. We ask the reader to keep in mind the notion of default ordering as just described during our discussion of Avgustinova's (1997) data.

Apart from direct object fronting, which results in OSV and OVS orders (for the sake of simplicity, we only consider transitive verbs here), there is one other word order that usually requires the DOC, namely VOS. According to Avgustinova (1997), VOS is possible with either $V_{FOCUS}O_{topic}S_{topic}$ or $VO_{topic}S_{focus}$.[33] Here, $VO_{topic}S_{focus}$ is predicted by FSP default word ordering working on a non-configurational VP. The same reasoning applies to $V_{FOCUS}O_{topic}S_{topic}$. Given this, we should expect $V_{FOCUS}S_{topic}O_{topic}$ to be equally acceptable, if the object and the subject are equally

[31] Note that this is not uncommon at all. It has long been known that scrambling in languages like e.g. German or Japanese is sensitive to the above-mentioned categories. Furthermore, especially topicality of phrases has been shown to play a role in determining the word order in several languages (cf. Choi 1999 for German and Korean; Ishihara 2000 for Japanese).

[32] Topic-before-comment ordering holds in 91.0%, subject-before-object in 89.5%, and agent-before-patient in 88.3% of the cases. The correlation between the three scales explains why the numbers are so close.

[33] Recall our convention to use capital letters for DFs (as part of the f-structure) and lowercase letters for IS-roles. Since Avgustinova (1997) does not make a comparable distinction, the annotation is our translation of her classification.

topical. Indeed, Avgustinova's data set contains examples for this word order. For both cases of a FOCUSed verb, topical subject and topical object, the clitic marks which of the NPs is the object. Furthermore, for both SOV and VSO the clitic is at least possible (if not preferred) *if and only if* the object is part of the topic.

Thus, in addition to what we said above, the proposal presented here accounts for the experimentally elicited word orders listed in Avgustinova (1997). Although a detailed syntactic analysis of all possible word orders has to be left to further research, we have sketched an analysis of the DOC in CD and its interaction with word order and information structure. We will refer to this use as '(direct object) topic agreement marker' usage. This label makes reference to Bresnan & Mchombo (1987) who distinguish grammatical and anaphoric agreement markers. We now turn to a second function of the DOC, its use as 'default' pronoun, and then show that our proposal makes the right predictions about the occurrences of those two different functions of the DOC.

V The DOC as default pronoun

Although this is not a salient topic in the literature on the Bulgarian object clitics (for an exception see Vakareliyska 1994:125), there is no doubt that the DOC has another use as the default pronoun. To further clarify what we mean by *default* and to illustrate the relation between the two types of pronouns, consider the following dialogue, where (44) but not (45) is a possible continuation of (43) if no contrast is intended:

(43) "Karl sreštna onazi tancjorka včera"
Karl met$_{3.SG}$ that dancer yesterday
Karl met that dancer yesterday.

(44) "Ivan sâšto ja (*neja/*NEJA) poznava."
Ivan too DOC$_{3.SG.FEM}$ her/HER knows$_{3.SG}$
Ivan knows her, too.

(45) #"Ivan sâšto poznava neja/NEJA."
Ivan too knows$_{3.SG}$ her/ HER
Intended: Ivan knows her too.

Since this has not been done by others, we tested for the possibility that all cases of the DOC as alleged default pronoun might be due to (topic) object drop. For some more details on the test, we refer the reader to our handout (Jaeger & Gerassimova 2002:10). Here we will just mention that, like English, Bulgarian allows specific and unspecific object drop (depending on the verb, cf. Fillmore 1986). We found that there are still cases left where object pro-drop is not possible and the DOC is the only realization of the object in the sentence. Thus we are forced to assume that there is one variant of the DOC with a PRED PRO. For a formal LFG analysis, this raises the question whether there are two entries for each DOC or one with an optional PRED PRO. Consider the hypothesis that there is one DOC with an optional PRED PRO. In that case, the default pronoun use of the DOC would always result in the object (i.e. the clitic itself) being marked as topic. It is not clear whether this is desirable, although one could argue that all pronouns have to be topical in some sense anyway, since their referent is 'salient' (cf. Chafe 1976) most of the time (in order to be identifiable). For now, it may be better to think of two separate lexical entries for the DOC, one with an optional PRED PRO (the default pronoun) and one with the topic equation. Again, this is illustrated for *ja*.

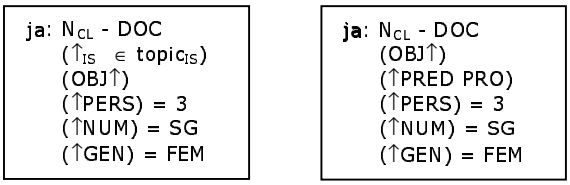


Figure 4 Revised lexical entries for the DOC *ja*.

The existence of two lexical entries poses the question of how our proposal can guarantee the right use of DOC for a given sentence. So far, because of the functional control established by the TOPIC position, the optional PRED PRO use is ruled out by UNIQUENESS whenever a fronted (object) constituent sits in a TOPIC position. Whenever the object is realized within the VP, UNIQUENESS again rules out two PRED values for the object, since the DOC *defines* the object (instead of just constraining it). With no other object constituent being realized, the DOC is interpreted as object (pronoun). In our account, this is guaranteed by (EXTENDED) COHERENCE. Next, we show that data from topicalization out of islands further support this analysis.

VI Island data: When can which type of DOC occur?

In this section, we show how our proposal makes the right predictions about the distribution of the two uses of the DOC (i.e. as default pronoun and as topic agreement marker). Rudin (1985) shows that NPs and PPs (whether complex or not) are islands to any kind of extraction in Bulgarian. Most of the other classical islands, however, do not seem to be islands in Bulgarian. This is supported by the preliminary results of our still ongoing online experiment (see above). Consider, for example, the following cases of topicalization:

(46) Todor e jasno, [$_{CP}$ če Ivan *(go) e vidjal].
Todor is clear that Ivan DOC$_{3.SG.MASC}$ is seen
Todor it is clear that Ivan has seen him.

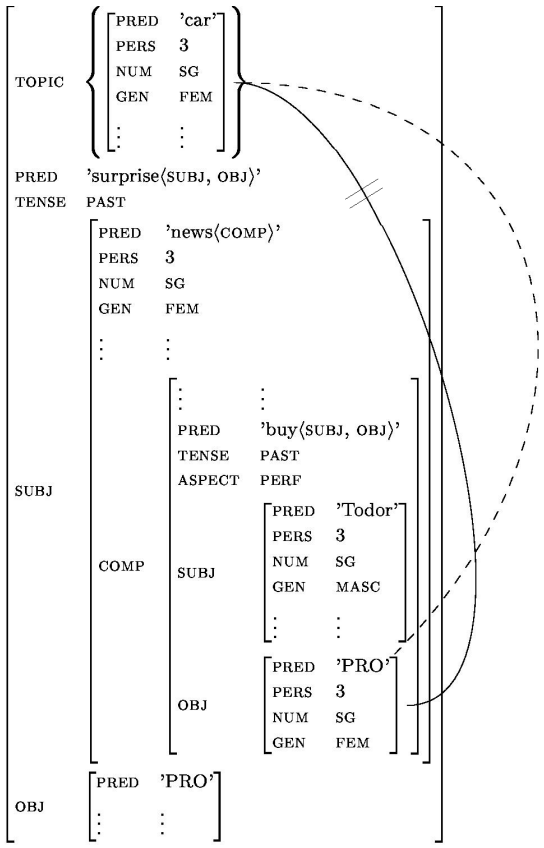
(47) Jabălkite [$_{NP}$ čovekyt, [$_{CP}$ kojto *(gi) donese]], e pilot.
apples$_{DEF}$ man$_{DEF}$ who DOC$_{3.SG.NEUT}$ brought is pilot
The apples the man who brought (them) is a pilot.

Sentences (46) and (47) show that topicalization out of a sentential subject, in (46), and a relative clause, in (47), is possible. Just as in the case of simple fronting, the DOC is obligatory. Now consider topicalization out of an island (here, an NP):

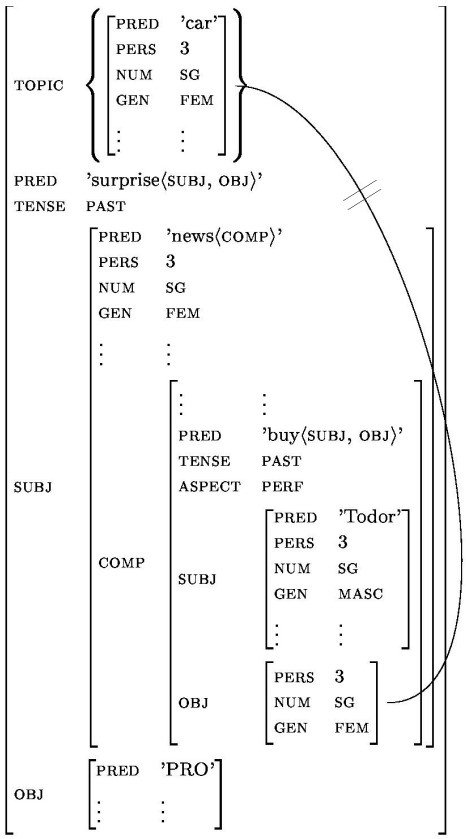
(48) *Kolata [$_{NP}$ novinata, [$_{CP}$ če Todor (ja) e kupil]], ni učudi.
car$_{DEF}$ news$_{DEF}$ that T. DOC$_{3.SG.NEUT}$ is bought us surprised
Intended: The car the news that Todor has bought (it) surprised us.

Regardless of whether the DOC is realized, sentence (48) is ungrammatical. According to Bresnan & Grimshaw (1978), filler-gap dependencies (i.e. functional control within LFG), but not anaphoric binding, obey island constraints. The DOC does not repair island-violations. In our account, the ungrammaticality of (48) is explained as follows. The fronted constituent can only satisfy the outside-in functional uncertainty equation (and thereby EXTENDED COHERENCE) if it is functionally controlled by a GF-bearing constituent further down in the f-structure. The fronted object cannot be functionally controlled by a constituent with a PRED value because this would violate UNIQUENESS. The DOC cannot be realized in the embedding sentence to bind the fronted object since the object function of the embedding sentence already has an object (with a PRED value). Finally, the DOC with a PRED PRO (i.e. the default pronoun) could be realized in the embedded clause in order to satisfy COMPLETENESS and COHERENCE. However, the outside-in functional uncertainty equation of the fronted object would still have to be resolved. This is not possible since the embedded GF (i.e. the direct object) is not accessible – it is in an island (cf. f-structure 2)[34]. F-structure 3 is out for the same reason – because the functional control violates the island condition (cf. above, Bresnan & Grimshaw 1978).

[34] Anaphoric binding is indicated by dotted lines, functional control by solid lines. The doubled crossed line stands for an island violation, which results in an invalid f-structure.



f-structure 2



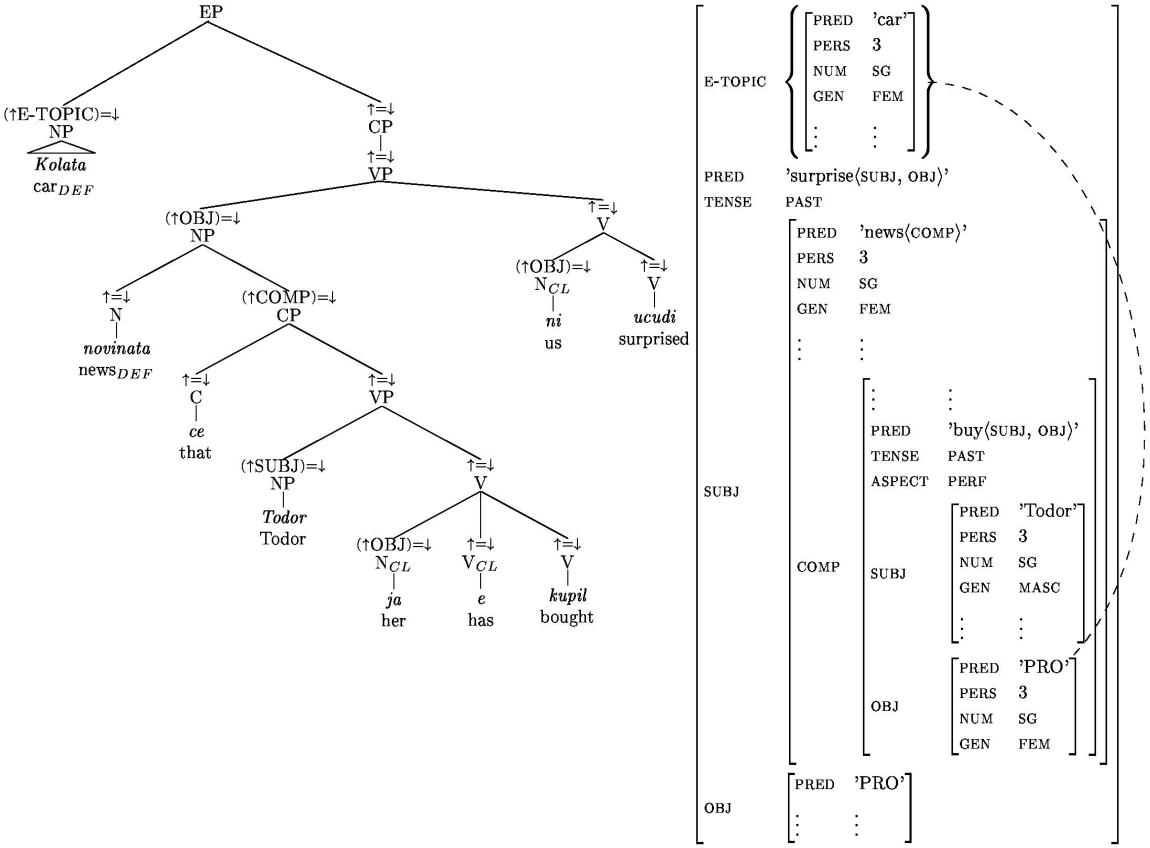
f-structure 3

The grammaticality of (46) and (47) is predicted, too. The fronted object is functionally controlled by the DOC, which has to be realized because it is the only constituent that agrees in person, number, and gender with the fronted object. The DOC with a PRED PRO cannot be chosen because this would violate UNIQUENESS.

There is one more phenomenon that supports our analysis: EXTERNAL TOPICs. Example (49) – if uttered with a clear pause between the fronted object and the following sentence – is grammatical. This kind of a detached constituent fulfills the criterion of a hanging topic (cf. Cinque 1977) or EXTERNAL TOPIC.

(49) Kolata PAUSE novinata, če Todor ja e kupil, ni učudi.
car$_{DEF}$ PAUSE news$_{DEF}$ that Todor DOC$_{3.SG.NEUT}$ is bought us surprised
The car, .. the news that Todor has bought (it) surprised us.

In the account presented here, the grammaticality of (49) follows from the fact that the phrase structure rule for EXTERNAL TOPICs, see (26) above, is not annotated with an outside-in functional uncertainty equation. Therefore, no functional control violates the island constraint and the DOC with a PRED PRO is realized in the embedded clause. To satisfy EXTENDED COHERENCE, the PRED PRO anaphorically binds the EXTERNAL TOPIC. This is illustrated by the c- and f-structure given below.



c-structure 4 **f-structure 4**

To sum up, the island data presented above is not only compatible with our theory but also predicted by it. In the next and final section, we summarize our analysis and list some open questions.

VII Conclusions and Outlook

We have shown how two functions/uses of the DOC interact. The DOC is a grammatical (direct object) agreement marker and the default pronoun of Bulgarian. In contrast to the object marker in Chicheŵa (cf. Bresnan & Mchombo 1987: 745), the Bulgarian DOC does not mark an f-structure TOPIC but an IS-topic. This insight helps to position the DOC within a typology of (object) markers. The DOC's object topic-marking function, in interaction with the proposed annotated phrase structure rules (i.e. especially the functional control of fronted topic), accounts for both obligatory TOPICalized object doubling and optional doubling of topical objects in general.

Our account stresses that linguistic forms can have several (independent) functions. This is even more evident when we consider that the DOC has a third function as intrusive pronoun, as mentioned in the introduction. First results of an ongoing online experiment on the intrusive pronoun DOC in extractions support our analysis. Those results will have to be fully incorporated into a complete account of the DOC.

The optionality of CD in many contexts shows that speakers have different options of coding e.g. a topical object depending on the register, genre and maybe other factors (see Leafgren 1997a, 2001, 2002 for a similar thought). Possible generalizations relating the choice of forms to their functions and other factors, such as register, merit further investigation. For example, does the absence of intonation in written language enforce the use of alternative linguistic means (such as more strict case marking in the case of otherwise optional case marking, or more strict word order) to identify GFs and DFs.

Also, we have provided one further example of a (non-dependent-marking) language which seems to compensate lack of GF-configurationality by morpho-syntactic means (head-marking). Although subject to further testing, the presented analysis is supported by a broad empirical basis. In addition to the native speaker intuitions of one of the authors (V.G.), the analysis accounts for data from Leafgren's (1997a, 1997b) corpus-based studies, Avgustinova's (1997) elicited question-answer data (more than 20 word order-prosody mappings for a transitive verb), and the data collected in our online experiment. To the best of our knowledge, unlike all other formal accounts so far (e.g. Rudin 1997, 1996, 1990/1991, 1985, Dyer 1992, Avgustinova 1997, Dimitrova-Vulchanova & Hellan 1998, Franks & King 2000), the account presented above predicts the obligatory CD in the case of fronted objects and provides a possible explanation for the optionality of CD in other cases. For example, Rudin's (1997) MP analysis of the DOC as a AgrO-head cannot predict why the DOC is obligatory in certain cases yet optional in others. Furthermore, we explicitly addressed the relatively free word order of Bulgarian and predicted the resulting word order depending on the IS-roles assigned to the different phrases. Although empirically attested, many of the word orders discussed at the end of section IV are ignored in most of the theoretical literature on Bulgarian.

While our analysis accounts for all observed word orders (including predictions about prosody via proposed constraints on the IS, e.g. via IPC, cf. (38) in section IV), it does not predict spurious parses or ambiguities arising from the lexical ambiguity of those two uses. The account presented here could therefore close the gap between the work on DF-configurationality and free word order in Bulgarian. It also is a first step to resolve the mismatch between the broad-coverage empirical work on Bulgarian and the literature on formal aspects.

Further research is necessary in order to see how the different functions of the DOC relate to each other. We also think that it is worth to investigate if there are further restrictions on the optional or obligatory presence of the DOC in certain contexts. For example, there are still possible mismatches in the observations made by Avgustinova (1997) and Leafgren (1997a,b) regarding the question in exactly which contexts the DOC is obligatory. Once we have a better picture of all the factors that determine the possible word orders for a given context, a (stochastic) OT account may be able to combine those factors into a formal description of the data. Related to this, it is very interesting that those dimensions which are strict factors in Bulgarian CD (i.e. specificity and topicality), seem to have occurred subsequently in the diachronic development of the much more general Macedonian CD and show up as statistic preferences in contemporary Macedonian CD (as a careful reading of Čašule 1997 suggests). Finally, another phenomenon that needs further research is the CD of quantified NPs. While we have shown how quantified NPs confirm that [-specifics] cannot be doubled (cf. section III), the details of CD of quantified NPs are yet to be worked out.

VIII References

AG, ACADEMY GRAMMAR (edited by Popov, Konstantin)

- (1983): *Grammatika na sâvremennija bâlgarski knižoven ezik, t. 3. sintaksis*; Sofia, Bâlgarska Akademija na Naukite.

AISSEN, JUDITH

- (1992): *Topic and Focus in Mayan*; in: *Language 68,* p. 43-80.

ALEXANDROVA, G.

- (1997): *Pronominal Clitics as g(eneralized) f(amiliarity)-licensing AGR^0*; in Browne, W. / Dornisch, E. / Kondrashova, N. / Zec, D. (eds.): *Formal Approaches to Slavic Linguistics: The Cornell Meeting 1995*, pp. 1-31; Ann Arbor, Michigan Slavic Publications.

ARNAUDOVA, OLGA

- (2001): *Prosodic movement and information focus in Bulgarian*; in: *Proceedings of Formal Approaches to Slavic Linguistics 9*; Michigan Slavic Publications.

AVGUSTINOVA, TANIA

- (1997): *Word Order and Clitics in Bulgarian*. in: *Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 5*. DFKI, Saarbrücken.

AVGUSTINOVA, TANIA & BISTRA ANDREEVA
- (1999): *Thematic intonation patterns in Bulgarian clitic replication*; in: *ICPhS99, San Francisco*, pp. 1501-4.

BAKER, MARK C.
- (1991): *On some subject/object asymmetries in Mohawk*; in: *Natural Language and Linguistic Theory 9*, pp. 537-576.

BILLINGS, LOREN A.
- (2000): *Phrasal Clitics*; in: *Slavic Linguistics 9(2), special issue: Festschrift for Leonard Babby*.

BRESNAN, JOAN
- (2001): *Lexical Functional Syntax*; Blackwell Publishing, Cambridge.

BRESNAN, JOAN / GRIMSHAW, JANE
- (1978): *The Syntax of Free Relatives in English*; in: *Linguistic-Inquiry* 9(3), pp. 331-391.

BRESNAN, JOAN / MCHOMBO, SAM A.
- (1987): *Topic, Pronoun, and Agreement in Chichewa*; in: *Language, Volume 63, Number 4*, pp. 741-782.

ČAŠULE, ILIJA
- (1997): *Functional Load of Short Pronominal Forms*, in: *Journal of Slavic Linguistics 5(1)*. 3-19.

CHAFE, WALLACE, L.
- (1976): *Givenness, Contrastiveness, Definiteness, subjects, Topics, and Point of View*; in Li, Ch. N. (ed.): *Subject and Topic*, pp. 27-55; New York, Academic Press.

CHOI, HYE-WON
- (1999): *Optimizing Structure in Context. Scrambling and Information Structure*; in: *Dissertations in Linguistics*; Stanford, CSLI Publications.

CINQUE, GUGLIELMO
- (1977): *The Movement Nature of Left Dislocation*; Linguistic Inquiry, 8, 2, pp. 397-412.

CYXUN, GENADIJ A.
- (1962): *Mestoimennata enklitika i slovoredât v bâlgarskoto izrečenie*; in: *Bâlgarski ezik 12(4)*, pp. 283-91.

DIMITROVA-VULCHANOVA, MILA / HELLAN, LARS
- (1998): *Introduction*; in Dimitrova-Vulchanova & Hellan (eds.): *Topics in South Slavic Syntax and Semantics*, pp. ix-xxvii; Amsterdam & Philadelphia, John Benjamins Publishing.

DYER, DONALD L.
- (1993): *Determinedness and the Pragmatics of Bulgarian Sentence Structure*; in: *Slavic and East European Journal, Volume 37, Number 3*, pp. 273-292.
- (1992): *Word Order in the simple Bulgarian Sentence: A Study in Grammar, Semantics and Pragmatics*; Amsterdam – Atlanta, Rodovi B. V.

EMBICK, DAVID / IZVORSKI, ROUMYANA
- (1994): *On long head movement in Bulgarian*; in: *Proceedings of the Eleventh Eastern States Conference on Linguistics 1994*, pp. 104-115; Ithaca , Cornell University.

FILLMORE, CHARLES J.
- (1986): *Pragmatically Controlled Zero Anaphora*; in: *BLS 12*, pp. 95-107.

FRANKS, STEVEN / KING, TRACY HOLLOWAY
- (2000): *A Handbook of Slavic Clitics*; Oxford, Oxford University Press.

FRASER, BRUCE
- (1988): *Types of English discourse markers*; in: *Acta Linguistica Hungarica 38*, pp. 19-33.

GEORGIEVA, ELENA
- (1974): *Slovored na prostoto izrečenie v bălgarskija knižoven ezik*; Sofia, Bălgarska Akademija na Naukite.

GERASSIMOVA, VERONICA A. / JAEGER, T. FLORIAN
- (2002): *Configurationality and the Direct Object Clitic in Bulgarian*; in Nissim, M. (ed.): *Proceedings of the Seventh Student Session of the ESSLLI 2002 in Trento, Italy, August 5th – 16th, 2002*, Chapter 6.

GIVÓN, TALMY
- (1992): *On Interpreting Text-Distributional Correlations. Some Methodological Issues*; in: Payne, D. L. (ed.): *Pragmatics of Word Order Flexibility*, pp. 305-320; Amsterdam & Philadelphia: John Benjamins Publishing.
- (1987): *The Pragmatics of Word-Order: Predictability, Importance and Attention*; in Hammod, M. et al (eds.): *Studies in Syntactic Typology* (=*Typological Studies in Language 17*), pp. 243-284; Amsterdam: J. Benjamins Publishers.
- (1976): *Topic, Pronoun, and Grammatical Agreement*; in Li, Ch. N. (ed.): *Subject and Topic*, pp. 149-188; New York & London, Academic Press Inc.

GUENTCHEVA, ZLATKA
- (1994): *Thématisation de l'objet en bulgare*; Bern: Peter Lang.

ISHIHARA, SHINICHIRO
- (2000): *Stress, Focus, and Scrambling in Japanese*; in: *MIT Working Papers In Linguistics 39*, pp. 142-175.

IVANČEV, SVETOMIR
- (1978): *Prinosi v bălgarskoto i slavjanskoto ezikoznanie*; Sofia, Nauka i izkustvo.
- (1968): *Problemi na aktualnoto členenie na izrečenieto*; in Ivančev (1978:173-84).
- (1957): *Nabljudenija vârxu upotrebata na člena v bâlgarski ezik*; in Ivančev (1978:128-52).

IZVORSKI, ROUMYANA
- (1994): *On the Semantics of the Bulgarian "Indefinite Article"*; in: *Formal Approaches to Slavic Linguistics: The MIT Meeting 1993*, pp. 235-254; Ann Arbor: Michigan Slavic Publications.

JAEGER, T. FLORIAN
- (2002): *Multiple Wh-questions in Bulgarian*; draft, available at http://www.stanford.edu/~tiflo/ as by 10/2002.

JAEGER, T. FLORIAN & VERONICA A. GERASSIMOVA
- (2002): *Bulgarian word order and the role of the direct object clitic in LFG*; handout for the *LFG02, Athens, July 3rd-5th, 2002*, available at http://www.stanford.edu/~tiflo/ as by 10/2002.

KAZAZIS, KOSTAS & JOSEPH PENTHERADOUKIS
- (1976): *Reduplication of indefinite direct objects in Albanian and Modern Greek*; in: *Language 52(2)*, pp. 398-403.

KING, TRACEY H.
- (1995): *Configuring Topic and Focus in Russian*. in: *Dissertations in Linguistics*; Stanford: CSLI Publications.

KISS, KATALIN E.
- (2001): *Discourse configurationality*; in Haspelmath, M. / Koenig, E. / Oesterreicher, W. / Raibler, W. (eds.): *Language Typology and Language Universals*, pp. 1442-1455; Berlin & New York, de Gruyter.
- (1995): *Introduction* in Kiss, K. E. (ed.): *Discourse Configurational Languages*, pp. 3-27; Oxford, Oxford University Press.
- (1994): *Sentence Structure and Word Order in Kiefer*; in F. / Kiss, K. E. (eds.): *Syntax and Semantics 27: The Syntactic Structure of Hungarian*, pp. 1-84; London, Academic Press.

LAMBOVA, MARIANA
- (2002): *Multiple topicalization wh-fronting in Bulgarian and the fine structure of the left-periphery*; in: *The 26th Penn Linguistics Colloquium.*

LAMBRECHT, KNUD
- (1994): *Information structure and sentence form. Topic, focus, and the mental representations of discourse referents*; Cambridge: Cambridge University Press.

LEAFGREN, JOHN R.
- (2002): *Register, Mode, and Bulgarian Object Placement*; presented at the Balkan Conference 2002, source: leafgren@email.arizona.edu.
- (2001): *Patient Packaging in Informal and Formal, Oral and Written Bulgarian*; presented at the AATSEEL 2001, New Orleans, source: leafgren@email.arizona.edu.
- (1998): *Object Reduplication in Spoken Bulgarian*; presented at the AATSEEL 1998, San Francisco, source: leafgren@email.arizona.edu.
- (1997c): *Topical Objects, Word Order, and Discourse Structure in Bulgarian*; presented at the AAASS 1997, source: leafgren@email.arizona.edu.
- (1997b): *Definiteness, Givenness, Topicality and Bulgarian Object Reduplication*; in: *Balkanistica 10*, pp. 296-311.
- (1997a): *Bulgarian Clitic Doubling: Overt Topicality*; in: *Journal of Slavic Linguistics, Volume 5, Number 1*, pp. 117-143.

MINČEVA, ANGELINA
- (1969): *Opit za interpretacija na modela na udvoenite dopâlnenija v bâlgarskija ezik*; in: Andrejčin, L. et at (eds.): *Izvestija na instituta za bâlgarski ezik 17*, pp. 3-50; Sofia: Bâlgarskata Akademija na naukite.

NORDLINGER, RACHEL
- (1998): *Constructive Case: Evidence from Australian languages*; in: *Dissertations in Linguistics*; Stanford, CSLI Publications.

PENČEV, JORDAN
- (1993): *Bŭlgarski Sintaksis: Upravlenie I Svŭrzvane*; Plodiv, Plovdivsko Universitetsko Izdatelstvo.
- (1984): *Stroež na bâlgarskoto izrečenie*; Sofia, Nauka i izkustvo.

POPOV, KONSTANTIN P.
- (1973): *Po njakoi osnovni vâprosi na bâlgarskija knižoven ezik*; Sofia: Narodna prosveta.
- (1963): *Sâvremenen bâlgarski ezik: Sintaksis*: Sofia: Nauka i izkustvo, 2nd edition.

POPOV, KONSTANTIN P. & VENČE S. POPOVA
- (1975): *Vâprosi na azikovata stilistika*; Sofia: Narodna prosveta.

RUDIN, CATHERINE
- (1997b): *Kakvo li e* li: *Interrogation and Focusing in Bulgarian*; in: *Balcanistica 10*, pp. 335-346.
- (1997a): *AgrO and Bulgarian Pronominal Clitics*; in: *Annual workshop on formal approaches to Slavic linguistics. The Indiana meeting 1996*, pp. 224-252; Michigan, Michigan Slavic Publications.
- (1996): *On Pronominal Clitics*; in Dimitrova-Vulchanova, M. / Hellan, L. (eds.): *Papers from the First Conference on Formal Approaches to South Slavic Languages, Volume 28*, pp. 229-246; University of Trondheim Working Papers in Linguistics.
- (1994): *On focus position and focus marking in Bulgarian questions*; in: *Papers from the Fourth annual meeting of the Formal Linguistic Society of Midamerica, April 15-18, 1993*, pp. 252-266, Iowa City.
- (1990/1991): *Topic and Focus in Bulgarian*; in: *Acta Linguistica Hungarica, Vol. 40 (3-4)*, pp. 429-449.
- (1989): *Multiple questions south, west and east: A Government-Binding approach to the typology of wh-movement in Slavic languages*; in: *International Journal of Slavic Linguistics and Poetics 39-40*.
- (1988b): *On multiple questions and multiple WH fronting*; in: *Natural Language and Linguistic Theory 6*, pp. 445-502.
- (1988a): *Multiple Question in South Slavic, West Slavic and Romanian*; in: *Slavic and East European Journal, Volume 32, Number 1*, pp. 1-24.
- (1985): *Aspects of Bulgarian Syntax: Complementizers and WH Constructions*; Columbus, Ohio: Slavica Publishers.

SCHIFFRIN, DEBORAH
- (1988): *Discourse markers*; Cambridge & New York: Cambridge University Press.

SELLS, PETER
- (1984): *Syntax and semantics of resumptive pronouns*; Ph.D. Thesis, University of Massachusetts at Amherst.

SIEWIERSKA, ANNA / UHLIŘOVÁ, LUDMILA
- (1998): *Word order in Slavic languages*; in Siewierska, A. (ed.): *Constituent Order in the Languages of Europe*, pp. 105-150; Berlin/New York, Mouton de Gruyter.

SGALL, PETR
- (1993): *The Czech Tradition*; in Jacobs, J. / von Stechow, A. / Sternefled, W. / Vennemann, T. (eds.): *Sytnax. Ein internationales Handbuch zeitgenoessisches Forschung – An international Handbook of Contemporary Research*, pp. 349-368; Berlin, New York, de Gruyter.
- (1975): *Conditions of the use of sentence and a semantic representation of topic and focus*; in Keenan, E.L. (ed.): *Formal semantics of natural language*, pp. 297-312; Cambridge: Cambridge University Press.

TOMIĆ, OLGA MIŠESKA
- (1997): *Non-initial as a default clitic position*; in: *Journal of Slavic Linguistics, Volume 5*, pp. 301-323.
- (1996): *The Balkan clausal clitics*; in: *Natural Language and Linguistic Theory, Volume 14*, pp. 811-872.

VAKARELIYSKA, CYNTHIA
- (1994): *"Na"-drop in Bulgarian*; in: *Journal of Slavic Linguistics, Volume 2, Number 1*, pp. 121-150.

VALLDUVÍ, ENRIC
- (1993): *Information Packaging: A Survey*; MS. Centre for Cognitive Science and Human Communication Research Centre , University of Edinborough.
- (1992): *The Informational Component*; New York, Garland.

PARTICIPLE-ADJECTIVE FORMATION IN MODERN GREEK

Valia Kordoni

Dept. of Computational Linguistics, University of Saarland,
P.O. BOX 15 11 50, D-66041 Saarbrücken, GERMANY
kordoni@coli.uni-sb.de

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

1 Introduction

This paper deals with Modern Greek (henceforward MG) words ending in *-menos*:

(1) To nifiko itan ra-**meno** apo ton rafti me
the wedding-dress.NEUT.N was sew.PRTC.NEUT.N by the tailor with
hrisi klosti.
golden thread
"The wedding dress was sewn by the tailor with golden thread".

Agreeing with the proposal of Markantonatou et al. (1996), according to which Modern Greek words ending in *-menos* should be considered to be participles rather than adjectives for reasons we present in Sections (3) and (3.2) below, our aim here is twofold:

1. to try to account for the fact that participles in *-menos* appear in the typical position of adjectives in Modern Greek (see Section (3.1)), and

2. to provide a formal account in LFG for participle-adjective formation in Modern Greek (see Section (5.2))

As far as the first of our aims is concerned, we show in Section (4) that participle-adjective formation in Modern Greek is better accounted for in the spirit of Bresnan's (1996) proposal for participle-adjective conversion in English, and not in the spirit of the predictions of Ackerman (1992) and Markantonatou (1995), which we also present briefly in the same section.

The main contribution of this paper, though, is the formalization in LFG of Bresnan's (1996) proposal for participle-adjective formation that we present in Section (5.2). The formalization we propose does not only cover the Modern Greek and English data at hand, but we are confident that it can easily be extended in order to account for the phenomenon of participle-adjective formation cross-linguistically.

2 Modern Greek words in *-menos*

Most of the literature todate has focused on the question whether the words ending in *-menos* in Modern Greek are adjectives or they bear a verbal nature, i.e., they are participles. The analyses proposed so far are split into two different classes as far as their conclusions are concerned.

Thus, according to the first class of analyses, Modern Greek words ending in *-menos* are adjectives. The analyses which take this assumption as their starting

point conclude that the expressions *ime... -menos* (be... -menos; cf., example (1) above) are phrase structures consisting of the auxiliary *ime* (be) and a complement (see Mozer (1994)).

According to the second class of analyses, though, Modern Greek words ending in *-menos* are participles. The analyses which take this assumption as their starting point conclude that the structures *ime... -menos* (be... -menos) are periphrastic expressions of the Passive Present Perfect (Present Perfect B'; see Veloudis (1990)).

Researchers who adopt the former view claim that the semantics of words ending in *-menos* is the same as the semantics of Modern Greek deverbal adjectives ending in *-tos*. That is, according to such views, *ine anigmenos/klismenos* and *ine anihtos/klistos* (be open/close) convey the same meaning (see Mozer (1994)).

3 Modern Greek words in *-menos*: Participles rather than Adjectives

Markantonatou et al. (1996), though, have shown that words ending in *-menos* bear more verbal characteristics than Modern Greek deverbal adjectives in *-tos.* That is, they have proposed clearly that words ending in *-menos* are participles rather than adjectives.

Look, for instance, at example (1), repeated here for convenience:

(2) To nifiko itan ra-**meno** apo ton rafti me
the wedding-dress.NEUT.N was sew.PRTC.NEUT.N by the tailor with
hrisi klosti.
golden thread
"The wedding dress was sewn by the tailor with golden thread".

In (2) above the expression *itan rameno* supports two complements, one denoting the "agent" (*ton rafti*) and the other denoting the "instrument" (*me hrisi klosti*). Both of these complements correspond to verbal complements, i.e., the subject and the instrument supported by the verbal head in example (3):

(3) O raftis erapse to nifiko me hrisi klosti.
the tailor.N sew.PAST.3S the wedding-dress.A with golden thread
"The tailor sewed the wedding dress with golden thread".

In contrast, deverbal adjectives ending in *-tos* do not permit the co-appearance of such complements in the same sentence (see example (4) below), showing thereby that their nature is "less verbal":

(4) To nifiko itan raf-**to** (*apo ton rafti)
the wedding-dress.NEUT.N was sewn.ADJ.NEUT.N (*by the tailor)
(*me hrisi klosti).
(*with golden thread)
"The wedding dress was sewn (*by the tailor) (*with golden thread)".

In the following some more examples are added in order to show clearly that the words ending in *-menos* bear more verbal characteristics than the Modern Greek deverbal adjectives in *-tos*:

(5) I porta itan anig-**meni** apo tus astinomikus me
the door.FEM.N was open.PRTC.FEM.N by the policemen with
losto.
metal-bar
"The door was opened by the policemen with a metal bar".

(6) Vrikan tin porta anig-**meni** me losto.
they-found the door.FEM.A open.PRTC.FEM.A with metal-bar
"They found the door opened with a metal bar".

(7) I porta itan anih-**ti** (*apo tus astinomikus) (*me
the door.FEM.N was open.ADJ.FEM.N (*by the policemen) (*with
losto).
metal-bar)
"The door was open (*by the policemen) (*with a metal bar)".

(8) Vrikan tin porta anih-ti (*me losto).
they-found the door.FEM.A open.ADJ.FEM.A (*with metal-bar)
"They found the door open (*with a metal bar)".

In example (5) the periphrasis *itan anig-meni* (was open) supports a complement which denotes the "agent" and a complement which denotes the "instrument". Both of these complements correspond to verbal complements, i.e., the logical subject and the instrument supported by the verbal head as shown in example (9) below:

(9) I astinomiki anixan tin porta me losto.
the policemen.N.PL open.PAST.3PL the door.A with metal-bar
"The policemen opened the door with a metal bar".

Even more evidence for the fact that words ending in *-menos* bear more verbal characteristics than Modern Greek deverbal adjectives in *-tos* comes from incorporation phenomena, as Markantonatou et al. (1996) have shown. That is, words

ending in *-menos* form compounds with adverbs of manner, such as *kala* (well), *kaka* (badly), *prohira* (off hand), etc., exactly like the corresponding verbs that they are derived from:

(10) Afta ta paputsia ine prohiroftiag-**mena**.
these the shoe.NEUT.N.PL be.3PL made-off-hand.PRTC.NEUT.N.PL
"These shoes seem to me to be made off hand".

(11) Tha ta prohirorapso ta paputsia tora ke tha ta
I-will cl.A.PL stich-off-hand.1S the shoe.A.PL now and I-will cl.A.PL
doso ston tsagari avrio.
give.1S to-the shoemaker tomorrow
"I will stich these shoes off hand now and I will give them to the shoemaker tomorrow".

The deverbal adjectives ending in *-tos*, though, cannot form compounds with adverbs of manner:

(12) *Afta ta paputsia ine prohiroraf-**ta**.
these the shoe.NEUT.N.PL be.3PL stiched-off-hand.ADJ.NEUT.N.PL
"These shoes seem to me to be stiched off hand".

We must underline here that the incorporation phenomena related to Modern Greek words ending in *-menos* persist even when the words ending in *-menos* appear in the typical position of adjectives. That is, for instance, after verbs like *fenete* (seems), *dihni* (shows), and so forth:

(13) Afta ta paputsia mu fenonte
these the shoe.NEUT.N.PL cl.G seem.3PL
prohiroftiag-**mena**.
made-off-hand.PRTC.NEUT.N.PL
"These shoes seem to me to be made off hand".

(14) *Afta ta paputsia mu fenonte
these the shoe.NEUT.N.PL cl.G seem.3PL
prohiroraf-**ta**.
stiched-off-hand.ADJ.NEUT.N.PL
"These shoes seem to me to be stiched off hand".

3.1 Modern Greek Participles in *-menos* as Adjectives

As Markantonatou et al. (1996) have also shown, Modern Greek participles in *-menos* may appear in the typical position of adjectives, as in example (6) above.

That is, in example (6) the participle in *-menos* appears as a complement in the typical position of an adjective. It is also very interesting to underline that the participle in *-menos* permits the co-appearance of another complement in the same sentence which denotes the "instrument", showing thereby its verbal nature.

In contrast, the deverbal adjectives ending in *-tos* do not licence the coappearance in the same sentence of such complements (see example (7) above), showing thereby clearly that their nature is "less verbal" than that of Modern Greek participles in *-menos.*

Below a few more example are given in order to show that participles in *-menos* have a verbal nature, but at the same time they can appear in the typical position of adjectives in Modern Greek:

(15) To buti ine poli psi-**meno** apo tin pano meria
the thigh.NEUT.N is very roast.PRTC.NEUT.N from the upper side
pu ekege o furnos ala apo tin kato ine shedon
where burn.PAST.3S the oven.N but from the bottom(side) is almost
apsi-**to**.
uncooked.ADJ.NEUT.N
"The thigh is overroasted on the upper side where the oven was burning but on the bottom side it is almost uncooked".

(16) To kotopulo itan pio psi-**meno** apo to
the chicken.NEUT.N was more roast.PRTC.NEUT.N than the
arni pu itan shedon apsi-**to**.
lamb.NEUT.N which was almost uncooked.ADJ.NEUT.N
"The chicken was cooked more than the lamb which was almost uncooked".

(17) To kotopulo mu fenete psi-**meno**.
the chicken.NEUT.N cl.G seem.3S roast.PRTC.NEUT.N
"The chicken seems done to me".

3.2 Overview

In the remaining of this section we are showing some more aspects of the verbal nature of Modern Greek participles in *-menos*, which differentiate them from Modern Greek deverbal adjectives in *-tos.* Some of the discussion here can also be found in Markantonatou et al. (1996; in Greek).

The syntactic realization of each predicate is assumed to be linked to its semantics, which consists of a group of semantic arguments that are related to each other by some logic variable. This assumption is deliberately very general in order to allow for flexibility as far as the semantic structure of predicates is concerned.

We are interested here neither in the variety of the semantic arguments, nor in their number.

Now, Modern Greek participles in *-menos* allow for a complement which denotes the "instrument" (see, for instance, examples (1), (5), and (6)). This does not hold for Modern Greek adjectives in *-tos* (see, for instance, examples (4), (7), and (8)). This complement, i.e., the "instrument", is related to the existence of a semantic argument which denotes *volitionality.*

But what do we mean by the term *volitionality*? Definitely something along the following terms: *Gianis* in example (18) below is a volitional participant in the event denoted by the verb, while *o aeras* (the air) in example (19) denotes only the (natural) cause that brings about the event described by the verb:

(18) O Gianis espase to parathiro.
the Gianis.N break.PAST.3S the window.A
"John broke the window".

(19) *O aeras espase to tzami me to tasaki.
the air.N break.PAST.3S the window.A with the ashtray
"*The air broke the window with the ashtray".

It seems that a semantic argument which denotes *volitionality* in the sense exhibited in examples (18) and (19) above is available to the participles in *-menos,* but not to the deverbal adjectives in *-tos*:

(20) Afta ta paputsia mu fenonte ra-**mena** me
these the shoe.NEUT.N.PL cl.G seem.3PL stich.PRTC.NEUT.N.PL with
spago.
string
"These shoes seem to me to be stiched with string".

(21) Afta ta paputsia mu fenonte raf-**ta**
these the shoe.NEUT.N.PL cl.G seem.3PL stiched.ADJ.NEUT.N.PL
(*me spago).
(*with string)
"These shoes seem to me to be stiched with string".

Moreover, in many cases Modern Greek participles in *-menos* allow for an *apo*(by)-PP as a complement. Among others, the *apo*(by)-PP in Modern Greek denotes the "agent" in passive sentences, as well as the "cause". In contrast, Modern Greek deverbal adjectives in *-tos* do not allow for such a complement:

(22) To spiti fenotan egataleli-**meno** apo tus
the house.NEUT.N seem.PAST.3S abandon.PRTC.NEUT.N by the
katikus tu.
inhabitant its
"The house seemed to be abandoned by its inhabitants".

(23) To spiti fenotan rimag-**meno** apo tin fotia.
the house.NEUT.N seem.PAST.3S destroy.PRTC.NEUT.N by the fire
"The house seemed to be destroyed by the fire".

(24) Vrika to fagito magire-**meno** apo tin Eleni.
I-found the food.NEUT.A cook.PRTC.NEUT.A by the Eleni
"I found out that Helen had cooked the food".

(25) To stifado fenete magire-**meno** apo kalo majira.
the stew.NEUT.N seem.3S cook.PRTC.NEUT.N by good cook
"The stew seems to be cooked by a good cook".

(26) *Ta papoutsia mu fenonte raf-**ta** apo kalo
the shoe.NEUT.N.PL cl.G seem.3PL stiched.ADJ.NEUT.N.PL by good
tsagari.
shoemaker
"The shoes seem to me to be stiched by a good shoemaker".

Similar phenomena can be found in English, too. According to Quirk, Greenbaum, Leech, and Svartvik (1985), the participles ending in *-ed* in English can co-occur with the adverb *very* and a *by*-PP, when the prepositional phrase denotes a "non-personal semi-agent":

(27) I am very disturbed by your attitude.

But personal agents are not excluded, either:

(28) ?I was very influenced by my college professors.

All the above shows that Modern Greek participles in *-menos* bear one semantic argument more than Modern Greek deverbal adjectives in *-tos*: the "agent" or the "cause" that brings about the action denoted by the verb. This semantic argument allows for the instantiation of the "instrument" syntactic argument.

Thus, it seems that the contrast between verb and adjective is not vertical in Modern Greek, but they are intermediate, transitional linguistic types. Modern Greek words in *-menos* seem to function both as participles and as adjectives with a more "dynamic" semantic dimension than their corresponding adjectives ending in *-tos*. Modern Greek has the *morphological means* to denote such a contrast.

4 Participle-Adjective Formation in Modern Greek

But why? That is, why can Modern Greek participles in *-menos* appear in the typical position of adjectives at all?

Here we support the view that the conversion of Modern Greek participles in *-menos* to adjectives, and consequently, their appearance in the typical position of adjectives in Modern Greek follows from the fact that they refer to the result state of the action denoted by the verb they are derived from.

Both *anig-meni* (open.PRTC) in example (5) (repeated here as example (29) for convenience) and *psi-meno* (roast.PRTC) in example (16) (repeated here as example (31) for convenience) refer to the result state of the action denoted by the verbs they are derived from (*anigo* (open) in example (9) (repeated here as example (30) for convenience) and *psino* (roast) in example (32) below, respectively):

(29) I porta itan anig-**meni** apo tus astinomikus me
the door.FEM.N was open.PRTC.FEM.N by the policemen with
losto.
metal-bar
"The door was opened by the policemen with a metal bar".

(30) I astinomiki anixan tin porta me losto.
the policemen.N.PL open.PAST.3PL the door.A with metal-bar
"The policemen opened the door with a metal bar".

(31) To kotopulo itan pio psi-**meno** apo to
the chicken.NEUT.N was more roast.PRTC.NEUT.N than the
arni pu itan shedon apsi-**to**.
lamb.NEUT.N which was almost uncooked.ADJ.NEUT.N
"The chicken was cooked more than the lamb which was almost uncooked".

(32) Epsisan to kotopulo pio poli apo to arni
they-roasted the chicken.NEUT.A more much than the lamb.NEUT.A
pu itan shedon apsi-**to**.
which was almost uncooked.ADJ.NEUT.N
"They roasted the chicken more than the lamb which was almost uncooked".

This view is in agreement with Bresnan's (1996) proposal for participle-adjective formation

> "...adjective conversion in general denotes a state derived from the semantics of the base verb. This seems to be true for all types of conversion, including the English present participles (*a smiling woman*)...The

> state denoted by the adjective appears to be the result state of the eventuality denoted by the participle" (Bresnan (1996, p. 12-13)).

To support her analysis of participle-adjective formation Bresnan uses English examples like the following:

(33) wilted lettuce, lettuce that has wilted
elapsed time, time that has elapsed
an escaped convict, a convict who has escaped

(34) *the run child, the child who has run
*an exercised athlete, an athlete who has exercised
*a flown pilot, a pilot who has flown
*a recently left woman, a woman who has left recently

She suggests that wilting in example (33) above involves an involuntary change of state, but even highly volitional eventualities such as having escaped can entail result states, such as freedom. She also points out that

> "it is strange to say *a run child*, because the activity of running lacks an inherent result state. But when the goal is supplied to the activity, a result state is defined, and conversion is possible (*a run-away child*)" (Bresnan (1996, p. 13)).

Thus, the converted adjectives of the following ergative past participles are all possible:[1]

(35) a run-away slave, a slave who has run away
an over-exercised athlete, an athlete who has exercised overly
a flown-away bird, a bird that has flown away
the widely-travelled correspondent, the correspondent who has travelled widely

whereas, in contrast, as Bresnan (1996, p. 13) explains, the verb *leave* in (34)[2] is bad because the predicate focuses on the source of motion, not on the goal, or result state.

Before drawing, though, our final conclusions for the phenomenon of participle-adjective formation in Modern Greek, we are going to explore the potential of Ackerman's (1992) and Markantonatou's (1995) proposals for participle-adjective formation when applied to the Modern Greek data at hand.

[1] Example (28) of Bresnan (1996, p. 13).
[2] Example (27b) of Bresnan (1996, p. 12).

Ackerman (1992) and Markantonatou (1995) predict that adjectival participles are related only to predicates which have a [-r] argument in their a(rgument)-structure. For Modern Greek, examples like the following illustrate clearly the predictions of Ackerman's (1992) and Markantonatou (1995)'s analyses:

(36) I giagia magirepse to fagito.
the grandmother.N cook.PAST.3S the food.A
"The grandmother cooked the food".

magirevo	<AGENT	THEME>
IC	-o	-r
Mapping Principles	SUBJ	OBJ

(37) To fagito ine magire-**meno**/*magiref-**to** apo tin giagia.
the food.N is cook.PRTC.N/cook.ADJ.N by the grandmother
"The food is cooked by the grandmother".

But examples like the following:

(38) O Gianis ipie poli krasi htes vradi sto parti.
the Gianis.N drank much wine.A yesterday night at-the party
"John drank too much wine at the party last night".

pino	<AGENT	THEME>
IC	-o	-r
Mapping Principles	SUBJ	OBJ

(39) O Gianis itan pio-**menos** htes vradi sto parti.
the Gianis.N was drink.PRTC.N yesterday night at-the party
"John was blind drunk at the party last night".

show that Ackerman's (1992) and Markantonatou's (1995) predictions that adjectival participles are related and refer **only** to a [-r] argument of predicates which contain such an argument in their a-structure, though not incorrect, do not cover all the cases of participle-adjective formation, at least in Modern Greek.

Our conclusion, thus, must be that Bresnan's (1996) proposal, which we have briefly shown earlier in the current section, is more reliable when it comes to participle-adjective formation in Modern Greek (see examples (29)-(32) above).

Moreover, employing the semantic concept of *result state* that Bresnan (1996) has proposed for participle-adjective conversions in English, we can also explain the restrictions on the formation of passive adjectives related to psychological predicates in Modern Greek without having to assume that the *experiencer* argument of Accusative Experiencer-Object Psych Verb Constructions (henceforward EOPVCs)

in Modern Greek bears the intrinsic classification feature [-r], as Markantonatou (1995) does.

Consider, for instance, examples (40)-(42)[3] below:

(40) O Gianis tromakse ton Kosta.
the Gianis.N frighten.PAST.3S the Kosta.A
"John frightened Kosta."

(41) O Kostas ine tromag-**menos**.
the Kostas.N be.3S frightened.ADJ.N
"Kostas is frightened."

(42) O Gianis ine tromag-**menos**.
the Gianis.N be.3S frightened.ADJ.N
"John is frightened."

(40) implies (41), but not (42). We agree with Markantonatou (1995, p. 291) that passive adjectives related to psychological predicates in Modern Greek refer to the experiencer semantic argument of Accusative EOPVCs. But this fact should not be assumed that it automatically entails in any way that this experiencer semantic argument must be considered to bear the Intrinsic Classification (IC) feature [-r].

The process of passive adjective formation, which is related to Accusative EOPVCs in Modern Greek, is not affected, though. Passive adjectives like the ones in (41) above are related to Modern Greek Accusative EOPVCs (cf., for instance, (40)) because these constructions clearly denote a *result state.* That is, the passive participles related to Accusative EOPVCs in Modern Greek denote a *result state,* and therefore, their conversion to adjectives is possible, according to Bresnan's (1996) predictions that we have seen above.

Our conclusion, then, is that passive adjectives related to Accusative EOPVCs in Modern Greek:

1. refer to the *experiencer* semantic argument of Accusative EOPVCs, and

2. their relation to these constructions is explained by the fact that Accusative EOPVCs clearly denote a *result state*. That is, the passive participles related to these constructions also denote a *result state;* thus, their conversion to passive adjectives is possible, according to Bresnan's (1996) predictions, which as we have already shown at the beginning of the current section explain the phenomenon of participle-adjective formation in Modern Greek in its entirety (for more details see Kordoni (2002)).

[3]Examples (74), (75), and (76) of Markantonatou (1995, p. 291).

5 Formalization of Participle-Adjective Formation in LFG

In the remaining we provide a formalization of participle-adjective formation in Modern Greek. This we are doing by formalizing in LFG the semantic concept of *result state* that Bresnan (1996) has proposed for participle-adjective conversions in English.

The formal proposal we present in this section is inspired amongst others by some aspects of the analysis of Grimshaw (1990) for derived nominals (see Grimshaw (1990, Chapter 3)).

In brief, we assume that Modern Greek words in *-menos* behave like result deverbal nominals. This, as we will show below, leads to the assumption that Modern Greek words in *-menos* have the Lexical Conceptual Structure (LCS) representation of verbal predicates, with one of their variables bound to the *R* argument, which according to Grimshaw is a non-thematic argument appearing at the level of a(rgument)-structure. According to her, this *R* argument is originally postulated to capture the predication or referentiality of nominal expressions. It serves as the external argument of nouns, but it is distinct from thematic arguments in that it is not realized in the syntactic representation.

The formalization of participle-adjective formation in Modern Greek that we propose in this section is based on the assumptions above.

5.1 Lexical Representations of Modern Greek Verbs and Nouns

Before moving onto the lexical representations of Modern Greek participles in *-menos*, though, we will present briefly the lexical representations of ordinary nouns and verbs in Modern Greek. The semantic and syntactic representations of Modern Greek nouns and verbs that we present here are in the spirit of the LFG analysis of Ohara (2001) for Japanese verbal nouns.

So the lexical representation of an ordinary noun in Modern Greek is as follows:

	LCS	vazo'(x)
(43)	a-structure	vazo <R(=x)>
	f-structure	(↑PRED) = 'vazo'

That is, the lexical representation of the ordinary Modern Greek noun *vazo* (vase) in (43) above includes a Lexical Conceptual Structure (LCS) level, an a(rgument)-structure level, and a predicate value at the f-structure level. The *R* argument is identified with a variable (x) at the level of LCS. Unlike a thematic argument, the *R* argument is not realized as a grammatical function at the f-structure level.

In addition, the lexical representation of an ordinary verb in Modern Greek is as follows:

	LCS	λyλxλe[spazo'(e) & θ(e,x), & θ(e,y)]
(44)	a-structure	spazo <[P-A], [P-P]>
	f-structure	(↑PRED) = 'spazo <SUBJ, OBJ>'

As shown in (44) above, the two participants of the event *spazo* (break) are linked to a Proto-Agent (P-A) and a Proto-Patient (P-P) argument at the level of argument structure (a-structure), and these arguments are in turn mapped to a subject and an object grammatical functions at the level of f-structure, respectively (in the spirit of Alsina (1996)).

We believe that the correspondence/linking between the levels of representation which describe the semantics and the syntax of ordinary nouns and verbs in Modern Greek is straightforward and can be extended so as to make the correct predictions about the relation between the semantics and the syntax of result deverbal nominals in Modern Greek.

Look, for instance, at example (45) below:

	LCS	λyλxλe[paratiro'(e) & θ(e,x), & θ(e,y)]
(45)	a-structure	paratirisi <R(=y)>
	f-structure	(↑PRED) = 'paratirisi'

In (45) above, the deverbal nominal *paratirisi* (observation) has the LCS representation of the Modern Greek verb *paratiro* (observe). Moreover, because it is a result deverbal nominal, its second participant is bound to the *R* argument at the level of argument structure (a-structure). This captures the fact that a result deverbal nominal refers to some concrete entity, which is associated with the event of the base verb it is derived from.

This way, result deverbal nominals are treated as having verbal information at the level of Lexical Conceptual Structure (LCS), which is mainly responsible for their conversion to nominals at the level of argument structure (a-structure) through the binding of a variable to the *R* argument of their a-structure.

For clarity, we need to add here that simple (i.e., non-result) deverbal nominals also have the Lexical Conceptual Structure (LCS) representation of the verb they are derived from. They also include the *R* argument in their a(rgument)-structure. But the variable bound to this *R* argument is different in reference than that of the result deverbal nominals we have just discussed above. Instead of binding the variable of a participant, the *R* argument of simple deverbal nominals in Modern Greek binds the variable of the whole event (e). This is shown in example (46) below:

	LCS	λyλxλe[proetimazo'(e) & θ(e,x), & θ(e,y)]
(46)	a-structure	proetimasia <R(=e)>
	f-structure	(↑PRED) = 'proetimasia'

Example (46) basically captures formally the fact that a simple deverbal nominal in Modern Greek, such as the nominal *proetimasia* (preparation), refers to the event itself, without looking at its internal structure. But since a simple deverbal nominal in Modern Greek has the *R* argument in its a(rgument) structure, it has referentiality, and behaves like a noun.

5.2 Lexical Representations of Modern Greek Participles in *-menos*

Turning to the case of Modern Greek words ending in *-menos* and as we have already mentioned at the beginning of Section (5.1) above, we assume that these behave like result deverbal nominals. That is, we assume that Modern Greek words in *-menos* have the LCS representation of verbal predicates, with one of their variables bound to the *R* argument, exactly like Modern Greek result deverbal nominals (see, for instance, example (45) in Section (5.1) above).

Look, for example, at (47) below:

	LCS	λyλxλe[magirevo'(e) & θ(e,x), & θ(e,y)]
(47)	a-structure	magiremeno <R(=y)>
	f-structure	(↑PRED) = 'magiremeno'

In (47), *magiremeno* (cook.PRTC) of example (37), which is repeated here for convenience:

(48) To fagito ine magire-**meno**/*magiref-**to** apo tin giagia.
the food.N is cook.PRTC.N/cook.ADJ.N by the grandmother
"The food is cooked by the grandmother".

has the LCS representation of the verb *magirevo* (cook) of example (36), which is also repeated here for convenience:

(49) I giagia magirepse to fagito.
the grandmother.N cook.PAST.3S the food.A
"The grandmother cooked the food".

Moreover, the second participant of the Modern Greek word *magiremeno* (cook.PRTC) as shown in (47) above is bound to the *R* argument at the level of argument structure (a-structure).

In this way, we capture the intuition that *magiremeno* (cook.PRTC), which refers to the result state of the action denoted by the verb *magirevo* (cook) that it is derived from, refers to some concrete entity, which is associated with the event of the base verb *magirevo* that the word *magiremeno* (cook.PRTC) is derived from.

Thus, the conclusion we are drawing here is that Modern Greek words ending in *-menos* have verbal information at the level of Lexical Conceptual Structure (LCS), exactly like Modern Greek result deverbal nominals that we have seen in Section (5.1) above (see, for instance, example (45)).

As in the case of Modern Greek result deverbal nominals, the fact that Modern Greek words ending in *-menos* are correctly treated as having verbal information at the level of Lexical Conceptual Structure (LCS; see examples (37), (48), and (47) above) is mainly responsible for their conversion to participles at the level of argument structure (a-structure) through the binding of a variable to the *R* argument of their a-structure. These participles, thus, are similar in nature to Modern Greek result deverbal nominals. A fact that justifies their appearance in the typical position of adjectives in Modern Greek, – in a typical position where a nominal category may appear, – as we have also shown in Section (4) in the previous.

Let us also take a closer look at example (39) that we have seen in Section (4) above, repeated here for convenience in (50) below:

(50) O Gianis itan pio-**menos** htes vradi sto parti.
the Gianis.N was drink.PRTC.N yesterday night at-the party
"John was blind drunk at the party last night".

The participle *piomenos* (drink.PRTC) in (50) has the Lexical Conceptual Structure (LCS) representation which is shown in (52) below. This is the LCS representation of the verb *pino* (drink) of example (38), repeated here for convenience in (51) below:

(51) O Gianis ipie poli krasi htes vradi sto parti.
the Gianis.N drank much wine.A yesterday night at-the party
"John drank too much wine at the party last night".

	pino	<AGENT	THEME>
	IC	-o	-r
	Mapping Principles	SUBJ	OBJ

(52)

LCS	λyλxλe[pino'(e) & θ(e,x), (& θ(e,y))]
a-structure	piomenos <R(=x)>
f-structure	(↑PRED) = 'piomenos'

The non-optional participant of the event denoted by the verb *pino* (drink)[4] is

[4]The notation (& θ(e,y)) denotes optionality.

bound to the *R* argument at the level of argument structure (a-structure) of the word *piomenos* (drink.PRTC) in (52) above.

In this way, we capture the intuition that *piomenos* (drink.PRTC) in (39) and (50), which refers to the result state of the action denoted by the verb *pino* that it is derived from, refers to the concrete entity which is associated to the non-optional participant of the event denoted by the base verb *pino*; that is, the participant denoted by the subject *o Gianis* in examples (38) and (51) above, in contrast to the predictions of Ackerman (1992) and Markantonatou (1995) that we have presented in Section (4) in the previous.

6 Conclusions

In this paper we have focused on Modern Greek words ending in *-menos*, which should be considered to be participles rather than adjectives, as Markantonatou et al. (1996) have shown (see Section (3)), since they bear one semantic argument more than Modern Greek deverbal adjectives ending in *-tos* (see Section (3.2)). This additional semantic argument of Modern Greek words ending in *-menos* is the "agent" or "cause" that brings about the action denoted by the verb which the words ending in *-menos* are derived from. Look, for instance, at examples (1)-(4) and (36)-(37), repeated here for convenience:

(53) To nifiko itan ra-**meno** apo ton rafti me
the wedding-dress.NEUT.N was sew.PRTC.NEUT.N by the tailor with
hrisi klosti.
golden thread
"The wedding dress was sewn by the tailor with golden thread".

(54) O raftis erapse to nifiko me hrisi klosti.
the tailor.N sew.PAST.3S the wedding-dress.A with golden thread
"The tailor sewed the wedding dress with golden thread".

(55) To nifiko itan raf-**to** (*apo ton rafti)
the wedding-dress.NEUT.N was sewn.ADJ.NEUT.N (*by the tailor)
(*me hrisi klosti).
(*with golden thread)
"The wedding dress was sewn (*by the tailor) (*with golden thread)".

(56) I giagia magirepse to fagito.
the grandmother.N cook.PAST.3S the food.A
"The grandmother cooked the food".

(57) To fagito ine magire-**meno**/*magiref-**to** apo tin giagia.
the food.N is cook.PRTC.N/cook.ADJ.N by the grandmother
"The food is cooked by the grandmother".

Moreover, in Section (3.1) of this paper we have shown that Modern Greek participles in *-menos* may also function as adjectives with a more enriched semantics than their corresponding adjectives in *-tos*.

Trying to account for the fact that participles in *-menos* appear in the typical position of adjectives in Modern Greek we first looked in Section (4) at the phenomenon of participle-adjective formation in Modern Greek. Our conclusion in the same section has been that participle-adjective formation in Modern Greek is better accounted for in the spirit of Bresnan's (1996) proposal for participle-adjective conversion in English, and not in the spirit of the predictions of Ackerman (1992) and Markantonatou (1995) that we have presented in the same section.

But this is not the only contribution of this paper as far as the phenomenon of participle-adjective formation is concerned.

In Section (5.2) we have presented a formalization in LFG of Bresnan's (1996) proposal for participle-adjective formation adapted, of course, to the Modern Greek data at hand. For this formalization we followed the analysis for the semantic and syntactic representations of Modern Greek verbs and nouns which we have presented in Section (5.1) and which is inspired by some aspects of the analysis of Grimshaw (1990) for derived nominals mainly in English.

With the formalization of Bresnan's (1996) analysis for participle-adjective formation that we have proposed in this paper at hand, it is more than interesting for future research to look at relevant data from languages other than English and Modern Greek in order for the current proposal to be extended accordingly and cover the phenomenon of participle-adjective formation cross-linguistically, as well as it currently does for Modern Greek and English.

References

Ackerman, F. (1992). Complex Predicates and Morphological Relatedness: Locative Alternation in Hungarian. In I. A. Sag and A. Szabolcsi (Eds.), *Lexical Matters. CSLI Lecture Notes no. 24*, pp. 55–84. Stanford, Calif.: CSLI Publications.

Alsina, A. (1996). *The Role of Argument Structure in Grammar: Evidence from Romance*. Stanford, CA: CSLI Publications.

Bresnan, J. (1996). Lexicality and Argument Structure. Invited paper given at the Paris Syntax and Semantics Conference, October 12-14, 1995. Corrected version: April 15, 1996. 27 pages. Available at:http://www-lfg.stanford.edu/lfg/bresnan/download.html.

Grimshaw, J. (1990). *Argument Structure*. Cambridge, Massachusetts: MIT Press.

Kordoni, V. (2002). *Psych Verb Constructions in Modern Greek: a semantic analysis in the Hierarchical Lexicon.* Ph. D. thesis, University of Essex, Colchester, U.K.

Markantonatou, S. (1995). Modern Greek deverbal nominals: an LMT approach. *Journal of Linguistics 31*, 267–299.

Markantonatou, S., V. Kordoni, V. Boubourekа, A. Kalliakostas, and V. Stavrakaki (1996). Modern Greek deverbal adjectives in -tos: a lexical semantic approach. *Studies in Greek Linguistics, Proceedings of the 17th Annual Meeting of the Department of Linguistics of the University of Thessaloniki.*

Mozer, A. (1994). The interaction of Lexical and Grammatical Aspect in Modern Greek. In *Themes in Modern Greek Linguistics.* Benjamins.

Ohara, M. (2001). *An Analysis of Verbal Nouns in Japanese.* Ph. D. thesis, University of Essex, Colchester, U.K.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. Longman.

Veloudis, G. (1990). The meta-linguistic character of the Perfect: Perfect B'. *Studies in Greek Linguistics 11, Proceedings of the 11th Annual Meeting of the Department of Linguistics of the University of Thessaloniki.*

Corpus-based Learning in Stochastic OT-LFG – Experiments with a Bidirectional Bootstrapping Approach

Jonas Kuhn

Stanford University* Department of Linguistics

University of Texas at Austin Department of Linguistics

jonask@mail.utexas.edu

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

*The reported work was conducted at Stanford University. The author was supported by a postdoctoral fellowship of the German Academic Exchange Service (DAAD).

Abstract

This paper reports on experiments exploring the application of a Stochastic Optimality-Theoretic approach in the corpus-based learning of some aspects of syntax. Using the Gradual Learning Algorithm, the clausal syntax of German has to be learned from learning instances of clauses extracted from a corpus. The particular focus in the experiments was placed on the usability of a bidirectional approach, where parsing-directed, interpretive optimization is applied to determine the target candidate for a subsequent application of generation-directed, expressive optimization. The results show that a bidirectional bootstrapping approach is only slightly less effective than a fully supervised approach.

1 Introduction

In Optimality Theory (OT), learning of a language amounts to determining the ranking relation over a given set of constraints. Under the target ranking, the observed language data have to be predicted as optimal (most harmonic) among the realization alternatives for the underlying meaning, or input. The fact that one alternative and not another is observed provides indirect negative evidence, which is exploited in learning algorithms (triggering a constraint re-ranking). A robust alternative to the original OT learning algorithm of Tesar and Smolensky (1998) is provided by Boersma (1998), Boersma and Hayes (2001):[1] the Gradual Learning Algorithm (GLA), which assumes a continuous scale for the constraint ranks. With a stochastic component in the determination of the effective constraint ranks, grammars can reflect variation in the training data, while effectively displaying categorical behaviour for most phenomena. This property has been exploited in the analysis of variation in syntax (Bresnan and Deo 2001, Koontz-Garboden 2001, Dingare 2001, Bresnan et al. 2001), based on the OT-LFG framework which uses LFG representations for the candidates, with the f-structures corresponding to the input and (mainly) the c-structure and lexical contribution differing across candidates (Bresnan 2000, Sells 2001b, Kuhn 2001a, forthcoming).

Experimental applications of GLA have so far adopted the idealization that not only the surface form of learning data is known, but the full analysis, including the input (and thus the entire candidate set). With this information, misinterpretations of the evidence for re-rankings are excluded, however a plausible learning approach cannot keep up this idealization. Furthermore, most studies have applied the GLA on a carefully controlled data set, focusing on variation in a small set of phenomena (i.e., keeping other choices fixed by design).

In this paper, I explore the application of GLA for learning clausal syntax, essentially from free corpus data (in the present study from a newspaper corpus of German). The candidate generation grammar is kept highly general, with the only inviolable restrictions being an extended X-bar scheme (in which all positions are

[1] An implementation of the GLA is included in the Praat program by Paul Boersma and David Weenink: `http://fonsg3.let.uva.nl/praat/`

optional). Crucially, I do not assume full syntactic analyses of the learning data as given. I make the weaker, and arguably much more plausible assumption that the learner can use language-independent evidence to narrow down the space of possible semantic representations for an observed form. In the corpus-based learning experiment this narrowing-down is simulated as follows: as training data I use individual clauses (main clauses or subclauses) extracted from a treebank, with a given underlying predicate-argument structure and the argument and modifier phrases pre-bracketed as fixed chunks, as shown in (1).

(1) [So streng] [sind] [auf den Gipfeln] [die Sitten und die Gesetze der Eitelkeiten]
So strict are on the summits the customs and the rules of vanities

With the clause boundaries and dependent phrases fixed, experiments with a bootstrapping approach building on a **bidirectional learning** scheme become possible. Under the bidirectionality assumption[2], the same constraint ranking that determines the grammatical form in expressive optimization (based on a fixed underlying meaning) is used in interpretive optimization: for a given string, the most harmonic parsing analysis is taken to be correct. Even though the space of possible interpretations is narrowed down, parsing with the liberal underlying grammar yields an average of more than 16 analyses for short sentences (with four or less "chunks"), so the interpretive optimization is not trivial.

2 OT Syntax background

This paper builds on the OT-LFG framework (Bresnan 1996, 2000, Kuhn forthcoming), in which an Optimality-Theoretic grammar for syntax is formalized based on LFG representations. The OT-LFG architecture is sketched for an example in the diagram in figure 1 and is introduced informally in the following. The small grammar fragment used for this illustration is essentially Bresnan's OT-LFG reconstruction of Grimshaw (1997) (Bresnan 2000, sec. 2).

A highly general LFG grammar G_{inviol} is assumed that constrains the set of universally possible c-structure/f-structure pairs, i.e., it encodes a basic (extended) X-bar scheme, but is very unrestrictive. In the standard expressive optimization, the set of *candidate structures* is defined as those G_{inviol}-analyses (c-structure/f-structure pairs) which share the semantically interpreted part of the f-structure (the "*input*"). So, there are different potential syntactic realizations of the same meaning to choose from. OT constraints (such as the ones in (2)) are structural descriptions of subparts of a c-structure, an f-structure or of both structures (related through the projection function ϕ). Subparts of the actual candidate structures may violate some of the descriptions/constraints, so the constraint set defines a *constraint violation profile* for each candidate structure.

[2](Smolensky 1996), for discussion in OT-LFG (outside the learning context) see Lee (2001), Kuhn (2000, 2001b).

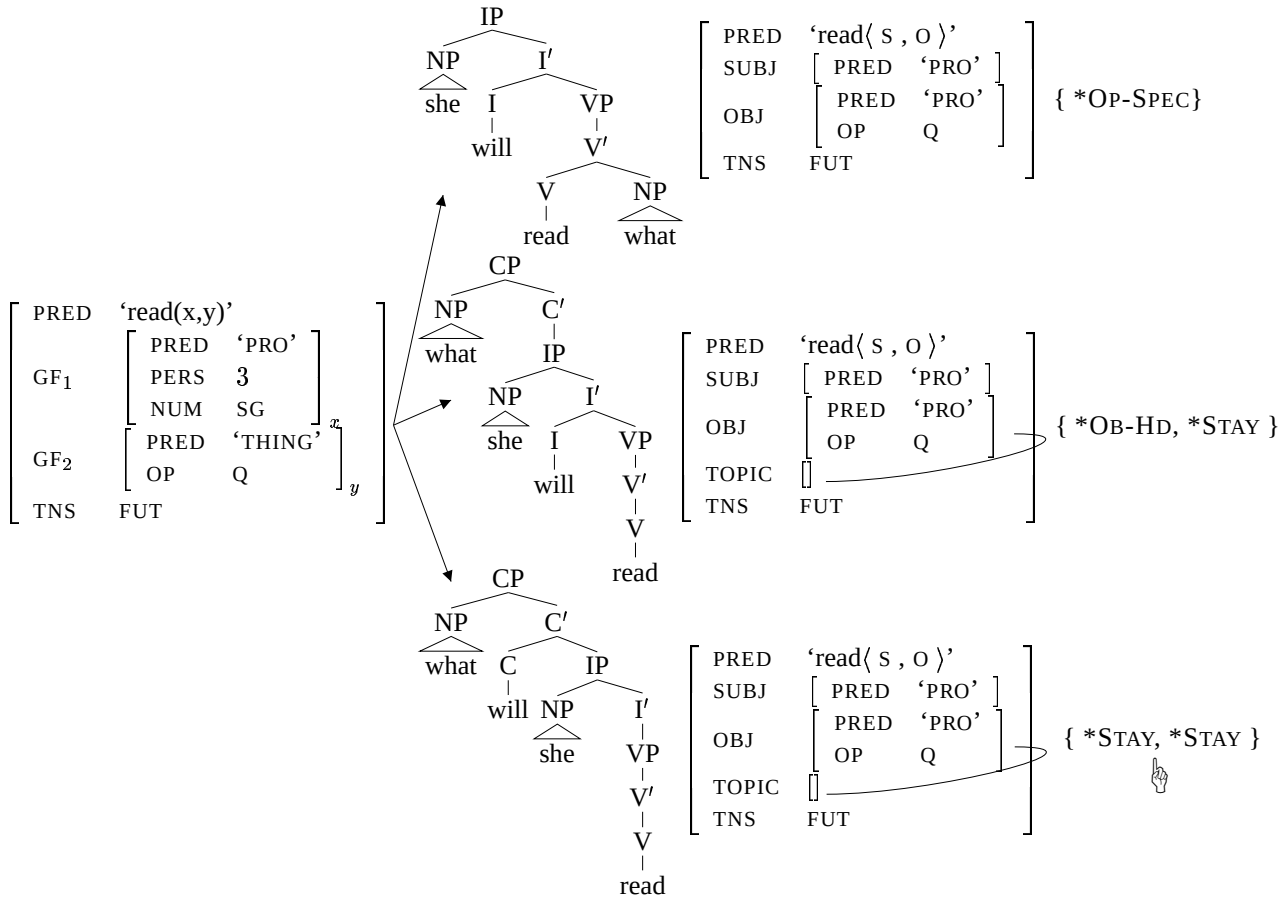


Figure 1: A sketch of the OT-LFG architecture (expressive optimization)

(2) a. OP-SPEC (Bresnan 2000)
An operator must be the value of a DF [discourse function] in the f-structure.

b. OB-HD (Bresnan 2000, (21))
Every projected category has a lexically filled [extended, JK] head.

c. STAY (Bresnan 2000, (24))
Categories dominate their extended heads.

Given the language-specific ranking of constraint importance, different structures from the set of candidates arise as optimal in the sense of violating the fewest of the most important constraints (see section 3 for some more discussion). In English (3), it is more important to mark the scope of *wh*-elements overtly than to realize arguments in their canonical position; in a *wh*-in situ language the situation is different: (4). Only the optimal candidate is defined to be a grammatical realization of the underlying part of the f-structure ("input").

(3) a. **R1:** OP-SPEC ≫ OB-HD ≫ STAY: *English*

b.

Candidate set:	OP-SPEC	OB-HD	STAY
$[_{IP}$ she will $[_{VP}$ read what]]	*!		
$[_{CP}$ what $[_{IP}$ she will $[_{VP}$ read]]]		*!	*
☞ $[_{CP}$ what **will** $[_{IP}$ she $[_{VP}$ read]]]			**

(4) a. **R2:** STAY≫ OP-SPEC ≫ OB-HD: *wh* in situ language

b.

Candidate set:	STAY	OP-SPEC	OB-HD
☞ $[_{IP}$ "she" "will" $[_{VP}$ "read" "what"]]		*	
$[_{CP}$ "what" $[_{IP}$ "she" "will" $[_{VP}$ "read"]]]	*!		*
$[_{CP}$ "what" **"will"** $[_{IP}$ "she" $[_{VP}$ "read"]]]	*!*		

Interpretive optimization The general architecture of standard expressive optimization is easily adapted to a slightly different formal system (Kuhn forthcoming, ch. 5): if the set of competing candidate structures is not defined by a common semantic representation, but by a common surface string, we get a system of *interpretive optimization*. Rather than choosing from different potential syntactic realizations of a meaning, the OT evaluation now chooses from different syntactic structures (many of which differ in semantic interpretation too) for a given surface string.

This "reverse" formal system has been adapted for a variety of linguistic modeling tasks, in particular for a derivation of the discrepancy between production and comprehension in language acquisition (Smolensky 1996), and in syntax to model word order freezing effects (Lee 2001, Kuhn 2001b). Interpretive optimization may also be assumed in the learning procedure for a standard expressive OT grammar, which will be discussed in section 5.1.

3 Ranking vs. weighting

The previous discussion—like most of the linguistic work in OT—took a central OT assumption for granted: The relative importance of the constraints for a specific language is determined by a *strict ranking*. This means that violating a high-ranking constraint is worse than arbitrarily many violations of some lower-ranking constraint. The ranking scheme is more restrictive than a summation over weighted constraints would be (which one might have chosen as a more general way of computing the joint

effect of constraints of different importance, and which is for instance underlying the predecessor of OT, Harmony Grammar).

The OT hypothesis of strict ranking is motivated for the very reason of making the system more restrictive, such that clearly testable typological predictions of the system follow from the assumption of a particular set of constraints. To illustrate this point, let us briefly compare the way predictions are grounded in a ranking scheme and how this compares to a weighting scheme.

If we have a constraint violation profile as in tableau (5) (with the ranking of the constraints open) and we observe candidate A in the data, we know that CONSTR. 3 must outrank the other two constraints: CONSTR. 3 $\gg$ { CONSTR. 1, CONSTR. 2 }—else candidate A would be the winner. This kind of configuration is called a *ranking argument*. The fact that candidate B incurs three violations of CONSTR. 3 and not just one is irrelevant: the only way that B will lose against A is when CONSTR. 3 is ranked highest.

(5)

Candidate set:	CONSTR. 1	CONSTR. 2	CONSTR. 3
candidate A	*	*	
candidate B			***

Now, if in addition to (5), we observe the A′ and A″ candidate of (6-b) and (6-c) data for the same language, we get an inconsistency: (6-b) and (6-c) are ranking arguments for CONSTR. 2 $\gg$ CONSTR. 3, and CONSTR. 1 $\gg$ CONSTR. 3, respectively.

(6) *Under the ranking hypothesis, (a) is incompatible with (b) and (c)*

(a) Candidate set:	CONSTR. 3	CONSTR. 1	CONSTR. 2
☞ candidate A		*	*
candidate B	***		

(b) Candidate set:	CONSTR. 1	CONSTR. 2	CONSTR. 3
☞ candidate A′			*
candidate B′		*	

(c) Candidate set:	CONSTR. 1	CONSTR. 2	CONSTR. 3
☞ candidate A″			*
candidate B″	*		

So, a small set of clear data is already very informative about an OT account, based on the ranking hypothesis. If we do observe all the data in (6) in a single language, we know that the constraint set assumed was inadequate; maybe an additional constraint or an entirely different set of constraints is needed.

Now, under a *constraint weighting* regime no such clear conclusion about the symbolic part of the theory—the constraints and the candidate representations—can be drawn. The (a) type of data may be compatible with (b), with (c), with both, or none. Examples (7)–(9) illustrate this with different negative weights assumed for

the constraints (the winner is defined to be the candidate with the greatest weighted sum over violation marks, e.g., (7-b,A′) wins over (7-b,B′) since $-4 < -3$). In all cases the (a) data are correctly predicted, since $w(\text{CONSTR. } 1) + w(\text{CONSTR. } 2) > 3 \times w(\text{CONSTR. } 3)$. Note however that absolute constraint weights would lead to different rankings in each of the cases, as is suggested by the order of notation (which of course has no technical effect, since we are looking at a weighting system).

(7) *(a) is compatible with (b), but not with (c)*

not compatible

(a) Candidate set:	CONSTR. 1 −4	CONSTR. 3 −3	CONSTR. 2 −1
☞ −5 cand. A	*		*
−9 cand. B		***	

(b) Candidate set:	CONSTR. 1 −4	CONSTR. 3 −3	CONSTR. 2 −1
☞ −3 cand. A′		*	
−4 cand. B′	*		

(c) Candidate set:	CONSTR. 1 −4	CONSTR. 3 −3	CONSTR. 2 −1
☞ −3 cand. A″		*	
−1 cand. B″			*

(8) *(a) is compatible with (b) and (c)*

(a) Candidate set:	CONSTR. 1 −6	CONSTR. 2 −5	CONSTR. 3 −4
☞ −11 cand. A	*	*	
−12 cand. B			***

(b) Candidate set:	CONSTR. 1 −6	CONSTR. 2 −5	CONSTR. 3 −4
☞ −4 cand. A′			*
−6 cand. B′	*		

(c) Candidate set:	CONSTR. 1 −6	CONSTR. 2 −5	CONSTR. 3 −4
☞ −4 cand. A″			*
−5 cand. B″		*	

(9) *(a) is compatible with neither (b) nor (c)*

not compatible *not compatible*

(a) Candidate set:	CONSTR. 3 −3	CONSTR. 1 −2	CONSTR. 2 −1
☞ −3 cand. A		*	*
−9 cand. B	***		

(b) Candidate set:	CONSTR. 3 −3	CONSTR. 1 −2	CONSTR. 2 −1
☞ −3 cand. A′	*		
−2 cand. B′		*	

(c) Candidate set:	CONSTR. 3 −3	CONSTR. 1 −2	CONSTR. 2 −1
☞ −3 cand. A″	*		
−1 cand. B″			*

As the example illustrated, the constraint weighting scheme has an undesirable property if we are interested in finding a linguistically motivated set of constraints for predicting a typological spectrum of languages: the effect of picking a particular constraint set is underdetermined—a readjustment of the constraint weights may have the

same effect as a modification of the constraint set, i.e., the symbolic part of the theory. This motivates the OT assumption of a constraint ranking regime. The strong interpretation of the OT constraint set assumes that the constraint set reflects innate restrictions on possible grammars (i.e., it formalizes Universal Grammar).

Limitations due to the ranking hypothesis Related to its restrictiveness, the ranking hypothesis has the effect that the phenomenon of optionality or variability of output forms for a single underlying input becomes almost impossible to derive. The strict constraint ranking differentiates between any two candidates with a different constraint profile, predicting all but one candidate to be ungrammatical.[3]

There are different possible ways of overcoming the limitations: one could assume a more fine-grained input representation, distinguishing between cases of optionality; the selection of this input representation itself could be modelled by a contextually controlled process, which may not be fully deterministic. A different modification of the strict OT system is the assumption of a less fixed ranking of the constraints (Anttila 1997, Boersma 1998, Boersma and Hayes 2001). The stochastic OT system proposed by Boersma will be discussed in more detail in the next section. Yet another option might be to assume a weighting scheme where the weights are typically widely separated, so the emerging behavior is almost that of a ranking scheme.

It is fairly difficult to find independent criteria for deciding between the various choices in the architecture of such a modified OT system: applying the systems for a non-trivial learning task, as is attempted in this paper, is one way of assessing their adequacy (although this alone may not lead to a conclusive answer).

4 Learning

The learning procedures that have been proposed for Optimality Theory are essentially *error-driven*. This means that during learning, a hypothetical constraint ranking is applied to the learning data. Under a simplifying assumption (which will be challenged in section 5.1), the learner has access to the underlying input representation for an observed piece of learning data; with the candidate set being defined in terms of G_{inviol} and the input, the learner has thus access to the full set of candidates. The learner will then need some monitoring ability, in order to be able to compare its own predictions of the output/winner, based on the hypothetical constraint ranking, with the output in the actual data. Whenever there is a mismatch, this is evidence that the hypothetical ranking cannot be (fully) correct.

For instance, in (10) the hypothetical ranking CONSTR. 1 $\gg$ CONSTR. 2 $\gg \ldots \gg$ CONSTR. 5 would predict candidate A to be the winner. But the observed output structure is candidate B. Hence, the assumed ranking must have been incorrect: CONSTR. 3

[3]Of course more than one candidate can have the same constraint profile, but with a realistic constraint set, this is no modelling option for most cases of optionality.

should outrank CONSTR. 1.

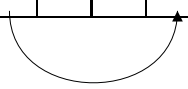
(10) *Detecting an error in the learner's system*

Candidate set:	CONSTR. 1	CONSTR. 2	CONSTR. 3	CONSTR. 4	CONSTR. 5
candidate A		*	*		
observed: candidate B	*!	*			*

In the Constraint Demotion Algorithm (Tesar and Smolensky 1998), this type of ranking argument is exploited to make conservative modifications of the ranking, which guarantee that learning will converge (on noise-free data). Constraints violated by both the predicted winner (A) and the observed output (B) and constraints violated by neither of the two are ignored in a learning step. Of the remaining constraints, the ones violated by observed output are demoted just below highest-ranking constraint violated by putative winner. So CONSTR. 1 is demoted just below CONSTR. 3:

(11) *Constraint demotion*

Candidate set:	CONSTR. 1	CONSTR. 2	CONSTR. 3	CONSTR. 4	CONSTR. 5
candidate A		*	*		
observed: candidate B	*	*			*



(12) *Constraint ranking after learning step*

Candidate set:	CONSTR. 2	CONSTR. 3	CONSTR. 1	CONSTR. 4	CONSTR. 5
candidate A	*	*!			
observed: candidate B	*		*		*

The Gradual Learning Algorithm (GLA) Since the Constraint Demotion Algorithm was developed for the strict OT ranking architectire, it cannot be used to learn from data displaying optionality/variation. Also, the algorithm is not robust; i.e., a single instance of data incompatible with the target ranking may corrupt the intermediate

ranking in a way from which the learner cannot recover. Boersma (1998), Boersma and Hayes (2001) propose an alternative learning algorithm, the Gradual Learning Algorithm (GLA), based on a modified ranking architecture, which is robust and can deal with optionality.

In the modified architecture—stochastic OT—the constraint ranking is no longer discrete, but the constraints are ranked on a continuous scale: the rank or strength of a constraint is represented by a numerical value. (However, we still have a ranking and not a weighting, i.e., just the relative strengths of constraints are relevant; there is no summation over the values of the violated constraints.) As the candidates in a tableau are evaluated, some random noise with a normal distribution is added to the constraint strength. This can have the effect of reversing the effective order of the constraint and thus leads to a variable behavior of the system.

Diagram (13) is a schematic illustration of a set of constraints ranked on the continuous scale, with strength decreasing from left to right. When the constraint strengths (i.e., the means of the normal distribution) are sufficiently far apart—as for CONSTR. 3 vs. CONSTR. 4—a reversal will effectively never happen, so we have a categorical effect like with a discrete ranking. For constraints with a similar strength (like CONSTR. 4 and CONSTR. 5), we will however find both orders, depending on the noise at evaluation time.

(13) 

In the GLA, designed for stochastic OT, a learning step (triggered by an observed error like in the Constraint Demotion Algorithm) does not lead to a readical change in the constraint ranks. Rather, a slight adjustment of the constraint ranks is made, promoting the constraints violated by the erroneous winner, and demoting the constraints of the observed output:

(14) *Promotion/demotion in the GLA*

Candidate set:	CONSTR. 1	CONSTR. 2	CONSTR. 3	CONSTR. 4	CONSTR. 5
candidate A		*	*		
observed: candidate B	*	*			*
	→		←		→

Data types occurring with sufficient frequency will cause a repeated demotion/promotion, so a quasi-categorical separation of the constraint strengths can result; noise in the data will have only a temporary effect. In variability phenonema, opposing tendencies of constraint demotion/promotion will ultimately balance out in a way that

reflects the frequencies in the data (assuming a large enough sample is presented to the learner).

As applications of the GLA in phonology and syntax (see the citations in section 1) have shown, the algorithm is able to adjust the constraint strengths for the linguistic constraint sets posited in these studies in an appropriate way: the behavior of the stochastic model indeed replicates the frequency distribution of the data types in the learning data.[4] However, so far GLA applications have focused on relatively small, clear-cut grammar fragments.

5 Experiments

The experiments reported in this paper address the following questions: (i) Can GLA be used for an exploratory analysis of a more complex cluster of interacting phenomena? (ii) What is the amount of target information required to control the error-based learning scheme?

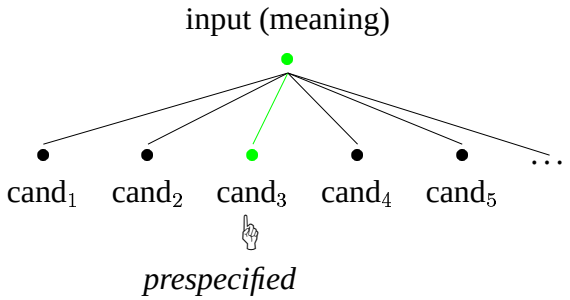
Methodologically, the idea was to start out with a certain set of linguistically well-understood constraints, and to add further constraints in order to explore interactions. The set of phenomena to be chosen for this investigation was supposed to display variation, but at the same time clearly obey certain language-specific principles. Under these criteria, the clausal syntax of German is a well-suited target for learning: the system is confronted with a high degree of word order variation in the relative order of argument phrases in the *Mittelfeld* (the area following the finite verb in matrix clauses), but the verb position in the various clause types is fixed and has to be learned as categorical facts. The exact way of representing the training data from a corpus was motivated by considerations concerning the "degree of supervision" in learning (question (ii)), which is discussed in the following subsection.

5.1 Target information in learning

How much information should be provided to the learner with the learning data? Previous studies of learning in OT—both for the constraint demotion algorithm and for the GLA—have assumed the following idealization: the learner is presented with the full candidate set (which is constructable from the exact input), plus the exact target output candidate (compare the diagram in (15)). This means that an error in the predictions of the learner's system can be very reliably detected—if any other candidate than the target output is more harmonic, we have an error.

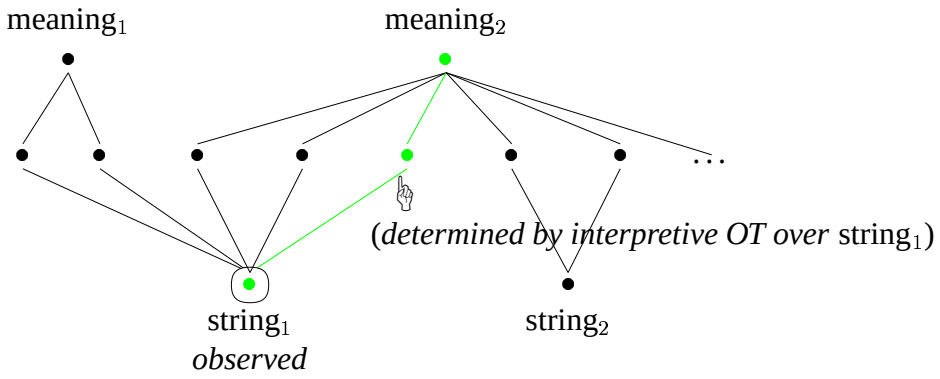
[4] Keller and Asudeh (2001) observe that for certain constraint sets that have been assumed in the linguistic literature, the GLA does not converge; however this may indicate that the assumed constraints are insufficient for an adequate description of the data.

(15) *Full target annotation (schematic)*



Of course, the only direct observation that a human learner has access to is the surface form (of utterances made by adult speakers). There may be many different underlying inputs for a given surface form, and even for the same combination of input and surface string, there may be differences in the syntactic analysis. In theoretical OT work, a process of *robust interpretive parsing* is assumed, which the learner applies to "guess" what the underlying input for an observed string is (Tesar and Smolensky 1998). The current constraint ranking is simply applied on the set of candidates defined by a common surface string (parsing-based or interpretive optimization). Based on the underlying input determined in this way, the standard generation-based or expressive optimization is applied as the basis for the actual learning (compare (16)).

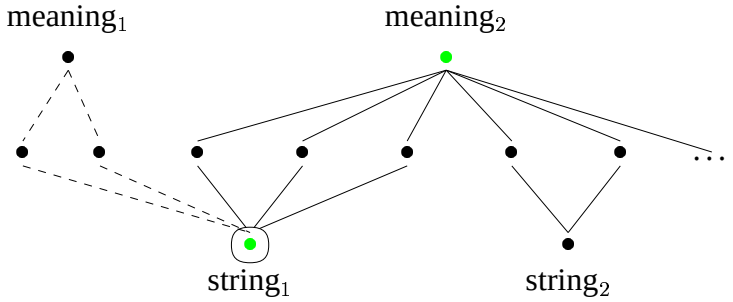
(16) *Determining the target for expressive optimizarion by interpretive optimization*



Hence, the mentioned idealization in the presentation of the target structure is not hard-wired into the OT architecture. A bidirectional of optimization (robust interpretive parsing, plus expressive optimization) works without this assumption. In the long run, one may hope that corpus-based learning experiments can apply the general bidirectional strategy. However, based exclusively on linguistic material, a corpus-based learner has a considerable disadvantage: the human learner can exploit semantic information and background knowledge, and this way the choices in interpretive parsing are often narrowed down considerably. In the present experiments, I tried to simulate this effect by providing the full predicate-argument structure (i.e., the full underlying

input) for the learning instances. This still leaves open which of the syntactic analyses for the observed string is the right target winner.

(17) *Narrowed down set of choices in interpretive optimization*



5.2 Experimental set-up

The training data were extracted from the TIGER treebank, a syntactically annotated newspaper corpus of German (cf. Brants et al. (2002), Zinsmeister et al. (2002)). The treebank includes full categorial and functional annotations, but this information was of course only partially exploited for training data (as far as justified by non-syntactic information available to the human learner).

The data was split up into single clauses, i.e., either matrix clauses or embedded clauses (presented as separate training instances). Since the focus was on the learning of clausal syntax, embedded argument/modifier phrases (NPs, PPs, etc.), were pre-bracketed, and their grammatical functions were provided. No syntactic information was provided about verbal constituents, i.e., verbs and auxiliaries were left as separate, unconnected units.

For example, sentence (18) would give rise to two training instances (19)—one for the matrix clause, including a single "chunk" for the embedded complement clause, and one for the internal structure of the complement clause.

(18) Der Vorstand der Firma hat gefordert, daß der Geschäftsführer entlassen
the board of the company has demanded that the managing director laid off
wird.
is

(19) a. [Der Vorstand der Firma] hat gefordert, [daß . . .]

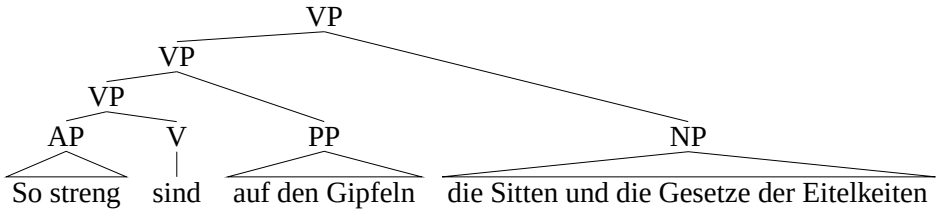
b. daß [der Geschäftsführer] entlassen wird

The candidate analyses The set of candidates was generated by a highly under-restricted LFG grammar (G_{inviol}), approximating the OT hypothesis that all universally possible structures should be included in this set. Reflecting inviolable principles, an extended X-bar scheme is encoded in the LFG grammar; the scheme is very general however, all positions are optional, functional projections (IP, CP) can be freely filled

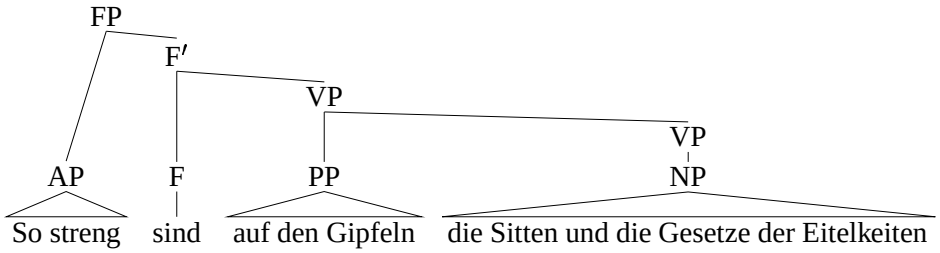
with verbs, auxiliaries, complementizers. The grammar was written and applied with Xerox Linguistic Environment (XLE).[5] As an illustration of the broad range of analyses licensed by the underlying grammar, consider the sample structures in (21) for sentence (20).

(20) [So streng] [sind] [auf den Gipfeln] [die Sitten und die Gesetze der Eitelkeiten]
So strict are on the summits the customs and the rules of vanities

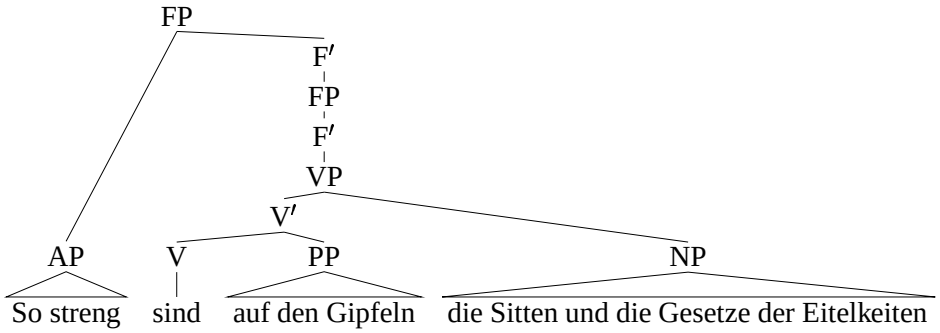
(21) a.



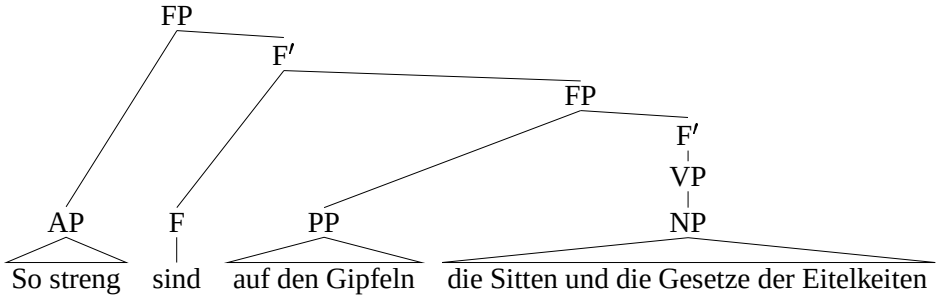
b.



c.



d.



The OT constraints The constraints were also encoded using XLE (compare Frank et al. (2001)). The core constraints adopted were inspired by OT accounts of clausal

[5]For technical reasons, a generation-based application of the grammar was simulated by parsing all permutations of the string. In future experiments, it should be possible to use the XLE generator.

syntax (Grimshaw 1997, Sells 2001a); further constraints were added to ensure distinguishability of candidates. A total of about 90 constraints was used—based on X-bar configurations, precedence relations of grammatical functions/NP types (pronominal vs. full), etc.

Due to computation-intensive preprocessing routines for the learning data, required after each change in the assumed constraint set, the learning experiments were only performed on small training sets. The reported results are from a specific sequence of experiments based on 195 training sentences.

5.3 Learning schemes

The corpus-based learning was performed with the GLA (using a simple Prolog implementation) in generation-based optimization. As discussed in sec. 4, the GLA is an error-based learning algorithm, i.e., at each state, the learner applies its present, hypothetical ranking. If the predicted winner matches the target output, no adjustment is necessary; if a different output is the target, the constraints violated only by the predicted winner have to be promoted, while those violated only by the target output are demoted.

As discussed in sec. 5.1, a realistic approach should compute the target winner based on a bidirectional approach. In order to test the feasibility of such an account—within the limits of the assumptions discussed above—three different learning schemes were compared in the experiment:

1. The "fully supervised" scheme:
   The exact target structure for the training clauses was manually annotated (based on the standard analysis of German clause structure).

2. The "string-as-target" scheme:
   No manual annotation was made; all candidates with the right word order count as target winners (no interpretive optimization is performed). Only predicted winners with an incorrect surface order count as errors—i.e., constraints violated by *any of the target winners* (and not the predicted winner) are demoted.

3. The bidirectional optimization (or "bootstrapping") scheme:
   The current ranking is used to determine a target winner among parsing alternatives for the observed string. All other candidates (possibly with correct surface order) count as errors.[6]

[6]For the bidirectional scheme, two variants were compared: one, in which the same effective ranking—i.e., the ranking after addition of noise—was used in generation and parsing; and another one, in which the initial parsing-based optimization was sampled several times (leading to different effective rankings), in order to determine a larger set of target winners. The evaluation showed that both variants lead to a very similar behavior.

5.4 Results

Evaluation schemes It is not straightforward how to best evaluate the performance of a generation-based optimization system. Demanding that the word string predicted for an unseen underlying predicate-argument structure be an exact match of the actual string in the corpus would be too strict, since there are many cases of real optionality: even in the concrete given context, several orderings are perfectly natural. Instead of evaluating how often the exact string in the corpus is predicted for unseen generation tasks, the main evaluation measure is based on a manual annotation of the acceptable permutations for a set of 100 evaluation sentences, which had not been presented as training data. All natural-sounding permutations in the given context were annotated as possible generation alternatives. No inter-subject comparsion of the annotations was made, so the raw percentage numbers for the various learning schemes should be treated with some caution. The focus of the experiments was on a *comparison* of the different schemes.

Besides this main evaluation measure, a variation of the bidirectional optimization technique was applied: the ranking that the learner came up with (through generation-based learning, possibly with a parsing-based determination of the target winner) is used in a disambiguation task. For sentences with ambiguous case marking on the argument phrases, a theory of word order preferences predicts how likely the individual readings are (compare the discussion of word order freezing in bidirectional OT in Kuhn (2001b), Lee (2001)). A corpus example of such an ambiguous case marking is shown in (22): both bracketed NPs can be either nominative or accusative. 50 such unseen examples from the corpus were used for the second evaluation measure, counting how often the intended reading was matched by the system's prediction.

(22) daß [die Bundesregierung] [die militärische Zusammenarbeit] wiederbelebt
that the federal government the military cooperation revitalized
hat
has

Results The evaluation results (for a specific series of experiments) are shown in (23). The left-most column shows the results for the initial ranking (with all constraints ranked the same).

(23) a. *Percent acceptable orderings on unseen data*

initial ranking	"string-as-target"	bidirectional	"supervised"
34%	66%	87%	90%

b. *Disambiguation of unseen parsing ambiguities*

initial ranking	"string-as-target"	bidirectional	"supervised"
54%	76%	84%	83%

Note that in (23a), the bidirectional approach leads to a significant improvement over the "string-as-target" scheme. For the disambiguation task (23b), the bidirectional

scheme is as good as the supervised approach.[7] So both measures indicate that the bidirectional bootstrapping approach is very promising.

6 Discussion

While the question about the usefulness of bidirectional optimization in learning can be answered positively, it is not entirely clear what conclusions can be drawn for the other question: can the GLA be used straightforwardly in an exploratory analysis with a large number of constraints? The large set of constraints seems to make the analysis of linguistic effects somewhat opaque. However, this may be due to a lack of analytical tools.

As I discussed in Kuhn (2002), there are certain cases in which the GLA is not able to deal with conflicting (statistical) ranking arguments. It is possible that the data sets contained such cases. A small experiment using a weighting-based model on the training data from the fully supervised scheme indicated that a better fit on the training data is possible (in this experiment, I used the log-linear model that Johnson et al. (1999) developed for disambiguation of parses with a large-scale LFG grammar[8]).

For deciding what is an adequate linguistically restricted learning model to deal with a larger number of interacting phenomena, further experiments are required. The learning instances should be kept more controlled, without having to move away from the use of real corpus data. A promising approach might be to use a (slightly relaxed) classical large-coverage grammar to produce the learning material.

[7] The fact that it is even slightly better may be an effect of the small size of the training data; it was easier for the bidirectional approach to come up with (potentially incorrect) generalizations over the data, whereas the supervised approach was confronted with the linguistically motivated target annotations, for which there may not have been enough support in the data.

[8] I would like to thank Mark Johnson for providing the learning code.

References

Anttila, Arto. 1997. *Variation in Finnish Phonology and Morphology*. PhD thesis, Stanford University.

Boersma, Paul. 1998. *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32:45–86.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Bresnan, Joan. 1996. LFG in an OT setting: Modelling competition and economy. In M. Butt and T. H. King (eds.), *Proceedings of the First LFG Conference*, CSLI Proceedings Online.

Bresnan, Joan. 2000. Optimal syntax. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford University Press.

Bresnan, Joan, and Ashwini Deo. 2001. Grammatical constraints on variation: 'be' in the Survey of English Dialects and (Stochastic) Optimality Theory. Ms., Stanford University.

Bresnan, Joan, Shipra Dingare, and Christopher Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In M. Butt and T. H. King (eds.), *Proceedings of the LFG 01 Conference*. CSLI Publications.

Dingare, Shipra. 2001. The effect of feature hierarchies on frequencies of passivization in English. Master's thesis, Stanford University.

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.

Grimshaw, Jane. 1997. Projection, heads, and optimality. *Linguistic Inquiry* 28:373–422.

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.

Keller, Frank, and Ash Asudeh. 2001. Probabilistic learning algorithms and Optimality Theory. Ms., Saarbrücken, Stanford University.

Koontz-Garboden, Andrew. 2001. A stochastic OT approach to word order variation in Korlai Portuguese. paper presented at the 37th annual meeting of the Chicago Linguistic Society, Chicago, IL, April 20, 2001.

Kuhn, Jonas. 2000. Faithfulness violations and bidirectional optimization. In M. Butt and T. H. King (eds.), *Proceedings of the LFG 2000 Conference, Berkeley, CA*, CSLI Proceedings Online, pp. 161–181.

Kuhn, Jonas. 2001a. *Formal and Computational Aspects of Optimality-theoretic Syntax*. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Kuhn, Jonas. 2001b. Generation and parsing in Optimality Theoretic syntax – issues in the formalization of OT-LFG. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality-theoretic Syntax*, pp. 313–366. Stanford: CSLI Publications.

Kuhn, Jonas. 2002. Extended constraint ranking models for frequency-sensitive accounts of syntax. Slides for a presentation at the Workshop *Quantitative Investigations in Theoretical Linguistics* (QITL), 3-5 October 2002, Osnabrück, Germany.

Kuhn, Jonas. forthcoming. *Optimality-Theoretic Syntax: a Declarative Approach*. Stanford, CA: CSLI Publications.

Lee, Hanjung. 2001. Markedness and word order freezing. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality-theoretic Syntax*, pp. 63–128. Stanford: CSLI Publications.

Sells, Peter. 2001a. *Alignment Constraints in Swedish Clausal Syntax*. Stanford: CSLI Publications.

Sells, Peter (ed.). 2001b. *Formal and Empirical Issues in Optimality-theoretic Syntax*. Stanford: CSLI Publications.

Smolensky, Paul. 1996. On the comprehension/production dilemma in child language. *Linguistic Inquiry* 17:720–731.

Tesar, Bruce B., and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Zinsmeister, Heike, Jonas Kuhn, and Stefanie Dipper. 2002. Utilizing LFG parses for treebank annotation. In *LFG 2002, Athens*.

CONTROL AND COMPLEX EVENT NOMINALS IN HUNGARIAN

Tibor Laczkó

Department of English Linguistics, Debrecen, Hungary

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

1. Introduction

In this paper I will offer a new LFG-account of control phenomena exemplified in (2b). Compare the sentences in (1) and (2).

(1) a. A fiú újra elkezd-ett kiabál-ni.
the boy.nom again start-PAST.3SG.INDEF shout-INF
'The boy started to shout again.'

b. A fiú újra elkezd-t-e a kiabál-ás-t.
the boy again start-PAST-3SG.DEF the shout-DEV-acc
'The boy started the shouting again.'

(2) a. A fiú újra elkezd-t-e énekel-ni a dal-t.
the boy.nom again start-PAST-3SG.DEF sing-INF the song-acc
'The boy started to sing the song again.'

b. A fiú újra elkezd-t-e a dal énekl-és-é-t.
the boy.nom again start-PAST-3SG.DEF the song.nom sing-DEV-its-acc
'The boy started the singing of the song again.'

(1a) and (2a) contain an infinitival complement of the matrix verb. In the spirit of Bresnan (1982) we can assume that it has the open (XCOMP) function and its unexpressed subject argument is functionally controlled by the subject of the matrix predicate. This control analysis can be naturally extended to (1b), which contains a complex event nominal derived from an intransitive verb. We can claim that the unexpressed POSS (or, depending on the details of our analysis, SUBJ) argument is controlled in the same way by the matrix predicate. The nature of the problem (2b) poses is as follows. Here the nominal has been derived from a transitive verb. The argument corresponding to the object of the input verb is mapped onto the POSS (or SUBJ) function and, thus, there is no obvious grammatical function onto which the agent argument which we would like to be realized by PRO could be mapped. So the parallel between (1a) and (2a) cannot be naturally maintained in the case of (1b) and (2b).

One easy way out would be to assume, following Grimshaw (1990), for instance, that at least in the transitive situation the agent argument is suppressed as a result of nominalization (cf. the standard analysis of passivization). However, Szabolcsi (1990) has proved that the unexpressed subject is not a suppressed argument, instead, it has all the major characteristics of a PRO argument (in a GB sense). For example, it can be controlled in the usual fashion, and when it is not controlled, its interpretation is "arbitrary" with the [+human] feature (as opposed to a suppressed argument, which has no such [+human] feature).

In the first part of the paper I will give a brief critical overview of the two salient strategies for addressing this problem that have been proposed so far. The first one is to extend the domain for the control of PRO to the argument structure of derived nominal predicates, cf. Szabolcsi (1990) in a GB framework and Laczkó (1995) in an LFG framework. The other strategy has been outlined by Komlósy (1998), also in an LFG-framework. His fundamental idea is that in the nominal domain, too, there are two distinct semantically unrestricted grammatical functions: POSS and SUBJ. The former is always realized by the possessor constituent and the latter is always unexpressed: it is realized by a PRO as a rule. In the paper I will argue for the first strategy but at the same time I will point out some major problems with the previous two analyses in this vein.

The most important aspects of my new account are as follows.

1) It leaves the standard LFG assumptions about argument structure, the system of grammatical functions available in the nominal domain, and LMT intact.
2) It holds that the unexpressed "subject" argument of a derived nominal argument can only be anaphorically controlled.

3) Just like my earlier account in Laczkó (1995), it postulates a POSS PRO argument for nominals derived from intransitive verbs.
4) In the case of nominals derived from transitive verbs, it assumes that the highest [–o] argument is associated with the zero GF symbol (∅). However, contrary to the general view, it claims that this symbol is ambiguous: in addition to triggering the existential interpretation of the (suppressed) argument it is associated with, it has another function. It can also invoke a "PRO-interpretation" of the argument in question.

2. Previous accounts

In section 2.1 I discuss two approaches similar in spirit. One of them has been proposed by Szabolcsi (1990) in a GB framework and the other by Laczkó (1995) in an LFG framework. Their common feature is that in the case of the type exemplified by (2b) they radically extend the domain for the control of PRO: they insert a PRO argument in the lexical form of the derived nominal.[1] In section 2.2 I summarize Komlósy's (1998) analysis. Its essence is that in the nominal domain, too, the SUBJ grammatical function is available in addition to POSS. However, only PRO arguments can be mapped onto SUBJ.

2.1. PRO in the lexical form of derived nominal predicates

Szabolcsi (1990) assumes a hierarchical lexical structure for derived nominals. This is fundamentally similar to an ordinary GB-style syntactic structure. She inserts the agent PRO argument in the subject position in this lexical structure and she further assumes that it cannot be projected into a syntactic position.[2]

In Laczkó (1995), I offer a rough LFG counterpart of this kind of analysis. In the intransitive case, there is an LFG-style PRO argument in the construction (present in the lexical form of the nominal and in f-structure) and it is mapped onto the POSS function. In the transitive case, I assume that PRO is inserted, without any grammatical function, in the argument structure of the complex event nominal derived from a transitive verb. The obligatory patient argument is mapped onto the POSS function. The PRO is controlled at the level argument structure. The reason for this locus of control is that this PRO has no grammatical function (by the help of which we could capture the control relation at the level of f-structure in the customary fashion).

Three general remarks are in order here.

1. As opposed to Szabolcsi (1990), I do not treat the intransitive and the transitive cases in a uniform manner.[3]

2. Both Szabolcsi's solution and mine handle control relations, at least in the problematic transitive case, in a rather marked fashion: the PRO argument to be controlled is not represented at the usual level: s-structure and f-structure, respectively. Instead, it is inserted into lexical structure and argument structure, respectively.

3. Szabolcsi's solution is a degree less marked than mine in the light of the standard principles of control, because it postulates that the PRO in lexical structure is in the subject position. By contrast, my PRO in the transitive case has no grammatical function at all.

2.2. PRO subject in the f-structure of the DP

[1] The most fundamental aspect of this extension is that control here operates over lexical structure as opposed to syntactic structure.

[2] One of the main motivations for Szabolcsi to introduce the notion of PRO insertion in lexical structure is that at least in the transitive case (that is, in DPs containing a noun head derived from a transitive verb) there is no syntactic position available to PRO, given that the only likely position is always occupied by the possessor constituent. Although in the intransitive case this position would, in theory, be available, Szabolcsi opts for the same lexical PRO insertion device. For details and criticism, see Laczkó (1995).

[3] In the new solution to be proposed in section 3.3, I will maintain this split and I will discuss my motivation for it.

This strategy has been outlined by Komlósy (1998), also in an LFG-framework. His fundamental idea is to describe control phenomena with the well-known tools in an invariant manner. In order to achieve this, however, he has to modify, rather radically, the inventory of grammatical functions available to arguments of derived nominals. In particular, he claims that in the nominal domain, too, there are two distinct semantically unrestricted grammatical functions: POSS and SUBJ. The former is always realized by the possessor constituent and the latter is always unexpressed phonetically: it is realized by a PRO as a rule. Consider the following examples and the grammatical functions Komlósy assumes.[4]

(3) a. a kiabál-ás
the shout-DEV
'the shouting' (SUBJ PRO)

b. a fiú kiabál-ás-a
the boy.nom shout-DEV-his
'the boy's shouting' (the boy: POSS)

(4) a. a dal énekl-és-e
the song.nom sing-DEV-its
'the singing of the song' (SUBJ PRO, the song: POSS)

b. a dal Edit által-i énekl-és-e
the song.nom Edith by-AFF sing-DEV-its
'the singing of the song by Edith' (the song: POSS, by Edith: OBL)

I would like to make the following three remarks on this analysis.

1. It is an unquestionable advantage of this solution, as opposed to the previous two discussed in section 2.1, that it is compatible with the principles and rules of LFG-style control theory, because the PRO to be controlled is mapped onto the SUBJ function in both the transitive and the intransitive cases.

2. At the same time, I think it is a rather serious disadvantageous feature that, in order to respect control theory, it makes the array of grammatical functions available in the DP domain considerably more complex by introducing the lexically always unexpressible SUBJ function. This move, as far as I can tell, has no independent motivation. Furthermore, if we compare the intransitive cases in (3) with non-finite participial constructions in Hungarian, we can observe a surprising discrepancy. In addition to the PRO SUBJ possibility, some participles also admit the incorporated subject pronoun option, realized by inflectional elements, and they also allow their subject argument to be overtly expressed. Consider the following infinitival constructions.

(5) a. Győz-ni kell.
win-INF must
'It is necessary to win.'

b. Győz-ni-e kell.
win-INF-3SG must
'It is necessary for him to win.'

c. János-nak/Neki győz-ni/győz-ni-e kell.
John-dat/he.dat win-INF/win-INF-1PL must
'It is necessary for John/him to win.'

[4] Note that in his system there is no SUBJ Condition in the DP domain, cf. (3b) and (4b). This contrasts with my POSS Condition in Laczkó (1995).

In (5a) there is an uninflected infinitive with a PRO subject. (5b) contains an incorporated pronominal subject. As (5c) demonstrates, the subject of the infinitive can also be expressed by a DP in the dative case. On such occasions agreement marking on the infinitive is optional. What I intuitively find surprising in Komlósy's system is that when the agent argument of the nominal derived from an intransitive verb is expressed by an incorporated pronoun or by a lexical DP in the nominative (or dative), he assumes that this argument is mapped onto the POSS and not the SUBJ function. On the basis of the pattern exhibited by Hungarian infinitival constructions, I think either of the following two alternative solutions would be more in line with this pattern attested in another non-finite domain, and hence more appropriate. One could allow a PRO argument also to be mapped onto POSS, or they could allow the subject to be lexically realized under certain circumstances. In the former case, both in (3a) and in (3b) we would uniformly have the POSS function, while in the latter case we would uniformly have the SUBJ function. Thus, it is strange that Komlósy's account postulates a PRO SUBJ but when the nominal is inflected (the inflection either solely marking person and number agreement or expressing an incorporated pronoun) an entirely different function is assumed: POSS. My suspicion is that Komlósy has been forced to employ this counter-intuitive solution in order to be able to keep the control principles intact.

3. In Laczkó (2000) I postulate that the POSS function in the DP domain is the true counterpart of the SUBJ function in the verbal domain. On these grounds, the verbal domain LMT principles can be naturally adapted to the nominal domain. Only two straightforward assumptions have to be made. A) Because of the intransitive nature of nominals, [+o] functions are unavailable. B) Because there is only one [–r] function available, the POSS, the mapping principles follow the ergative strategy (just like certain Hungarian participles). Its essence is that the default rule maps the [–r] argument onto POSS (in the unaccusative and transitive cases) and there is an elsewhere condition which maps the highest [–o] argument onto POSS in the absence of [–r] in the argument structure (in the unergative case). Although Komlósy (1998) is not explicit on this point, it is obvious that in his system the mapping principles will have to be made rather stipulative and peculiar to the DP domain. In particular, the choice between the SUBJ and the POSS functions will be dependent on the inflected vs. uninflected nature of the noun head.

3. PRO at a different level of representation

The accounts briefly presented in sections 2.1 and 2.2 have one trait in common: in order to capture the control relations of nominals derived from transitive verbs, they introduce a marked feature in one of the components of grammar. The former (Szabolcsi (1990) and Laczkó (1995)) extend the use of PRO and the scope of control relations to the lexical structure (in LFG: the argument structure) of nominals derived from transitive verbs, while the latter (Komlósy (1998)) keeps the treatment of control relations intact, however, it introduces an otherwise unmotivated SUBJ grammatical function. It should be obvious from the foregoing discussion that, because of the nature of the relevant phenomena, it is inevitable for any approach to employ some marked device. Consequently, our fundamental choice between the two main alternatives has to be determined by our preference as to which component of grammar should be affected by the introduction of a special device. My choice both in Laczkó (1995) and here is leaving the well-attested system of grammatical functions in the DP domain untouched (and making the minimally required, intuitively plausible changes in applying the principles of LMT to this domain) and extending the treatment of control phenomena to other components of grammar. The basic motivation for this is that in my view the former area is definitely and entirely in the scope of syntax proper while the latter (the system of various coreference relationships) is not necessarily.

It is important to note that the accounts in both Szabolcsi (1990) and Laczkó (1995) (which extend capturing control phenomena to lexical or argument structure) are far from being fully developed. Moreover, the insertion of a PRO without a grammatical function in the argument structure in the transitive case raises a significant theory internal problem: it is not clear how the standard notion of the completeness constraint can be satisfied. For this reason in what follows I will outline a modified account which handles the "transitive PRO element" at a different level of representation and I will also sketch the way in which I propose control should work.

The essence of the new approach is as follows. Contrary to the standard LFG view, I postulate that the ∅ grammatical function symbol admits more than one interpretation of the argument it is associated with. I claim that in addition to the argument's being existentially bound, this symbol can also trigger a "PRO-like" interpretation of this argument under certain specifiable circumstances. As a consequence, control relations on this account are not treated at the level of f-structure but in semantic structure. The reason for this is that a "PRO-like" element such as this, without any grammatical function, cannot even appear in f-structure and, consequently, control theory as conceived of so far cannot target it. The actual format of semantic structure is not crucial for this proposal as long as it can capture certain basic relationships, for instance the existentially bound nature of an argument associated with the ∅ function symbol. For the sake of concreteness, I will make use of the relevant aspects of Halvorsen's (1983) classical semantic component.

3.1. Halvorsen's (1983) semantic structure

In his model, when linking is made between f-structure and semantic structure, the following operations take place. Base expressions are replaced by their semantic translations, quantifiers acquire their scope, variables are introduced for representing controlled expressions. In addition, embeddings in f-structure as well as purely formal syntactic elements irrelevant to semantics (e. g., marking for case, number and gender) are deleted.[5]

Consider the following pair of examples from Halvorsen (1983).[6]

(6) a. John kicked Pluto.

b.
$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'John'} \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{TENSE} & \text{PAST} \\ \text{PRED} & \text{'kick < (SUBJ) , (OBJ) >'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'Pluto'} \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

c.
$$\begin{bmatrix} \text{PREDICATE} & \textit{kick'} \\ \text{ARG1} & \lambda \textit{PP\{j\}} \\ \text{ARG2} & \lambda \textit{PP\{p\}} \\ \text{TENSE} & \textit{H} \end{bmatrix}$$

(6b) shows the f-structure of (6a) and (6c) demonstrates the corresponding semantic structure. Given that two proper names are involved, the semantic structure does not contain number features solely required for morphosyntactic agreement in f-structure. Furthermore, in semantic structure the two arguments have no grammatical functions.

From our present perspective, another important ingredient of Halvorsen's system is that when in a passive sentence there is no *by*-phrase, that is when the agent argument is associated with

[5] In Halvorsen's model this semantic structure serves as input to the intensional logical level of representation, which, in turn, is input to model theoretic interpretation. These additional levels of his semantic component are irrelevant for my present purposes, so in the remainder of this paper I will be concerned with his semantic structure.

[6] I have made some insignificant changes in the f-structural representation in order for it to be formally similar to the other f-structures in this paper.

the ∅ grammatical function symbol, Halvorsen offers the following semantic translation of this ∅ function.[7]

(7) $\lambda P \exists x[P\{x\}]$

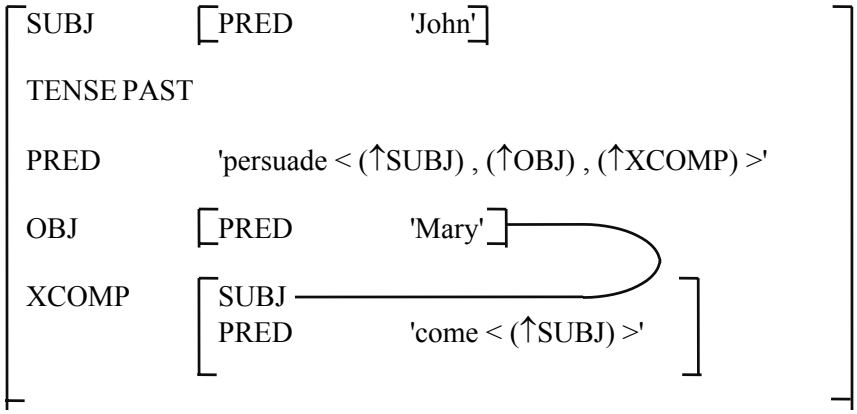
This technically means that argument *x* is existentially bound (∃) in the semantic structure of the given predicate (*P*).

Now let me briefly illustrate how Halvorsen analyses control phenomena through an example of lexically induced functional control.

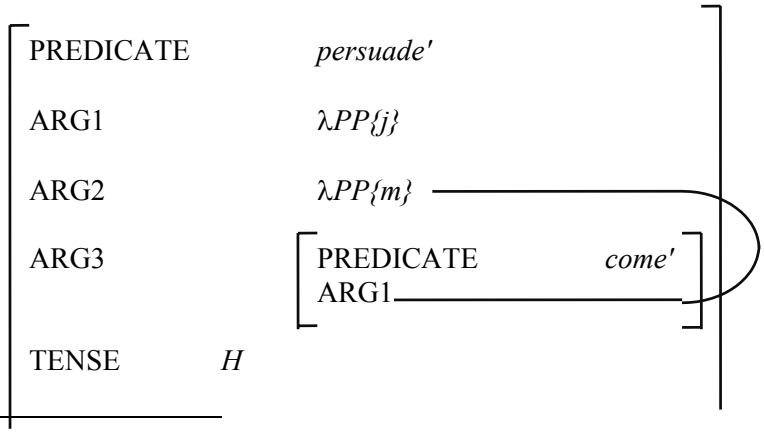
(8) a. John persuaded Mary to come.

b. lexical form: V, 'persuade < (↑SUBJ) , (↑OBJ) , (↑XCOMP) >'
(↑OBJ) = (↑XCOMP SUBJ)

c. f-structure:



d. semantic structure:



[7] As is well-known, it is not only passivization that can optionally associate an argument with the ∅ symbol in LFG. For instance intransitivization is classically treated along similar lines:
(i) eat_1, V 'EAT < (SUBJ) , (OBJ) >'
(ii) eat_2, V 'EAT < (SUBJ) , ∅ >'

3.2. Control relations in Hungarian DPs and Komlósy's (1998) approach

In this section I will offer an overview of what kinds of LFG-style control relations[8] we can assume when in a DP an (agent-like) argument of a derived nominal is controlled from outside (or it has an arbitrary [+human] interpretation) and I will also discuss the relevant aspects of Komlósy's (1998) analysis.

A) When the DP realizes one of the arguments of a matrix verb, we can consider two possibilities: (i) lexically induced functional control and (ii) anaphoric control. The reason why structurally induced functional control is unavailable is that according to the relevant LFG principles in such a relation the argument to be controlled is the subject of an adjunct. From the discussion of Komlósy (1998) above it should be obvious that his choice is (i). Now let us take a closer look at the consequences of this view. It will turn out that it would lead to some significant modifications in the classical version of LFG's control theory. Consider the following examples.

(9) a. A fiú elkezd-t-e a kocog-ás-t.
the boy.nom start-PAST-3SG.DEF the jog-DEV-acc
'The boy started (the) jogging.'

b. A fiú elkezd-t-e a levél ír-ás-á-t.
the boy.nom start-PAST-3SG.DEF the letter.nom write-DEV-its-acc
'The boy started (the) writing of the letter.'

(10) a. A lány rábeszél-t-e a fiú-t a kocog-ás-ra.
the girl.nom talk.into-PAST-3SG.DEF the boy-acc the jog-DEV-onto
'The girl talked the boy into (the) jogging.'

b. A lány rábeszél-t-e a fiú-t a levél ír-ás-á-ra.
the girl.nom talk.into-PAST-3SG.DEF the boy.acc the letter.nom write-DEV-its-onto
'The girl talked the boy into (the) writing of the letter.'

(11) a. A lány ráerőltet-t-e a fiú-ra a kocog-ás-t.
the girl.nom force-PAST-3SG.DEF the boy-onto the jog-DEV-acc
'The girl forced (the) jogging onto the boy.'

b. A lány rábíz-t-a a fiú-ra a levél ír-ás-á-t.
the girl.nom entrust-PAST-3SG.DEF the boy-onto the letter.nom write-DEV-its-acc
'The girl entrusted (the) writing of the letter to the boy.'

In the spirit of Komlósy (1998), in the relevant DPs in all these examples (whether intransitive or transitive) we have to assume the presence of a controlled subject. The DPs in turn have to be taken to be mapped onto XCOMP, because in lexically induced functional control, as a rule, it is the subject of an XCOMP that is controlled. This looks like a plausible assumption in the case of constructions exemplified in (10), given that Komlósy (1992) analyses several oblique case-marked DPs as mapped onto the XCOMP function with certain matrix predicates. Consider:

(12) Kati darab-ok-ra tör-t-e a játék-a-i-t.
Kate.nom piece-PL-onto smash-PAST-3SG.DEF the toy-her-PL-acc
'Kate smashed her toys to pieces.'

[8] Cf. Bresnan (1982) and Bresnan (2001).

There is, however, a significant difference between (10) and (12). In (12) and, generally, when the *-rA* oblique case marked constituent is an argument of the verb *tör* 'smash' it is always mapped onto the XCOMP function. Compare (12) and (13).

(13) *Kati a padló-ra tör-t-e a játék-a-i-t.
Kate.nom the floor-onto smash-PAST-3SG.DEF the toy-her-PL-acc
'*Kate smashed her toys onto the floor.'

By contrast, the corresponding argument of *rábeszél* 'talk somebody into something' in (12) can clearly have an ordinary (OBL) function when it is not expressed by a DP containing a complex event nominal. Consider:

(14) A lány rábeszél-t-e a fiú-t a zöld sapká-ra.
the girl.nom talk.into-PAST-3SG.DEF the boy.acc the green cap-onto
'The girl talked the boy into the green cap.'

Furthermore, as I will discuss in detail below, the relevant DPs in (12) could naturally be regarded as mapped onto OBL if Komlósy's (1998) functional control assumption did not force the XCOMP function. This means that in such cases, in order for the principles of this control type to be applicable in an unmarked manner, Komlósy has to associate predicates like *rábeszél* 'talk somebody into something' with two lexical forms with partially different arrays of grammatical functions: <(SUBJ), (OBJ), (OBL)> and <(SUBJ), (OBJ), (XCOMP)>. I find it even more counter-intuitive that on Komlósy's account in (9) and (11) one has to consider DPs in the accusative to be mapped onto XCOMP. On the one hand, as far as I can see, no other phenomena in Hungarian trigger (or motivate) this move. On the other hand, we can argue in the case of (11), too, that there is simply no need for an additional lexical form for the predicates *ráerőltet* 'force something onto somebody', *rábíz* 'entrust something to somebody', except for Komlósy's functional control assumption.

In my opinion Komlósy's account would be rather appealing if it could leave the classical LFG-style control principles entirely intact (at the above-mentioned expense of complicating the inventory of grammatical functions, mapping principles and lexical forms in the DP domain). However, there is one crucial respect in which it is inevitably forced to make a radical change: it has to assume that arguments with certain [+r] functions, namely OBLiques, can also function as controllers, as opposed to the standard [–r] assumption, cf. (11). In my view, this fact weakens the initial appeal of the account and reinforces its marked features.

Partially on the basis of these considerations, in section 3.3 I will propose that in the relevant instances we are dealing with (an extended notion of) anaphoric control rather than lexically induced functional control. My main motivation, however, will be that, given my assumptions about grammatical functions in the Hungarian DP, the transitive case simply cannot be handled in terms of functional control even if we try to make some sensible modifications. By contrast, the extension of anaphoric control relations appears to be much more plausible.

B) A DP containing a controllee can also be part of an adjunct. Komlósy (1998) is not explicit about which of the two possible analyses he would opt for. Consider the following examples.

(15) a. A kocog-ás után a fiú iv-ott egy kólá-t.
the jog-DEV after the boy.nom drink-PAST.3SG.INDEF a coke-acc
'After (the) jogging the boy drank a coke.'

b. A levél meg-ír-ás-a után a fiú iv-ott egy kólá-t.
the letter.nom PERF-write-DEV-its after the boy.nom drink-PAST.3SG.INDEF a coke-acc
'After (the) writing of the letter the boy drank a coke.'

One possibility is to assume structurally induced functional control in both the intransitive and the transitive cases. In Komlósy's framework we could claim that the DPs functioning as the

complements of the postposition *után* 'after' in both (15a) and (15b) contain an agent argument mapped onto the SUBJ function and, according to the classical LFG-style control principles, the functional structure of this argument is identical to that of a constituent in the matrix clause, *a fiú* 'the boy' in these two examples. The other possibility is the use of the LFG version of a PRO subject, appearing in the lexical form of the nominal predicate and in the f-structure of the matrix DP. In this case the relevant relationship between the controller and the controllee is anaphoric.

C) When there is no controller present in a construction, that is when, in GB terms, we are dealing with arbitrary control, the classic LFG solution is to assume a PRO element in the lexical form of the predicate and in f-structure (but not in c-structure) with the special [+human] interpretation, cf. Bresnan (1982).

(16) A dokumentum[9] megsemmisít-és-e nagyon fontos.
the document.nom destroy-DEV-its very important
'The destruction of the document is very important.'

Although Komlósy is not explicit about such constructions, either, it seems natural that he would adopt the same solution.

In the context of this construction type, I think it is a further peculiar aspect of Komlósy's approach that, if it aims at being consistent, it has to assume that *all* DPs containing a complex event nominal head have to be mapped onto either the XCOMP or the COMP function in accordance with the standard LFG control principles.[10] Compare the following examples.

(17) a. A dokumentum megsemmisít-és-e mindenki-t meglep-ett.
the document.nom destroy-DEV-its everybody-acc surprise-PAST.3SG.INDEF
'The destruction of the document surprised everybody.'

b. A dokumentum-nak a fiú által-i megsemmisít-és-e mindenki-t meglep-ett.
the document-dat the boy by-AFF destroy-DEV-its everybody-acc surprise-PAST.3SG.INDEF
'The destruction of the document by the boy surprised everybody.'

c. A hír mindenki-t meglep-ett.
the news.nom everybody-acc surprise-PAST.3SG.INDEF
'The news surprised everybody.'

The DPs in (17a) and (17b) must be taken to be mapped onto a closed function, given that both arguments of the nominal *megírás* 'writing' are realized in them. In the former the agent argument is

[9] On most accounts (e. g., Szabolcsi (1990), Laczkó (1995) and Komlósy (1998)) the possessor constituent in the nominative and in the dative (the two forms are, as a rule, in complementary distribution) is mapped onto one and the same grammatical function. For different views, see É. Kiss (2000) and Chisarik–Payne (2001).

[10] On the basis of evidence from Balinese, Arka–Simpson (1998) propose that in addition to the customary functional control of XCOMP SUBJ, the control of SUBJ SUBJ should also be incorporated into LFG's control principles. Such a modification would help Komlósy's account to solve the problem posed by (17a) and (18a). Note, however, that for this account to be capable of handling all the relevant cases, it would have to modify the control principles to a much larger extent: it would also have to allow the control of the following additional argument types: OBL SUBJ and OBJ SUBJ, cf. (10) and (11), respectively. I would also like to point out that Dalrymple (2001) analyses certain control relationships in English as anaphoric as opposed to the classical lexically induced functional control account, cf. Bresnan (1982) and (2001). For instance, she postulates the following lexical entries for *try* and *convince* (Dalrymple (2001: 327)).

(i) try V (↑PRED)= 'TRY < SUBJ , COMP >'
(↑COMP SUBJ PRED)= 'PRO'

(ii) convince V (↑PRED)= 'TRY < SUBJ , OBJ , COMP >'
(↑COMP SUBJ PRED)= 'PRO'

It seems to me that if, on the basis of the diagnostics discussed by Dalrymple (2001), Komlósy's (1998) account were recast in terms of anaphoric control, it would get rid of several rather marked features.

an LFG-style PRO mapped onto the SUBJ function, and, depending on the context in which the sentence occurs, it is involved in anaphoric control or it receives arbitrary interpretation. On the basis of the logic of Komlósy's account and the general LFG assumptions about the relationship between XCOMP and COMP, the most likely function to be chosen for the DPs in (17a) and (17b) is COMP. This, however, leads to at least two unfavourable consequences.

First, a predicate like *meglep* 'surprise' has to have the alternative lexical form shown in (18a) in addition to that in (18b), which is needed because of (17c).

(18) a. meglep, V 'SURPRISE < (COMP) , (OBJ) >'
b. meglep, V 'SURPRISE < (SUBJ) , (OBJ) >'

Again, this duplication of the lexical forms of such predicates would not be necessary otherwise. As I will point out in section 3.3, it can be naturally postulated that DPs containing complex event nominal heads are never mapped onto the XCOMP and COMP functions, instead, they are always mapped onto "nominal" functions like SUBJ, OBJ and OBL.

Second, the mapping pattern in (18a) is rather exceptional because it violates the Subject Condition. It is true that, just like in several other languages, in Hungarian there is a set of predicates that lack an argument structure, and, therefore, they must be exempted from the Subject Condition; however, all other verbal predicates, whether finite or non-finite, can be shown to observe it. Thus, (18a) would cause a special problem in this respect. This problem could be avoided by assuming, in a somewhat incoherent manner in Komlósy's system, that *meglep* 'surprise' as used in (17a) and (17b) also has the lexical form shown in (18b). This solution would be incoherent because in Komlósy's functional control pattern the DP with its unexpressed subjects must have the XCOMP function.

3.3. Towards an alternative theory of control in Hungarian DPs

As I have already pointed out, the new analysis I will outline here is similar in spirit to Szabolcsi's (1990) and Laczkó's (1995) accounts, and it contrasts with Komlósy's (1998) approach, inasmuch as it fundamentally proposes an extended domain and mechanism for the treatment of control phenomena and leaves the other relevant components of LFG intact (e. g., LMT principles and the inventory of grammatical functions in DPs). In fact, it is a considerably revised version of Laczkó (1995).

My first general assumption both in Laczkó (1995) and here is that at least Hungarian DPs containing an argument to be controlled from outside are, as a rule, involved in anaphoric control and not in functional control.[11] In Laczkó (1995) I give some justification for this and in section 3.2 I have

[11] As a matter of fact, in Hungarian even the possibility of realizing a propositional argument by an infinitival construction mapped onto the XCOMP function and involved in functional control according to the classical control principles of LFG is rather severely restricted. For instance, there are very few raising (to subject or object) predicates, and even they are not compatible with all kinds of intransitive XCOMP predicates and they are typically incompatible with transitive XCOMP predicates with [+definite] objects. Compare:

(i) János fárad-ni látsz-ott.
John.nom get.tired-INF appear-PAST.3SG.INDEF
'John seemed to be getting tired.'

(ii) ??János fut-ni látsz-ott (a kert-ben).
John.nom run-INF appear-PAST.3SG.INDEF the garden-in
'John seemed to be running (in the garden).'

(iii) Lát-t-am János-t könyv-et olvas-ni.
see-PAST-1SG John-acc book-acc read-INF
ca. 'I saw John involved in book-reading.'

(iv) ??Lát-t-am János-t ez-t a könyv-et olvas-ni.
see-PAST-1SG John-acc this-acc the book-acc read-INF
'I saw John reading this book.'

(v) *Hisz-em János-t haldokol-ni / szeret-ni a zené-t.
believe-PRES.1SG John-acc be.dying-INF like-INF the music-acc
'I believe John to be dying / to like music.'

offered a more detailed discussion, criticizing various aspects Komlósy's (1998) functional control analysis.

The treatment of intransitive constructions, that is, DPs containing nominal heads derived from intransitive verbs, on this new account is the same as in Laczkó (1995), because this type has never raised any theoretical problems. I still assume that in the DP domain the POSS function is semantically unrestricted and it is the true counterpart of the SUBJ function in the verbal domain, with the [–r, –o] feature specification.[12] When the nominal head denotes a complex event and no possessor constituent is present in the DP, I postulate that the nominal's lexical form is associated with the (↑POSS PRED) = 'PRO' equation. The DP can function as either a propositional argument or a propositional adjunct of the matrix predicate, and in both cases the POSS PRO in it is anaphorically controlled. Otherwise the interpretation of this PRO argument is arbitrary with the [+human] feature. In order for anaphoric control to work in these instances, two assumptions have to be made. A) A POSS PRO argument can also be a controllee.[13] B) Anaphoric control is also allowed into DPs with a variety of grammatical functions: SUBJ, OBJ, OBL and ADJ.[14,15]

As I briefly discussed in the introductory part of section 3, the control analysis of the transitive case in Laczkó (1995) raised some important theory-internal problems. Now I propose to eliminate them along the following lines. It is standardly assumed that the argument associated with the ∅ grammatical function symbol has an existential interpretation. The essence of my proposal is that under clearly identifiable circumstance yet another interpretation can be associated with such an argument. This is a "PRO-like" interpretation. In (7) above I have already shown the semantic

---

There are several "intransitive" equi-verbs but practically no "transitive" ones with infinitival XCOMPs. Instead of such infinitival phrases nominal constituents are used, cf.:

(vi) János megpróbál-t fut-ni / level-ek-et ír-ni.
John try-PAST.3SG.INDEF run-INF / letter-PL-acc write-INF
'John tried to run / write letters.'

(vii) *Az igazgató utasít-otta János-t fut-ni / level-ek-et ír-ni.
the manager.nom order-PAST.3SG.DEF John-acc run-INF / letter-PL-acc write-INF
'The manager ordered John to run / to write letters.'

(viii) Az igazgató utasít-otta János-t a fut-ás-ra / level-ek ír-ás-á-ra.
the manager.nom order-PAST.3SG.DEF John-acc the run-DEV-onto / letter-PL-acc write-DEV-their-onto
'The manager ordered John to run / to write letters.'

It is also noteworthy in this connection that Rappaport (1983) argues that in English DPs the "unexpressed" subject arguments of the propositional arguments of nominal predicates are always anaphorically controlled. Compare:

(ix) The captain (SUBJ) ordered the private (OBJ) to leave (XCOMP).

(x) the captain's (POSS) order to the private (OBL) to leave (COMP)

In Hungarian this can be even more straightforwardly assumed, given that practically none of the verbal predicates taking infinitival XCOMP arguments can be nominalized, and the nominal predicates that exist correspond to verbal predicates whose propositional argument, as a rule, is realized by DPs and not infinitival constituents.

On the basis of all these considerations it appears to be plausible to assume in a uniform manner that DPs in Hungarian are involved in anaphoric control relationships in two respects: a) when they express a propositional argument, and b) when they contain a nominal head which has a propositional argument (and this argument is never realized by an infinitival construction, as opposed to the English counterparts).

[12] For an extensive discussion, see Laczkó (2000).

[13] This cannot be either a general or a theory-specific problem, because in several fundamental respects POSS in DPs can be considered a true counterpart of SUBJ in clauses, cf., for instance, Bresnan (1982), Bresnan (2001), Laczkó (1995) and Laczkó (2000). Moreover, in principle it is not implausible in an analysis along these general lines to assume that POSS is actually SUBJ in the DP domain (however, for considerations supporting the POSS view, see Laczkó (2000)). This POSS → SUBJ replacement would only be problematic in Komlósy's (1998) framework, which employs both grammatical functions in Hungarian DPs.

[14] Note in this connection that, as has already been pointed out above, Arka–Simpson (1998) propose that even functional control should be allowed into SUBJ (in addition to XCOMP).

[15] On the formalism encoding anaphoric control, see the discussion of the transitive case.

translation Halvorsen (1983) offers for the customary existential treatment of the ∅ function. For convenience, I repeat it as (19) below.

(19) $\lambda P \exists x[P\{x\}]$

In this vein, I suggest that the "PRO-like" function of this ∅ symbol should translated as follows.

(20) $\lambda P \pi x[P\{x\}]$

The novelty of (20) is that it alternatively replaces the ∃ symbol, indicating existential binding, by π.[16] This symbol prescribes that the argument associated with it has to be handled in semantic structure in the same way as ordinary syntactic PRO arguments are treated at this level of representation: either it has to be (anaphorically) controlled or it has to receive arbitrary interpretation with the [+human] specification.

As is well-known, the ∅(∃) symbol can only be associated with an argument with a negative intrinsic feature: [–o] or [–r]. On the basis of the general characteristics of "syntactic" PRO, my proposal is that the ∅ symbol in this alternative function (∅(π)) can only target the highest negatively specified argument in the argument structure. Consider (21).

(21) ∅
- a. ∅(∃): $\lambda P \exists x[P\{x\}]$, condition: $x = {}^{\wedge}\Theta_{[-o]}$ or $\Theta_{[-r]}$
- b. ∅(π): $\lambda P \pi x[P\{x\}]$, condition: $x = {}^{\wedge}\Theta_{[-o]/[-r]}$

Note that although both rules make it possible to target the argument structure of intransitive predicates in addition to that of transitive ones, certain general LFG principles will rule out the unwanted intransitive cases. When a verb is transitive, (21a) can apply to its argument structure in two different ways. A) It can associate the ∅(∃) symbol with the $^{\wedge}\Theta_{[-o]}$ argument in the course of passivization. B) It can associate this symbol with the $\Theta_{[-r]}$ argument in the course of intransitivization. Both the unergative and the unaccusative intransitive cases will be filtered out by the Subject Condition: if the sole [–o] or [–r] argument is associated with the ∅(∃) symbol, this condition simply cannot be met. (21b) is my newly introduced function attributed to ∅. Although in theory it allows the association of the ∅(π) symbol with the highest [–o] argument in the argument structure of transitive and unergative verbs and also with the [–r] argument of unaccusative verbs, the unergative and the unaccusative cases are ruled out by my Possessor Condition. Apparently, the only additional assumption we need to make is that it depends on the derivational affix (or process) in question whether it employs the ∅(∃) or the ∅(π) function. As a first approximation we can say that passivization and intransitivization makes use of the former while nominalization, at least in Hungarian, utilizes the latter.[17]

At this point the following question arises. How can we reconcile this proposal with the standard LFG control principles? In my answer I will refer to Halvorsen's (1983) model for concreteness and expository reasons, but I think my general assumptions could be shown to carry over to more recent alternative semantic approaches within LFG. Recall that in Halvorsen's analysis functional control relations represented in f-structure are inherited by semantic structure, cf. (8c) and (8d). Likewise, the coindexation, in f-structure, of constituents involved in anaphoric control is also inherited by semantic structure. My claim is that these ordinary cases are supplemented by a special instance of obligatory anaphoric control. Its domain is semantic structure solely. The two relevant aspects of this subtype of control are encoded by the ∅(π) symbol associated with the designated "PRO-like" argument of the derived nominal and the specification in the lexical form of the matrix predicate to the effect that it is an obligatory (anaphoric) control predicate. It is also indicated in the

[16] This function name is mnemonic: π ~ P(RO).

[17] In this paper I have no space to discuss the status and possible analysis, in this framework, of "by-phrases" in nominal, as opposed to passive verbal, constructions. I intend to do this elsewhere.

lexical form which argument of this predicate is a potential controller. For instance, for the matrix predicates in (9), (10) and (11) I propose the lexical forms in (22a), (22b) and (22c), respectively.

(22) a. elkezd, V 'START $< \Theta_1 , \Theta_2 >$'
[–o] [–r]
{+AC}

b. rábeszél, V 'TALK-INTO $< \Theta_1 , \Theta_2 , \Theta_3 >$'
[–o] [–r] [–o]
{+AC}

c. ráerőltet, V 'FORCE-ONTO $< \Theta_1 , \Theta_2 , \Theta_3 >$'
[–o] [–r] [–o]
{+AC}

The interpretation of {+AC} is as follows. If in the f-structure in which the matrix predicate occurs there is a PRO/π argument, the {+AC} argument must understood as the anaphoric controller of this argument. We need the *if*-clause in this description because the majority of the matrix predicates in question can also have a non-propositional argument, in which case there is no control, cf. (14).[18] In the case of PRO, coindexation already takes place in f-structure (and it is inherited by semantic structure), while in the case of π, it takes place in semantic structure.

In my view it is semantic structure that is the most appropriate level for (ultimately) checking control relationships. After all, the complete identity or the coreferentiality of arguments is most naturally handled in semantic terms. It is noteworthy that the scenario which allows certain (but not all) control relationships to be encoded in f-structure and which also requires that all these relationships be checked in semantic structure can be likened to the GB treatment of WH-movement. The scope of WH-expressions is checked in Logical Form. Languages vary as to whether in the course of generating multiple WH-questions they move one (English) or all (Hungarian) WH-phrases into the scopally appropriate positions in S-Structure or they move the rest (English) or all (Chinese) of these constituents at the level of LF.

Finally, I would like to point out that the extended treatment of control phenomena proposed here may provide a more flexible tool for capturing apparently obligatory control relations in cases when, at least according to Grimshaw's widely adopted view, the "controllee does not appear in an argument structure, only in the lexical-conceptual structure of a nominal. Consider the following examples.

(23) a. A professzor elkezd-te a beteg operáció-já-t.
the professor.nom start-PAST.3SG.DEF the patient.nom operation-his-acc
'The professor started the patient's operation.'

b. A professzor elkezd-te az operáció-t.
the professor.nom start-PAST.3SG.DEF the operation-acc
'The professor started the operation.'

c. A professzor elkezd-te az előadás-t.
the professor.nom start-PAST.3SG.DEF the lecture-acc
'The professor started the lecture.'

On the basis of Grimshaw's (1990) generalized diagnostics and also Szabolcsi's (1990) Hungarian-specific tests, we can safely say that *operáció* 'operation' and *előadás* 'lecture' are "simple event" or "result" nominals with a lexical conceptual structure but without an argument structure. However, a

[18] Or, alternatively, we can postulate two lexical forms for these matrix predicates. Note that this is only a possible option on this account, while it is absolutely necessary in Komlósy's (1998) model.

participant in these LCS-s can be taken to be obligatorily controlled by the subject argument of the matrix verb. I think cases like these can be, in principle, more appropriately captured in the framework proposed here, provided that the details of the analysis are fully and consistently developed. It is interesting to note that Williams (1987) uses a similar example to demonstrate that a controlled argument (or participant) does not necessarily have a grammatical function and it is best treated as a kind of an implicit argument, cf.:

(24) The professor performed Mary's operation.

4. Concluding remarks

In this paper I have offered a considerably modified analysis of control phenomena in Hungarian DPs containing complex event nominal heads. Its most important aspects are as follows.

1) It leaves the standard LFG assumptions about argument structure, the system of grammatical functions available in the nominal domain intact, and supplements LMT in Hungarian DPs in a principled manner.
2) It holds that the unexpressed "subject" argument of a derived nominal argument can only be anaphorically controlled.
3) It postulates a POSS PRO argument for nominals derived from intransitive verbs.
4) In the case of nominals derived from transitive verbs, it assumes that the highest [–o] argument is associated with the zero GF symbol (∅) which triggers a "PRO-interpretation" of the argument in question in semantic structure.
5) Thus, the PRO interpretation of an argument comes from two sources: a) from the appearance of a PRO argument with a grammatical function in the lexical form of a predicate and consequently in the f-structure, and b) from the appearance of the π functor associated with the argument in question in the semantic structure.

Finally, let me make two general comments on this approach.

- As I have already mentioned in passing, the treatment of the transitive type calls for a marked solution in some component of the grammar on any account. Given that in my view the system and the operation of grammatical functions clearly belong to (morpho-)syntax, while control relations are best regarded as being in the scope of both syntax and semantics, my motivation was to handle control by such means as do not interfere with solely (morpho-)syntactic phenomena. That is why I set out to explore a possible approach in semantic structural terms.
- The reason why I treat "the transitive PRO" and the "intransitive PRO" differently is that I consider LFG's principles pertaining to argument structure and grammatical relations of primary importance. Therefore, I only employ grammatical functions for the existence of which in a particular construction type we have (independent) evidence. If a grammatical function is available then it must be present – associated with an ordinary or a PRO argument. This principle is especially relevant in the "intransitive case" because the following question arises. What motivates the use of a POSS PRO? Why not postulate the same ∅(π) strategy as in the transitive case? My answer is twofold. On the one hand, in my LMT approach to Hungarian DPs I assume the POSS Condition, which is the mirror image of the verbal domain SUBJ Condition. From this it follows that the sole argument of an unergative or unaccusative derived nominal predicate cannot be associated with the ∅(π) symbol. On the other hand, I think we can claim that certain economical considerations also point in this direction. It can be argued that a suppression process, that is the association of an argument with the ∅(∃) or ∅(π) symbol, is more costly than mapping this argument onto an available grammatical function, which is POSS in this case.

References

Arka, Wayan – Jane Simpson (1998) Control and complex arguments in Balinese In: Butt, Miriam – Tracy H. King, (eds.) *Proceedings of the LFG '98 Conference*. Berkeley: University of California,

Berkeley. pp. 18. (On-line publication: 1998 CSLI Publications, ISSN 1098-6782, http://www-csli.stanford.edu/publications/LFG2/lfg98.html
Bresnan, Joan (1982) Control and complementation. In: Bresnan, Joan (ed.) *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: The MIT Press, 282–390.
Bresnan, Joan (2001) *Lexical-Functional Syntax*. London: Blackwell.
Chisarik, Erika – John Payne (2001) Modelling possessor constructions in LFG: English and Hungarian. In: Butt, Miriam – Tracy H. King, (eds.) *Proceedings of the LFG '00 Conference*. Berkeley: University of California, Berkeley. pp. 18. (On-line publication: 2001 CSLI Publications, ISSN 1098-6782, http://www-csli.stanford.edu/publications/LFG2/lfg00.html)
Dalrymple, Mary (2001) *Lexical Functional Grammar*. New York: Academic Press.
É. Kiss, Katalin (2000) The Hungarian noun phrase is like the English noun phrase. In: Kenesei, István – Alberti, Gábor (eds.) *Approaches to Hungarian. Volume 7. Papers from the Pécs Conference.* Szeged: JATEPress, 119–149.
Grimshaw, Jane (1990) *Argument Structure*. Cambridge, Mass.: The MIT Press.
Halvorsen, Per-Kristian (1983) Semantics for lexical-functional grammar. *Linguistic Inquiry* 14, 567–615.
Komlósy András (1998) *A nomen actionis argumentumainak szintaktikai funkcióiról* [On the syntactic functions of the arguments of the nomen actionis]. Manuscript. Budapest: MTA, Nyelvtudományi Intézet.
Laczkó, Tibor (1995) *The Syntax of Hungarian Noun Phrases – A Lexical-Functional Approach. Metalinguistica 2*. Frankfurt am Main: Peter Lang.
Laczkó, Tibor (2000) Derived nominals, possessors and lexical mapping theory in Hungarian DPs. In: Butt, Miriam – Tracy H. King (eds.) *Argument Realization*. Stanford: CSLI Publications, 189–227.
Rappaport, Malka (1983) On the nature of derived nominals. In: Levin, Lori – Malka Rappaport – Annie Zaenen (eds.) *Papers in LFG*. Bloomington, Ind.: Indiana University Linguistic Club, 113-142.
Szabolcsi, Anna (1990) Suppressed or PRO subjects? The argument structure of event nominals in Hungarian. In: Kenesei, István (ed.) *Approaches to Hungarian, Vol. 3. Structures and Arguments*. Szeged: JATE, 147–181.
Williams, Edwin (1987) Implicit arguments, the binding theory, and control. *Natural Language and Linguistic Theory* 5, 151-180.

Infinitival complements in Norwegian and the form - function relation

Helge Lødrup

University of Oslo

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

0. Abstract

This paper attempts to give a description and an explanation of the fact that infinitival complements in Norwegian raising sentences are often introduced by an empty preposition. The preposition is shown to head a PP with the infinitive phrase as a complement, and it is argued that it is a functional head. The use of the preposition is explained on the basis of the form - function relation. The canonical function of infinitive phrases is a parameter of variation in the world's languages. In Norwegian, their canonical function is object. In raising sentences, the prepositional head makes the complement a PP, thus avoiding an infinitive phrase in a marked function.

1. Introduction[1]

Raising has been a favorite topic in generative grammar for decades. Even so, there is almost nothing written about raising in Norwegian. (The exception is the "complex passive"; see Christensen (1991), Engh (1994).) The standard literature (for example Faarlund et al. 1997) gives the impression that raising from infinitival complements is a marginal phenomenon in Norwegian, with a very small number of raising verbs. This is far from being true, however; Norwegian allows raising with a rather large number of verbs. The infinitival complement in Norwegian raising sentences is often introduced by a preposition. This preposition is in most cases *til* 'to', as in (1)-(2).

(1) Han ser ut til å sove
    he seems (particle) to (PREP) to (INF.MARKER) sleep
    'He seems to sleep'
(2) Vi fikk ham til å sove
    we got him to (PREP) to (INF.MARKER) sleep
    'We made him sleep'

This property distinguishes Norwegian raising sentences from their counterparts in related languages like English, German and Icelandic[2]. It has never been discussed in the literature, however.

[1] I am grateful to the audience at The 7th International LFG Conference for comments and discussion, especially Ida Toivonen. Parts of this work have been presented at The Oslo Syntax Seminar, The Ninth Meeting on the Norwegian Language (MONS 9, Oslo 2001) and The 19th Scandinavian Conference of Linguistics (Tromsø 2002). I am grateful to the audiences for comments and discussion, especially Kersti Börjars.

[2] It is possible to find prepositions in English raising sentences, cf. (i)-(ii).
(i) He prevented there from being a riot (Postal 1974:159)
(ii) I regard Max as being incompetent (Postal 1974:240)
These cases are different from the ones to be discussed here, however. The preposition is required by the verb independently of the category of the following phrase.

Section 2 gives a description of the use and non-use of the preposition. My data is mostly authentic, found in corpora and on web pages. (Some authentic example sentences have been slightly edited.)

2. Raising with(out) the preposition

2.1 Active raising to subject verbs

With active raising to subject verbs, the use or non-use of the preposition *til* is determined by the governing verb. Two groups of verbs usually take the preposition: some aspectual verbs (*ta til* 'begin', *komme til* 'be going to', *skulle til* 'be about to') and some verbs that mean 'seem' or are hyponyms to 'seem' (*late til* 'seem', *se ut til* 'seem', *tegne til* 'seem', *høres ut til* 'sound', *kjennes ut til* 'feel'). Cf. (3) - (4).

(3) Det kommer **til** å skje en ulykke
there comes to (PREP) to (INF.M) happen an accident
An accident is going to happen'
(4) Han ser ut **til** å sove
he seems (particle) to (PREP) to (INF.M) sleep
'He seems to sleep'

There are also verbs that take the preposition optionally; the ones in my dialect are *råke* 'happen', *slumpe* happen', *tendere* 'tend'.

A number of raising to subject verbs do not usually take the preposition. These are the modal verbs (Lødrup 1994, 1996), aspectual verbs like *begynne* 'begin' and *slutte* 'stop', and some verbs which mean 'seem' or something similar (for example *forekomme* 'seem', *vise seg* 'turn out', *virke* 'seem', *synes* 'seem'[3]).

2.2 Active raising to object verbs

The Norwegian raising to object verbs that are usually mentioned in the literature are the perception verbs (*se* 'see', *høre* 'hear', *føle* 'feel', *kjenne* 'feel') and the verb *la* 'let'. These verbs have properties that make them rather atypical raising to object verbs (Barron 1999), and they will be put aside here. There are many other verbs that allow raising to object, however. The verbs in (5) have been found in raising sentences with the preposition *til* 'to'[4]; examples are (6)-(7).

(5) *anslå* 'estimate', *anta* 'assume', *beregne* 'estimate', *bestemme* 'decide', *dedusere* 'deduce', *erklære* 'state', *estimere* 'estimate', *fastslå* 'ascertain', *finne* 'find', *foreslå* 'propose', *forstå* 'understand, *forvente* 'expect', *få* 'make', *mene* 'mean', *oppfatte*

[3] In authentic texts, there is some variation concerning the use and non-use of the preposition. For example, the last three verbs (*vise seg* 'turn out', *virke* 'seem', *synes* 'seem') can be found with the preposition, and most of the verbs that are mentioned above (3) can be found without the preposition.

[4] Some raising verbs can also be used as equi verbs (for example *anslå* 'estimate'). This is not important for my purposes (see section 6), and will not be discussed further here.

'understand', *oppgi* 'state', *oppleve* 'experience', *regne* 'consider', *rapportere* 'report', *spå* 'foretell', *stipulere* 'stipulate', *tippe* 'guess', *vedta* 'decide', *vurdere* 'evaluate'

(6) Vi antar temperaturen **til** å være 10 º C (authentic)
we assume the-temperature to (PREP) to (INF.M.) be 10 º C
'We assume the temperature to be 10 º C'
(7) Vi forventer dekket **til** å være klart (authentic)
we expect the-deck to (PREP) to (INF.M.) be ready
'We expect the deck to be ready'

A handful of verbs can take another preposition than *til*, as in (8).

(8) Jeg holder det **for** å være selvinnlysende at … (authentic)
I take it for to (INF.M.) be obvious that …
'I take it to be obvious that …'

Many verbs that take raising to object with the preposition *til* also allow raising without the preposition, as in (9)-(10).

(9) som dere på forhånd antok det å være (authentic)
like you in advance assumed it to be
'like you in advance assumed it to be'
(10) Ingen forventer lærere å være perfekte... (authentic)
nobody expects teachers to be perfect
'Nobody expects teachers to be perfect'

Some verbs have only been found with the preposition in raising sentences (for example *få* 'make', *mene* 'mean', *rapportere* 'report'). Only a couple of verbs have only been found without the preposition (*hevde* 'claim', *påstå* 'claim', in addition to the perception verbs and the verb *la* 'let').

There is nothing peculiar about the set of raising verbs in Norwegian. Considering the meaning of the verbs, they are rather similar to the set of raising verbs in for example Swedish (Teleman et al. 1999:573, 576-79) or English (Postal 1974).

2.3 Passive raising verbs

In passive raising sentences, the preposition *til* seems to be optional. I know only one verb that takes it obligatorily (*fås* 'make-PASS'). Verbs that are always found with the preposition in active raising sentences can be found with or without the preposition in passive raising sentences. An example is *rapportere* 'report', as in (11)-(12).

(11) Arbeidsmengden rapporteres **til** å være tre til fire timer (authentic)
the-workload report-PASSIVE to (PREP) to (INF.M.) be three to four hours
'The workload is reported to be three to four hours'
(12) Latexallergi rapporteres å være økende (authentic)
latex-allergy report-PASSIVE to (INF.M.) be increasing
'Latex allergy is reported to be increasing'

Norwegian has a number of verbs that allow raising to subject in the passive, even if they do not allow raising to object in the active (Engh 1994:77-87). Many of these verbs can be found with or without the preposition, as in (13)-(14).

(13) De antydes **til** å være 25 000 - 30 000 år gamle (authentic)
they suggest-PASS to (PREP) to (INF.M.) be 25 000 - 30 000 years old
'They are suggested to be 25 000 - 30 000 years old'
(14) Investeringsrammen antydes å være hele 10,5 milliarder (authentic)
the-investment-frame suggest-PASS to (INF.M.) be all 10.5 billions
'The investment frame is suggested be as much as 10.5 billions'

3. Constituency

What is the constituent structure of raising sentences with the preposition *til*? Standard criteria like topicalization and proforms are unavailable, for reasons to be given below. Even so, it is possible to give some arguments that the preposition and the infinitive phrase are one constituent.

First argument: The preposition and the infinitive phrase can never be separated. This is especially striking in a sentence like (15), in which there is a PP between the infinite main verb and the preposition.

(15) Den er erklært av komitéen **til** å være en integrert del (authentic)
it is declared by the-committee to (PREP) to (INF.M) be an integrated part
'It is declared by the committee to be an integrated part'

Second argument: Some raising verbs can take an AP as an alternative to the preposition and the infinitive phrase. With an AP, there can be no preposition, cf. (16)-(17). This indicates that the preposition and the infinitive phrase are one constituent

(16) Han ser (***til**) snill ut
he seems (*to) kind (particle)
'He looks nice'
(17) Vi fikk ham (***til**) glad
we got him (*to) happy
'We made him happy'

Third argument: Raising sentences are often more acceptable when the raised object is topicalized, cf. (18)-(19).

(18) ?Han erklærer verdien å være fem millioner
he declares the-value to (INF.M) be five millions
'He declares the value to be five millions
(19) Verdien erklærer han å være fem millioner
the-value declares he to (INF.M) be five millions
'The value, he declares to be five millions

This property of raising to object sentences is well known from several languages, but not really understood (see Postal 1974, Kayne 1981, Rooryck 1997, Boskovic 1997). However, it only concerns sentences in which the raised object is a sister of an infinitive phrase. (20)-(21) are fully acceptable.

(20) Vi erklærte ham skyldig
we pronounced him guilty
'We pronounced him guilty'
(21) Han erklærer sin markedsverdi **til** å være minst 5 millioner (authentic)
he declares his market-value to (PREP) to (INF.M) be at-least five millions
He declares his market value to be at-least five millions'

The reason a raised object is fully acceptable in the ordinary object position in a sentence like (21) must be that it is not a sister of an infinitive phrase, but of a PP. This also gives an argument that the preposition and the infinitive phrase are one constituent

It might be suggested that the reluctance of a raised object to be a sister of an infinitive phrase could give a functional explanation for the use of the preposition. But it does not, since the preposition can also be used when the object is topicalized, and with raising to subject.

The question is still why Norwegian uses a preposition in raising sentences. My best suggestion is to look at the relation between form and function.

4. Form - function

A syntactic theory must have rules saying what formal categories can be assigned what grammatical functions. These rules are not much discussed. The reason is probably that they are looked upon as trivial, but they do raise some interesting questions. This kind of rules should account for the unmarked cases of the form - function relation. There will always be cases that must be considered marked, for example PP subjects (Bresnan 1994). The rules will to a large extent be universal, but there is some variation. Examples of such rules are (22)-(25).

(22) The unmarked function of DP is SUBJ, OBJ, OBJ_{theta} (the core functions)
(23) The unmarked function of AP is XCOMP, XADJ (the open functions)
(24) The unmarked function of non-finite VP is XCOMP
(25) The unmarked function of PP is (roughly) anything except the core functions

PPs are different from the other categories in that it is difficult to pick out one or two unmarked functions. They can have most functions, and the best generalization would probably be to say what their unmarked functions are not.

The rule for finite CP is a parameter of variation. Dalrymple and Lødrup (2000) discuss the grammatical functions of finite complement clauses. We propose that UG gives two options for realizing a clausal complement: as an object or a COMP. The object complement clauses alternate with DP objects, and behave syntactically like grammatical objects, in the sense that they

topicalize, correspond to a subject in the passive, etc. The COMP complement clauses lack these object properties. We claim that some languages have object clausal complements, some have COMP clausal complements, and some have both, in the sense that different predicates take complement clauses with different functions. Norwegian is mentioned as a language that has object clausal complements (but the existence of exceptions is noted).

Another parameter of variation is the function of infinitive phrases. In Norwegian, their unmarked function is object[5] (Lødrup 1991, see also Andrews 1982 on Icelandic). The infinitival complement of most equi verbs[6] alternates with a DP object, it can topicalize, and it can be realized as a subject in the passive. Cf. (26)-(29).

(26) Vi har akseptert å betale skatt
we have accepted to pay taxes
'We have accepted to pay taxes'
(27) Vi har akseptert dette
we have accepted this
'We have accepted this'
(28) Å betale skatt har vi akseptert
to pay taxes have we accepted
'To pay taxes, we have accepted'
(29) Å betale skatt er blitt akseptert
to pay taxes has been accepted
'To pay taxes has been accepted'

The passive (29) also shows that the infinitive phrase can occur without a realized controller. This means that control is anaphoric, and not functional (Bresnan 1982, Bresnan 2001:267-301).

Differences between lexical categories in taking infinitive phrase complements give an important argument that the unmarked function of Norwegian infinitive phrases is object. The situation is the following:

Adjectives and nouns do not take infinitive phrases as complements[7], while prepositions take infinitive phrases freely. Cf. (30)-(32).

(30) *et forsøk å finne en vei ut
an attempt to find a way out
(31) *stolt / ivrig å gjøre dette
proud / eager to do this

[5] I assume that being a subject is a marked function, since an infinitive phrase subject cannot be in the canonical subject position; it must topicalize.

[6] Equi verbs are often called control verbs; I will use the term equi verbs to avoid confusion with the LFG notion of (functional and anaphoric) control.

[7] There is an exception to this claim; an adjective can take an infinitival COMP in the "tough movement" construction. Besides, a couple of clarifications might be in place. First, a handful of adjectives can take an object (cf. *ulikt ham* 'unlike him', see Platzack (1982) on Swedish). This object can be realized as an infinitive phrase; cf. *ulikt å være kunstner* 'unlike to be an artist'. Second, certain nouns can take an infinitive phrase, as in *kunsten å fiske* 'the art to fish' (i.e. 'the art of fishing'). These are not complements, however, but appositions cf. Faarlund et al. (1997:1011-1014).

(32) ved / fra å svømme
by / from to swim
'by / from swimming'

With adjectives and nouns, an infinitival complement must be the object of a preposition, as in (33)-(34).

(33) et forsøk på å finne en vei ut
an attempt on to find a way out
'an attempt to find a way out'
(34) stolt over / ivrig etter å gjøre dette
proud over / eager after to do this
'proud / eager to do this'

The generalization is that the transitive lexical categories can take infinitive phrases, while the intransitive ones cannot. This is an important argument that the unmarked function of Norwegian infinitive phrases is object.

It is not impossible for a Norwegian infinitive phrase to be COMP or XCOMP, but they must be considered marked functions. In English, on the other hand, COMP or XCOMP do not seem to be marked functions for infinitive phrases.

We are now in a position to answer the question why Norwegian uses an empty preposition in raising sentences. In LFG, raising is basically a case of functional control where the controller does not get a thematic role from its governing verb. The complement in a raising sentence is an XCOMP. This is a marked function for an infinitive phrase in Norwegian. Using the preposition gives us a PP instead.

The preposition makes it possible to avoid a marked form - function assignment, and to get a formal category that is suited for the function XCOMP. One could ask, then, why Norwegian doesn't always use a preposition in a raising sentence. The answer is that other constraints, like economy of expression and full interpretation, pull in the opposite direction.

5. The nature of *til*

In sentences like (35)-(36), the preposition *til* introduces a PP oblique with an infinitive phrase object.

(35) Jeg vil gjerne bidra til å redde salamanderen
I will gladly contribute to (PREP) to (INF.M) save the-newt
'I will gladly contribute to save the newt'
(36) Vi har overtalt ham til å betale
we have persuaded him to (PREP) to (INF.M) pay
'We have persuaded him to pay'

There are several important syntactic differences between a PP XCOMP in a raising sentence, and a PP oblique with an infinitive phrase object.

First difference: In raising sentences, control must be functional. In sentences like (35)- (36), control must be anaphoric. This follows from the fact that the

infinitive phrases are objects of prepositions. That control is not functional can be seen from (37), the passive version of (36), where there is no syntactically realized controller.

(37) Det ble bidratt til å redde salamanderen
it was contributed to (PREP) to (INF.M) save the-newt
'People contributed to save the newt'

The normal, and expected, situation is that an infinitive phrase has anaphoric control when it is the complement of a preposition. Functional control is not available for an object. Besides, functional control is a lexical property of the verb. The verb cannot specify functional control of the complement of an oblique, which is too far "down" in the functional structure.

Second difference: Related to the difference in control is the fact that the infinitive phrase object in a PP oblique alternates with a DP object, cf. (38)-(39), while the infinitive phrase in a raising sentence does not, cf. (40)-(41).

(38) Jeg vil gjerne bidra til dette
I will gladly contribute to this
'I will gladly contribute to this'
(39) Vi har overtalt ham til dette
we have persuaded him to this
'We have persuaded him to this'
(40) *Han ser ut til dette
he seems (particle) to this
(41) *Vi fikk ham til dette
we got him to this

In some cases, like (42), the infinitive phrase in a raising sentence might seem to alternate with a DP.

(42) De anslår antallet til en million
they estimate the number to one million
'They estimate the number to one million'

However, sentences like (42) are not grammatically parallel to the raising sentences discussed here; this is shown in section 7.

Third difference: A PP oblique can topicalize, cf. (43)-(44), while the PP in a raising sentence can not, cf. (45)-(46)[8].

(43) Til å redde salamanderen vil jeg gjerne bidra
to (PREP) to (INF.M) save the-newt will I gladly contribute
'To save the newt, I will gladly contribute '
(44) Til å betale har vi overtalt ham
to (PREP) to (INF.M) pay have we persuaded him
'To pay, we have persuaded him'

[8] If (43) and (44) sound slightly unnatural, the reason is probably that Norwegian prefers preposition stranding wherever possible.

(45) ***Til** å sove ser han ut
to (PREP) to (INF.M) sleep seems he (particle)
(46) ***Til** å bli glad fikk vi ham
to (PREP) to (INF.M) be happy got we him

This difference can also be related to control. It follows from an independent generalization which says that an XCOMP with a verbal (f-structure) head cannot enter into an unbounded dependency[9]. This restriction is in some way part of the classical "Higgins' generalization", which also prohibits a COMP to topicalize (cf. Dalrymple and Lødrup 2000, Lødrup 2001, and references there).

Fourth difference: The infinitive phrases show a corresponding difference concerning unbounded dependencies. The infinitive phrase of a PP oblique can topicalize, as expected, cf. (47)-(48), while the infinitive phrase of a PP in a raising sentence can not, cf. (49)-(50).

(47) Å redde salamanderen vil jeg gjerne bidra til
to (INF.M) save the-newt will I gladly contribute to (PREP)
'To save the newt, I will gladly contribute '
(48) Å betale har vi overtalt ham til
to (INF.M) pay have we persuaded him to (PREP)
'To pay, we have persuaded him'
(49) *Å sove ser han ut **til**
to (INF.M) sleep seems he (particle) to (PREP)
(50) *Å bli glad fikk vi ham **til**
to (INF.M) be happy got we him to (PREP)

The fact that the infinitive phrase in a raising sentence cannot topicalize indicates that the preposition is a functional head (see for example Corver and van Riemsdijk 2001:2-3). Being the complement of a functional head, the infinitive phrase has no grammatical function. In LFG, only a phrase that has a grammatical function can enter into an unbounded dependency, since unbounded dependencies are accounted for at the level of functional structure. This means that the ungrammaticality of (49)-(50) follows when the preposition is a functional head.

The assumption that the preposition is a functional head also accounts for the fact that functional control takes place "across" the preposition. Functional control is accounted for in the functional structure, in which a functional head is not present. Structure sharing will therefore take place as if there was no preposition.

The preposition *til* in raising sentences is a head with a minimal meaning, which is there to satisfy the need for a head in the PP. Seen this way, the preposition *til* could be compared to English *do* in *do*-insertion (Bresnan 2000).

An important fact to be accounted for is that the need for the preposition *til* varies with the choice of verb. As shown in section 2, the pattern is rather

[9] This generalization gives correct predictions for complements of Norwegian raising verbs, except for complements of auxiliaries (Lødrup 1996). It is not universal, however, Wurmbrand (2001:159) says that topicalization is possible in German raising sentences.

complicated, with some verbs requiring it, some verbs not allowing it, and some verbs taking it optionally. However, verbs that take an XCOMP are known to have differing requirements for the formal category of their XCOMP. (Cf. for example *Kim grew poetical / *a success, Kim ended up poetical / a success*. See Pollard and Sag 1987:122-23.) It is therefore necessary that a verb that takes an XCOMP specifies the formal category of this XCOMP (Falk 2001:129-30).

6. Equi sentences

So far, only raising verbs have been discussed. But if my analysis is correct, there is no reason the functional head *til* should be restricted to raising sentences. In LFG, an equi verb takes a complement with functional control (an XCOMP) or a complement with anaphoric control (a COMP or an OBJ). An equi verb that takes an XCOMP differs syntactically from a raising verb only in giving a thematic role to the controlling argument. My analysis is therefore compatible with the existence of equi verbs that take an XCOMP with the functional head *til*.

A clear case is the verb *tenke* 'think, intend'. This verb can take an infinitival complement or a PP with the preposition *til,* as in (51)-(52). The two seem to be synonymous. The verb can also take a PP with the preposition *på* 'on', as in (53), which gives a slightly different meaning.

(51) Jeg har tenkt å gjøre det
I have thought to (INF.M.) do it
'I intend to do it'
(52) Jeg har tenkt **til** å gjøre det
I have thought to (PREP) to (INF.M.) do it
'I intend to do it'
(53) Jeg har tenkt på å gjøre det
I have thought on to (INF.M.) do it
'I have thought about doing it'

The PP with the preposition *på* 'on' is an ordinary oblique, with the expected properties. The PP with the preposition *til* is an XCOMP; it has the same properties as the corresponding PP in a raising sentence. (54)-(58) show the differences concerning alternation with DP objects, topicalization and passivization.

(54) Jeg har tenkt på / ***til** oppdraget
I have thought on / to the-assignment
'I have thought about the assignment'
(55) På / ***til** å gjøre det har jeg tenkt lenge
on / to (PREP) to (INF.M) it have I thought long
'About doing it, I have thought for a long time'
(56) Å gjøre det har jeg tenkt på / ***til** lenge
to do it have I thought on / to long
'Doing it, I have thought about for a long time'

(57) Å gjøre det er blitt tenkt på / ***til*** lenge
to do it has been thought on / to long
'Doing it has been thought about for a long time'
(58) Det er blitt tenkt lenge på / ***til*** å gjøre det
it has been thought long on / to (PREP) to (INF.M) do it
'It has been thought for a long time about doing it'

Especially striking is the contrast in (57) and (58). The verb *tenke* 'think, intend' can passivize when it takes the preposition *på*, allowing both the pseudopassive (57) and the impersonal passive (58). With the preposition *til*, however, it does not allow passivization at all. The reason must be that the complement with the preposition *til* is an XCOMP and needs a syntactically realized controller[10].

Other cases include some inherently reflexive verbs. The verb *bestemme seg* 'decide' can take a PP with either the preposition *for* 'for' or the preposition *til*. The differences between the PPs are as expected cf. (59)-(62).

(59) De bestemte seg for / **til** å gjøre det
they decided (reflexive) for / to (PREP) to (INF.M.) do it
'They decided on doing it'
(60) De bestemte seg for / ***til*** dette
they decided (reflexive) for / to this
'They decided on this'
(61) For / ***til*** å gjøre det må vi bestemme oss
for / to (PREP) to (INF.M.) do it must we decide (reflexive)
'On doing it, we must decide'
(62) Å gjøre det må vi bestemme oss for / ***til***
to (INF.M.) do it must we decide (reflexive) for / to (PREP)
'Doing it, we must decide on'

A problem with *bestemme seg* 'decide' is that passivization is unavailable because the verb is inherently reflexive. This makes it difficult to establish that control is functional.

There are also a number of verbs with a thematic object that seem to take an XCOMP with the preposition *til*. An example is *spesifisere* 'specify', as in (63).

(63) Han spesifiserte utgiftene **til** å gjelde kost og losji
he specified the-expenses to (PREP) to (INF.M.) concern food and lodging
'He specified the expenses to concern food and lodging'

Neither the PP nor the complement can topicalize, cf. (64)-(65).

(64) ***Til** å gjelde kost og losji spesifiserte han utgiftene
to (PREP) to (INF.M.) concern food and lodging specified he the-expenses

[10] Nominalizations give more evidence for this difference in control. The nominalization *tanke* 'thought' can take a PP with *på*, but not a PP with *til*. (Cf. *tanken på/****til*** *å gjøre det* 'the-thought on / to (PREP) to (INF.M.) do it'.) This follows from the fact that nouns cannot induce functional control.

(65) *Å gjelde kost og losji spesifiserte han utgiftene **til**
to (INF.M.) concern food and lodging specified he the-expenses to (PREP)

Verbs with this option include the ones in (66).

(66) *akseptere* 'accept', *bedømme* 'judge', *beskrive* 'describe', *betrakte* 'regard', *etablere* 'establish', *forandre* 'change', *presisere* 'make clear', *spesifisere* 'specify', *tolke* 'interpret', *utpeke* 'appoint', *utnevne* 'appoint'

Again, it is difficult to establish that there is functional control. However, it seems to be impossible to leave out the controlling object with these verbs, this might be taken as an indication that control is functional.

7. Another XCOMP with *til*

A possible parallel to the use of the preposition *til* discussed here can be found in sentences like (67)-(69), which contain a a DP object and a PP XCOMP with the preposition *til*.

(67) Han pratet meg til nervevrak
he chatted me to nervous-wreck
'He chatted me into a nervous wreck'
(68) De anslår antallet til en million
they estimate the number to one million
'They estimate the number to one million'
(69) Vi har spesifisert målsettingen til en million
we have specified the-aim to one million
'We have specified the aim to one million'

The use of the preposition *til* in sentences like (67)-(69) can be explained in the same way as in sentences with infinitive phrases. Again, the point is the relation between form and function. Being an XCOMP is a marked function for a DP. Norwegian only has a handful of verbs that allow a DP XCOMP (for example *være* 'be', *bli* 'become', and *hete* 'be-called'). Both with DPs and infinitive phrases, the preposition *til* is used to avoid a marked form - function assignment, and to get a formal category that is suited for the function XCOMP. However, there are arguments that the PPs in (67)-(69) do not have the same analysis as the ones with infinitival complements.

First argument: With a DP complement, the preposition *til* seems to have some meaning. There is at least a tendency that it is used only with non-stative verbs. Stative verbs usually take the preposition *som* 'as' instead (Eide and Åfarli 1999:170), as in (70).

(70) De betrakter ham som / *til en ydmyk mann
they regard him as / to a humble man
'They regard him as a humble man'

With an infinitival complement, on the other hand, the preposition *til* is also used with stative verbs (see examples (1), (6), (7)).

Second argument: The preposition does not behave like a functional head in sentences like (67)-(69). Its complement can topicalize, cf. (71).

(71) En million har vi spesifisert den til
a million have we specified it to
'A million, we have specified it to'

This fact precludes an analysis of *til* as a functional head with a DP complement. (See for example Corver and van Riemsdijk (2001:2-3). This kind of elements have been called semi-lexical, cf. Eide and Åfarli (2001), Rafel (2001).) The topicalized complement must have some grammatical function, which in turn requires an analysis in which the head is lexical.

If the non-functional preposition *til* takes an object, there is no reason this object should not be realized as an infinitive phrase. (72) is an example.

(72) Vi har spesifisert målsettingen til å unngå tap
we have specified the-aim to (PREP) to (INF.M.) avoid losses
'We have specified the aim to avoid losses'

This infinitive phrase does not behave like the infinitival complement of the functional head *til*. It has the syntactic properties of an object of a preposition. It can topicalize, as in (73), and the whole PP can topicalize, as in (74).

(73) Å unngå tap har vi spesifisert den til
to (INF.M.) avoid losses have we specified it to (PREP)
'To avoid losses, we have specified it'
(74) Til å unngå tap har vi spesifisert den
to (PREP) to (INF.M.) avoid losses have we specified it
'To avoid losses, we have specified it'

Control of the infinitival subject is not functional in (72). The infinitival subject could be understood to be either somebody who is not mentioned in the sentence, or the subject. The subject cannot be a functional controller, however, since it can be left out in the passive, as in (75).

(75) Målsettingen ble spesifisert til å unngå tap
the-aim was specified to (PREP) to (INF.M.) avoid losses
'The aim was specified to avoid losses'

There must be a PRO subject for the infinitive phrase in (72)[11] [12].

[11] A difficult question is what is obligatorily controlled in the PP XCOMP in (72). It cannot be the subject position of the infinitive phrase, it seems to be the PP as a whole, or the infinitive phrase as a whole.

[12] A complication is that some stative verbs allow the preposition *som* 'as' as a functional head. This means that even with *som*, there can be a contrast between an infinitive phrase with anaphoric control, as in (i), and an infinitive phrase with functional control, as in (ii).
(i) Dette må betraktes som å bli degradert
this must regard-PASS as to be degraded
'This must be regarded as being degraded' [This note continues on the next page]

8. Other strategies

Norwegian also has other ways to get a less marked form – function relation with XCOMPs, even if they are less important in terms of the number of verbs involved.

A small number of verbs take a complement without the infinitival marker as an XCOMP, both raising and equi verbs. Cf. (76)-(77).

(76) Vi så ham svømme
we saw him swim
'We saw him swim'
(77) Vi ba ham gjøre det
we asked him do it
'We asked him to do it'

I assume that an infinitive phrase is an IP with the infinitival marker in I. Without an infinitival marker there is no IP, only a VP. The unmarked function of a non-finite VP is XCOMP (see section 4). This means that not using the infinitival marker can be seen as a strategy for providing a form that is suitable for the function XCOMP[13].

A number of raising verbs can sometimes be found with a VP XCOMP in texts. This sounds somewhat strange to me, but sentences like (78) - (79) are not uncommon.

(78) Det synes være kommet nye folk der (authentic)
there seem be come new people there
'There seem to have come new people there'
(79) Produksjonen 2001 forventes bli på ca. 4-500 tonn (authentic)
the-production 2001 expect-PASS be about 4-500 tons
The production in 2001 is expected to be about 4-500 tons'

Norwegian has still another strategy to realize an XCOMP. (80)- (81) are so-called pseudocoordinations.

(80) Det sitter en mann på kontoret og skriver dikt
there sits a man in the-office and writes poems
'A man is writing poetry in the office'
(81) Det driver og blir varmere (authentic)
it carries-on and gets warmer
'It is getting warmer'

(ii) De blir betraktet **som** å likne den gitte situasjonen (authentic)
they are regarded as to resemble the given situation
'They are regarded as resembling the given situation'
The differences concerning topicalization are as expected.

[13] It is an old insight that phrases without the infinitival marker are often non-nominal. It should be mentioned, however, that the distribution of the infinitival marker in Norwegian is more complicated, see Johannessen (1998).

In Lødrup (2002), I show that most pseudocoordinations are really subordinations, in which an XCOMP copies the morphosyntactic features of its governing verb. Most verbs that take pseudocoordinations are equi verbs, a couple are raising verbs. This means that what seems to be the second coordinated VP in (80)- (81) is really an XCOMP. This could be seen as another way of avoiding an infinitive phrase as an XCOMP.

9. Conclusion

The functional head *til* introduces complements with functional control both with raising verbs and equi verbs. This requires a theory in which the complement of an equi verb can have (but does not need to have) the same syntactic properties as the complement of a raising verb. LFG's theory of control and complementation gives the framework needed.

LFG's theory of control and complementation was proposed twenty years ago (Bresnan 1982), and it has been remarkably stable over the years (see Bresnan 2001:267-301). When it was proposed, its distinction between anaphoric and functional control and its grammatical function XCOMP had no parallels in competing theories. In traditional Chomskyan syntax, the raising - equi distinction was taken to be decisive for the syntactic properties of controlled complements. It is striking that recent Chomskyan syntax is diminishing the difference between raising and equi (Hornstein 1999, Manzini and Roussou 2000), thus approaching a view that is more similar to LFG's.

LITERATURE

Andrews, Avery 1982 The representation of case in modern Icelandic. In Joan Bresnan (ed.) *The Mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press.

Barron, Julia 1999 *Perception, volition and reduced clausal complementation*. Dissertation. Department of Linguistics, University of Manchester.

Boskovic, Zeljko 1997 *The Syntax of Nonfinite Complementation: An Economy Approach*. MIT Press.

Bresnan, Joan 1982 Control and complementation. In Joan Bresnan (ed.) *The Mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press, 282-390.

Bresnan, J. 1994 Locative inversion and the architecture of Universal Grammar. *Language* 70, 72 - 131.

Bresnan, Joan 2000. Optimal Syntax. In *Optimality Theory: Phonology, Syntax and Acquisition*, edited by Joost Dekkers, Frank van der Leeuw and Jeroen van de Weijer. Pp.334--385. Oxford: Oxford University Press.

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

Christensen, Kirsti Koch 1991 Complex passives reanalyzed. *Working Papers in Scandinavian Syntax* 48, 45-75.

Corver, Norbert and Henk van Riemsdijk 2001 Semi-lexical categories. In Norbert Corver and Henk van Riemsdijk (eds.) *Semi-lexical Categories*. Berlin: Mouton de Gruyter. Pp. 1-19.

Dalrymple, Mary and Helge Lødrup 2000 The grammatical functions of complement clauses. In Miriam Butt and Tracy Holloway King (eds.) *Proceedings of the LFG00 Conference*. CSLI Publications. http://csli-publications.stanford.edu/LFG/

Eide, Kristin M. and Tor A. Åfarli 1999 The Syntactic Disguises of the Predication Operator. *Studia Linguistica* 53, 155-181.

Eide, Kristin M. and Tor A. Åfarli 2001 Semi-lexical heads in a semantically charged syntax. In Norbert Corver and Henk van Riemsdijk (eds.) *Semi-lexical Categories*. Berlin: Mouton de Gruyter. Pp. 455-473.

Engh, Jan 1994 *Verb i passiv fulgt av perfektum partisipp: Bruk og historie*. Oslo: Novus.

Faarlund, Jan Terje et al. 1997 *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Falk, Yehuda N. 2001 *Lexical-Functional Grammar: An Introduction to Parallel Constraint-based Syntax*. Stanford, CA: CSLI publications.

Hornstein, Norbert 1999 Movement and control. *Linguistic Inquiry* 30, 1, 69-96.

Johannessen, Janne Bondi 1998 Negasjonen ikke: Kategori og syntaktisk posisjon. In Jan Terje Faarlund, Brit Mæhlum and Torbjørn Nordgård (eds) *MONS7: Utvalde artiklar frå det 7. Møtet Om Norsk Språk*. Novus forlag, Oslo.

Kayne, Richard S. 1981 On certain differences between French and English. *Linguistic Inquiry* 12, 3 349-371.

Lødrup, Helge 1991 Clausal complements in English and Norwegian. *Norsk lingvistisk tidsskrift*, 105-136. (Oslo: Novus.)

Lødrup, Helge 1994 "Surface proforms" in Norwegian and the definiteness effect. In M. Gonzalez (ed.) *Proceedings of the North East Linguistics Society 24*. Amherst: GLSA, Department of Linguistics, University of Massachusetts, 1994. Pp. 303-315.

Lødrup, Helge 1996 Properties of Norwegian auxiliaries. In Kjartan G. Ottósson et al. (eds.) *Proceedings of The Ninth International Conference of Nordic and General Linguistics*. Oslo: Novus, 1996. Pp. 216-228.

Lødrup, Helge 2001 Clausal Arguments and Unbounded Dependencies. In Arthur Holmer, Jan-Olof Svantesson and Åke Viberg (eds). *Proceedings of the 18th Scandinavian Conference of Linguistics*. Travaux de l'Institut de Linguistique de Lund. Lund University, Sweden

Lødrup, Helge 2002 The Syntactic Structures of Norwegian Pseudocoordinations. *Studia Linguistica* 56, 2, 121-143.

Manzini, M. Rita and Anna Roussou 2000 A minimalist theory of A-movement and control. *Lingua* 110, 409-447.

Platzack, Christer 1982 Transitive Adjectives in Swedish: A Phenomenon with Implications for the Theory of Abstract Case. *Linguistic Review* 2, 1, 39-56.

Pollard, Carl and Ivan A. Sag 1987 *Information-based Syntax and Semantics*. Volume 1. Stanford, CA: CSLI publications.

Postal, Paul M. 1974 *On Raising*. Cambridge, Mass: MIT Press.

Rafel, Joan 2001 As for *as* / *for*, they are semi-lexical heads. In Norbert Corver and Henk van Riemsdijk (eds.) *Semi-lexical Categories*. Berlin: Mouton de Gruyter. Pp.475-503.

Rooryck, Johan 1997 On the interaction between raising and focus in sentential complementation. *Studia Linguistica* 51, 1-49.

Teleman, Ulf, Staffan Hellberg and Erik Andersson. 1999. *Svenska akademiens grammatik*. Stockholm: Svenska Akademien.
Wurmbrand, Susanne 2001 *Infinitives: Restructuring and Clause Structure*. Berlin: Mouton de Gruyter.

Helge Lødrup
University of Oslo
Department of Linguistics
Pb. 1102, Blindern
N-0317 Oslo, Norway
helge.lodrup@ilf.uio.no
http://folk.uio.no/helgelo/home.html

Prominence Mismatches and Differential Object Marking in Bantu*

Yukiko Morimoto
Universität Düsseldorf

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

* I am grateful to Peter Sells and Dieter Wunderlich for valuable feedback on an earlier draft, and to the audience at LFG-02 for useful questions and comments. Thanks also to Barbara Stiebels for the discussion of the Leixcal Decomposition Grammar formalism incorporated in the present OT analysis. All remaining errors or misrepresentations are my own. This research was carried out as part of the project in Sonderforschungsbereich (SFB) 282 'Theorie des Lexikons', supported by the German Science Foundation (DFG).

Abstract

Majority of Bantu languages encode subjects by head-marking and objects by positional licensing. This reflects a point in the historical process whereby positional licensing of objects becomes obligatory due to the loss of inflecctional morphology. What we observe in synchronic grammar is considerable variation both across and within languages in the use of head-marking morphology for objects. This paper examines this variation under the general concept of DIFFERENTIAL OBJECT MARKING (DOM). I show that an Optimality-Theoretic LFG account of DOM in Bantu enables us to provide a unified account of differential marking of objects across typologically diverse languages—realized by case, agreement, or by lexical choice—which is conditioned by the same semantic/pragmatic factors (animacy and definiteness/specificity). The present analysis also illustrates that cross-linguistic variation and language-internal variation (= 'optionality') operate within a single typological space made available by the system of universal, violable constraints.

1 Differential Object Marking

DIFFERENTIAL OBJECT MARKING (DOM), in which only some direct objects are case marked due to their semantic and pragmatic properties, has been extensively documented in functional and typological studies on case marking languages (e.g. Silverstein 1976, Comrie 1979, 1980, Croft 1988, among others). According to earlier studies, DOM takes many forms. For example in Sinhalese, animate-referring objects may be optionally case-marked (Gair 1970). In Hebrew definite objects are obligatorily case marked (Givón 1978). In Romanian, object case marking is obligatory for animate-referring personal pronouns and proper nouns, optional for others, and excluded for a third set (Farkas 1978). In recent work, Aissen (2000) proposes a single generalization of these seemingly disparate facts, and provides a systematic account of previously documented instances of DOM within Optimality Theory (Aissen 1999, Bresnan 2000, Sells 2001a,b). Aissen's key generalization is stated in (1).

(1) The higher in prominence a direct object the more likely it is to be overtly case marked—where the dimensions along which prominence is assessed include animacy (1a) and definiteness (1b).

   a. Animacy: Human > Animate > Inanimate

   b. Definiteness: Pronoun > Name > Definite > Indef. Specific > Non-specific

Despite the impressive body of work on DOM, this phenomenon has received relatively little attention outside case marking languages. In this paper, I present data from Bantu languages, which are primarily head marking, and argue that DOM in case marking languages and previously observed variation in the use of object agreement in some Bantu languages are one and the same phenomenon conditioned by the single generalization in (1).

The present discussion proceeds as follows. In section 2, I present the core facts on object marking in Bantu. The two crucial points will be the following: (i) object marking in Bantu is conditioned by animacy and definiteness, just as DOM is in case marking languages, and hence deserves a unified explanation; and (ii) we find considerable variation both across and within Bantu languages as to whether, and/or when, object marking is (not) used. The discussion in section 3 identifies theoretical issues raised by the observed facts in Bantu object marking: the cross-linguistic variation and optionality, and notions of 'iconicity' and 'economy' that are central in Aissen's (2000) OT analysis of DOM. In section 4, I first outline briefly a set of theoretical assumptions adopted in my analysis for morphosyntactic realization of arguments before turining to my OT analysis. The final section includes a summary of the findings and a brief discussion of potential extension of the present approach to DOM beyond case marking and agreement languages.

2 Object Marking in Bantu

Bantu languages are characterized primarily as head marking languages, where the subject and object marker on a verb cross-reference the verb's arguments by agreeing in person, number and gender. The nature of object marking, however, is rather complex: majority of Bantu languages make use of word order rather than agreement for licensing objects, and object marking on the verb appears only when it is topic-anaphoric (like English pronouns). In some of these languages, however, both object agreement and positional licensing are required for particular types of object. In this section, I present these facts from some representative samples of the Bantu family.

No Object Agreement

Bresnan and Mchombo (1987), whose detailed study of the head-marking morphology focuses on Chicheŵa subject and object markers, show conclusively that the subject marker (SM) functions either as a topic-anaphoric pronoun or an agreement marker coindexing a clause-internal, non-topical subject NP, whereas the object marker functions only as a topic-anaphoric pronoun, being in complementary distribution with a clause-internal, non-topical object NP. In their theory of agreement developed within the LFG framework (Bresnan 1982, 2001, Dalrymple et al. 1995, Falk 2001), the subject marker is said to be ambiguous between ANAPHORIC and GRAMMATICAL AGREEMENT, while the object marker is unambiguously anaphoric agreement. One piece of evidence for their claim about these markers comes from word order. In simple transitive sentences, the object must immediately follow the verb when the verb contains no OM while the subject can be freely re-ordered (Bresnan and Mchombo, p.744–745). This is illustrated in (2), where only (2a) and (2b) with the V-O order are acceptable.

(2) a. SVO: Njûchi zi-ná-lúm-a alenje. Chicheŵa
bees SM-PAST-bite-INDIC hunters
'The bees bit the hunters.'

b. VOS: Zinálúma alenje njûchi.
c. OVS: *Alenje zinálúma njûchi.
d. VSO: *Zinálúma njûchi alenje.
e. SOV: *Njĉhi alenje zináluma.
f. OSV: *Alenje njûchi zináluma.

On the other hand, when the OM is present, all the word order permutations become acceptable, as shown in (3).

(3) a. SVO: Njûchi zi-ná-wá-lúm-a alenje. Chicheŵa
bees SM-PAST-OM-bite-INDIC hunters
'The bees bit them, the hunters.'

b. VOS: Zináwálúma alenje njûchi.
c. OVS: Alenje zináwálúma njûchi.
d. VSO: Zináwálúma njûchi alenje.
e. SOV: Njĉhi alenje zináwáluma.
f. OSV: Alenje njûchi zináwáluma.

Bresnan and Mchombo argue that the contrast between (2) and (3) can be explained under the following assumptions: (i) the object NP must be inside VP requiring strict adjacency with V, (ii) the OM functions only as a incorporated pronominal argument, and (iii) the object NP appearing with the OM in (3) is a floating topic which is outside the minimal clause containing the OM. A number of tests Bresnan and Mchombo present clearly show that the OM is systematically prohibited to co-occur with an object NP that cannot be a topic, such as a *wh*-phrase, a non-referential object that is part of a verb-object idiom (e.g. *a-ku-nóng'ónez-a bôndo* 'whisper-to his knee' meaning 'feeling remorse' in Chicheŵa), and a focused object (e.g. in cleft). The SM, on the other hand, co-occurs with all such elements. For example, when the object of a verb-object idiom (= a non-referential object) is passivized, the subject marker co-occurs with it.

Sensitivity to Animacy

Bresnan and Mchombo (1987) further note the following variation on object marking across the Bantu family: in the Imithupi dialect of Makua studied by Stucky (1981, 1983), the OM is obligatory for the human classes (classes 1 & 2) even when the overt object NP is not topical. This is best illustrated by the example in (4), in which the focus of *wh*-question is the object, and the OM is obligatory.

(4) a. Aráárima a-n-líh-íre mpáni? Makua
Araarima SM-OM-feed-T/A who
'Who did Araarima feed?

b. *Aráárima a-líh-íre mpáni?
Araarima SM-feed-T/A who human object

In KiSwahili, the OM is optional when the object NP is inanimate, but obligatory when it is animate (originally noted by Bokamba 1981; also Wald 1979). In (5), the OM agrees with *watoto* 'children'. We see in (5b) that in KiSwahili, the object can be questioned in situ, and co-occurs with the agreeing OM. The point about the optionality of object marking with inanimates will be returned to shortly.

(5) a. Bakari a-na-wa$_i$-som-e-a watoto$_i$ hadithi maktaba-ni. KiSwahili
Bakari SM-PRES-OM-read-APPL-INDIC children stories library-LOC
'Bakari is reading stories to/for the children in/at the library.'

b. Bakari a-na-wa$_i$-some-e-a nani$_i$ hadithi maktaba-ni?
Bakari SM-PRES-OM-read-APPL-INDIC who stories library-LOC
'To/for whom is Bakari reading stories in/at the library? human object

The sentences in (6) more clearly illustrate the animate-inanimate (rather than the human-nonhuman) opposition; they exemplify the presence of object marking with a non-human animate object but not with an inanimate object (Vitale 1981:123–124, (16a) & (19a)).

(6) a. Juma a-li-m-piga risasi tembo jana usiku. Swahili
Juma SM-PST-OM-hit bullet elephant yesterday night
'Juma shot an/the elephant last night.' animate object

b. risasi i-li-piga mti karibu na sisi.
bullet SM-PST-hit tree near us
'The bullet struck the tree near us.' inanimate object

Sensitivity to Definiteness

In addition to the effects of animacy on object marking, other Bantu languages display sensitivity to definiteness. Bresnan and Moshi (1993:52) note that in Kichaga, the object marker is obligatory when the object NP is an independent pronoun—the highest element in the definiteness hierarchy shown earlier in (1b). In (7a), the beneficiary (class 1) is pronominalized and triggers the class 1 object agreement *m*; in (7b), the theme (class 7) is pronominalized and co-occurs with class 7 object agreement *kí*; in (7c), both theme and beneficiary are pronominalized and co-occurs with their respective agreement markers.

(7) a. N-ä-ï-m̀-lyì-á k-èlyá ò — OM^i ... NP$_{pro}{}^i$
FOC-1S-PR-1O-eat-AP-FV 7-food 1PRO
'He/she is eating food for/on him/her.'

b. N-ä-ï-kì-lyì-à m-kà kyô — OM^i ... NP$_{pro}{}^i$
FOC-1S-PR-7O-eat-AP-FV 1-wife 7PRO
'He/she is eating it for/on the wife.'

c. N-ä-ï-kì-ḿ-lyï-à òó kyò — OM^i OM^j ... NP$_{pro}{}^j$ NP$_{pro}{}^i$
FOC-1S-PR-7O-1O-eat-AP-FV 1PRO 7PRO
'He/she is eating it for/on him/her.' Kichaga

Along the dimension of definiteness, facts in Kiyaka reveal another pattern: Kidima (1987) reports that Kiyaka requires object agreement when the object NP is a personal name—the second highest element in the definiteness hierarchy, as exemplified in (8). The ungrammaticality of (8b) shows that object marking is obligatory (and not optional) with a personal name.

(8) a. tu-n-telelé Maafú. — b. *tu-telelé Maafú. — Kiyaka
2SM-1OM-call.PAST Maafú — 2SM-call.PAST Maafú — Proper Name
'We called Maafú.'

Object marking in Kiyaka is optional, however, when the object NP is definite, as illustrated in (9) (Kidima, p.180). Without the OM, the object can be interpreted as either definite or indefinite, as indicated by the translation in (9b). Put differently—and more accurately—when the OM is present, the object cannot be interpreted as indefinite.[1]

(9) a. ba-aná ba-n'-súumb-idi khoomboó — Kiyaka
2child 2SM-1OM-buy-P 1goat
'The children bought the goat.'

b. ba-aná ba-suúmb-idi khoomboó
2child 2SM-buy-P 1goat
'The children bought a/the goat.' Definite object

As also noted by Bresnan and Mchombo (1987), Takizala (1973) reports that in Kihung'an the OM is used for definite objects, as exemplified in (10) (Takizala 1973, (11a) & (19)). Example (10a) is without the OM, and the object receives the indefinite interpretation; in (10b) on the other hand, the presence of the OM induces the definite reading of the object. The same contrast is reported in Zulu (Wald 1979).

[1]It should be noted that the Kiyaka facts cited here are only part of a much more complex picture of object marking in this language (Kidima 1984, 1987). In addition to object prefixes, which are instantiated only for classes 1 and 2 (animate singular and plural respectively), there are enclitics and full pronouns that may also co-occur with object NPs. Which form of coindexing is used and whether coindexing is obligatory or optional are apparently determined, in part, by interaction of the semantic and person hierarchy. Another conditioning factor seems to be discourse prominence. A clear picture of the complex interaction of these factors in Kiyaka object marking is yet to emerge.

(10) a. Kipese ka-swiim-in kit zoon. Kihung'an
Kipese SM-buy-PST chair yesterday
'Kipese bought a chair yesterday.'

b. Kipese ka-**ki**-swiim-in kit zoon.
Kipese SM-OM-buy-PST chair yesterday
'Kipese bought the chair yesterday.'

Conflicting Data in Swahili

KiSwahili presents conflicting data with respect to definiteness. Bresnan and Mchombo (1987:760) note that KiSwahili shows sensitivity to definiteness in addition to the effects of animacy: an indefinite object does not require the OM while a definite object does. For example in (11a) there is no OM, and the object NP has the indefinite reading. As shown in (11b) the presence of an OM induces the definite reading of the co-occurring (clause-internal) object NP, displaying the same pattern as Kihung'an (cf. (10)). Additional Swahili data showing that the presence of the OM induces the definite reading of the object NP are also found, for example, in Vitale (1981) and more recently in Zwart (1997).

(11) a. U-me-let-a kitabu? Swahili
you-PERF-bought-INDIC book
'Have you bought a book?' indefinite reading

b. U-me-ki-let-a kitabu?
you-PERF-OM-bought-INDIC book
'Have you bought the book definite reading

On the other hand, Wald (1979) cites examples that contain indefinite objects co-occurring with the agreeing OM (discussed in Nicolle 2000:682).

(12) a. a-ka-m-kuta mzee mwingine ndugu wa yule. Swahili
SM-ASP-OM-meet old.person other sibling of that.one
'then she met another old lady, sister of the first one.'

b. si-ja-ki-ona chochote.
SM-NEG-OM-see anything
'I haven't seen anything.'

In (12a), the object NP *another old lady* is indefinite specific; in (12b) the object NP *anything* is indefinite non-specific. These definiteness values are the second lowest and the lowest elements in the definiteness hierarchy respectively.

Seidl and Dimitriadis (1997, hereafter S&D) argue that variable object marking observed in Swahili is conditioned by the information status of the object—in the sense of Prince (1992),[2] rather than definiteness or animacy. According to their findings, object markers coreferential with animate objects

[2]In their study, Seidl and Dimitriadis (1997) adopt the following cross-classification of discourse referents proposed by Prince (1992):

represent 72% of the total of 312 sentences, and those coreferential with inanimates represent 12% (see Table 4 of Seidl and Dimitriadis 1997). When classified according to their hearer status, hearer-new objects are rarely (only 5%) pronominalized or doubly marked by agreement (see tables 5 of S&D, p.379). Based on these figures, S&D conclude that hearer status is a more significant than animacy status.

While S&D's findings highlight the importance of information status in the grammar of Swahili—a notion that has been found to figure prominently in Bantu grammar (cf. Morimoto 2000), one crucial element is missing in their picture of object marking—namely, the functional distinction between the bound pronominal object and object agreement. They state that the difference between 'object agreement' and 'object pronouns' is only "terminological", and not a morphological (formal) one. Indeed, as Bresnan and Mchombo (1987) point out, the crucial distinction between the two is not a formal one but a functional one: a topic-anaphoric pronoun and object agreement are one and the same morphological category (= affix) with two distinct morphosyntactic functions. That there need not be one-to-one correspondence between form and meaning is captured straightforwardly by LFG's parallel architecture, for the categorial information and morphosyntactic content of linguistic elements are represented independently of each other, and are only related through corresponding principles. Thus, based on an analysis which conflates topic-anaphoric and grammatical agreement, it is difficult to assess just how significant information status is in predicting the presence/absence of object agreement.

On the other hand, these data are perhaps not enough to conclude that Swahili object marking is sensitive to the definiteness hierarchy in the same way as, for example, Kihung'an is. Note that it is possible to account for these data by reference to animacy. For example in (12a), the object NP *another old lady* is human, and co-occurs with the OM; the object NPs in (11a,b) and (12b) are inanimates and, as we see, they appear with or without the corresponding OM. We might then conclude that these data can be attributed to the animacy effects: the OM co-occurs with the object NP when the object is human, but is optional when the object is inanimate. We already saw earlier in (6) that with inanimates, the OM is optional.

Optionality of object marking with lower elements on a prominence hierarchy is in fact quite commonly observed. In a study of a Bantu language Chi-Mwi:ni, which is closely related to Swahili and displays the same agreement pattern for objects, Kisseberth and Abasheikh (1977:182, fn.3) also note that although the language permits object marking for both animate and inanimate objects, the OM is not commonly used for inanimates.

Now it is possible that effects of both animacy and definiteness collectively determine the presence/absence of Swahili object marking. But until more data becomes available, I will assume that Swahili object marking can be explained solely in terms of animacy. It is nonetheless important to note that when the OM is optionally present, there is a strong preference for the definiteness reading of the object NP; the question of how this is brought about in the synchronic grammar of Swahili is worth considering (discussed in section 3).

To summarize, based the core data presented in this section, we can draw the following generalizations: (i) Bantu object marking is conditioned by animacy and definiteness, just as in DOM in case

(i) Information statuses (Prince 1992)

	Discourse-new	**Discourse-old**
Hearer-new	Brand-New	—
Hearer-old	Unused	Evoked

Hearer-new information includes newly mentioned discourse entity as well as entities not mentioned by can be inferred from the discourse. Hearer-old information includes discourse entities that are newly introduced but already familiar to the hearer (e.g. *President of the United States*) and those that have already been evoked in previous discourse.

marking languages, and hence these phenomena deserve a unified account; (ii) we observe considerable variation in the presence/absence of object marking both across and within Bantu languages; (iii) the (optional) presence of OM preferentially induces definite reading of the object NP. The generalization about the cross-linguistic variation and optionality in (ii) and the apparent definiteness effect noted in (iii) above can together be better understood if these relevant data are viewed from a diachronic perspective, as also suggested by Nicolle (2000). In the next section I briefly consider the source of variability and definiteness reading of optional object marking.

3 Characterizing DOM in Head-Marking Languages

Having presented the core facts in DOM across the Bantu family and descriptive generalizations about conditioning factors, I now turn to the relevant theoretical notions that allow us to provide a coherent analysis of these facts. First I consider the above data from a historical perspective in an attempt to better understand the nature of optionality and definiteness effects. I will then discuss the concept of 'iconicity' and 'economy' that figure importantly in Aissen's analysis.

3.1 From Topic-Anaphoricity to Agreement: the Source of Variability

According to typologists, agreement systems evolve from a topic construction, in which a full (morphologically unreduced) pronoun is used to refer to the topic NP anaphorically. The anaphoric pronoun is then reduced to a clitic-like element and then to a morphologically dependent affix, with the subsequent loss of the pronominal content to a mere agreement marker (cf. Givón 1979). Givón (1979) refers to the process whereby the topic anaphoric pronoun develops into grammatical agreement as 'de-marking'. The process of de-marking in which the subject pronouns in topic-shift (TS) constructions are reanalyzed as obligatory subject agreement markers is schematized in (13) (Givón, p.155). Givón notes that this reanalysis process is widely attested in French and non-standard dialects of English, as well as pidgins and creoles derived from the vocabulary of either French or English.

(13) Topic agreement — Subject agreement

The man, he came	⇒	The man he came
TOP PRO		SUBJ AGR

Givón further proposes that the development of object agreement follows essentially the same process, as schematized in (14). He notes that in SVO languages, an after-thought (AT) construction may have played a role as an extra step in the development of topic anaphoric pronouns to object agreement.

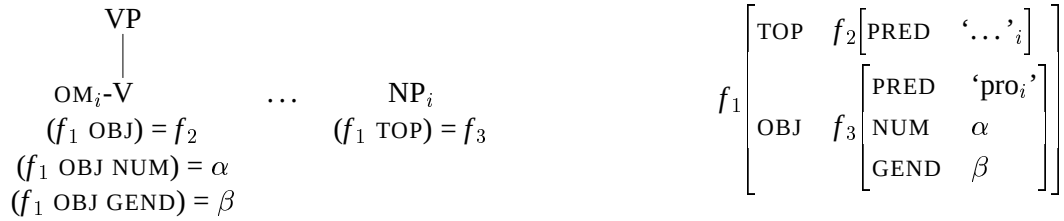
(14) Topic Shift ("marked") — Afterthought ("semi-marked") — Neutral ("demarked")

the man, I saw him	→	I saw him, the man	→	I saw-him the man
TOP PRO		PRO TOP		AGR

As noted at the beginning of section 2, Bresnan and Mchombo (1987) argue that the Chicheŵa object marker is unambiguously topic-anaphoric, while the subject marker is ambiguous between being topic-anaphoric and grammatical agreement. Following Givón's proposal, they suggest that Bantu subject agreement has followed the same path as that shown in (13), and that the same seems to be happening for object markers in some Bantu languages (in particular with respect to Swahili data they cite).

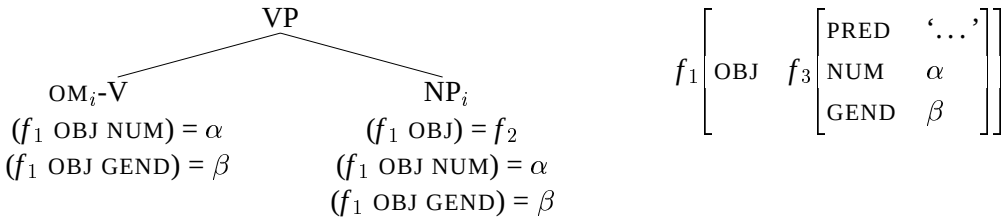
We can characterize more concretely the changes in the morphosyntactic properties of the OM using f-structure representation. The f-structure in (15) represents the situation in which the bound pronominal on the verb is topic-anaphoric to the dislocated topic object; the grammaticized discourse function TOPIC (f_2) anaphorically bounds the OBJ function (f_3).

(15) OM = topic-anaphoric pronoun



The f-structure in (16) reflects the situation in which the bound pronominal on the verb has lost the pronominal content and functions as grammatical agreement (16): it contributes morphosyntactic information such as number and gender, which unifies with the information contributed by the VP-internal object NP.

(16) OM = grammatical agreement



The structure in (17) represents the morphosyntactic information contributed by the object NP internal to VP. Note that the f-structure representation corresponding to the VP with object agreement (16) and the one without object agreement (17) are identical. That is, object agreement in (16) only provides redundant number and gender information that the object NP contributes to its f-structure. This naturally leads to the loss of the agreement marker that contributes the redundant information.

(17) non-topical object (without the OM)

If we accept the process of historical change described by Givón and take the representations shown in (15)–(17) to be different stages in the historical process, it is rather easy to see how the definiteness reading is induced by the optional presence of an OM: historically, the OM was used only topic-anaphorically (as in the current state of Chichewa). Even if topic-anaphoricity was lost in the course of the change in the morphosyntactic properties of the OM, it could still be induced in the synchronic

grammar by the optional presence of the OM.

Viewing these facts from the historical perspective also helps us understand that the transitory stages in the process display high variability both across the Bantu family and within individual languages. Nicolle (2000:683) also reports, reexaming the results of Seidl and Dimitriadis (1996), that in written texts (novels, plays, journals), animate objects co-occur with object agreement nearly 100%, while in spoken sources, co-occurrence of animate objects with object agreement is far less frequent. The register variation observed here is typical of what happens during language change. Given that most linguistic changes are gradual,[3] the old and new linguistic forms naturally co-exist, and they do so in a predictable fashion: written language is generally more resistant to change. Thus the older form persists longer in written language while the new form replaces the old in spoken language; the older forms are preferred for formal speech, and the newer forms are reserved for informal speech. When such distributional patterns of new and old forms across different registers (written vs. spoken language; formal vs. informal speech) are of course also observed across geographic boundaries, which we identify as dialectal variation. I have not been able to identify any dialectal variation in the use of object marking within an individual Bantu language. What we clearly observe, however, is the variation *across* the languages of the close-knit Bantu family that resembles dialectal variation within a single language.

3.2 On 'Iconicity' vs. 'Economy'

One of the key ideas expressed in Aissen's analysis is that DOM represents a tension between ICONICITY and ECONOMY. ICONICITY generally favors iconic relations between form and content (or function). In DOM (in case marking languages), the form refers to morphological case, and the content refers to the semantic properties of objects (e.g. human objects, definite objects). Iconicity constraints thus favor marked object types to be formally marked by morphological case.

Relative markedness of object types is expressed through HARMONIC ALIGNMENT of the relational hierarchy (18a) either with the animacy hierarchy (18b) or the definiteness hierarchy (18c). The left-most (= the most prominent) element in the animacy and definiteness hierarchy is most marked for objects, and the right-most (= the least prominent) element in the respective hierarchies is the least marked, or the most prototypical objects.

(18) a. Su(bject) > Non-Subject (or Obj(ect) for simplicity)

b. Hum(an) > Anim(ate) > Inan(imate)

c. Pron(oun) > Name > Def(inite) > Indef(inite) Spec(ific) > Non-spec(ific)

In order to derive the relevant markedness constraints to account for DOM from these hierarchies, what we need is the alignment of Obj, the lower element in the relational hierarchy with the animacy/definiteness hierarchy. This yields the markedness subhierarchies in (19). The order of the markedness constraints in each subhierarchy is universally fixed.

[3]—with the exception of the development of pidgin languages, in which we observe rapid simplification and reduction of contact languages (e.g. Holm 1988).

(19) Markedness Subhierarchies

a. *Hum-O $\gg$ *Anim-O $\gg$ *Inan-O

b. *Pron-O $\gg$ *Name-O $\gg$ *Def-O $\gg$ *IndefSpec-O $\gg$ *NSpec-O

The constraint subhierarchy in (19a) more severely penalizes Hum(an) O(bject) than Anim(ate) O(bject), because the constraint penalizing the former, *Hum-O is always ranked above the one penalizing the latter, *Anim-O. Similarly the top-most constraint in (19b), *Pron-O penalizes Pro(nominal) O(bject), and the second highest constraint penalizes Personal Name Object, and the violation of the former is universally worse than the violation of the latter.

In DOM, however, marked object types that are dispreferred by the markedness constraints in (19) are nonetheless admitted so long as they are morphological marked. In Aissen's analysis, morphological case on objects is forced by a constraint against absence of morphological case, given in (20). The subscripted *c* indicates CASE.

(20) *$\emptyset_c$ 'Star Zero': Penalizes the absence of a value for the feature CASE. (Aissen 2000)

The relation between markedness and morphological complexity is formally expressed by LOCAL CONJUNCTION[4] of the markedness constraints and the Star Zero constraint. For example, the constraint subhierarchy in (21) illustrates the local conjunction of Star Zero and the markedness constraints on the animacy dimension. The locally conjoined constraints are violated only if both elements of the conjunct are not satisfied. The constraints are thus satisfied by all object types as long as they have overt case morphology.

(21) Local conjunction of *$\emptyset_c$ with the animacy subhierarchy

*Hum-O & *$\emptyset_c$ $\gg$ *Anim-O & *$\emptyset_c$ $\gg$ *Inan-O & *$\emptyset_c$

Now, DOM arises precisely because only some objects are obligatorily case-marked. To capture this, Aissen proposes an ECONOMY constraint against having morphological case, shown in (22). Forms without morphological case are less complex and thus more economical (in terms of the morphological structure) than those with case marking. The cross-linguistic variation is derived by interpolating the economy constraint at different points in the markedness constraints in (21).

(22) *STRUC$_c$: penalizes a value for the morphological category case.

While the notions of iconicity and economy that motivate her analysis are quite general and intuitive, the formal expressions of these notions in her OT analysis (*$\emptyset$ for iconicity and *STRUC for economy), also adopted by much of the subsequent OT work on markedness, are problematic and undermine the nature of constraints and their interaction that OT is designed to explain.

First, note that the Star Zero constraint and the economy constraint *STRUC state exactly the opposite conditions. The two types of constraints familiar in OT, faithfulness and markedness, are of course often in conflict. However, given that grammatical phenomena and cross-linguistic variation are explained solely in terms of resolutions of constraint conflict, and given that each constraint must be grounded

[4]—originally proposed by Smolensky 1995; see for example early work by Arstein (1998) and Aissen (1999) for application of local conjunction to syntactic problems.

conceptually and/or typologically, it is not desirable to propose constraints that directly contradict with each other. In order to circumvent this problem, Aissen (1999, 2000) stipulates that the use of Star Zero is restricted within local conjunction. But technically (and conceptually), if a constraint is used as an element of local conjunction, then it must exist independently in the constraint system.

Aissen (2000:9, footnote 10) makes an interesting point that Star Zero that enforces presence of case marking does not fall into a faithfulness constraint. The primary role of case marking is to signal grammatical function. Thus, the motivation for the constraint requiring case marking must be sought in the listener-oriented functional principle "Minimization of Perceptual Confusion" (Boersma 1998). However it is not difficult to find configurations in which there is no potential "perceptual confusion", yet case (or head) marking is still required. For example, sentences like those in (23) obviously cause no perceptual confusion as to the (default) argument-function mapping. But in terms of animacy, (23b) exhibits the marked animacy configuration (inanimate subject, human object).

(23) a. The knife cut the bread — S = Inan; O= Inan

b. the needle pierced the child — S = Inan; O = Hum

Thus in languages that require case marking for human objects, (23b) would still require case marking. Along the dimension of definiteness, it is not entirely clear how the notion of "minimizing perceptual confusion" explains obligatory case marking. In short, listener-oriented functional principles like Minimization of Perceptual Confusion, at best, offers only partial explanation for DOM, and seems to be providing rather shaky grounding for the constraint $*\emptyset_C$.

In recent work, Grimshaw (2001) takes up the question of whether the notion of economy should be expressed as a constraint, or it is a by-product of the system of constraints on structure. Through the discussion of a word order typology, she shows that the right set of constraints on phrase structure will yield a more economical structure as an optimal output among a set of possible structures without positing additional 'economy constraints'.

In the present work I propose a simpler analysis that eliminates $*\emptyset$ and *STRUC without losing the basic insight articulated in Aissen's analysis. The current analysis exemplifies the fundamental way in which conflicts are resolved by interaction of markedness and faithfulness constraints. Economy is indeed a by-product of markedness constraints.

4 An OT Account of DOM in Bantu

As represented by Aissen's work, previous OT syntax work has shown that prominence hierarchies play an important role in determining the forms of expression in various domains of grammar, and it demonstrated that OT successfully models both the universality of prominence hierarchies and variability in the effects of those hierarchies.[5] In this respect, the idea explicated in the present work for Bantu object marking is not a novel one. Rather, the central argument is that the OT analysis analogous to that of DOM in case marking languages proposed by Aissen (2000) allows us to highlight the striking parallelism between these phenomena and to illuminate a more general picture that what has been taken to be independent instances of variable object marking across the Bantu family is in fact a specific way in which languages structurally mark non-prototypicality, using the resources available in the language(s) in question.

[5]For earlier representative work dealing with the role of prominence hierarchies, see Sells (2001).

4.1 Hierarchy of Morphological Features for Argument Roles

In this section I motivate the use of the argument hierarchy and use of binary features for expressing argument roles rather than the relational hierarchy as in Aissen's analysis.

Researchers working in Lexical Decomposition Grammar (LDG; Joppen and Wunderlich 1995, Kaufmann 1995, Wunderlich 1997a,b, 2000, 2001, Kaufmann and Wunderlich 1998, Stiebels 1999, 2000) have developed a principled approach to argument structure and a typology of argument structure linking. LDG provides a means of systematically deriving the ARGUMENT HIERARCHY, in which argument roles are strictly ordered, from the 'Semantic Form' (SF). The SF is a level of representation that serves as the interface between morphosyntactic structure and semantics on the one hand, and semantic structure and conceptual structure on the other. It includes the semantic information of a lexical item in the form of a set of lexically-decomposed primitive predicates, as well as the information that is relevant for deriving the argument structure of the predicate. The argument structure is derived by means of λ-abstraction of the argument variables in the SF, as shown in (24).

(24)

			Semantic Form
a.	sleep:	λx	SLEEP(x)
b.	kiss:	λy λx	KISS(x,y)
c.	give:	λz λy λx	{ACT(x) & BECOME POSS(y,z)}

The λ-abstracted argument roles are assigned ABSTRACT CASE FEATURES [$\pm$hr] ("there is a/no higher role") and [$\pm$lr] ("there is a/no lower role"), as illustrated in (25).

(25)

		Theta Str	Semantic Form
a.	sleep:	λx [$-$hr] [$-$lr]	SLEEP(x)
b.	kiss:	λy λx [+hr] [$-$hr] [$-$lr] [+lr]	KISS(x,y)
c.	give:	λz λy λx [+hr] [+hr] [$-$hr] [$-$lr] [+lr] [+lr]	{ACT(x) & BECOME POSS(y,z)}

The abstract case features are linked to structural case features, which may be realized as morphological case, agreement, or by position. The structural cases are specified in terms of the same set of features [$\pm$hr] and [$\pm$lr], shown in (26).

(26)

Nominative/Absolutive (NOM/ABS)	[ ]
Accusative (ACC)	[+hr]
Ergative (ERG)	[+lr]
Dative (DAT)	[+hr, +lr]

According to these feature classifications of structural cases, Nominative is the least marked case, and Dative the most marked. Linking of the abstract case to structural case is achieved by unification of

compatible features, yielding three canonical case patterns for an accusative and ergative system, as shown in (27).

(27)

	a. Intransitives	b. Transitives	c. Ditransitives
	λx	λy λx	λz λy λx
	[−hr]	[+hr] [−hr]	[+hr] [+hr] [−hr]
	[−lr]	[−lr] [+lr]	[−lr] [+lr] [+lr]
ACC-system:	NOM	ACC NOM	ACC DAT NOM
ERG-system:	ABS	ABS ERG	ABS DAT ERG

For intransitive predicates, the sole argument is encoded by the features [−hr, −lr]—there is no higher or lower role, and these features are compatible only with the Nominative case. For transitive predicates, x is the higher role and is specified as [−hr, +lr]—there is no higher role, and there is a lower role. In an accusative system, the argument will be realized by NOM, and the lower argument, specified as [+hr, −lr] (there is a higher role and no lower role), is realized by ACC. In an ergative system, the higher role maps to ERG, and the lower role to ABS. For ditransitive predicates, the medial argument in the SF is specified as [+hr, +lr]—there is a higher role and a lower role, the most marked specifications, and is mapped to Dative.[6]

Within this framework, Stiebels (2000) proposes the markedness scale of the relation between argument roles and their morphosyntactic realizations, shown in (28). It is important to note that [+hr] > [+lr] is not simply equivalent to "lowest argument" > "highest role" as in the (reversed) argument role hierarchy, or to "accusative" > "ergative" as in the case hierarchy (e.g. Comrie 1989, Wierzbicka 1981). It is precisely as it is stated: "accusative marking of the lowest role" > "ergative marking of the highest role". I will refer to the hierarchy in (28) as the HIERARCHY OF MORPHOLOGICAL FEATURES for argument roles, or simply 'argument feature hierarchy'.

(28) [+hr] > [+lr]

Read as: "Accusative marking of the lowest role is less marked than ergative marking of the highest role."

The system of argument linking adopted here therefore directly links argument roles and their morphosyntactic realizations.

4.2 Markedness Constraints

Given the hierarchy of morphological features in (28) we can now formerly express the relative markedness of the relation between animacy/definiteness and morphological marking. Harmonic alignment of the argument feature hierarchy with the animacy hierarchy, again repeated here in (29), produces the markedness hierarchies shown in (30).

(29) Animacy: Hum(an) > Anim(ate) > Inan(imate)

[6]Other non-canonical case patterns (e.g. passive and antipassive, one of the arguments being lexically marked) are also possible under this theory of argument linking, and have been rigorously discussed in earlier work (cited at the beginning of this section) with a wide range of cross-linguistic data.

(30) Harmonic Alignment of (28) with the Animacy scale

$H_{[+h]}$: [+hr]/Hum $\succ$ [+hr]/Anim $\succ$ [+hr]/Inan

$H_{[+l]}$: [+lr]/Inan $\succ$ [+lr]/Anim $\succ$ [+lr]/Hum

The harmonic alignment of [+hr] with the animacy scale, referred to as $H_{[+h]}$, states that [+hr]—accusative marking of the lower role—that is human is less marked (more harmonic, prototypical) than accusative marking of the lower role that is animate, and that is less marked than accusative marking of the lower role that is inanimate. In other words, human objects are marked, and as such, overt morphological (or morphosyntactic) marking of these marked objects is expected.

Conversely, the harmonic alignment of [+lr] with animacy, referred to as $H_{[+l]}$, states that [+lr]—ergative marking of the higher role—that is inanimate is less marked than ergative marking of the higher animate role, and that is less marked than ergative marking of the highest human role. In the present discussion, only the harmonic alignment $H_{[+h]}$ will be relevant.

The constraint subhierarchies derived through the harmonic alignment in (30) are given in (31). Starting from the top (left-most), the constraints in the subhierarchy $C_{[+h]}$ is interpreted as "avoid [+hr]—accusative marking of the lower role—that is inanimate". In other words, an unmarked, prototypical object type such as an inanimate object should not be overtly marked. Given the fixed subhierarchy, accusative marking of inanimate [+hr] will be the most severely penalized, and accusative marking of animate [+hr] is more severely penalized than accusative marking of human [+hr]. Conversely, ergative marking of the higher role should be avoided when it is human, the most prototypical subject type. The constraint subhierarchies for [+lr] thus describes split ergativity in morphologically ergative languages: ergative case marking is avoided for subjects when they are high in animacy (cf. Dixon 1979, 1994).

(31) Constraint subhierarchies on the dimension of animacy

$C_{[+h]}$: *[+hr]/Inan $\gg$ *[+hr]/Anim $\gg$ *[+hr]/Hum

$C_{[+l]}$: *[+lr]/Hum $\gg$ *[+lr]/Anim $\gg$ *[+lr]/Inan

These constraint subhierarchies thus express the same form-function relations as those proposed by Aissen. But while Aissen's constraint system imposes a positive constraint—that marked objects must be formally marked, the present system imposes a negative constraint: unmarked object type must not be formally marked.

Along the dimension of definiteness, repeated again in (32), we arrive at the harmonic alignment in (33).

(32) Definiteness: Pronoun > Name > Definite > IndefSpec > Non-specific

(33) Harmonic Alignment of (28) with the Animacy scale

$H_{[+h]}$: [+hr]/Pro $\succ$ [+hr]/Name $\succ$ [+hr]/Def $\succ$ [+hr]/IndefSpec $\succ$ [+hr]/NSpec

$H_{[+l]}$: [+lr]/NSpec $\succ$ [+lr]/IndefSpec $\succ$ [+lr]/Def $\succ$ [+lr]/Name $\succ$ [+lr]/Pro

Here, the harmonic alignment of [+hr] with definiteness ($H_{[+h]}$) expresses that [+hr]—accusative marking of the lower role is most expected (= most harmonic) when it is a pronoun, because it is the most marked object type. The markedness increases as [+hr] descends the scale of definiteness. For the higher role, the least marked situation is to be marked by ergative, the marked case, when non-specific because that is the most marked subject type. It is most marked for the lower role to be marked by ergative in a pronoun, the least marked subject type. Reversing the harmonic alignment derives the markedness constraints shown in (34).

(34) Constraint subhierarchies on the dimension of definiteness

$C_{[+h]}$: *[+hr]/NSpec $\gg$ *[+hr]/IndefSpec $\gg$ *[+hr]/Def $\gg$ *[+hr]/Name $\gg$ *[+hr]/Pro

$C_{[+l]}$: *[+lr]/Pro $\gg$ *[+lr]/Name $\gg$ *[+lr]/Def $\gg$ *[+lr]/IndefSpec $\gg$ *[+lr]/NSpec

The constraint subhierarchy $C_{[+h]}$ most severely penalizes accusative marking of the lowest argument that is non-specific. The lowest constraint in the subhierarchy penalizes the least marked situation in which there is accusative marking of the lowest argument that is pronominal. Conversely, $C_{[+l]}$ most severely penalizes the highest argument that is marked as ergative when it is a pronoun.

As we see, the constraint subhierarchies in (31) and (34) express the same markedness generalizations as those proposed by Aissen (1999, 2000) without use of $\emptyset$ or local conjunction. Yet these constraints also expresses the iconic relation between form and content: the unmarked object types should not be marked morphologically. Note that the absence of overt morphological marking for unmarked objects is not only iconic but also more economical. Thus, as argued by Grimshaw for constraints on phrase structure, the proposed constraints also derives economy without an additional economy constraint.

Furthermore, as noted by Stiebels (2000), the constraints based on the argument feature hierarchy [+hr] $>$ [+lr] solve problems for ergative languages that Aissen's analysis suffers, which Aissen (1999) herself also recognizes.

4.3 Deriving a Typology of DOM in Bantu

In Aissen's system, the constraint subhierarchies penalizing objects without morphological mark interact with the economy constraint *STRUC to yield DOM and cross-linguistic variation. In the present proposal, the constraint hierarchies penalizing violation of iconicity (no extra marking for unmarked objects) just discussed interact with a constraint on input-output faithfulness, given in (35). The faithfulness constraint in (35) expresses the idea that the argument roles in the input must be realized and marked overtly in the output. The subscripted *agr* forces realization of [+hr] by agreement.

(35) Input-Output Faithfulness

MAX(+hr)$_{agr}$: The [+hr] role in the input must be realized by agreement.

It is assumed here that MAX constraints include various specific instantiations such as MAX$_{agr}$, MAX$_{case}$ and MAX$_{pos}$ (for syntactic position). This allows, for example MAX$_{agr}$ and MAX$_{case}$ to have different effects in languages with both dependent marking and agreement (e.g. some Australian languages, Amharic, Hungarian), or MAX$_{agr}$ and MAX$_{agr}$ and MAX$_{pos}$ to have different effects in languages that use both agreement and positional licensing (e.g. Bantu). Amharic, for example, employs both case marking and verbal agreement: while the object marker on the verb is used exclusively for

topic-anaphoricity (as a bound pronominal argument), case marking is used to mark definite objects (Hudson 1997). In contrast in Hungarian, which also employs case marking and agreement, verbal object agreement is used only when the object is definite (see for example, Horvath 1986, Puskás 2000).[7]

We now interpolate the faithfulness constraint in the constraint subhierarchies in (31) to derive the typology of DOM in Bantu along the dimension of animacy. The treatment of the optionality of object marking we observed earlier will be discussed shortly.

(36) Dimension of animacy

	←	MAX(+hr)$_{agr}$	
*[+hr]/Inan			
|	←	MAX(+hr)$_{agr}$	Swahili (5)–(6)
			Chi-Mui:ni (Kisseberth and Abasheikh 1977)
*[+hr]/Anim			
|	←	MAX(+hr)$_{agr}$	Makua (4)
*[+hr]/Hum			
	←	MAX(+hr)$_{agr}$	Chicheŵa (Bresnan and Mchombo 1987)

Any object type shown above MAX(+hr)$_{agr}$ cannot be marked by object agreement, as it means that form-content iconicity expressed by the markedness constraint is more important than satisfying input-output faithfulness and realizing the lower role by agreement. For example in Chicheŵa, the faithfulness constraint is ranked below all the markedness constraints against object agreement. Thus, object agreement cannot co-occur with any clause-internal object NP (e.g. Bresnan and Mchombo 1987). Interpolating the MAX(+hr) between *[+hr]/Hum and *[+hr]/Anim derives Makua (4), which does not trigger agreement for animate or inanimate objects, and object marking is observed only with human objects. Promoting MAX(+hr) above *[+hr]/Anim but below *[+hr]/Inan derives Swahili (cf. (5)–(6)) and Chi-Mui:ni, in which object agreement is obligatory with humans and animates (in writing) but optional with inanimates.

DOM along the dimension of definiteness is characterized in (37).

(37) Dimension of definiteness

	←	MAX(+hr)$_{agr}$	
*[+hr]/NonSpec			
|	←	MAX(+hr)$_{agr}$	
*[+hr]/IndefSpec			
|	←	MAX(+hr)$_{agr}$	Kihung'an (10), Zulu (Wald 1979)
*[+hr]/Def			
|	←	MAX(+hr)$_{agr}$	Kiyaka (8)
*[+hr]/Name			
|	←	MAX(+hr)$_{agr}$	Kichaga (7)
*[+hr]/Pro			
	←	MAX(+hr)$_{agr}$	Chicheŵa

Here again Chicheŵa ranks MAX(+hr) below the constraint subhierarchy of definiteness and hence

[7]Thanks to Aaron Broadwell for providing a description of the relevant facts and reference on Amharic, and to Andrew Spencer for bringing my attention to the facts in Hungarian.

no object NPs trigger object agreement. Interpolating MAX(+hr) between *[+hr]/Pro and *[+hr]/Name yields Kichaga (7), in which object marking is observed for pronominals regardless of its animacy status. Promoting MAX(+hr) one step up in the hierarchy of the markedness constraints yields Kiyaka (8), in which object agreement is present when the object is a personal name. Placing the faithfulness constraint further up between *[+hr]/Def and *[+hr]/IndefSpec derives Kihung'an (10) and Zulu (Wald 1979), in which all definite objects co-occur with object marking.

Modeling Optionality: Stochastic OT

Within this system, optionality in the use of object marking in a single language is easily accommodated by use of stochastic ranking. This allows us to articulate the relation between cross-linguistic variation, optionality, and diachronic change in a uniform fashion.

In stochastic OT (cf. Boersma 1997, 1998, Forthcoming, Keller 1998, Asudeh 2001, Boersma and Hayes 2001, Bresnan and Deo 2001, Bresnan, Dingare, and Manning 2001, among others), constraints are distributed along a continuous scale, each with a ranking value. At every constraint evaluation, the ranking is perturbed by a random variable (either in a positive or negative direction). At a given constraint evaluation, when two constraints are relatively farther away from each other, the results is a categorical ranking of these constraints. Optionality arises when two constraints have close ranking values. Optionality in Bantu object marking can be modeled by the stochastic ranking of MAX(+hr)$_{agr}$ and markedness constraints. For example, the data from Swahili in (6)–(11) and (12b) and from Chi-Mui:ni (Kisseberth and Abasheikh 1977) suggest that MAX(+hr) and *[+hr]/Inan float and have relatively close ranking values. Optional marking of inanimate objects in these languages arises when MAX(+hr) is promoted above *[+hr]/Inan. The data from Kiyaka in (9) suggest that MAX(+hr) and *[+hr]/Def are closely ranked and can float. Optional marking of definite objects arises when MAX(+hr) is promoted above *[+hr]/Def.[8]

By implementing stochastic ranking, the present analysis allows us to understand cross-linguistic variation and optionality as closely interrelated phenomena, essentially being two sides of the same coin: the analysis shows that the typological space provided by the system of constraints is the space that also allows variation in a single language. A similar point is made by Bresnan and Deo (2001) in their study of *be* across English dialects and inter-speaker variation. In pre-OT generative syntax, much effort has been put into characterizing universal properties as well as cross-linguistic variation. Variation within a single language, however, has not been seriously considered as part of a grammatical system. But once we understand how variation across languages (= typology) and variation within languages (= optionality) exist within one single 'typological' space, it becomes obvious that they are both expected to be systematic and hence equally deserve a principled account in linguistic theory.

5 Concluding Remarks

To summarize, in this paper I hope to have shown that first, object agreement in Bantu languages is sensitive to animacy or definiteness of the object, and the cross-linguistic data can be given a unified account under the general concept of differential object marking. Secondly, DOM is highly variable both across and within languages, but the variability is expected once we situate the synchronic states in the context of diachronic change and view the variability in synchronic grammar as unstable transitory stages in a diachronic process (cf. Greenberg 1969, 1978, Croft, Denning, and Kemmer 1990). Thirdly,

[8] An on-going closer examination of Swahili object marking using a corpus of Swahili made available by the University of Helsinki attempts to articulate the analysis of optionality briefly outlined here by stochastic OT.

the present analysis highlights the idea that cross-linguistic variation and language-internal variation operate within an identical typological space provided by a system of violable, universal constraints. By implementing stochastic ranking, the present analysis can potentially provide a more precise formal model of language-internal variation.

Further Extension

Before I conclude this paper, I briefly mention a recent study reported by Schwenter and Silva (2002) on null vs. overt objects in Brazilian Portuguese, which suggests that differential object marking extends beyond languages with case marking or agreement. Brazilian Portuguese allows null objects, but the choice of null vs. overt expression is conditioned by an interaction of animacy and specificity.

First, the object cannot be null if it is both animate and specific as shown in (38). Here the human object *João* is mentioned in the first utterance, and hence *ele* in the second sentence is both animate and specific, and hence cannot be null.

(38) Você lembra do João? A gente viu **?*∅/ele** na festa da Carla ontem.
'Do you remember J.? We saw him yesterday at C's party.' [+anim, +spec, (+def)]

On the other hand, the object can be optionally be null if it is either animate or specific. In (39a), *a driver* is human but indefinite specific. As a result, in both the original utterance and English translation an indefinite pronoun *one* (*um*) is used. In BP, the pronoun becomes optional. In (39b), *a beautiful old house* is inanimate, but it is specific. Again the object *it* in the subsequent utterance can be either null or overt.

(39) a. Ela me disse que precisava de um motorista para levar as meninas ao colégio. Contratei **∅/um.**
'She told me that she needed a driver to take the girls to school. I hired one.' [+anim, −spec, (−def)]

b. Na minha rua tem uma casa antiga linda, mas eles vão derrubar **∅/?ela**.
'On my street there's a beautiful old house, but they're going to knock it down.' [−anim, +spec, (−def)]

Finally the object cannot be overtly expressed if it is both inanimate and non-specific—the most prototypical type. In (40), the relevant object is *a ticket*, which is inanimate and non-specific. As a result, it cannot be referred to by the definite pronoun *it* the subsequent sentence (in both BP and English).

(40) Estava procurando um ingresso para o teatro e finalmente encontrei **∅/?*ele.**
'I was looking for a ticket for the theater and finally I found one/*it.' [−anim, −spec, (−def)]

These data further corroborate the idea that DOM is only a particular manifestation of the more general tendency observed across languages to mark non-prototypicality overtly by whatever means available in a given language (type)—by case marking, head marking, or lexically.

References

Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. *Natural Language & Linguistic Theory* 17(4), 673–711.

Archangeli, Diana, and D. Terene Langendoen (eds.). 1997. *Optimality Theory: An Overview*. Oxford, UK, Blackwell Publishers.

Artstein, Ron. 1998. Hierarchies. Available via the Rutgers Optimality Archive at `http://ruccs.rutgers.edu/roa.html`, May 31 version.

Asudeh, Ash. 2001. Linking and optionality in Marathi: An Optimality Theory analysis. In Sells (Sells 2001), 257–312.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. University of Amsterdam. Available on-line at `http://ruccs.rutgers.edu/roa.html` (ROA-221-109).

Boersma, Paul. 1998. *Functional Phonology*. The Hague, Holland Academic Graphics.

Boersma, Paul. Forthcoming. Phonology-semantics interaction in OT, and its acquisition. In Robert Kirchner, Wolf Wikeley, and Joe Pater (eds.), *University of Alberta Papers in Experimental and Theoretical Linguistics 6*. December 18, 1999 version. Available on-line at `http://www.fon.hum.uva.nl/paul`.

Bokamba, Eyamba Georges. 1981. *Aspects of Bantu Syntax*. Preliminary edition. Department of Linguistics, University of Illinois, Urbana-Champaign, Illinois.

Bresnan, Joan (ed.). 1982. *The Mental Representation of Grammatical Relations*. Cambridge, MA, The MIT Press.

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

Bresnan, Joan. In Press. Optimal syntax. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, Syntax and Acquisition*. Oxford University Press.

Bresnan, Joan, Shipra Dingare, and Christopher Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In Tracy Halloway King and Miriam Butt (eds.), *The On-line Proceedings of LFG 2001*, Stanford, CA. CSLI Publications. `http://csli-publications.stanford.edu/LFG/6/lfg01.html`.

Bresnan, Joan, and Lioba Moshi. 1993. Object asymmetries in comparative Bantu syntax. In Sam Mchombo (ed.), *Theoretical Aspect of Bantu Grammar*, 47–91. Stanford, CA, CSLI Publications.

Croft, William, Keith Denning, and Suzanne Kemmer (eds.). 1990. *Studies in Typology and Diachrony: Papers Presented to Joseph H. Greenberg on his 75th Birthday*. Amsterdam, John Benjamins.

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

Deo, Ashwini, and Joan Bresnan. 2001. 'be' in the *Survey of English Dialects*: A stochastic OT account. MS. Stanford University.

Dixon, Robert M.W. 1979. Ergativity. *Language* 55(1), 59–138.

Dixon, Robert M.W. 1994. *Ergativity*. Cambridge, MA, Cambridge University Press.

Falk, Yahuda N. 2001. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Stanford, CA, CSLI Publications.

Farkas, Donka. 1978. Direct and indirect object reduplication in Romanian. In Donka Farkas, Wesley M. Jacobsen, and Karol W. Todrys (eds.), *Papers from the fourteenth regional meeting of the Chicago Linguistic Society*, 88–97, Chicago. Chicago Linguistic Society.

Greenberg, Joseph H. 1968. Some methods of dymanmic comparison in linguistics. In Jan Puhvel (ed.), *Substance and Structure of Language*, 147–203. Berkeley, CA, University of California Press.

Greenberg, Joseph H. 1978. Diachrony, synchrony and language universals. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik (eds.), *Universals of Human Language*, Vol. 1, 61–92. Stanford, CA, Stanford University Press.

Grimshaw, Jane. 2001. Economy of structure in ot. MS. Rutgers University. Available via the Rutgers Optimality Archive at `http://ruccs.rutgers.edu/roa.html`, May version.

Holm, John. 1988. *Pidgins and Creoles Volume I: Theory and Structure*. Cambridge, UK, Cambridge University Press.

Horvath, Julia. 1986. *FOCUS in the Theory of Grammar and the Syntax of Hungarian*. Dordrecht, Foris Publications.

Hudson, Grover. 1997. Amharic and Argobba. In Robert Hetzron (ed.), *The Semitic Languages*, 457–486. London, Routledge.

Joppen, Sandra, and Dieter Wunderlich. 1995. Argument linking in Basque. *Lingua* 97, 123–169.

Kager, René. 1999. *Optimality Theory*. Cambridge, Cambridge University Press.

Kaufmann, Ingrid. 1995. What is an (im-)possible verb? Restrictions on Semantic Form and their consequences for argument structure. *Folia Linguistica* 29, 67–103.

Kaufmann, Ingrid, and Dieter Wunderlich. 1998. Cross-linguistic patterns of resultatives. *Working papers SFB Theorie des Lexikons*. University of Düsseldorf.

Keller, Frank. 1998. Gradient grammaticality as an effect of selective constraint re-ranking. In M. Catherine Gruber, Dirrick Higgins, Kenneth Olson, and Tamara Wysocki (eds.), *Papers from the 34th Annual Meeting of the Chicago Linguistic Society*, Vol. 2: The Panels, Chicago, Illinois.

Kidima, Lukowa. 1984. *Objects and Object Agreement in Kiyaka*. University of Pittsburgh.

Kidima, Lukowa. 1987. Object agreement and topicality hierarchies in Kiyaka. *Studies in African Linguistics* 18, 175–209.

Kisseberth, Charles W., and Mohammad Imam Abasheikh. 1977. The object relationship in Chi-Mwi:ni, a Bantu language. In Peter Cole and Jerald Sadock (eds.), *Grammatical Relations*, Vol. 8 of *Syntax and Semantics*, 179–218. New York, Academic Press.

Morimoto, Yukiko. 2000. *Discourse Configurationality in Bantu Morphosyntax.* Doctoral dissertation, Stanford University, Stanford, CA.

Nicolle, Steve. 2000. The Swahili object marker: syntax, semantics, and mythology. In H.E. Wolff and O. Gensler (eds.), *Proceedings of Word Congress African Linguistics, Leipzig*. Köln, Rüdiger Köppe.

Prince, Alan, and Paul Smolensky. 1993. *Optimality Theory: Constraint interaction in Generative Grammar*. RuCCS Technical Report No.2. Piscataway, NJ: Rutgers University Center for Cognitive Science. To appear, Cambridge, Mass: MIT Press.

Prince, Ellen F. 1992. The ZPG letter: Subjects, definiteness, and information-status. In William C. Mann and Sandra A. Thompson (eds.), *Discourse Description: Diverse Linguistic Analyses of Fund-Raising Text*, 295–325. Amsterdam, Benjamins.

Puskás, Genovena. 2000. *Word Order in Hungarian: The Syntax of A-Positions.* Vol. 33 of *Linguistik Aktuell/Linguistics Today*. Amsterdam/Philadelphia, John Benjamins.

Schwenter, Scott A., and Gláucia Silva. 2002. Overt vs. null direct objects in spoken brazilian portuguese: A semantic/pragmatic account. MS. The Ohio State University.

Seidl, Amanda, and Alexis Dimitriadis. 1996. The Discourse Function of Object Marking in Swahili. MS. University of Pennsylvania.

Sells, Peter (ed.). 2001. *Formal and Empirical Issues in Optimality Theoretic Syntax*. Stanford, CA, CSLI Publications.

Smolensky, Paul. 1995. On the internal structure of constraint component *con* of UG. Handout of the talk given at UCLA, April 7. Available via the Rutgers Optimality Archive at `http://ruccs.rutgers.edu/roa.html`.

Stiebels, Barbara. 1999. Noun-verb symmetries in Nahuatl nominalizations. *Natural Language & Linguistic Theory* 17, 783–836.

Stiebels, Barbara. 2000. Linker inventories, linking splits and lexical economy. In Barbara Stiebels and Dieter Wunderlich (eds.), *Lexicon in Focus*, 213–247. Berlin, Akademie Verlag.

Stucky, Susan. 1981. *Word Order Variation in Makua: A Phrase Structure Grammar Analysis.* Doctoral dissertation, University of Illinois, Urbana, IL.

Stucky, Susan. 1983. Verb phrase constituency and linear order in Makua. In Gerald Gazdar (ed.), *Order, Concord, and Constituency*, 75–94. Dortrecht, Foris.

Takizala, Alexis. 1973. Focus and relativization: The case of Kihung'an. In John P. Kimball (ed.), *Syntax and Semantics*, Vol. 2, 123–148. New York, Academic Press.

Vitale, Anthony J. 1981. *Swahili Syntax.* Dordrecht/Cinnaminson, Foris Publications.

Wald, Benji. 1979. The development of the Swahili object marker: A study of the interaction of syntax and discourse. In Talmy Givón (ed.), *Discourse and Syntax*, Vol. 12 of *Syntax and Semantics*, 505–524. New York, Academic Press.

Wierzbicka, Anna. 1981. Case marking and human nature. *Australian Journal of Linguistcs* 1, 43–80.

Wunderlich, Dieter. 1997a. Argument extension by lexical adjunction. *Journal of Semantics* 14, 95–142.

Wunderlich, Dieter. 1997b. Cause and the structure of verbs. *Linguistic Inquiry* 28(1), 27–68.

Wunderlich, Dieter. 2000. Predicate composition and argument extension as general options—a study in the interface of semantic and conceptual structure. In Barbara Stiebels and Dieter Wunderlich (eds.), *Lexicon in Focus*, 249–272. Berlin, Akademie Verlag.

Wunderlich, Dieter. 2001. On the interaction of structural and semantic case. *Lingua* 277–418.

Zwart, Jan-Wouter. 1997. Rethinking subject agreement in Swahili. MS. NWO/University of Groningen.

Heinrich-Heine Universität Düsseldorf
Institute für Sprache und Information
morimoto@phil-fak.uni-duesseldorf.de
http://www-csli.stanford.edu/~morimoto

CLITICS AND PHRASAL AFFIXATION IN CONSTRUCTIVE MORPHOLOGY*

Rob O'Connor
University of Manchester

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

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

* I thank Kersti Börjars, Thomas Klein, Tracy Holloway King, Rachel Nordlinger and Andrew Spencer for comments and advice; Marko Gregović, Bosiljka Janjusević, Adisa Lokmić, Alexandra Perović and Drazen Simlesa for grammaticality judgements. Any remaining shortcomings are my responsibility alone. This work has been funded by AHRB award no. 00/3687.

Abstract

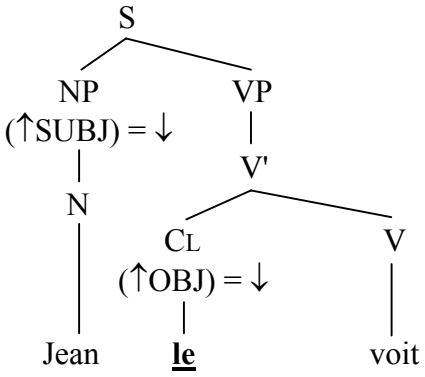
To date LFG approaches to clitics have viewed them as independent c-structure entities, hence representing them as separate terminal nodes. However, the literature on clitics also includes work like that of Anderson (1992, 2000) and Legendre (2000), among others, in which clitics are treated as phrasal affixes. In this paper I apply the idea of clitics as phrasal affixes to Serbian auxiliary and pronominal clitics and adapt the phrasal affix approach to LFG through the use of Constructive Morphology (Nordlinger 1998). In this way grammatical function and other information associated with clitics is contributed to the clause at the right level of c-structure while avoiding the need to represent clitics (phrasal affixes) as separate c-structure nodes.

1. Introduction

Within LFG clitics have, on the whole, been represented as syntactically transparent entities, that is as independent terminal nodes. There have been two versions of this approach which differ only in the c-structure labelling of the node dominating the clitic. Firstly, Grimshaw (1982) and some more recent LFG representations of clitics such as Bresnan (2001) and Schwarze (2001) treat Romance pronominal clitics as daughters of a CL node, as in (1b).[1]

(1) (a) Jean **le** voit.
Jean DO.3.M.SG see
'Jean sees him.'

(b)



(Grimshaw 1982: 93)

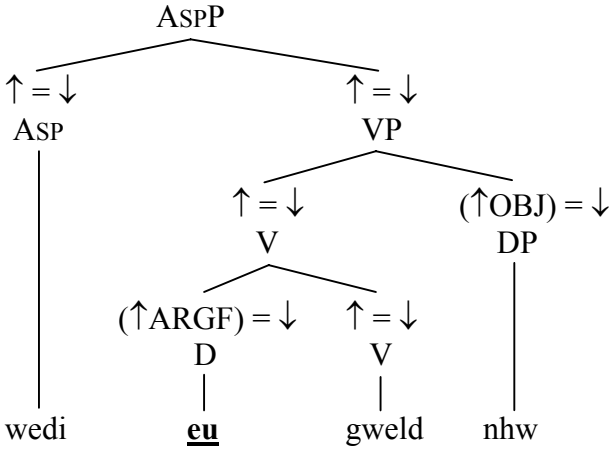
A CL node implies that all clitics can be grouped together within a single syntactic category. However, since clitics include, in addition to pronominals, elements as disparate as auxiliaries, discourse particles and grammatical particles, it is impossible to sustain a unified syntactic category corresponding to such a CL node.

In the second version of this approach clitics are dominated by nodes representing categories that reflect their varied grammatical functions. For instance, in King (1995) the Russian *yes*/*no* question clitic, *li*, is treated as category C, while in Sadler (1997) Welsh pronominal clitics are treated as category D, as in (2b).

[1] Throughout this paper clitics in the examples are in boldface and are underlined. The following abbreviations are used in the interlinear glosses in the examples: 1/2/3 – first/second/third person; ACC – accusative case; AOR – aorist; ASP – aspect; AUX – auxiliary; CL – clitic; DO – direct object; F – feminine; INF – infinitive; IO – indirect object; M – masculine; N – neuter; NOM – nominative case; PL – plural; PPT – participle; PRES – present; PRN – pronominal; SG – singular.

(2) (a) ...wedi **eu** gweld nhw.
ASP CL.3.PL see.PPT PRN.3.PL
'…has seen them.'

(b)



(Sadler 1997: 9)

King (1995) and Sadler (1997) argue for the syntactic transparency of these clitics. The establishment of such transparency supports their representations of clitics as c-structure nodes.

However, in the case of Serbian auxiliary and pronominal clitics, there is evidence to suggest that these are not syntactically transparent elements, and hence should not be represented as syntactic terminals, whether as daughters of CL, or daughters of a node like D corresponding to a specific syntactic category. An alternative approach is to treat Serbian clitics as examples of phrasal affixation (e.g. Anderson 1992).[2] Under this view cliticisation is regarded as the morphology of phrases and is analogous to word-level affixation.[3] The aim of this paper then is to outline how such an approach to clitics can be accommodated to LFG.

This paper is organised as follows: in section 2 I provide some background including arguments in favour of treating Serbian clitics as phrasal affixes; in section 3 I turn to Constructive Morphology (Nordlinger 1998) as a means of accommodating the phrasal affixation approach to cliticisation into LFG; section 4 contains some concluding remarks.

2. Background

In this section I present an overview of Serbian auxiliary and pronominal clitics (section 2.1); briefly discuss Serbian phrase structure (section 2.2) and set out some arguments in favour of a phrasal affix approach, focussing on morphological idiosyncratic behaviour of some clitics as well as the lack of a syntactic relationship between clitics and the elements adjacent to them (section 2.3).

2.1 Serbian Auxiliary and Pronominal Clitics

Serbian clitics occupy the second position in their clause and encliticise prosodically to the element that precedes them. The syntactic and/or prosodic nature of the host element and the mechanisms by which the clitics come to be in second position are not the focus of the present paper. These aspects of Serbian clitics have been widely discussed elsewhere in the literature – see, for example, Progovac

[2] See also Luis, A., Sadler, L. & Spencer, A. (this volume) for a paradigm function approach to Portuguese clitics.

[3] This paper considers only morphological aspects of Serbian clitics but their placement is also influenced by prosodic factors. See O'Connor (2002) for a discussion of these factors.

(1996), Radanović-Kocić (1996), Anderson (2000), Bošković (2000), Franks & King (2000) and O'Connor (2002), among many others.

Auxiliary clitics, given in (3), are used in the formation of the past, conditional and future. The past is formed from the present tense clitic forms of *biti*, 'to be', plus past participle, as in example (4a); the conditional is formed from the aorist clitic forms of *biti* plus past participle, as in example (4b); and the future is formed from the present clitic forms of *hteti*, 'to want', plus infinitive, as in example (4c).

(3)

		biti, present tense	*biti*, past tense	*hteti*, present tense
Singular	1	sam	bih	ću
	2	si	bi	ćeš
	3	je	bi	će
Plural	1	smo	bi/bismo	ćemo
	2	ste	bi/biste	ćete
	3	su	bi	će

(4) (a) Devojk-a **<u>je</u>** oborila drv-o.
girl-NOM AUX.3.SG.PRES chop.PPT.F.SG tree-ACC
'The girl chopped the tree.'

(b) Devojk-a **<u>bi</u>** oborila drv-o.
girl-NOM AUX.3.SG.AOR chop.PPT.F.SG tree-ACC
'The girl would chop the tree.'

(c) Devojk-a **<u>ću</u>** oboriti drv-o.
girl-NOM AUX.3.SG.PRES chop.INF tree-ACC
'The girl will chop the tree.'

Serbian pronominal clitics, given in (5), include direct object clitics, as in example (6a), and indirect object clitics, as in (6b). There is also a reflexive clitic with the form *se* for all numbers and persons.

(5)

			direct object clitics	indirect object clitics
Singular	1		me	mi
	2		te	ti
	3	masc./neut.	ga	mu
		fem.	je/ju	joj
Plural	1		nas	nam
	2		vas	vam
	3		ih	im

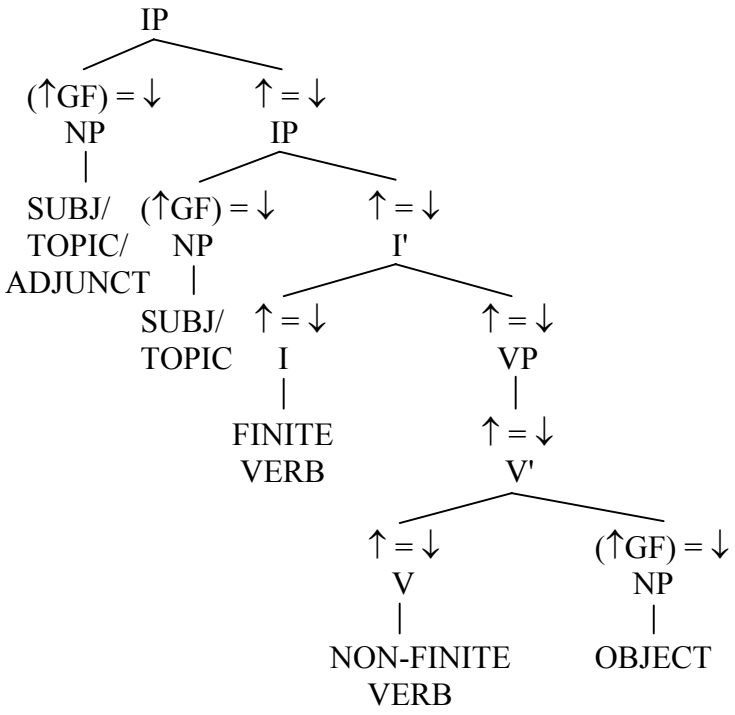
(6) (a) Devojk-a **<u>ga</u>** obori.
girl-NOM DO.3.SG.N chop.3.SG.PRES
'The girl is chopping it.'

(b) Devojk-a **<u>mi</u>** **<u>je</u>** dala knjig-u.
girl-NOM IO.1.SG AUX.3.SG.PRES give.PPT.F.SG book-ACC
'The girl gave me the book.'

2.2 Serbian C-Structure

For the purposes of the discussion in this paper I assume the c-structure in (7) for Serbian (which is based on that discussed in King 1995 and Bresnan 2001 for Russian).

(7)



In (7), any element in specifier position of IP is a discourse function – e.g. subject or topic. Although SVO is the usual Serbian word order, this is not fixed and fronting of other elements according to topic-comment considerations is very common, as illustrated by examples (9b, c) and (11) below. Such fronting is allowed for by the adjunction structure in (7). I take inflected verbs to be in I^0. In Serbian, while most instances of the past, conditional or future feature a clitic auxiliary, the occurrence of a phonologically strong form of the auxiliary is also possible in special circumstances. I take these strong auxiliaries to be in I^0 and hence the non-finite verb form, whether accompanied by a clitic or full form auxiliary, to be in V^0.

2.3 Serbian Clitics as Phrasal Affixes

Evidence for the phrasal affix status of clitics consists of both morphological and syntactic arguments. Morphologically, three aspects of Serbian clitics bear a close resemblance to the behaviour of word level affixes.

Firstly, clitic clusters in Serbian exhibit a rigid internal ordering which is somewhat at odds with the free word order found elsewhere in the language. This rigidity is shown in (8) and exemplified in (9) and (10).[4]

(8) *li* – AUX (except *je*) – IO – DO – *je*, *se*

(9) (a) Marija **mu** **je** da.
Marija IO.3.SG.M DO.3.SG.F give.3.SG.PRES
'Marija is giving it to him.'
(b) *Marija **je** **mu** da.

[4] The Serbian *yes*/*no* question particle, *li*, is also a second position enclitic like the auxiliary and pronominal clitics. In the present paper considers neither the syntactic/phrasal affixal status of *li* nor the means by which its associated information is contributed at the level of the clause.

(10) (a) Jovan **mi** **ih** **je** dao.
Jovan IO.1.SG DO.3.PL AUX.3.SG.PRES give.PASTP.M.SG
'Jovan gave them to me.'
(b) *Jovan **je** **mi** **ih** dao.
(c) *Jovan **mi** **je** **ih** dao.
(d) *Jovan **ih** **mi** **je** dao.

Secondly, as indicated in (8), when auxiliary *je* occurs in a clitic cluster, it is constrained to follow the pronominal clitics. By contast all other auxiliary clitics have to precede the pronominals. Example (11) demonstrates this idiosyncratic behaviour of auxiliary *je*.

(11) (a) Dala **mu** **je** knjigu.
give.PPT.F.SG IO.3.SG.M AUX.3.SG.PRES book
'She gave him a book.'
(b) Dala **sam**/**si** **mu** knjigu.
give.PPT.F.SG AUX.1.SG/2.SG.PRES IO.3.SG.M book
'I/you.SG/we/you.PL/they gave him a book.'

Thirdly, the feminine singular direct object clitic, *je*, has a morphophonological alternation, appearing as *ju*, when followed by auxiliary *je*, as shown in (12).

(12) U Beogradu **ju**/***je** **je** Marija kupila
In Beograd DO.3.SG.F AUX.3.SG.PRES Marija buy. PPT.F.SG
'Marija bought it in *Beograd*.'

Such behaviour – rigid ordering of elements; idiosyncratic ordering of a specific element; morphophonological alternation of a specific element when juxtaposed with some other element – is more reminiscent of affixal morphology than of syntactically independent elements. This suggests that treating Serbian clitics as phrasal affixes may be more appropriate than treating them as syntactic terminals.

In syntactic terms Serbian clitics precede and follow such a variety of elements that no syntactically consistent pattern of placement is apparent. The clitics must occur in a fixed order in second position in the clause, as in (13a) (alternatively, attached to an initial prosodic element of some kind – O'Connor (2002) discusses attachment to both an initial prosodic word and an initial phonological phrase). It is impossible for the clitics to occur in any other position (13b, c).

(13) (a) Marija **mu** **je** da.
Marija IO.3.SG.M DO.3.SG.F give.3.SG.PRES
'Marija is giving it to him.'
(b) ***Mu** **je** Marija da.
(c) *Marija da **mu** **je**.

While clitics are subject to strict placement and ordering, non-clitic elements can be relatively freely ordered resulting in various discourse effects. In (14a), with 'neutral' word order, the clitic follows the subject and precedes the verb. In (14b), a word order which places some degree of emphasis on *knjigu*, the clitic is no longer adjacent to the verb. This is also true in (14c), with even greater emphasis on *knjigu* while, in addition, the clitic precedes the subject.

(14) (a) Jovan **je** čitao knjigu.
Jovan AUX.3.SG.PRES read.PPT.M.SG book
'Jovan read the book.'
(b) Jovan **je** knjigu čitao
(c) Knjigu **je** Jovan čitao.

Whether a clause is CP or IP, clitics are nevertheless restricted to second position. Example (15a) follows the same pattern as (14a), but in (15b) the clitic is again separated from the verb by the subject, and precedes that subject.

(15) (a) Marija **je** kupila knjigu.
Marija AUX.3.SG.PRES buy. PPT.F.SG book
'Marija bought a book.'
(b) Šta **je** Marija kupila?
what AUX.3.SG.PRES Marija buy. PPT.F.SG
'What did Marija buy?'

Clitics also follow fronted adverbial material, as in (12), repeated as (16), and which resembles the pattern in (14c).

(16) U Beogradu **ju** **je** Marija kupila
In Beograd DO.3.SG.F AUX.3.SG.PRES Marija buy. PPT.F.SG
'Marija bought it in *Beograd*.'

There is also an alternation in clitic placement when the initial syntactic constituent is an NP with adjectival premodification. The clitic either follows the whole NP, as in (17a), or the first modifier, as in (17b).

(17) (a) Mladi čovek **je** čitao knjigu.
young man AUX.3.SG.PRES read.PPT.M.SG book
(b) Mladi **je** čovek čitao knjigu.
young AUX.3.SG.PRES man read.PPT.M.SG book
'The young man read a book.'

Examples (13)-(17) indicate that, if Serbian clitics are to be considered syntactically transparent, then they can occupy a great variety of positions which, in terms of the c-structure in (7), can be summarised as in (18). However, there is no consistent characteristic linking either the varied positions supposedly occupied by the clitics or the adjacent constituents with which they supposedly form some syntactic relationship.

(18)

Position	Follows	Precedes	Examples
Between SPECIP and V^0	Subject	Main verb	(14a), (15a), (17a)
Between SPECIP and I^0	Subject	Main verb	(13a)
Between upper and lower SPECIP	Subject	Object	(14b)
Between upper and lower SPECIP	Fronted object/ adverbial	Subject	(14c)/(16)
Between CP and SPECIP	WH-element	Subject	(15b)
Between AP/DP and N^0 (in SPECIP)	Premodifier	N^0 head	(17b)

The strict association of Serbian clitics with second position, no matter where that is in phrase structure terms, and their lack of an apparent syntactic relationship with any element such as the main verb, together point to the conclusion that they are syntactically opaque. Hence, in the remainder of this paper, I consider phrasal affixation, as described in Anderson (1992) and outlined in (19) and (20), to be a more appropriate representation for Serbian clitics.

(19) (a) The clitic is located within some syntactic constituent (S vs. VP vs. NP, etc.) which constitutes its domain.
(b) The clitic is located by reference to the {first vs. last vs. head} element of a specified sort within the constituent in which it appears.
(c) The clitic {precedes vs. follows} this reference point.

(20) (a) The affix is located in the scope of some constituent which constitutes its domain. This may be either a morphological constituent (the word-structural head vs. entire word) or a prosodic one (prosodic word).
(b) The affix is located by reference to the {first vs. last vs. main stressed} element of a given type within the constituent in which it appears.
(c) The affix {precedes vs. follows} the reference point.

Given syntactic opacity and phrasal affixation, I treat Serbian clitics, not as syntactic terminals, but as affixes and hence the sequence prosodic host + clitic as a single syntactic entity. In the next section I propose how this approach can be implemented within LFG.

3. A Constructive Morphology Approach

Constructive Morphology (Nordlinger 1998), a sub-theory of LFG, represents how case morphemes can contribute grammatical function information to f-structure. It has been especially successfully applied to non-configurational dependent marking languages such as Wambaya and a number of other Australian languages. Constructive Morphology works through two mechanisms: 'inside out' function application and morphological composition. These are dealt with in turn.

The designator (SUBJ↑) in the lexical entry for the Serbian feminine singular nominative suffix in (21) is an example of an inside out designator. The assignment of such designators to a case morpheme associates the element to which that morpheme is affixed with the relevant grammatical relation, in this case subject.

(21) *-a*: (SUBJ↑)
(↑CASE) = NOM
(↑NUM) = SG

(SUBJ↑) denotes the higher f-structure containing the SUBJ attribute, i.e. f' in (22).

(22) $_{f'}$[SUBJ $_{f}$[ ]]

In other words, the inside out designator allows the affix to build the f-structure in (22). This, in effect, is a type of shorthand encompassing the fact that the affix *-a* carries the information that there is an f-structure, f' in this case, and that f-structure contains a subject attribute.

The Principle of Morphological Composition in (23) (Nordlinger 1998: 102) allows for the sequential contribution of affixal information. The affix may contain an inside out designator as is the case in some of Nordlinger's work on Australian languages. The Serbian case morphemes to be dealt with in this section are followed by clitics or phrasal affixes (AFF_{XP}) which, for present purposes, can be considered not to contain inside out designators.[5]

(23) Stem AFF ⇒ Stem AFF
(GF^n↑) (↑x) (GF^n↑) ((GF^n↑) x)

(where x represents a sequence of attributes)

Applied to Serbian host-clitic sequences, the host/stem's inside out designator, (SUBJ↑) in (24), is inserted for ↑ in the clitic's lexical entry

(24) X-*a* AFF_{XP} ⇒ Stem AFF_{XP}
(SUBJ↑) (↑PRED) (SUBJ↑) ((SUBJ↑)PRED)
(↑…) ((SUBJ↑)…)

This has the consequence that the f-structure containing the SUBJ attribute also has the PRED feature and other features associated with AFF_{XP}. In the remainder of this section morphological composition is applied to sequences in Serbian which contain a subject NP hosting a pronominal clitic (section 3.1), an auxiliary clitic (section 3.2), and a clitic cluster (section 3.3). Other patterns are dealt with in sections 3.4 and 3.5.

3.1 Serbian Auxiliary Clitics

Consider example (4a), repeated as (25) below.

(25) Devojk-a **je** oborila drv-o.
girl-NOM AUX.3.SG.PRES chop.PPT.F.SG tree-ACC
'The girl chopped the tree.'

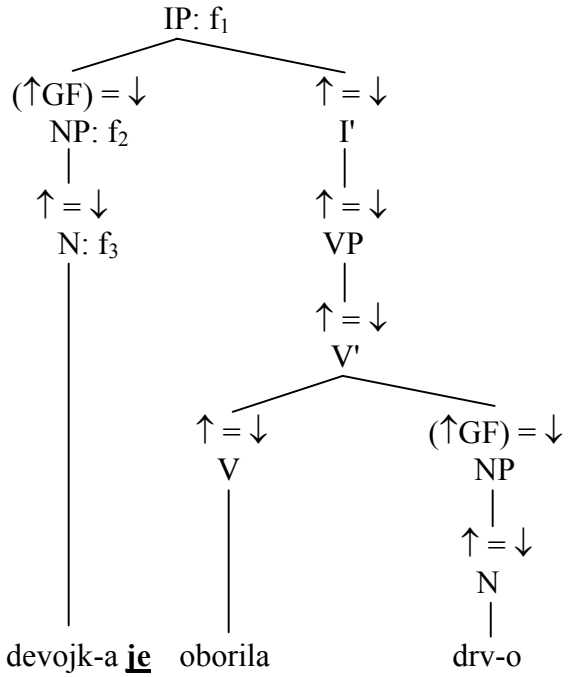
In the following I illustrate how Constructive Morphology works for the host-clitic sequence *devojk-a je*. As stated at the end of section 2 I treat this sequence as a single syntactic entity which is associated with the lexical information in (26).

(26) (a) *devojk-*: N (↑PRED) = 'girl'
(↑GEND) = F
(↑PERS) = 3
(b) *-a*: AFF_N (SUBJ↑)
(↑CASE) = NOM
(↑NUM) = SG
(c) *je*: AFF_{XP} (↑TENSE) = PAST
(↑SUBJ PERS) = 3
(↑SUBJ NUM) = SG

[5] See section 3.4 for a modification whereby clitics have an associated optional inside out designator.

Example (25) has the c-structure in (27) which, by means of the functional equations in (28), is mapped to the f-structure in (29).

(27)



(28) (a) $(f_1\ GF) = f_2$
(b) $f_2 = f_3$
(c) $(f_3\ PRED) =$ 'girl'
$(f_3\ GEND) = F$
$(f_3\ PERS) = 3$
(d) $(SUBJ\ f_3)$
(e) $(f_3\ CASE) = NOM$
$(f_3\ NUM) = SG$
(f) $((SUBJ\ f_3)\ TENSE) = PAST$
$((SUBJ\ f_3)\ SUBJ\ PERS) = 3$
$((SUBJ\ f_3)\ SUBJ\ NUM) = SG$

(29)

$$f_1\begin{bmatrix} \text{TENSE} & \text{PAST} \\ \text{SUBJ} & f_2, f_3\begin{bmatrix} \text{PRED} & \text{'girl'} \\ \text{GEND} & \text{F} \\ \text{PERS} & 3 \\ \text{CASE} & \text{NOM} \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

In particular, the inside out designator in the first functional equation in (28f) – i.e. $((SUBJ\ f_3)\ TENSE) = PAST$ – allows the TENSE feature contributed by the auxiliary clitic, *je*, to be associated with the outer f-structure, f_1 (i.e. the f-structure that contains the SUBJ attribute and its value, the f-structure f_3). This f-structure corresponds to the whole clause, which is the level at which the TENSE attribute is relevant. This is despite the fact that *je* itself is treated as affixed to the subject rather than as an independent syntactic terminal.

3.2 Serbian Pronominal Clitics

A sequence containing a pronominal clitic like (6a), repeated as (30), is treated in a similar fashion.

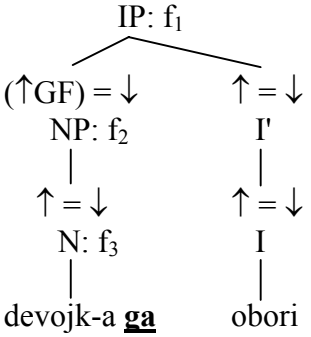
(30) Devojk-a **ga** obori.
girl-NOM DO.3.SG.N chop.3.SG.PRES
'The girl is chopping it.'

Devojk-a ga has the lexical information in (31).[6]

(31) (a) *devojk-*: N (↑PRED) = 'girl'
(↑GEND) = F
(↑PERS) = 3
(b) *-a*: AFF_N (SUBJ ↑)
(↑CASE) = NOM
(↑NUM) = SG
(c) *ga*: AFF_{XP} (↑OBJ PRED) = 'PRO'
(↑OBJ PERS) = 3
(↑OBJ NUM) = SG
(↑OBJ GEND) = N

Example (30) has the c-structure in (32) which is mapped via (33) to the f-structure in (34).

(32)



(33) (a) (f_1 GF) = f_2
(b) f_2 = f_3
(c) (f_3 PRED) = 'girl'
(f_3 GEND) = F
(f_3 PERS) = 3
(d) (SUBJ f_3)
(e) (f_3 CASE) = NOM
(f_3 NUM) = SG
(f) ((SUBJ f_3) OBJ PRED) = 'PRO'
((SUBJ f_3) OBJ PERS) = 3
((SUBJ f_3) OBJ NUM) = SG
((SUBJ f_3) OBJ GEND) = N

[6] For an indirect object clitic the associated lexical information would take the form (↑OBJ_Θ PRED) = 'PRO', and so on.

(34)
$$\begin{bmatrix} \text{SUBJ} & \overset{f_2,\, f_3}{\begin{bmatrix} \text{PRED} & \text{'girl'} \\ \text{GEND} & \text{F} \\ \text{PERS} & 3 \\ \text{CASE} & \text{NOM} \\ \text{NUM} & \text{SG} \end{bmatrix}} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{PERS} & 3 \\ \text{CASE} & \text{NOM} \\ \text{GEND} & \text{N} \end{bmatrix} \end{bmatrix}_{f_1}$$

In this case the effect of the inside out designator is to ensure that the information associated with the clitic *ga* ends up in the correct part of the f-structure for the whole clause. Without the inside out designator the OBJ attribute and its f-structure value would have ended up inside the f-structure f_2/f_3 just like the information associated with the case affix *-a*.

3.3 Clitic Clusters

Clitic clusters such as the auxiliary-pronominal sequence *ga je* in (35) can also be handled by this approach.[7] As in the previous cases involving a single clitic, the sequence of host plus clitic cluster, *devojk-a ga je*, is to be treated as a single syntactic entity.

(35) Devojk-a **ga** **je** oborila.
girl-NOM DO.3.SG.N AUX.3.SG.PRES chop.PPT.F.SG
'The girl chopped it.'

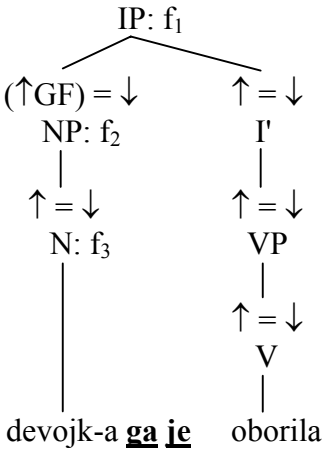
Devojk-a ga je has the associated lexical information in (36).

(36) (a) *devojk-*: N (↑PRED) = 'girl'
(↑GEND) = F
(↑PERS) = 3
(b) *-a*: AFF_{N} (SUBJ ↑)
(↑CASE) = NOM
(↑NUM) = SG
(c) *ga*: AFF_{XP} (↑OBJ PRED) = 'PRO'
(↑OBJ PERS) = 3
(↑OBJ NUM) = SG
(↑OBJ GEND) = N
(d) *je*: AFF_{XP} (↑TENSE) = PAST
(↑SUBJ PERS) = 3
(↑SUBJ NUM) = SG

[7] For a sequence of affixes morphological composition as originally put forward in Nordlinger (1998) – see (23) and (24) above – operates on each affix in turn. See Nordlinger & Sadler (this volume) for a reinterpretation of morphological composition in terms of paradigm function morphology whereby a complete affix sequence is generated before affixation to a stem. This could also be applied to a sequence of clitics/phrasal affixes.

Example (35) has the c-structure in (37) which is mapped via the equations in (38) to the f-structure in (39).

(37)



(38) (a) $(f_1\ GF) = f_2$
(b) $f_2 = f_3$
(c) $(f_3\ PRED) =$ 'girl'
$(f_3\ GEND) = F$
$(f_3\ PERS) = 3$
(d) $(SUBJ\ f_3)$
(e) $(f_3\ CASE) = NOM$
$(f_3\ NUM) = SG$
(f) $((SUBJ\ f_3)\ OBJ\ PRED) =$ 'PRO'
$((SUBJ\ f_3)\ OBJ\ PERS) = 3$
$((SUBJ\ f_3)\ OBJ\ NUM) = SG$
$((SUBJ\ f_3)\ OBJ\ GEND = N$
(g) $((SUBJ\ f_3)\ TENSE) = PAST$
$((SUBJ\ f_3)\ SUBJ\ PERS) = 3$
$((SUBJ\ f_3)\ SUBJ\ NUM) = SG$

(39)

$$\begin{bmatrix} \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \overset{f_2,\ f_3}{\begin{bmatrix} \text{PRED} & \text{'girl'} \\ \text{GEND} & \text{F} \\ \text{PERS} & 3 \\ \text{CASE} & \text{NOM} \\ \text{NUM} & \text{SG} \end{bmatrix}} \\ \text{OBJ} & \overset{f_1}{\begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \\ \text{GEND} & \text{N} \end{bmatrix}} \end{bmatrix}$$

As was the case previously in (29) and (34) the information associated with each clitic appears in the appropriate place in the f-structure corresponding to the whole sequence *devojk-a ga je*.

3.4 Other Hosts

So far I have considered only instances where the clitic host is a subject NP. However, it is not always the case in Serbian that the host element is the subject since pro-drop and free word order effects mean that elements other than subject NPs are frequently in first position and therefore fulfill the role of clitic host.

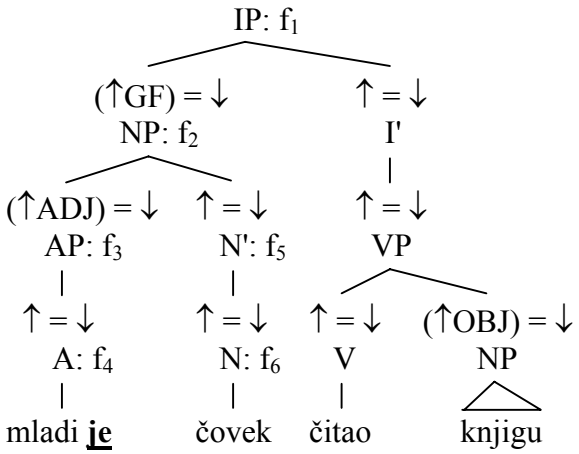
The most straightforward situation is that in which the host element is an NP marked for a case other than nominative. Such case markers carry an alternative inside out designator – e.g. (OBJ↑) for accusative, (OBJ_θ↑) or (OBL_θ↑) for dative, (OBL_θ↑) for locative and instrumental, (POSS↑) for genitive. As before, these allow the information carried by the clitic to be contributed at the correct level of f-structure.

Less straightforward are situations in which the host is a premodifier within the noun phrase, as in (17b), repeated as (40) below.

(40) Mladi **je** čovek čitao knjigu.
young AUX.3.SG.PRES man read.PPT.M.SG book
'The young man read a book.'

Serbian adjectives, like nouns, are marked for case and thus adjectival affixes can be regarded as carrying the same inside out designators as their nominal counterparts. However, this in itself does not produce the desired result, as the following demonstrates. The c-structure corresponding to example (40) is given in (41).

(41)



The lexical information associated with *mladi je* and *čovek* is contained in (42) and (43) respectively. The result of unifying this information is given in the f-structure in (44) which is ungrammatical in a number of respects.

(42) (a) *mlad-*: A (↑PRED) = 'young'
(b) *-i*: AFF_A (SUBJ ↑)
(↑CASE) = NOM
(↑NUM) = SG
(c) *je*: AFF_{XP} (↑TENSE) = PAST
(↑SUBJ PERS) = 3
(↑SUBJ NUM) = SG

(43) (a) *čovek-*: N (↑PRED) = 'man'
(↑GEND) = M
(↑PERS) = 3
(b) ∅: AFF_N (SUBJ ↑)
(↑CASE) = NOM
(↑NUM) = SG

(44) *
$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{ADJ/SUBJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'young'} \\ \text{NUM} & \text{SG} \\ \text{CASE} & \text{NOM} \\ \text{PERS} & 3 \end{bmatrix}_{f_3,\, f_4} \right\} \\ \text{PRED} & \text{'man'} \\ \text{GEND} & \text{M} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \\ \text{CASE} & \text{NOM} \\ \text{TENSE} & \text{PAST} \end{bmatrix}_{f_2,\, f_5,\, f_6} \end{bmatrix}_{f_1}$$

The main problems with this structure is that SUBJ information occurs in both of the inner f-structures, $f_2/f_5/f_6$ and f_3/f_4, and that the f-structure f_3/f_4 is the value of both an ADJ attribute (from the annotation (↑ADJ) = ↓ on the AP node in (41)) and a SUBJ attribute (from the (SUBJ ↑) designator in the lexical entry for the adjectival affix *-i* in (42b)). The remedy adopted by Nordlinger (1998: 99), and which I will follow in this paper, is to assume that modifiers such as adjectives have an (ADJ ↑) designator in their lexical entries. Thus (42a) above becomes (45).

(45) *mlad-*: A (ADJ ↑)
(↑PRED) = 'young'

The effect of this is to allow the adjectival suffix *-i* with its associated (SUBJ ↑) designator to build f-structure outside of that built by the adjectival root *mlad-*, as illustrated below. The annotations on the c-structure in (41) together with the lexical information in (42b, c), (43) and (45) produce the functional equations in (46), associated with *mladi je*, and (47), associated with *čovek*. These equations provide the mapping from (41) to the correct f-structure in (48).

(46) (a) (f_1 GF) = f_2
(b) (f_2 ADJ) = f_3
(c) $f_3 = f_4$
(d) (ADJ f_4)
(f_4 PRED) = 'young'
(e) (SUBJ (ADJ f_4))
((ADJ f_4) CASE) = NOM
((ADJ f_4) NUM) = SG
(f) ((SUBJ (ADJ f_4)) TENSE) = PAST
((SUBJ (ADJ f_4)) SUBJ PERS) = 3
((SUBJ (ADJ f_4)) SUBJ NUM) = SG

(47) (a) $(f_1\ GF) = f_2$ (as above)
(b) $f_2 = f_5 = f_6$
(c) $(f_6\ PRED) =$ 'man'
$(f_6\ GEND) = M$
$(f_6\ PERS) = 3$
(d) $(SUBJ\ f_6)$
$(f_6\ CASE) = NOM$
$(f_6\ NUM) = SG$

(48)
$$\begin{bmatrix} SUBJ & \begin{bmatrix} PRED & 'man' \\ CASE & NOM \\ NUM & SG \\ PERS & 3 \\ GEND & M \\ ADJ & \{f_3, f_4 [PRED\ 'young']\} \end{bmatrix} \\ & f_2, f_5, f_6 \\ f_1 \begin{bmatrix} TENSE & PAST \end{bmatrix} \end{bmatrix}$$

Finally there are cases in which the element which hosts the clitics has no case marking. For instance, in (49) the host is a verbal participle while in (50) it is an adverb.

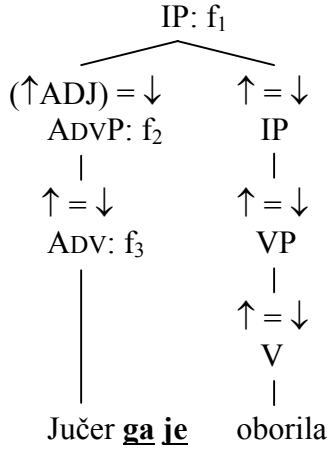
(49) Oborila **ga** **je**.
chop.PPT.F.SG DO.3.SG.N AUX.3.SG.PRES
'She chopped it.'

(50) Jučer **ga** **je** oborila.
Yesterday DO.3.SG.N AUX.3.SG.PRES chop.PPT.F.SG
'Yesterday she chopped it.'

In example (49) the host does not require an inside out designator since, as a verbal participle, it contributes information directly to the topmost or clause-level f-structure, in contrast with previous examples in which the hosts contribute information to a lower f-structure (such as the f-structure value for the SUBJ attribute) embedded within the main f-structure.

In example (50), however, the host is an adjunct and is therefore an embedded f-structure within the main f-structure representing the whole clause. As for *mlad-* above, *jučer* is a modifier and can therefore be regarded as carrying its own (ADJ ↑) designator. Thus from the c-structure in (51) and the lexical information in (52) the f-structure in (54) can be built via the equations in (53).

(51)



(52) (a) *jučer*: N (ADJ↑)
(↑PRED) = 'yesterday'
(b) *ga*: AFF_{XP} (↑OBJ PRED) = 'PRO'
(↑OBJ PERS) = 3
(↑OBJ NUM) = SG
(↑OBJ GEND) = N
(c) *je*: AFF_{XP} (↑TENSE) = PAST
(↑SUBJ PERS) = 3
(↑SUBJ NUM) = SG

(53) (a) $(f_1\ \text{ADJ}) = f_2$
(b) $f_2 = f_3$
(c) $(\text{ADJ}\ f_3)$
$(f_3\ \text{PRED})$ = 'yesterday'
(d) $((\text{ADJ}\ f_3)\ \text{OBJ PRED})$ = 'PRO'
$((\text{ADJ}\ f_3)\ \text{OBJ PERS}) = 3$
$((\text{ADJ}\ f_3)\ \text{OBJ NUM}) = \text{SG}$
$((\text{ADJ}\ f_3)\ \text{OBJ GEND}) = \text{N}$
(e) $((\text{ADJ}\ f_3)\ \text{TENSE}) = \text{PAST}$
$((\text{ADJ}\ f_3)\ \text{SUBJ PERS}) = 3$
$((\text{ADJ}\ f_3)\ \text{SUBJ NUM}) = \text{SG}$

(54)
$$f_1\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \\ \text{GEND} & \text{N} \end{bmatrix} \\ \text{TENSE} & \text{PAST} \\ \text{ADJ} & \{f_2, f_3 \begin{bmatrix} \text{PRED} & \text{'yesterday'} \end{bmatrix}\} \end{bmatrix}$$

4. Conclusion

In this paper I have shown that a treatment of Serbian auxiliary and pronominal clitics as phrasal affixes, along the lines proposed by Anderson (1992), is compatible with an LFG approach to phrase structure. These clitics/phrasal affixes are not represented as independent syntactic terminal elements in Serbian phrase structure. Nevertheless, the information that they carry can still be contributed to well formed f-structures in spite of the variety of grammatical functions associated with the elements to which these clitics are attached. The means by which this can be achieved is Constructive Morphology as advocated in Nordlinger (1998). In particular, inside out function assignment and the principle of morphological composition allow clitic information to be associated not with the embedded f-structure belonging to the host element, but with the immediately containing f-structure belonging to the clause itself.

References

Anderson, S. (1992). *A-morphous morphology*. Cambridge: CUP.

Anderson, S. (2000). Toward an optimal account of second-position phenomena. In Dekkers, J. et. al. (eds.), 302-333.

Bošković, Ž. (2000). Second position cliticization: syntax and/or phonology? In Beukema, F. & den Dikken, M. (eds.), *Clitic phenomena in European languages*. Amsterdam: Benjamins. 71-119.

Bresnan, J. (2001). *Lexical-functional syntax.* Oxford: Blackwell.

Dekkers, J., van der Leeuw, F. & van der Weijer, J. (eds.), *Optimality theory: syntax, phonology and acquisition*. Oxford: OUP.

Franks, S. & King, T. (2000). *A handbook of Slavic clitics*. Oxford: O.U.P.

Grimshaw, J. (1982). On the lexical representation of Romance reflexive clitics. In Bresnan, J. (ed.) *The mental representation of grammatical relations.* Cambridge, MA: MIT Press. 87-148.

Halpern, A. & Zwicky, A. (eds.) (1996). *Approaching second: second position clitics and related phenomena*. Stanford: C.S.L.I. Publications.

King, T. (1995). *Configuring topic and focus in Russian*. Stanford: CSLI Publications.

Legendre, G. (2000). Morphological and prosodic alignment of Bulgarian clitics. In Dekkers, J. et. al. (eds.), 423-462.

Nordlinger, R. (1998). *Constructive case: evidence from Australian languages.* Stanford: CSLI Publications.

O'Connor (2002). The placement of enclitics in Bosnian, Croatian and Serbian. Ms. University of Manchester (available at: http://roa.rutgers.edu/view.php3?roa=521.)

Progovac, L. (1996). Clitics in Serbian/Croatian: comp as the second position. In Halpern, A & Zwicky, A. (eds.), 411-428.

Radanović-Kocić, V. (1996). The placement of Serbo-Croatian clitics: a prosodic approach. In Halpern, A. & Zwicky, A. (eds.), 429-445.

Sadler, L. (1997). Clitics and the structure-function mapping. In Butt, M. & King, T., *Proceedings of the LFG97 conference.* (available at: http://www-csli.stanford.edu/publications/.)

Schwarze, C. (2001). On the representation of French and Italian clitics. In Butt, M. & King, T., *The proceedings of the LFG '01 conference.* (http://cslipublications.stanford.edu/LFG/6/lfg01.html.)

Case Marking and Subject Extraction in Danish[1]

Bjarne Ørsnes
Copenhagen Business School

Proceedings of the LFG02 Conference
National Technical University of Athens, Athens
Miriam Butt and Tracy Holloway King (Editors)
2002
CSLI publications
HTTP://CSLI-PUBLICATIONS.STANFORD.EDU/

[1]For helpful discussion, I am grateful to Mary Dalrymple, Line Mikkelsen, Sten Vikner and the audience at LFG02.

1 Introduction

The empirical domain of case theory involves four types of phenomena ([Lee, 2002b]): The marking of core grammatical functions, semantically induced case marking, dependency effects and domain effects. Common to the first three of these phenomena is that case is determined on the basis of local head-dependent relationships, but also "case Stacking" which is given as an illustration of a domain effect in [Lee, 2002b], can be argued to involve a local head-dependent relationship between a nominal and its nominal adjuncts. In this paper I discuss another kind of domain effect where case assignment, however, can not be determined on the basis of a head-dependent relation. Instead, case assignment is determined by the specific syntactic construction. The case in point is the accusative marking of a (non-locally) extracted pronominal subject in Danish.

In Danish pronominal subjects are assigned nominative case, but non-locally extracted pronominal subjects (subjects extracted across a clause boundary) are assigned accusative case. The use of nominative for an extracted subject inevitably brings out a reading of the pronominal as the matrix subject. It is argued that this use of the accusative for an extracted pronoun has a clearly disambiguating function. Accusative signals that the fronted constituent is not the subject of the matrix clause (but possibly the subject of an embedded clause). The paper shows how this generalization can be represented in LFG and, concomitantly, how this kind of constructional case assignment can be accommodated in a lexicalist framework such as LFG.

A further challenge is to account for the distribution of the personal pronouns in a language which otherwise does not employ morphological function specification. Only the personal pronouns exhibit a morphological distinction between nominative and accusative. From a monolingual point-of-view, it is thus empirically inadequate to postulate the existence of nominative and accusative case for other kinds of nominals. A crucial point in the present analysis is to restrict the account to the relevant class of lexical items. I demonstrate how the *Constructive Case*-approach of [Nordlinger, 1998] is capable of accomplishing exactly this.

The accusative marking of non-locally extracted subjects is observed in an informal register of Danish while standard Danish does not seem to allow extraction of pronominal subjects at all. The use of an accusative pronoun to mark an extracted subject is associated with a certain stylistic effect (a colloquial register), and this use is not accepted by all speakers. I will show how this pattern of variation can be accounted for by a small difference in the lexical entries of the personal pronouns. In addition, it is shown how the observed variation in extraction of pronominal subjects can be accommodated within the *Constructive Case*-approach.

Central to the discussion is the question whether an extracted subject is identified in terms of c-structure properties or in terms of f-structure properties. This discussion sheds important light on the syntax of extraction in Danish and I conclude that the extracted subject can only be identified in f-structure terms. It emerges from the analysis that an extracted object is associated with an empty category while an extracted subject is not.

The present analysis has been implemented in XLE (Xerox Linguistic Environment) as part of a broad-coverage LFG-grammar for Danish.[2]

2 The Distribution of the Personal Pronouns

Apart from genitive which is only used to mark nominal attributes, case no longer serves to identify grammatical functions in Danish. Grammatical functions are identified on a configurational basis:

[2]This work is supported by a grant from Nordisk Ministerråd as part of the language technology programme.

objects occur in VP or in I ("object-shift", [Sells, 2001]) while the subject canonically occurs in the specifier position of IP. However, Danish is a V2 language and allows for the topicalization of almost any kind of constituent. In these cases, grammatical functions can frequently only be determined on the basis of selectional restrictions. Morphological function specification is only observed in conjunction with the personal pronouns whose distribution is sensitive to their grammatical function.[3] A distinction between nominative and accusative is observed in the 1st and 2nd person pronouns. Among the 3rd person singular pronouns, the natural-gender specific pronouns, *han*/'he' and *she*/'she', have an accusative form and only take human antecedents. The grammatical-gender specific pronouns *det*/'it' and *den*/'it' have no accusative form and only take non-human antecedents. The 3rd person plural pronoun *de*/'they' has an accusative form and take both human and non-human antecedents.

The basic generalization about the distribution of the personal pronouns in Danish is that the nominative forms are used for subjects (SUBJ) while the accusative forms are used for all other grammatical functions OBJ, OBJ2, XCOMP and in "case-less" positions, i.e. positions which are not associated with an argument function (e.g. "adjoined-to"-positions).

(1) *han* vinder (SUBJ)
*he.*NOM wins

(2) det er *ham* der vinder (XCOMP)
it is *him.*ACC who wins
'he is the one who is winning'

(3) hun giver *ham* præmien (OBJ2)
she gives *him.*ACC the award

(4) hun giver præmien til *ham* (OBJ)
she gives the award to *him.*ACC

(5) *dig,* du kan gå din vej ("case-less" position)
*You.*ACC, you.NOM can go your way
'As for you, you can go your own way'

Contrary to the proposed markedness-hierarchies of case forms in [Woolford, 2001] and [Lee, 2002a], the accusative form seems to be the most unmarked form in Danish. The accusative form is used unless there is good reason not to. In this respect the Danish personal pronouns pattern with the English personal pronouns. [Hudson, 1990] claims that case no longer exists in English. Instead, we find a (closed) class of lexical items whose distribution is determined by their grammatical function. The generalization in Hudson is that the nominative forms only occur with finite verbs. Since the nominative forms impose a constraint on their governing head, i.e. that it be finite, Hudson analyses the distribution of the nominative pronouns as head-marking. In Danish the use of the nominative forms is even more restricted: nominative forms are only used when the pronominal is a local subject. A non-local subject is realized with its accusative form giving rise to a kind of movement paradox:

(6) Peter tror *han* vinder
Peter thinks *he* wins
'Peter thinks he is going to win'

[3]The distribution of the reflexive pronouns could also be determined by grammatical function. The present analysis can straight-forwardly be extended to cover these pronouns as well. Since, however, their distribution also can be claimed to follow from syntactic constraints on coreferentiality they are excluded from the present discussion.

(7) *ham* tror Peter *e* vinder
him thinks Peter wins
'he is the one of whom Peter believes that he is going to win'

(8) * *han* tror Peter *e* vinder
he thinks Peter wins
'he is the one of whom Peter believes that he is going to win'

In (7), the pronoun *ham*/'him' is the subject of the embedded verb *vinder*/'wins'. Example (8) is only possible on a reading with *han*/'he' as the matrix subject.[4]

In non-subject extraction configurations, the pronominal retains its case:

(9) *ham* tror jeg ikke jeg kender *e*
*him.*ACC think I not I know *e*
'As for him I don't think I know him'

The extraction of a pronominal subject is not accepted by all speakers. Speakers who reject (7) (i.e. the accusative marking of an extracted subject), also reject (8). For those speakers extraction of a pronominal subject is not possible at all. The accusative marking in (7) belongs to an informal register, but it is not restricted to spoken language. On the contrary, it is common on internet-pages from where the following examples are extracted.

(10) *Dem* håber jeg vil komme og besøge mig i det nordsjællandske.
them hope I will come and see me in North Zealand
'I hope they will come and see me in North Zealand'
(www.hjem.get2net.dk/vmf/side_7.htm)

(11) *dem* ved jeg er gode!
them know I are good
'I know they are good'
(home.worldonline.dk/∼ejstrups/kokkenhaven_juni.htm)

(12) *ham* ved jeg ikke lige hvem er
him know I not exactly who is
'I don't quite know who he is'
(strikeforce.boomtown.net/phpBB/viewtopic.php?topic=414&forum=2)

(13) *dem* ved jeg ikke hvordan smager
them know I not how taste
'I don't know how they taste'
(www.fyldepennen.dk/tekster/520)

3 Previous Approaches

3.1 Stylistic Variation: Hansen 1972

The use of the nominative and the accusative forms of the personal pronouns is subject to considerable variation. This is only to be expected given that case no longer serves to identify grammatical functions (a point made in [Nordlinger, 1998]). The most common contexts of variation are:

[4]The *e* indicates the intended reading of *han*/'he' as an extracted subject. It does not necessarily correspond to an empty category. Cf. the discussion in section 6.2.1.

- Coordinated structures

  (14) Peter og *jeg/mig* har været i vandet i dag
  Peter and I/me have been swimming today
  (example from [Hansen, 1972])

- Object of preposition

  (15) mange af de/dem udefra ved i virkeligheden bedre besked
  many of *they/them* from outside are better informed actually
  (example from [Hansen, 1972])

- Pronouns with restrictive modification

  (16) *de/dem* her ser da meget bedre ud
  they/them here look a lot better, don't they?
  (example from [Hansen, 1972])

To account for this variation, [Hansen, 1972] distinguishes two different registers associated with social connotations, register 1 being "standard" and register 2 being "informal/colloquial". These two registers employ different rules for the use of the pronominal forms. The main difference between the two registers is that pronouns in obligatorily stressed positions are in the accusative case in discourses in register 2, regardless of their grammatical function (i.e. also subjects as in (14) and (16) above).

Topicalization of constituents other than the matrix subject is always associated with stress, so the extraction configuration in (7) is in accordance with Hansen's generalization about discourses in register 2. However, there is a crucial differences between the extraction context in (7) (which is not discussed in Hansen) and the variation contexts in (14) through (16) above. The variation contexts permit a choice between the nominative and the accusative form, and the choice points to a certain register. In (7) there is no choice as to the form of the extracted pronominal. An extracted pronominal subject must be marked with the accusative form, or it can not be extracted at all.

Furthermore, there are exceptions to the generalization that the accusative form is used for obligatorily stressed pronouns. Pronominal subjects occurring with focus adverbials are obligatorily stressed, and yet only the nominative form is possible:

(17) kun/netop/også han/*ham kan klare det
only/exactly/also he/*him can do it

Hansen claims that this construction does not constitute a counter-example to the generalization about the use of the pronoun forms in register 2 on the grounds that pronouns with focus adverbials do not occur in register 2 discourses at all. For this reason no variation is observed in this construction.

The same kind of argument could be made about the extraction configuration: subject extraction does not occur in discourses in register 1, therefore no variation is observed. The picture is, however, more complicated. Subject extraction seems to occur in both registers, depending on the nature of the extracted element. Subject extraction involving lexical items without case marking, as in (18) and (19) below, is accepted by all speakers.

(18) *det* tror jeg ikke *e* passer
that think I not *e* is true
'I really don't think that is true'

(19) det er ham *som*/e jeg tror vinder
it is him *who.*REL I think wins
'he is the one who I think is going to win

In (18) the neuter subject pronoun *det*/'that' is extracted, and selectional restrictions prevent it from being interpreted as the matrix subject. (19) illustrates subject relativization with the conjunction *som*/'as' in C or an empty C position. In this case, the matrix subject is identified by means of its position.

Subject extraction as such does not seem to be restricted to different registers or discourses, even though there may be a difference in frequency. Rather the variation in acceptability between e.g. (18) and (7) stems from the case-marking of the extracted item. Some speakers do not allow the use of the accusative form to mark an extracted subject, and using the nominative form leads to a different interpretation than the intended one. Summing up,the use of the accusative form to mark an extracted subject can not be accounted for by appealing to different registers and an associated difference in the use of the pronominal forms in obligatorily stressed positions.

3.2 "Default Case": Schütze 2001

Schütze ([Schütze, 2001]) argues that the morphological case of the pronouns in English calls for a notion of default case. DPs are licensed by abstract Case (*structural licensing*) while morphological default case is assigned to DPs which fail to be assigned case otherwise. Languages vary as to which case counts as the default case. In English and Danish, accusative is the default case, as hinted at above. The most important environments where default case applies, are environments without a case assigner for the DP (as in (5) above), or where the pronoun does not occupy the head position D of the DP. On Schütze's analysis, default case is licensed on modified pronouns since modified pronouns do not occupy the head position D (example (16) above). Case spreading in coordinated structures is subject to parametric variation, and on the assumption that case does not spread in English and Danish coordinated structures, default case is licensed in these structures accounting for (14) above. Schütze's analysis essentially accounts for the variation contexts above, but his analysis does not extend to the accusative marking of a non-locally extracted subject. The extraction configuration in (7) is not a default-case environment on Schütze's account. Example (7) contains a case assigner for the accusative subject, i.e. the finite verb of the embedded clause, and the pronoun, being a simple bare pronoun, occupies the head position D. What gives this use of the accusative a flavour of being "default", is the fact that the accusative marks any grammatical function except for the subject, and the subjecthood of the preposed constituent in (7) is somewhat obscured by the fact that the pronoun receives its grammatical relation from an embedded predicate. This use of "default" is, however, not syntactic since there is no syntactic motivation for using a default case in (7). Rather the intuition behind the use of the accusative to mark a non-locally extracted subject, is that the constituent is not the subject of the closest finite verb. This is the intuition which is given a precise formulation in section 6.1.

3.3 Case Neutralization: Taraldsen 1981

[Taraldsen, 1981] develops an account of pronominal extraction in Norwegian within a transformational framework. In Norwegian, the extracted pronominal retains its case as shown in (20) below from [Taraldsen, 1981].

(20) Han hadde de trodd *e* ville komme forsent
He had they thought *e* would be late

Taraldsen, however, notes that only 3rd person pronouns can be extracted. 1st and 2nd person pronouns are barred from extraction: [5]

(21) + Jeg hadde de trodd *e* ville komme forsent
I had they thought *e* would be late

On Taraldsen's account, the extracted subject leaves a trace in its original position (within S, i.e. in the specifier of IP) and in its intermediate landing site (COMP in S', i.e. specifier of CP). Taraldsen gives the following structural representation:

(22) *han$_i$ hadde de trodd [$_{S'}$ t$_i$ [$_S$ t$_i$ ville komme forsent]]*

Taraldsen's analysis crucially relies on case theory. The trace of the topicalized subject is assigned nominative case by INFL in its canonical position, and the trace is assigned accusative case by the matrix verb *trodd*/'believed' in its landing site in S'. Conflicting case requirements are thus imposed on the preposed constituent in the matrix clause. The constituent has to spell out the feature matrix [+NOM,+ACC] which is only possible through a neutralized form. According to Taraldsen, the 3rd person pronouns are neutralized between nominative and accusative in modern Norwegian, while 1st and 2nd person pronouns are not. In this way, only the 3rd person pronouns can resolve the conflicting case requirements resulting from movement of the embedded subject, and the examples with 1st and 2nd person pronouns are ruled out.

Taraldsen's analysis is interesting in that it assumes the possibility of exceptional case marking from the matrix verb, thus accounting for the accusative form of the preposed constituent. But an account based on case theory faces several difficulties in Danish.

As will be shown in section 4, there is no evidence that the matrix verb in e.g. (7) assigns case to the subject of the embedded predicate.

More importantly, the extracted subject can not be associated with an empty category in the specifier position of the embedded CP since this position may be occupied by an overt *wh*-phrase. Cf. the examples in (12) and (13) repeated below for convenience.[6]

(23) *ham* ved jeg ikke lige *hvem* er
him know I not exactly *who* is
'I don't quite know who he is'
(strikeforce.boomtown.net/phpBB/ viewtopic.php?topic=414&forum=2)

(24) *dem* ved jeg ikke *hvordan* smager
them know I not *how* taste
'I don't know how they taste'
(www.fyldepennen.dk/tekster/520)

Finally, no neutralization is observed in the 3rd person pronouns in Danish (outside the specific variation contexts mentioned in section 3.1).

[5]Taraldsen uses the symbol '+' to indicate ungrammaticality.

[6]To account for crossover phenomena in German [Berman, 2000] also assumes that non-locally extracted constituents are associated with an empty category in the specifier of the embedded in CP. Again, this can not be the case in Danish, since the specifier may be occupied by an overt constituent.

(25) han/*ham vinder
he/*him wins

Interestingly, however, 1st and 2nd person pronominal subjects do not extract so easily in Danish either, though it does not seem to be impossible:

(26) mig forventer de laver det hele
me expect they do everything
'they expect me to do everything

This behaviour of the 1st and 2nd person pronouns can not be explained by appealing to neutralization, not even if the variation contexts in section 3.1 are considered reflections of neutralization. These patterns of variation pertain to 1st, 2nd and 3rd person pronouns alike and so do not point to an asymmetry in the behaviour of the pronouns under extraction.

I have no explanation for the apparent decrease in acceptability pertaining to extraction of 1st and 2nd person pronouns. It is, however, striking that this restriction pertains to the deictic pronouns as opposed to the anaphoric pronouns. Possibly this behaviour is related to discourse semantic factors, but this issue awaits further study.

To sum up the discussion of this section: Hansen's account of the variation in the use of nominative and accusative pronouns does not explain all the intricacies of the accusative marking of extracted subjects since this construction does not allow a choice between a nominative and an accusative form. Taraldsen case- and trace-based approach can not account for the Danish data as there can be no trace in the specifier of the embedded CP. The origin of the accusative form remains unexplained in the Danish data.

In the following section, I return to this question. Is the preposed constituent really an extracted constituent or does it receive its grammatical relation from the matrix verb as in raising constructions?

4 Subject Extraction or Object Raising

The phenomenon that a matrix constituent is coreferential with a subject of an embedded clause, is known from raising and equi contexts. In examples such as (7), the displaced constituent is not a semantic argument of the matrix verb, but we still need to consider the possibility that it is assigned its grammatical relation by the matrix verb as in raising-constructions. This would explain the presence of the accusative form, as hinted at in the analysis in [Taraldsen, 1981].

There are several differences between the extraction construction in (7) and raising constructions. First of all, embedded verbal predicates in raising constructions are generally infinite while the verbal predicate in (7) is finite.

Secondly, the extracted constituent can not occur in object position, neither within VP not in I, which is the position of unstressed pronominal objects ("object-shift") ([Sells, 2001]). In raising contexts, the raised constituent can occur in object position as shown below. Note that in (27b) the pronoun must be stressed.

(27) a. * Peter forventer [I' jo [VP ham [IP vinder]]]
Peter expects as you know him wins

b. Peter ser [I' jo [VP ham [IP flygte]]]
Peter sees as you know him escape

(28) a. * Peter forventer [I' [I ham] jo [VP [IP vinder]]]
Peter expects him as you know wins

b. Peter ser [I' [I ham] jo [VP [IP flygte]]]
Peter sees him as you know escape

The extracted constituent can only occur in the specifier position of the matrix CP:

(29) [CP ham [C forventer] Peter jo vinder]
him expects Peter as you know wins

Finally, the extracted constituent can not raise to subject under passivization while an accusative controller can:

(30) a. * han forventes vinder
he is_expected wins

b. han ses flygte
he is_seen escape

There is no evidence that the dislocated constituent receives its grammatical relation from the matrix verb or that it is is assigned case by Exceptional Case Marking. The use of the accusative form of the pronoun is constructionally determined and can not be the result of dependent-marking.

5 Two Hypotheses about the Use of the Nominative and Accusative Forms

Before stating the generalization about the use of the nominative and accusative forms of the personal pronouns, I am going to explore two alternative hypotheses. The first one is that the accusative forms of the personal pronouns serve to mark discourse functions (TOPic or FOCus). The second hypothesis is that case marking is tied to phrase structure positions, i.e. that case marking is a mere c-structure annotation. In the latter case, position in itself serves to identify grammatical functions and the use of morphological case thus constitutes a more or less redundant piece of information.

I will show that both of these hypotheses face serious theoretical and empirical difficulties, and that the use of the nominative and accusative forms serves a disambiguating function.

5.1 Accusative as a Marker of Discourse Functions

One of Hansen's ([Hansen, 1972]) insights about case marking in the colloquial register in Danish was that pronouns in obligatorily stressed positions are in the accusative case. Obligatory stress is an indication of discourse prominence. The use of the accusative form could thus in itself signal discourse prominence so that discourse functions are associated with accusative case (wherever this applies).

Further support for this hypothesis comes from the fact that pronouns left-adjoined to the matrix CP are generally in the accusative case, cf. (31) and (32) below.

(31) *Dig,* du kan gå din vej
*You.*ACC, you.NOM can go your own way

(32) *ham der*, han brokker sig altid
him there, he is always complaining
'that man, he is always complaining'

Example (32) furthermore illustrates the tendency to use the accusative form of a restrictively modified pronoun as in (33). This use is, however, subject to variation (cf. (16) above).

(33) *dem der kommer for sent*, skal sidde på første række
*those.*ACC *who are late*, have to sit in the first row

In LFG, this use could be accounted for by assuming an implicational statement to the effect that an f-structure is associated with accusative case if it is the value of a discourse function.

(DF↓) ⇒ (↓CASE)=ACC

This analysis faces a number of difficulties. First of all it seems counter-intuitive to use the most unmarked form to indicate discourse prominence. More importantly, the nominative form may also be associated with discourse prominence (signalled by (contrastive) stress) as in (34) below. The accusative form is excluded in this case.

(34) *HAN/*HAM* kommer i hvert fald ikke
*HE/*HIM* comes certainly not
'He certainly won't be here'

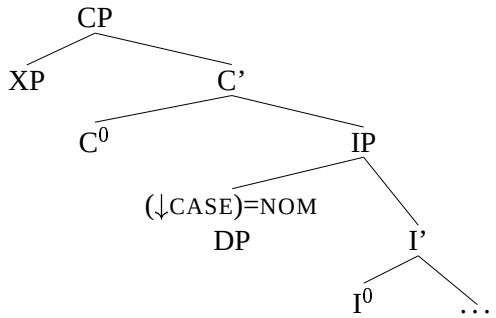
Finally, the nominative form may be associated with a focus adverbial as in (35).

(35) *kun hun* kan gøre det
Only she can do it
'she alone can do it'

Accusative as a marker of discourse functions only applies to constituents which are not simultaneously matrix subjects. This in itself suggests that this is not the right generalization.

5.2 Case-marking as a C-structure Annotation

As noted, case no longer serves to identify grammatical functions in Danish. The canonical position of the subject is the specifier position of IP while extracted constituents are in the specifier of CP. If the specifier position of IP is associated with nominative case, subjects are only associated with nominative in their canonical position. The relevant piece of c-structure would look as depicted below.



One disadvantage of this approach is that all DPs in the specifier position of IP are associated with nominative case in the f-structure. Since common nouns are unspecified for case in the lexicon, the association of nominative case with the specifier of IP does not prevent common nouns from occurring in this position. From a monolingual point of view, however, there is no empirical motivation for postulating nominative or accusative case for common nouns in Danish. What we need to ensure, is that common nouns may appear in the specifier of IP while only nominative personal pronouns may appear in this position. This can be accomplished by associating the following disjunction with the specifier of IP:[7]

$$\neg(\downarrow \text{CASE}) \vee (\downarrow \text{CASE})=_c \text{NOM}$$

This disjunction licenses a DP in the specifier of IP if the DP bears no case specification, or if it is associated with nominative case. Lexical entries for common nouns contain no case information while the lexical entry for the pronoun *han*/'he' specifies that it is nominative:

han D (↑CASE)=NOM
⋮

In this way, all common nouns but only nominative personal pronouns are licensed in subject position. Accordingly, other nominal c-structure positions must be constrained to require accusative case or no case specification.

This proposal crucially hinges on the split IP/CP-hypothesis, i.e. on the hypothesis that the subject always occurs in the specifier of IP, also in non-inverted declarative main clauses. This is the stance adopted in [Sells, 2001], while others argue that all V2-clauses are CPs (e.g.[Vikner and Schwartz, 1996]). On either account, a unique position for the subject can not be identified. On Sells' analysis constituents bearing a grammaticalized discourse function appear in the specifier of CP. Subjects in non-inverted main clauses may thus appear in the specifier of CP if they are associated with a discourse function (indicated with prosodic prominence, e.g. contrastive stress) (p. 35). In these cases, however, the subject retains its nominative case, as we have already seen:

(36) *HAN/*HAM* kommer i hvert fald ikke
*HE/*HIM* comes certainly not
HE certainly won't be here'

In light of examples such as (36), it is not possible to maintain that nominative pronouns are always in the specifier of IP, and that constituents associated with prosodic emphasis are always in the specifier of CP. The pronoun in (36) would have to be in two different positions at the same time. On Sells' account, the matrix subject may be in the specifier of either IP or CP (according to its status as a discourse function). On the account in [Vikner and Schwartz, 1996], the subject is always in the specifier of CP alongside all other kinds of topicalized constituents. On either account there is no unique position for the subject, and the hypothesis that case marking is an annotation associated with a specific c-structure position can not be maintained.

[7] This approach has been suggested to me by Mary Dalrymple (p.c.).

6 Accounting for the Distribution of the Nominative and Accusative Forms

6.1 Generalization about the Use of the Nominative and Accusative Forms

The discussion so far has established that the use of the nominative and accusative forms is not associated with discourse prominence. Nor is it a mere c-structure annotation.

The generalization that emerges from the data can be stated as follows (ignoring the variation contexts mentioned in section 3.1):

Nominative The DP is the subject of the immediately containing f-structure

Accusative The DP is *not* the subject of the immediately containing f-structure (but possibly the subject of an embedded f-structure)

The immediately containing f-structure, in turn, is defined as below.

Immediately containing f-structure The immediately containing f-structure is the f-structure associated with the closest functional projection (either CP or IP) dominating the DP

From a processing perspective, an initial accusative pronouns signals that the pronominal DP is not the subject of the matrix clause, and that it must be related to its canonical position later in the clause, either as an argument of the matrix predicate or as an argument of an embedded predicate. The nominative, in turn, signals the the DP is the subject of the matrix clause. In this way the use of the nominative and the accusative pronouns serves a disambiguating function. Note that case marking here is the only means to disambiguate the sentence. In main clauses there is no unique position for the subject, as detailed above. Furthermore, selectional restrictions do not suffice to disambiguate in most of these cases. "Bridge"-verbs are generally associated with human subjects, but an extracted personal pronoun evidently is also human apart from the 3rd person plural pronoun which may also (anaphorically) refer to non-humans. In far the most cases, the involved constituents consequently share the same relevant semantic features (e.g. [+HUMAN].). Nothing indicates that the default-association of TOP(ic) and matrix subject is overridden, except for the case marking. The accusative form prevents the topicalized constituent from being interpreted as the matrix subject.[8]

It follows from the generalization above that the case form is not determined by a local head-dependent relation alone, but also by the construction itself. The challenge is to account for constructionally determined case assignment, while restricting case distinctions to a specified subset of the nominals as detailed above. In the next section I address the question of how to identify a non-locally extracted subject. Two approaches present themselves: case assignment mediated through an empty category and *Constructive Case* ([Nordlinger, 1998]).

I conclude that *Constructive Case* is to be preferred on empirical as well as on theoretical grounds.

[8] A corollary of this analysis is that extraction involving two proper nouns can not be so disambiguated. Selectional restrictions are of no use since both nominals are human, and no case marking is available:

(1) Peter tror Louise kommer
Peter believes Louise is coming

(1) is (potentially) ambiguous between a reading where *Peter* is the matrix subject and the subject of the embedded predicate *kommer*/'is coming'. Out of context, only a reading where *Peter* is the matrix subject and the topic of the clause seems to be available. It does, however, remain to be investigated whether intonation and/or contextual information suffice to bring out the other reading in examples such as (1). Subject extraction involving nominals with no case marking and common semantic features, is strikingly rare. This seems to suggest that such constructions are generally avoided.

6.2 Case Assignment and Empty Categories

The discussion so far has revealed that the choice between the nominative or the accusative form of the pronoun is sensitive to the status of the subject as a local or a non-local subject. It has also been established that a non-locally extracted subject can not be identified in terms of phrase structure position since the specifier of CP is also the position of topicalized local subjects. Another way to identify extracted elements in terms of c-structure properties, is through their association with an empty category.

According to [Bresnan, 2001], non-local extraction always involves an empty category as a last resort to identify grammatical functions (p. 202). Languages vary, however, as to whether local extraction involves empty categories. In configurational languages such as English and Danish, local extraction of non-subjects involves empty categories while local extraction of subjects does not since the SUBJ(ect) is associated with the TOP(ic) by default. On this account, an extracted pronoun is associated with an empty category in exactly the cases where it appears in accusative case (i.e. extracted non-subjects and non-locally extracted subjects). Consequently we can impose the requirement on an empty category that it be identified with a discourse function bearing accusative case as shown below.

$$e$$
$$((\mathrm{x}\uparrow)\ \mathrm{DF})=\uparrow$$
$$((\mathrm{x}\uparrow)\ \mathrm{DF\ CASE})=\mathrm{acc}$$

This account, however, crucially hinges on the assumption that non-locally extracted subjects are associated with an empty category. To what extent is this assumption empirically justified?

6.2.1 Subject/Object-Asymmetry in crossover

In [Bresnan, 2001], binding in crossover is used as a diagnostic for detecting empty categories. Binding is associated with linear precedence in many languages, and linear precedence pertains to c-structure as do empty categories.

As the examples in (37a) through (38b) illustrate, subjects and objects behave differently in their binding potential in crossover in Danish.

(37) a. *Peter*$_i$ tror *han*$_{i/j}$ vinder
Peter thinks *he* wins

b. *ham*$_{*i/j}$ tror *Peter*$_i$ vinder
him thinks *Peter* wins

(38) a. *Peter*$_i$ tror ikke de får fat på *ham*$_{i/j}$
Peter thinks not they get hold of *him*
'Peter doesn't think they will get hold of him

b. *ham*$_{\%i/j}$ tror *Peter*$_i$ ikke de får fat på
him thinks Peter not they get hold of
'Peter doesn't think they will get hold of him'

In (37a), the matrix subject *Peter* may be coreferential with the subject pronoun *han*/'he' in the embedded clause. If the pronoun is extracted, as in (37b), coreference is excluded. Note further that the very same pattern is observed with two pronouns as in (39a) and (39b) below. The ungrammaticality of the co-referential reading of (37b), thus, does not reduce to a principle C violation:

(39) a. *han*$_i$ tror *han*$_{i/j}$ vinder
he thinks *he* wins

b. *ham*$_{*i/j}$ tror *han*$_i$ vinder
him thinks *he* wins

In (38a), the matrix subject *Peter* can be coreferential with the object pronoun in the embedded clause. If the object is extracted, as in (38b), coreference is still possible for most speakers.[9]

These data suggest that subjects and objects behave differently under non-local extraction, viz-à-viz binding. It is, however, a matter of discussion whether the binding patterns above are subject to syntactic constraints. [Bresnan, 2001] argues that coreference relationships of the kinds illustrated above, are not governed by syntactic principles, but rather by semantic or discourse semantic factors. In detecting empty categories, Bresnan instead relies on semantic binding, i.e. binding of pronouns by semantic operators such as quantifiers. She argues that semantic binding is governed by syntactic principles, and that binding relationships involving quantifiers can be used to detect empty categories.

In operator binding the very same pattern as above is observed. An extracted object may be bound by a matrix subject operator, but an extracted subject may not.

(40) a. * *dem selv*$_i$ tror *ingen*$_i$ bliver opdaget af skattevæsenet
themselves thinks *noone* will_be discovered by the tax office
'noone thinks they will be discovered by the tax office'

b. % *dem selv*$_i$ tror *ingen*$_i$ skattevæsenet får fat på
themselves thinks *noone* the tax office gets hold of
'noone thinks the tax office will get hold of them'

In Danish as in English, an operator must linearly precede and syntactically outrank the bindee. [Bresnan, 2001] gives the following definitions:

- The domain of a binder excludes any pronominal that f-precedes it
- The domain of a binder excludes any pronominal contained in a constituent that outranks the binder

F-precedence in turn is defined as below:

(41) F-precedence

f f-precedes *g* if the rightmost node in $\phi^{-1}(f)$ precedes the rightmost node in $\phi^{-1}(g)$.

The condition on operator binding are given in (42).

(42) **Condition on operator binding**

An operator must outrank and f-precede the bindee.

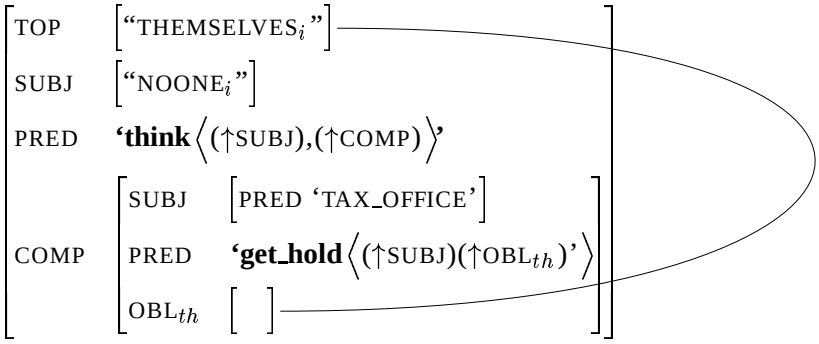
[9] As indicated, not all speakers accept the coreferential reading in (38b). It has been suggested to me by Sten Vikner (p.c.) that this variation may be associated with whether one allows preposing of reflexive anaphors as in (1).

(1) % sig selv elsker Peter mest af alle
himself loves Peter most of all

It seems to be the case that speakers who accept (1) also accept (38b). This issue awaits further study.

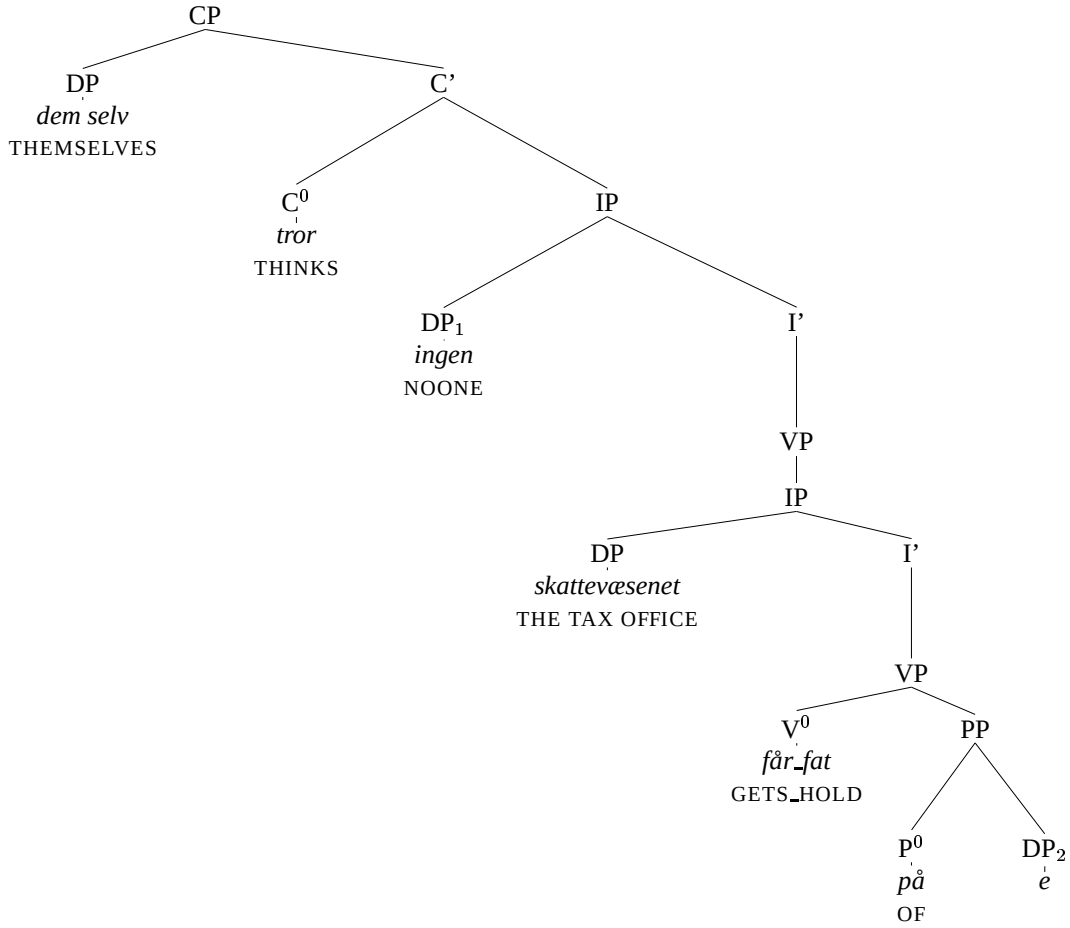
Lets us consider the binding conditions for each of the examples in (40a) and (40b) in turn, beginning with the case of object extraction.

The f-structure for (40b) is given below:



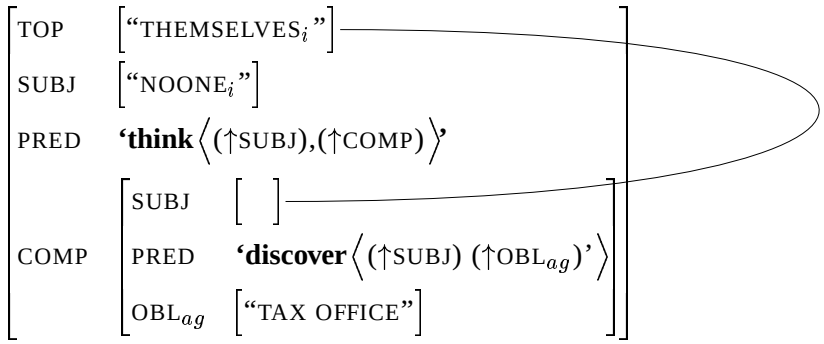
Since the SUBJ of the matrix clause, *ingen*/'noone', outranks the COMP containing the OBL$_{th}$, *dem selv*/'themselves', the requirement on syntactic rank is fulfilled.

Consider next the c-structure for (40b):



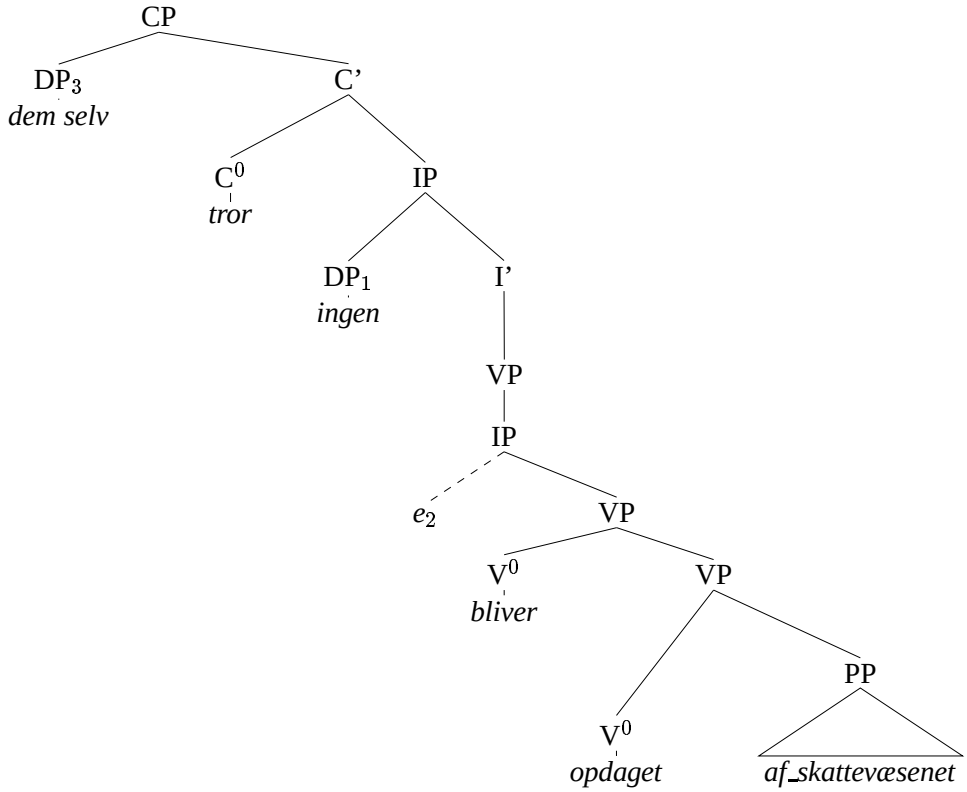
The right edge of the operator (the binder) is DP_1 in the figure above. The right edge of the bindee is DP_2, i.e. the DP dominating the empty category. The right edge of the binder linearly precedes the right edge of the bindee which means that the operator f-precedes the bindee. Since both requirements in (42) are met, the binding pattern is predicted to be acceptable.

Consider next the f-structure for (40a):



The SUBJ of the matrix clause outranks the COMP containing the SUBJ of the extracted constituent. The outranking condition is fulfilled.

The c-structure is given below.



If the extracted subject is associated with an empty category as suggested in [Bresnan, 2001], the right edge of the extracted constituent is the node (e_2), and DP_1 of the operator f-precedes the bindee. Since both requirements in (42) are met, binding is erroneously predicted to be acceptable. If, however, the extracted subject is not associated with an empty category, the right-edge of the pronoun is DP_3 and so the operator no longer f-precedes the bindee. This means that the second condition on operator binding is not met, and the example is correctly ruled out.

What these data suggest is that subject extraction is not associated with an empty category (neither locally nor non-locally), while object-extraction is associated with an empty category (locally and non-locally). These data thus lends support to the analysis of extraction in [Falk, 2001]. [Falk, 2001]

claims that subject extraction is not associated with an empty category and suggests that "bridge"-verbs are associated with an optional functional uncertainty equation identifying the discourse function of the clause with the subject function of a COMP.

6.2.2 A Comparison with a Traceless Account

[Dalrymple et al., 2001] develops an account to binding in weak cross-over which does not rely on the presence of empty categories.

On their account, f-precedence is defined as follows:

F-precedence f_1 f-precedes f_2 if and only if all c-structure nodes corresponding to f_1 precede all nodes corresponding to f_2.

In determining syntactic prominence and linear precedence, Dalrymple et al. invoke the notion of co-argumenthood: i.e. arguments of the same predicate. Syntactic prominence must hold between co-arguments, i.e. the co-arguments containing the operator and the pronominal, respectively. Linear precedence must hold between the co-argument of the operator and the pronominal itself. The relevant definitions are given below.

Let CoargOp and CoargPro be coargument f-structures such that CoargOp contains O and CoargPro contains P. Then:

Syntactic Prominence An operator O is more prominent than a pronoun P if and only if CoargOp is at least as high as CoargPro on the functional hierarchy

Linear Prominence An operator O is more prominent than a pronoun P if and only if CoargOp f-precedes P

In both example (40a) and (40b), the *CoargOp* is the matrix subject, *ingen*/'noone', and the *CoargPro* is the COMP of the matrix verb *tror*/'believes'. Since the SUBJ(ect) is more prominent than the COMP on the relational hierarchy, the syntactic prominence condition is fulfilled. The condition on linear prominence must hold between the the c-structure nodes corresponding to the *CoargOp* and the c-structure nodes corresponding to the pronoun P. In both (40a) and (40b), the pronoun linearly precedes the *CoargOp*. Since the condition on linear precedence is not met, the traceless account of Dalrymple et al. rules out both (40a) and (40b), contrary to fact.

6.2.3 Conclusion on Case Marking and Empty Categories

The lesson to learn from the lengthy discussion on empty categories is that an extracted subject can not be identified through the presence of an empty category since a non-locally extracted subject is not associated with an empty category. As a consequence, accusative marking of extracted constituents can not be enforced by means of an empty category. Since an extracted subject has been shown not to be associated with a unique c-structure position, it seems that an extracted subject can not be identified in c-structural terms at all.

6.3 *Constructive Case*

[Nordlinger, 1998] develops an account to morphological function specification where the case bearing constituent projects its f-structure environment directly. The formal means is that of an inside-out function equation, i.e. the case bearing element is associated with an equation stating that the f-structure of the constituent must be the value of a certain GF. On this approach the German singular, accusative, masculine determiner *den*/'the' is associated with the following (simplified) lexical entry:

den DET (OBJ↑)

This equation states that the f-structure of the determiner (and of its co-head nominal) must be the value of the OBJ attribute.

Two features of the *Constructive Case*-approach makes it particularly advantageous for the present analysis.

Constructive Case allows the morphological function specification to be lexically associated with the relevant lexical items. In the grammar of Danish, only the personal pronouns are associated with case-related constraints on their syntactic environment. No such constraints are associated with common nouns whose distribution is independent of their grammatical function.

Secondly, inside-out functional equations allow the use of functional uncertainty. It is possible to specify paths of variable length through the f-structure. Functional uncertainty thus allows for the implementation of constructionally dependent morphological function specification while maintaining a lexical association.

The most straight-forward lexical representation of the personal pronouns would be to associate the equation below with the accusative form of the personal pronouns stating the the pronoun is not a subject.

¬(SUBJ↑)

This approach, however, will not yield the desired result. An inside-out functional uncertainty may denote several f-structures, even with a fixed path ([Dalrymple, 2001]). For the example in (7), the equation above denotes the f-structure of the entire clause and the f-structure of the COMP. But the equation above states that there can be no f-structure in which the f-structure of an accusative pronoun is the value of the SUBJ attribute. Since the accusative pronouns is the SUBJ of the f-structure associated with the COMP, the equation above erroneously rules out example (7).

Instead we need a disjunctive statement saying that either the accusative pronoun is not a subject, or it is an extracted subject. In f-structural terms, a subject is (non-locally) extracted if the f-structure associated with the pronoun is structure-shared with the discourse function of an f-structure denoted by an inside-out functional equation starting in the f-structure of the SUBJ attribute and leading through at least one occurrence of the COMP attribute.[10] The pronoun *ham*/'him' is associated with the following lexical entry:

[10]Actually this account is somewhat simplified since the subject may also be extracted from the complement clause of a preposition in Danish. To enhance readability I have left out this part of the equation in the lexical entry above. The fully expanded equation takes the following format:

(1) (({OBJ|COMP}$^{+}$ SUBJ ↑) DF)=↑

I follow [Dalrymple and Lødrup, 2000] in assuming that the complement clause of a preposition bears the function OBJ(ect).

ham: N $(\neg(\text{SUBJ} \uparrow) \vee$
$((\text{COMP}^{+}\ \text{SUBJ} \uparrow)\ \text{DF}) = \uparrow)$
$(\uparrow \text{PRED}) = \text{‘PRO’}$
$(\uparrow \text{PERS}) = \text{3RD}$
$(\uparrow \text{SEX}) = \text{MASC}$
$(\uparrow \text{NB}) = \text{SING}$

The equations state that either the f-structure of the pronoun is not the value of a SUBJ attribute, or it is structure-shared with the discourse function of an f-structure in which the pronoun is the subject of a COMP attribute (which in turn may be contained in a COMP itself).

As far as the nominative pronoun *han*/‘he’ is concerned, we need an equation to the effect that the pronoun must be a subject and that it can not be an extracted subject. Thus we need two conjunctively specified equations (conjunction is implicit in lexical entries). The lexical entry is given below:

han: N $(\text{SUBJ} \uparrow)$
$\neg(((\text{COMP}^{+}\ \text{SUBJ}\uparrow)\ \text{DF}) = \uparrow)$
$(\uparrow \text{PRED}) = \text{‘PRO’}$
$(\uparrow \text{PERS}) = \text{3RD}$
$(\uparrow \text{SEX}) = \text{MASC}$
$(\uparrow \text{NB}) = \text{SING}$

The equations state that the f-structure of the nominative pronoun is the value of a SUBJ attribute and that it can not be an extracted subject, i.e. it can not structure-shared with the discourse function of an f-structure in which the pronoun is the subject of a COMP attribute. These lexical entries capture the generalization in section 6.1 and account for the data. Only the accusative pronoun can occur as a non-locally extracted subject, and the nominative pronoun can not be non-locally extracted at all.

Let us return briefly to the pattern of variation mentioned in the beginning. As noted, not all speakers accept extraction of a pronominal subject. This fact can easily be accommodated in the present analysis. These speakers simply lack the second half of the disjunctive functional uncertainty equation in the lexical entry for *ham*/’him’: $((\text{COMP}^{+}\ \text{SUBJ} \uparrow)\ \text{DF}) = \uparrow)$. In the absence of this second half of the equation, the entry states that an accusative pronoun can never be a subject. Since a nominative pronoun can not be an extracted subject (as shown in the lexical entry above), non-local extraction of a personal pronoun is blocked altogether. A grammatical-gender specific pronominal subject such as *det*/’that’ may still be extracted since the pronoun is associated with no constraints on its syntactic function. A possible stylistic restriction on the extraction of pronominal subjects (as hinted at in the discussion in (3.1)) is thus tied to the shape of the lexical entry for the accusative pronouns.

7 Conclusion

The present analysis has shown that the use of the nominative and accusative forms of the personal pronouns in Danish is constructionally determined (in addition to being determined by the grammatical function of the pronouns). The use of accusative to mark a non-locally extracted subject has a clearly disambiguating function since neither phrase structure position nor selectional restrictions serve to identify the matrix subject in these cases. The construction-specific constraints on the distribution of the personal pronouns can not be identified in terms of c-structure (neither through position in the phrase structure nor through association with an empty category). The constraints, however, can

be stated in terms of f-structure properties, represented in the lexical entries of the pronouns by means of inside-out functional uncertainty equations. This approach has the advantage that it associates morphological function specification directly with the relevant lexical items without postulating case marking in other parts of the grammar where it is not empirically justified. Furthermore it allows for a straight-forward representation of (stylistic) variation in this particular construction and it accounts for the observed subject/object-asymmetry of operator binding in crossover.

References

[Berman, 2000] Berman, J. (2000). *Topics in the Clausal Syntax of German*. PhD thesis, Universität Stuttgart.

[Bresnan, 2001] Bresnan, J. (2001). *Lexical-Functional Syntax*. Basil Blackwell.

[Dalrymple, 2001] Dalrymple, M. (2001). *Lexical-Functional Grammar*, volume 34 of *Syntax & Semantics*. Academic Press.

[Dalrymple et al., 2001] Dalrymple, M., Kaplan, R. M., and King, T. H. (2001). Weak Crossover and the Absence of Traces. In Butt, M. and King, T. H., editors, *Proceedings of the LFG01 Conference*. CSLI Publications.

[Dalrymple and Lødrup, 2000] Dalrymple, M. and Lødrup, H. (2000). The Grammatical Function of Complement Clauses. In Butt, M. and King, T. H., editors, *Proceedings of the LFG2000 Conference*. CSLI Publications.

[Falk, 2001] Falk, Y. (2001). *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Number 126 in CSLI Lecture Notes. CSLI Publications.

[Hansen, 1972] Hansen, E. (1972). Dr. Jekyll og Mr. Hyde i Dansk Grammatik. University of Copenhagen.

[Hudson, 1990] Hudson, R. (1990). *English Word Grammar*. Basil Blackwell.

[Lee, 2002a] Lee, H. (2002a). Crosslinguistic Variation in Argument Expression and Intralinguistic Freezing Effects. In Ionin, T., Ko, H., and Nevins, A., editors, *MIT Working Papers in Linguistics*, volume 43, pages 103–122. Cambridge, MIT Linguistics department.

[Lee, 2002b] Lee, H. (2002b). A Lexicalist OT Approach to Case. University of North Carolina at Chapel Hill.

[Nordlinger, 1998] Nordlinger, R. (1998). *Constructive Case: Evidence from Australian Languages*. Stanford, CA: CSLI Publications.

[Schütze, 2001] Schütze, C. T. (2001). On the Nature of Default Case. *Syntax*, 4(3):205–238.

[Sells, 2001] Sells, P. (2001). *Structure, Alignment and Optimality in Swedish*. Stanford Monographs in Linguistics. CSLI Publications.

[Taraldsen, 1981] Taraldsen, K. T. (1981). Case-conflict in Norwegian topicalization. In Burke, V. and Pustejovsky, J., editors, *NELS 11: Proceedings of the Eleventh Annual Meeting of the North Eastern Linguistics Society*, pages 377–398. Graduate Students Linguistics Association of the University of Massachussetts.

[Vikner and Schwartz, 1996] Vikner, S. and Schwartz, B. (1996). The Verb Always Leaves IP in V2 clauses. In Belletti, A. and Rizzi, L., editors, *Parameters and Functional Heads. Essays in Comparative Syntax*, pages 11–61. Oxford: Oxford University Press.

[Woolford, 2001] Woolford, E. (2001). Case Patterns. In Legendre, G., Grimshaw, J., and Vikner, S., editors, *Optimality-Theoretic Syntax*. Cambridge, Mass.: MIT Press.

Object Asymmetries in Korean [1]

Hyun-Ju Park
National University of Singapore
artp9338@nus.edu.sg

Proceedings of the LFG 02 Conference
National Technical University of Athens, Athens
Miriam Butt and Tracy Holloway King (Editors)
2002
CSLI Publications
http://www-csli.stanford.edu/publications/

The aim of this paper is to suggest that Korean exhibits object asymmetry by showing that there is only one argument that demonstrates the "primary object" syntactic properties of case marking, reciprocalisation and the passive, despite clauses with more than one argument bearing accusative marking, such as in the morphological causative construction.

1. Introduction

The aim of this paper[2] is to look into double accusative-marked objects in a single clause, and to argue that they are asymmetrical with respect to case marking, reciprocalisation and the passive in Korean. Let us begin with example (1)[3]:

(1) a. * nay-ka ***ku*** ***yeca-lul*** ***simpwulum-ul*** sikhi-ess-ta.
I-N the woman-A errand-A make-PA-DEC
'I made the woman do some errand.' (Yang 1998: 247)

[1] This paper has also appeared in the proceedings of the LSK International Conference held in Seoul, Korea, in August 2002, and has been revised. I thank the audience at the LSK conference for their feedback on the earlier version.

[2] I would like to thank Arto Anttila, Vivienne Fong, K.P. Mohanan, and my supervisor Tara Mohanan for their valuable comments and suggestions on an earlier version of this paper. All the shortcomings or mistakes are of course my own.

[3] Abbreviation: N: Nominative D: Dative A: Accusative I: Instrument PA: Past
DEC: Declarative CAUS: Causative morpheme KEY: -key complementizer
E: suffix –e NI: Nominalizer

b. Mary-ka ***ttal-lul*** ***sakwa-lul*** mek-I-ess-ta.
Mary-N daughter-A apple-A eat-CAUS-PA-DEC
'Mary fed (her) daughter an apple.'

Given (1), in which the objects of the simple triadic predicate *sikhi-* 'make some do', the causative verb formed with *mek-* 'eat', and the causative morpheme *-I-* bear double accusative marking, the following questions are raised: (i) What contexts does double accusative marking allow? (ii) Is double accusative marking in Korean morphological causatives analogous to the o-marked morphological causatives in Japanese (Mohanan 1988, Matsumoto 1992, and Manning, Sag and Iida 1996) involving clause-embeddedness? and (iii) Does the double accusative marking in (1) suggest that Korean, in which a (direct) object is marked with the accusative marker *–lul/ul*, is typologically a symmetrical object language like Kichaga, discussed in Bresnan and Moshi (1990) and Alsina (1996)? Based on these questions, let us consider object asymmetries in Korean in a clause with a non-derived predicate.

2. Object Asymmetries in Korean

2.1. Case Marking

The accusative marker in Korean marks not only a direct object but also a non-direct object. This is shown by example (2):

(2) a. nae-ka ***path-ey***[4] mwul-ul cwu-ess-ta.
I-N field-D water-A give-PA-DEC
'I watered the field.'

b. nae-ka ***path-ul*** mwul-ul cwu-ess-ta.
I-N field-A water-A give-PA-DEC
'I watered the field.' (Yang 1998: 245)

As suggested by Yang (1998: 238), examples (2a) and (2b) differ in meaning. In (2a), the argument *path* 'filed' with the dative marker is read as either being totally or partially watered, while in (2b) the

[4] The inanimate dative marker

accusative bearing argument *path* is watered as a whole. Such contrasting meaning may be illustrated by (3):

(3) a. nae-ka ***path-ey*** mwul-ul ***ilpwu*** cwu-ess-ta.
I-N field-D water-A in part give-PA-DEC
'I partially watered the field.'

b. * nae-ka ***path-ul*** mwul-ul ***ilpwu*** cwu-ess-ta.
I-N field-A water-A in part give-PA-DEC
INT: I partially watered the field.

(ibid.)

That the accusative-marked argument in (3b) is semantically incompatible with the adverb that denotes partiality, as argued in Yang, indicates that the accusative marker involves a semantic notion of 'total affectedness'. Given (2) and (3), we can say that the accusative maker in Korean does not necessarily associate with the grammatical function of direct object. However, a direct object in Korean is by default marked with the accusative marker *–ul/lul*. Examples are given in (4) and (5):

(4) a. Bill-i ***John-ul*** ttayli-ess-ta.
Bill-N John-A hit-PA-DEC
'Bill hit John.'

b. * Bill-i ***John-eykey*** ttayli-ess-ta.
Bill-N John-D hit-PA-DEC
'Bill hit John.'

(5) a. * nae-ka path-ey ***mwul-ey/lo*** cwu-ess-ta.
I-N field-D water-D/I give-PA-DEC

b. * nae-ka path-ul ***mwul-ey/lo*** cwu-ess-ta.
I-N field-A water-D/A give-PA-DEC

In (4a), the object of the simple dyadic predicate *ttayli-* 'hit' *John* bears accusative marking. (4b) shows us that the object argument cannot be marked with any case marker other than the accusative. Given (4), we can account for the ungrammaticality of example (5) because the primary object bears non-accusative marking. This means that a primary object cannot be marked with any case marker other than the accusative. With respect to case marking, the double accusative-bearing objects in (5) are asymmetrical in that the primary object necessarily bears accusative marking.

2.2. Reciprocalisation

In Korean, when the phrase *kak* 'each' occurs with the phrase *selo* 'each other', the former c-commands the latter, as in English. An example of Korean reciprocalisation is given below:

(6) a. nay-ka ***kak*** namca-eykey ***selo-uy*** kanpang-ul cwu-ess-ta.
I-N each man-D each other-G bag-A give-PA-DEC
'I gave each man each other's bag.'

b.* nay-ka ***selo-uy*** namca-eykey ***kak*** kanpang-ul cwu-ess-ta.
I-N each other-G man-D each bag-A give-PA-DEC
'I gave each other's man each bag.'

As demonstrated by (6), the indirect object is asymmetrical to the direct object in that it cannot be reciprocalised. The asymmetry between the indirect and direct object is invariant even if the indirect object is in the accusative case, as shown in (7):

(7) a. nay-ka ***kak*** ***namca-lul*** ***selo-uy*** kanpang-ul cwu-ess-ta.
I-N each man-A each other-G bag-A give-PA-DEC
'I gave each man each other's bag.'

b.* nay-ka ***selo-uy*** ***namca-lul*** ***kak*** kanpang-ul cwu-ess-ta.
I-N each other-G man-A each bag-A give-PA-DEC
INT: I gave each man each other's bag.

Reciprocalisation therefore suggests that there is one direct object among the accusative-marked arguments in Korean.

2.3. The Passive

The asymmetry between the objects is also shown by passivisation. Example of this are given in (8):

(8) a. nae-ka ***path-ul*** ***mwul-ul*** cwu-ess-ta.
I-N field-A water-A give-PA-DEC
'I watered the field.'

b. ***mwul-i*** (na-eyuhaye) ***path-ey*** cwu-e ci-ess-ta.
water-N I-by field-D give-*E* become-PA-DEC
LIT: Water was given to the field (by me).

c. * *mwul-i* (na-eyuhaye) ***path-ul*** cwu-e ci-ess-ta.
water-N I-by field-A give-*E* become-PA-DEC
INT: Water was given to the field (by me).

d. * ***path-i*** (na-eyuhaye) mwul-ul/i cwu-e ci-ess-ta.
field-N I-by water-A/N give-*E* become-PA-DEC
INT: The field was given water (by me).

In a double accusative-marked object construction, the second argument bearing accusative marking, as in (8a), becomes the passive subject, as shown in (8b), while the first argument bearing the accusative marker cannot be the passive subject, as in (8d). The double accusative marked objects are asymmetrical with respect to the passive such that only a primary object can be the passive subject. Note that, as in (8c), no accusative marking is allowed in the passive in Korean, which indicates that there is no primary object. To summarise, case marking, reciprocalisation and the passive show us that double accusative-marked objects are asymmetrical.

3. Morphological Causatives in Korean

3.1. Morphological Causatives

Now let us consider double accusative marking in Korean morphological causatives, which I refer to as MC from now on. A verb of which the logical subject is affected can be morphologically causativised in Korean, such as *mek-* 'eat', *ip-* 'wear', *ilk-* 'read', *nok-* 'melt', and so on. An example of MC is given in (9):[5]

(9) a. ttal-i sakwa-lul ***mek***-ess-ta.
Daughter-N apple-A eat-PA-DEC
'The daughter ate an apple.'

b. Mary-ka ***ttal-lul*** ***sakwa-lul*** mek-I-ess-ta.
Mary-N daughter-A apple-A eat-CAUS-PA-DEC
'Mary fed (her) daughter an apple.'

A simple dyadic predicate as in (9a) is causativised, as shown in (9b). Since the causative morpheme introduces an additional argument, the causer *Mary*, the causative verb in (9b) has three arguments. The causee can bear accusative marking. Firstly, I show that double accusative marking in MC does not involve clause-embeddedness, unlike the Japanese *–(s)as(e)* causatives (Mohanan 1988, Matsumoto 1992, and Manning, Sag and Iida 1996), providing evidence from subject honorification, the distribution of negative polarity items, the clause-bound reflexive, and control in a participle clause.

3.2. Monoclausality of MC

3.2.1. Subject Honorification

[5] In previous studies, it has been claimed that the morphological causative in Korean is idiosyncratic, given that it is irregular and unproductive (O'Grady 1991, and Y-M Park 1991, among many others), however I argue in my Ph.D. thesis (to be completed in 2003) that the base verbs of which the logical argument is affected allow the causative morpheme.

To show the monoclausality of MC, I place MC as an object complement of the dyadic predicate *cwuliki-* 'enjoy', and the clause boundary is indicated by square brackets, as in (10):

(10) a. Halapeci-kkeyse **[**sonca-lul sakwa-lul ***mek-I-si-ki*** **]**-lul
grandfather-H **[**grandson-A apple-A eat-CAUS-SH-NI**]**-A

cwulki-***si***-ess-ta.
enjoy-SH-PA-DEC
'Grandfather enjoyed feeding (his) grandson an apple.'

b. * Halapeci-kkeyse **[**sonca-lul sakwa-lul ***mek-I-ki*** **]**-lul
grandfather-H **[**grandson-A apple-A eat-CAUS-NI **]**-A

cwulki-***si***-ess-ta.
enjoy-SH-PA-DEC

In (10a), the causative verb as a whole has the subject honorific marker *–si-*. In (10b), the causative verb does not bear subject honorification morphology, and it is ungrammatical. The contrast between (10a) and (10b) in subject honorific marking can be accounted for if we assume that the grammatical subject of the causative verb is the *halapeci* 'grandfather' in the embedded clause, indicated by the square brackets. The fact that there is only one subject in MC suggests that it consists of a single clause, given that subject honorification is a clause-bound agreement between a subject and its verb in Korean.

3.2.2. Distribution of Negative Polarity Item

The distribution of negative polarity items also suggests MC has a single syntactic clause, as shown in (11):

(11) a. Mary-***pakkey*** ttal-lul sakwa-ul ***an*** mek-i-ess-ta.
Mary-except daughter-A apple-A NOT eat-CAUS-PA-DEC
'Only Mary fed (her) daughter an apple.'

b. Mary-ka ***ttal-pakkey*** sakwa-ul ***an*** mek-i-ess-ta.
Mary-N daughter-except apple-A NOT eat-CAUS-PA-DEC
'Mary did not feed anyone an apple except for (her) daughter.'

c. Mary-ka ttal-lul ***sakwa-pakkey*** ***an*** mek-i-ess-ta.
Mary-N daughter-A apple-except NOT eat-CAUS-PA-DEC
'Mary did not feed (her) daughter anything but an apple.'

In (11), the negated causative verb using the negative element *an* 'not' licenses the negative polarity item *–pakkey* 'except for' on any argument. This suggests that the grammatical functions and the verb are in the same clause, provided the locality condition of the negative polarity item and the negative element are in the same clause.

3.2.3. Reflexive

The clause-bound reflexive *casin* 'self' takes the causer as antecedent, as shown in (12). This indicates that MC has one subject, and thus a single clause.

(12) Mary***i***-ka ttal***j***-lul ***casin i*j-uy*** sakwa-ul mek-i-ess-ta
Mary***i***-N daughter***j***-A self***i*j***-G apple-A eat-CAUS-PA-DEC
'Mary ***i*** fed the daughter ***j*** self ***i*j***'s apple.'

3.2.4. Control

That the PRO in the participle clause *–myense* 'while' can be controlled by either a matrix subject or object if the matrix clause is biclausal, as shown in (13), while only the subject can be the controller in MC, as shown in (14), suggests that MC is monoclausal:

(13) a. [***PROij*** thelepi-lul po-myense]
PRO***ij*** TV-A see-while

Maryi-ka ttal***j***-lul sakwa-ul mek-key ha-ess-ta
Mary***i***-N daughter***j***-A apple-A eat-*KEY* do-PA-DEC
'Mary ***i*** fed the daughter ***j*** an apple while PRO ***ij*** watching TV.'

b. ***Maryi-ka*** ttal***j***-lul [PRO***ij*** thelepi-lul po-myense]
Mary***i***-N daughter***j***- A PRO***ij*** TV-A see-while

sakwa-ul mek-key ha-ess-ta
apple-A eat-*KEY* do-PA-DEC

(14) a. [***PROi*j*** thelepi-lul po-myense]
PRO***i*j*** TV-A see-while

Maryi-ka ttal***j***-lul sakwa-ul mek-I-ess-ta
Mary***i***-N daughter***j***-A apple-A eat-CAUS-PA-DEC
'Mary ***i*** fed (her) daughter***j*** an apple while PRO ***i*j*** watching TV.'

b. ***Maryi-ka*** ttal***j***-lul [PRO***i*j*** thelepi-lul po-myense]
Mary***i***-N daughter***j***- A PRO***i*j*** TV-A see-while

sakwa-ul mek-I-ess-ta
apple-A eat-CAUS-PA-DEC

Having shown that double accusative marking in MC involves no clause-embeddedness, I argue that MC is another source of object asymmetry in exactly the same way as non-causative double accusative-marked clauses.

4. MC and Object Asymmetries

4.1. MC and Case Marking

Firstly, objects in MC are asymmetrical with respect to case marking. As shown in (9b), *sakwa* 'apple' only bears accusative marking, while the causee can be marked with either the dative or the accusative marker. This implies that *sakwa* is the primary object in MC, due to the fact that the direct object gets the accusative marker by default in Korean:

(9) a. Mary-ka ttal-lul ***sakwa-ul*** mek-I-ess-ta.
Mary-N daughter-A apple-A eat-CAUS-PA-DEC
'Mary fed (her) daughter an apple.'

b. * Mary-ka ttal-lul ***sakwa-eykey*** mek-I-ess-ta.
Mary-N daughter-A apple-D eat-CAUS-PA-DEC
'Mary fed (her) daughter an apple.'

4.2. MC and Reciprocalisation

That only the object of the base verb can be reciprocalised suggests that the accusative-marked causee is not the direct object, as shown in (15) and (16):

(15) a. nay-ka **kak ai-eykey** ***selo-uy*** ***ppang-ul*** mek-I-ess-ta.
I-N each child-D each other-G bag-A eat-CAUS-PA-DEC
'I fed each child each other's bread.'

b.* nay-ka ***selo-uy*** ***ai-eykey*** ***kak*** ***ppang-ul*** mek-I-ess-ta.
I-N each other-G child-D each bread-A eat-CAUS-PA-DEC
'I fed each other' child each bread.'

(16) a. nay-ka ***kak ai-lul*** ***selo-uy*** ***ppang-ul*** mek-I-ess-ta.
I-N each child-A each other-G bag-A eat-CAUS-PA-DEC
'I fed each child each other's bread.'

b.* nay-ka ***selo-uy*** ***ai-lul*** ***kak*** ***ppang-ul*** mek-I-ess-ta.
I-N each other-G child-A each bread-A eat-CAUS-PA-DEC
'I fed each other' child each bread.'

The causee in dative case as shown in (15b) and in accusative case as shown in (16b) cannot be reciprocalised. This indicates that the accusative causee is not a direct object.

4.3. MC and The Passive

The passive also shows us that the double accusative-bearing objects in MC are asymmetrical such that the object of the base verb becomes the passive subject, as demonstrated by (15a), but not the causee, as shown in (17b):

(17) a. ***sakwa-ka*** (Mary-eyuhay) ttal-eykey mek-I-e ci-ess-ta.
Apple-N Mary-by daughter-D eat-CAUS-*E* become-PA-DEC
LIT: An apple was eaten by the daughter (by Mary).

b. * sakwa-ka (Mary-eyuhay) ***ttal-lul*** mek-I-e ci-ess-ta.
Apple-N Mary-by daughter-A eat-CAUS-*E* become-PA-DEC

c. * ttal-ka (Mary-eyuhay) ***sakwa-ka*** mek-I-e ci-ess-ta.
daughter-N Mary-by apple-N eat-CAUS-*E* become-PA-DEC

5. Concluding Remarks

The object asymmetries exhibited by a simple triadic predicate are consistently observed in MC because while the accusative-marked causee does not have the properties of a primary object, the accusative-marked argument of the base verb has because it obligatorily bears accusative marking, can be reciprocalised, and can be the passive subject; that is, there is only one primary object in MC.

The fact that Korean has only one primary object can be captured in the theory of AOP (Asymmetrical Object Parameter) in Bresnan and Moshi (1996) and Alsina (1996), because there is only one internal argument that is semantically unrestricted, having [-r] which maps onto a primary object.

The asymmetries of double accusative-marked objects can be accounted for if we assume that the semantically restricted argument associated with the semantic notion of [+affectedness] yields double accusative marking at the level of constituent structure, but its grammatical function remains invariant at the level of functional structure. That is, the accusative marker on the causee associates with the semantic notion, but does not associate with the grammatical function of direct object; case marking may vary in the given context, while case feature remains still.

References

Alsina, A. and S. Joshi. 1991. Parameters in Causative Construction. *Proceedings of the Chicago Linguistic Society* 27: 1-15.

Alsina, A. 1996. Passive Types and the Theory of Object Asymmetries. *NLLT* 14: 673-723.

Baker, M.C. 1988. *Incorporation*: *A Theory of Grammatical Function Changing.* Chicago: The University of Chicago Press.

Bresnan, J. and L. Moshi. 1990. Object Asymmetries in Comparative Bantu Syntax. *LI* 21: 145-185.

Hong, Ki-Sun. 1991. Argument Selection and Case Marking in Korean. Doctoral Dissertation. Stanford, CA: Stanford University.

Manning, D. Christopher, Ivan A. Sag and Masayo Iida. 1996. The Lexical Integrity of Japanese Causatives. Manuscript. (To appear in G. Green and R. Levin (eds), *Readings in HPSG,* Cambridge University Press).

Matsumoto, Yo. 1996. *Complex Predicates in Japanese: A Syntactic and Semantic Study of the Notion 'Word'*. CSLI and Kurosio Publishers.

Mohanan, T. 1988. Causatives in Malayalam. Stanford, CA: Stanford University. Department of Linguistics. MS.

Mohanan, T. 1990. Argument Structure in Hindi. Doctoral Dissertation. Stanford, CA: Stanford University.

Zaenen. A., J. Maling, and H. Thrainsson. 1990. Case and Grammatical Functions. *Syntax and Semantics* 24: *Modern Icelandic Syntax*: 95-134.

Yang, Insun. 1998. Object Marker in Korean. *The Twenty-Fourth LACUS Forum*: 238-250.

LFG AS A PEDAGOGICAL GRAMMAR

Johannes Thomann
University of Zurich

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

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

Abstract

The paper describes a presentation format of grammatical information for language teaching. c-structures and f-structures are represented as graphical annotation to a text. It is advocated that LFG-like concepts are preferable to traditional grammar rules.

After an anti-grammar movement in the 1980s we can recognize today a resurgence of interest in the role of grammar in language teaching (Hedge 2000: 143, Tyler 1994). The term Pedagogical Grammar denotes the types of grammatical analysis and instructions designed for the needs of second language students (Odlin 1994: 1). It comprises the notion of a set of rules (prescription), of an archive of variability (descritive), of a model of the internal competence (internal system) and of an abstract formal system (axiomatic). This hybrid nature makes it difficult to define a clear cut profile of its design and its application in practice. GB/PP, Relational Grammar, GPSG/HPSG and LFG have been proposed as hopeful candidates. As concerns LFG, not much work has been done on the question of its usability for Second Language Acquisition purposes (but see Pienemann 1998).

In a pedagogical grammar, rules should be traditionally: (1) concrete, (2) simple, (3) nontechnical, (4) cumulative, (5) close to popular/traditional notions and (6) in rule-of-thumb form (Odlin 1994). (1)-(3) and (6) concern primarily the form of presentation, and this will be the topic of this contribution. The samples included herein shown how LFG-structures could be described for the lay public. First tests were made in a one year introductory Arabic course at university level.

The idea is to visualize grammatical information as an annotation to a given text, not as an independent structure.

The matrix notation of f-structure is transformed into a directed graph. All PRED-values are lined up in the sequence of their corresponding words in the sentence, or phrase. The result is mathematically equivalent to the original matrix representation.

The tree notation of c-structures is transformed into colored boxes indicating the grouping of the linear structure and the categories of the constituents. Some functional annotations are added to the labels.

Both, c-structure and f-structure annotations, can be used in combination. The curved arrows are clearly distinguishable from the rectangular boxes.

In a traditional textbook the genitive construction (Iḍāfa) is explained the following way (Schulz 2000: 70-72):

- The governing word is in the so-called construct state; it does not take the article or *nunation*.
- All terms except the last in a genitive construction consisting of several terms (genitive chain) are in the construct state.
- Not more than one noun should constitute the 1st term of a genitive

construction - in good style.

- If the 2nd term of the Iḍāfa is definite, the 1st term, which is in the construct state, is also regarded as definite.
- Consequently, an adjectival attributive adjunct ascribed to the 1st term has to be construed with the article.
- However, as the terms of the genitive construction must not be separated ... , the attributive adjunct must either follow the whole genitive construction, ... or else it follows the 1st term, ... , and the 2nd term of the genitive construction which has been dissolved by now is added by means of *li-* .
- If the 2nd term of the Iḍāfa is indefinite, the 1st term in the construct state is regarded as indefinite. An adjectival attributive adjunct ascribed to the 1st term of Iḍāfa ... follows indefinite.

The learner has to memorize all these rules concerning genitive construction without any guiding concept.

In contrast to this traditional explanation, we can capture the entire set of phenomena by means of one recursive rule: An NP can consist of an N, followed by an NP in the genitive, receiving CAS, NUM etc. from the head N and DEF from the NP (POSS is omitted here and in the following examples for the sake of simplicity). This is analogous to the simple NP, consisting of an N and a Det.

$$\begin{array}{lcccc} \text{NP} & = & \text{N} & & \text{NP} \\ & & \uparrow{=}\downarrow & & (\uparrow \text{DEF}) {=}\downarrow \text{DEF} \\ & & & & \downarrow \text{CAS} = \text{gen} \end{array}$$

This form of notation is not acceptable in language teaching. It is particularily confusing for Arabic, where the writing direction is right to left. The following examples illustrate a pedagogical alternative :

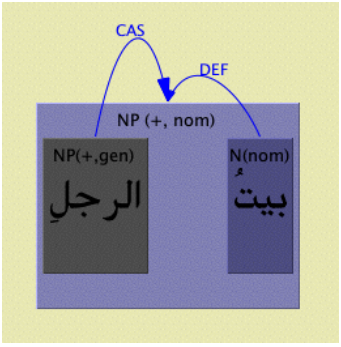


Figure 1: A simple genitive construction (‘the house of the man’)

The analogy to the simple NP is emphasised by the color pattern:

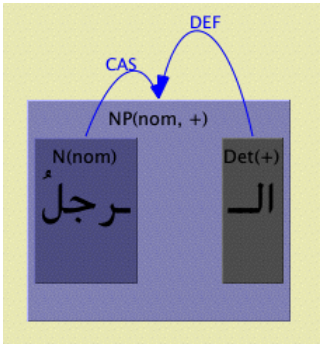


Figure 2: A simple definite nominal phrase (‘the man’)

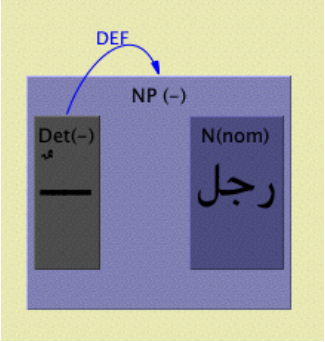


Figure 3: A simple indefinite nominal phrase (‘a man’)

The definition predicts a further regularity, not covered by the descriptive set of rules in the textbook: attributed adjectives follow in reverse order of their respective substantives (see figure below).

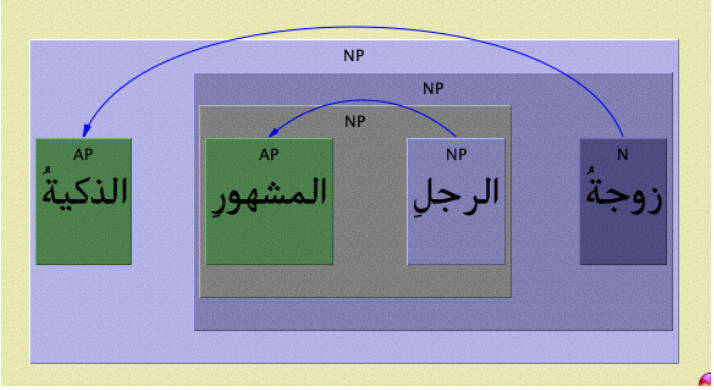


Figure 4: Attributed adjectives (‘the intelligent wife of the famous man’)

The above samples are the output of a Java application:

`http://www.ori.unizh.ch/lfg/`

In combination with modules for parsing and feature structure unification it is part of an authoring tool for the design of pedagogical grammars. Apart from the introductory Arabic course it will be used in the Arabic Papyrology School:

`http://www.ori.unizh.ch/aps/`

Further tests must decide, if this presentation form is usefull in the communication between learner on one side and teacher or teaching systems on the other side.

References

Hedge, Tricia. 2000. *Teaching and Learning in the Language Classroom.* Oxford: University Press.

Odlin, Terence (ed.). 1994. *Perspectives on Pedagogical Grammar.* Cambridge: University Press.

Pienemann, Manfred. 1998. *Language Processing and Second Language Development.* John Benjamin Publishing (I owe this reference to Veit Reuer, Osnabrueck)

Schulz, Eckehard. 2000. *Standard Arabic : An elementary – intermediate course.* Cambridge: University Press.

Tyler, Andrea, and Donne Lardier. 1996. Beyond consiousness raising : Re–examining the role of linguistics in language teacher training. In *Georgetown University Round Table on Languages and Linguistics 1996*, ed. James E. Alatis, Carolyn A. Straehle, Maggie Ronkin, and Brent Gallenberger. 270-287. Washington, D.C.: Georgestown University Press.

Aspects of the syntax of psychological verbs in Spanish

A lexical functional analysis

Henk Vanhoe

University of Gent

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

In Spanish, people generally distinguish three kinds of psychological verbs, those that are syntactically realized like *temer*, like *preocupar* and like *gustar*. One of the peculiarities of these verbs is that, despite their semantic relatedness, each type of verbs shows up a different correspondence pattern between thematic roles and grammatical functions. In this paper I develop a unified account of these empirical data, based on Lexical Mapping Theory. As the difference between *preocupar* and the other kinds of verbs seems to be mainly semantic, and more specifically aspectual, I propose to reformulate the thematic theory of Dowty (1991), in order to accommodate aspectual differences; more specifically, the proto-role linking of arguments is made dependent in part on the aspectual decomposition of the event denoted by a verb. In order to explain the syntactic differences between *temer* and *gustar*, I propose to modify the mapping theory, by introducing an optional rule operating on the thematic structure of the *gustar* verbs. These modifications give as an additional result a more consistent analysis of Spanish (and generally Romance) indirect objects and a preliminary analysis of the Spanish *leísmo* (through which an object, traditionally analyzed as a direct object, can be marked with dative morphology) as it operates in the case of the psychological verbs.

1 Three classes of psychological verbs[1]

In Spanish, as in other Romance languages, it is possible to distinguish three kinds of psychological verbs, those that are syntactically realized like the verb *temer* in (1), like *asustar* or *preocupar* in (2), and like *gustar* in (3):

(1) Juan teme el fuego.
John fears the fire
'John fears the fire.'

(2) El fuego asusta a Juan.
The fire frightens ACC John
'The fire frightens John.'

(3) El teatro le gusta a Juan.
The theater 3SG/DAT pleases DAT John
'John likes the theater.'

[1]I wish to thank the organizing committee, for making it possible to present the research developed in my doctoral thesis at the 2002 International LFG conference. I would also like to thank all those who provided me during the conference with stimulating questions, comments and advices: this experience allowed me to further refine certain ideas elaborated in my dissertation.

Despite the syntactic differences exemplified by these sentences, the three verbs seem to show up a thematic equivalence: in all three examples, there is an "experiencer" reacting emotionally to a "theme". The theme is realized as the subject in (2) and in (3), and as the direct object in (1), while the experiencer is realized as the subject in (1) and as the direct and the indirect object in (2) and (3) respectively. Thus, one of the puzzles concerning the analysis of these verbs goes as follows: how can it be explained that apparently equivalent thematic relations can be realized as three syntactically different constructions?

One of the oldest and most popular solutions to this puzzle was formulated within the transformational framework, and takes this unifying thematic factor as its starting point: the thematic equivalence between these three kinds of psychological verbs can be explained if we postulate an equivalent or at least a similar deep structure for all of them (Belletti and Rizzi 1988).[2] But a closer look at the semantic content of these verbs shows that there is a systematic difference between the verbs of the *temer* and *gustar* classes on the one hand and the class of *preocupar* on the other hand. Indeed, careful analysis of Spanish data shows that if the verbs of the *temer* and the *gustar* classes can best be characterized aspectually as states, the verbs of the class of *preocupar* are aspectually closer to "achievements". On the other hand, from an aspectual and generally semantic point of view the verbs of the *temer* and the *gustar* classes seem to be essentially equivalent. However, these two kinds of verbs present a *syntactically* differentiated behavior: the *temer* verbs behave like regular transitive verbs, while the *gustar* verbs have unaccusative characteristics.

This analysis of the empirical data suggests that the original puzzle should be decomposed into two new questions:

1. how can we explain the different syntactic configurations of the semantically equivalent verbs of the *temer* and *gustar* classes?

2. if we consider the verbs of the *temer* class to follow the thematically unmarked linking pattern Experiencer/SUBJ—Theme/DO (as is suggested for

[2] One of the first versions of the transformational proposal can be found in Postal (1971)'s Psych-Movement rule. Pesetsky (1995) further develops and refines the transformational approach.

instance by Grimshaw 1994), how can we derive the inverted syntactic configuration of the *preocupar* verbs from their semantic characteristics?

2 A reformulation of Dowty's proto-role theory

2.1 A new pair of proto-role properties

The answers to these two questions have to be of a very different nature. The explanation of the difference between the *temer* and the *gustar* verbs is essentially non-semantic. On the other hand, the difference between the *preocupar* and *temer* classes of verbs seems to be triggered by a semantic distinction. The correspondence between semantic roles and grammatical functions is the object of study of different kinds of "mapping theories", both inside and outside the framework of LFG. As is explained in Butt and Holloway King (2000), different approaches were developed within LFG, Lexical Mapping Theory being one of the most popular (Bresnan and Kanerva 1989, Bresnan 2001). Other theories were developed outside LFG (Joppen and Wunderlich 1995, Wechsler 1995) or from a perspective which is relatively neutral as to the syntactic theory one adopts (Ackerman and Moore 2001). In this paper, I will primarily follow Lexical Mapping Theory. The first step in my analysis is to integrate the aspectual difference between these two types of verbs into the mapping theory. Therefore, I take as a starting-point the thematic theory of Dowty (1991).[3] Dowty distinguishes two lists of properties, which can be used to characterize the two thematic roles ("proto-roles") he distinguishes:

(4) Contributing properties for the Agent Proto-Role:
   a. volitional involvement in the event or state
   b. sentence [sic] (and/or perception)
   c. causing an event or change of state in another participant
   d. movement (relative to the position of another participant)

[3] Several other authors working within an LFG framework also use Dowty's theory as the basis for their mapping theories: Ackerman and Moore (1999), Alsina (1996), Kelling (2002), Zaenen (1993). However, the proposal presented here is different in several respects from the approaches developed by these authors.

e. (exists independently of the event named by the verb)

(5) Contributing properties for the Patient Proto-Role:

a. undergoes change of state

b. incremental theme

c. causally affected by another participant

d. stationary relative to movement of another participant

e. (does not exist independently of the event, or not at all)

The linking of these proto-roles with the grammatical functions follows the Argument Selection Principle (Dowty 1991: 576):

> In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of Proto-Agent properties will be lexicalized as the subject of the predicate; the argument having the greatest number of Proto-Patient entailments will be lexicalized as the direct object.

Independently of my approach and of the data I want to account for, Dowty's lists of proto-role properties seem to be insufficient to account for all types of verbs. Indeed, Dowty himself presents these lists as only provisional. On the one hand the two lists of properties are rather heterogeneous, and on the other hand, they don't seem to cover all thematically relevant semantic distinctions. Other authors also have tried to extend Dowty's lists with new properties. Ackerman and Moore (1999, 2001), for instance, add the property of being a telic entity to the list of proto-patient properties; as a matter of fact, my modification of Dowty's theory will resemble to a certain extent that of Ackerman and Moore.

However, it is difficult to exactly classify the different kinds of psychological verbs from an aspectual point of view, or from the point of view of their "Aktionsart": at least the *frighten*-type verbs don't seem to fit exactly in none of the aspectual classes distinguished for instance by Vendler (1967). Although many authors classify these verbs either as achievements or as accomplishments, sometimes implicitly, by characterizing them as causative or as telic verbs, they don't

behave like typical examples of these categories.[4] For instance, with respect to the standard telicity-test (the standard test for determining achievement-hood or accomplishment-hood), compatibility with a delimiting complement, these verbs show very heterogeneous results:

(6) ?* En cinco minutos, el problema de cambiar de casa me preocupó.
'In five minutes, the problem of moving preoccupied me.'

(7) ? Que pensaras así me enfadó en cinco minutos.
'That you thought so angered me in five minutes.'

(8) En cinco minutos, fascinó a todo el mundo con su labia.
'In five minutes, he fascinated everybody with his volubility.'

This problem of classification is also reflected in the bibliography, where one can find all kinds of aspectual classifications for these verbs (cf. full references in Vanhoe 2002: 135–139).

However, most analyses seem to agree to consider the *frighten* verbs as telic verbs, while the other two types of verbs are generally analyzed as atelic verbs: although they don't have a consistent behavior with respect to their compatibility with a delimiting complement, they are telic with respect to other telicity-tests. Most importantly, they are compatible with complements indicating a gradual change over time:[5]

(9) Poco a poco, el problema de cambiar de casa me preocupó.
'Little by little, the problem of moving preoccupied me.'

[4]Kelling (2002) also notices this fact in French, but reaches different conclusions with it. More particularly she distinguishes two aspectual classes within the class of *frighten* verbs, a class of telic verbs and a class of atelic verbs, by using the two tests of compatibility with a durative complement (*for X time*) and with a delimiting complement (*in X time*). However, at least in Spanish, it seems that all *frighten* verbs are relatively acceptable with a durative complement, while, as we will show presently, compatibility with a delimiting complement varies from very bad to acceptable; thus, with respect to this test, it is not really possible to distinguish two discrete subclasses within the *frighten* class.

[5]Tenny (1994: 66) applies the same argument to English data. The other test introduced by Tenny, reference to an "endstate entailment", doesn't seem readily applicable to Spanish data, as the kind of resultative construction she uses doesn't exist in Spanish.

(10) Poco a poco, me enfadó que pensaras así.
'Little by little, it angered me that you thought so.'

(11) Gradualmente, fascinó a todo el mundo con su labia.
'Gradually, he fascinated everyone with his volubility.'

Therefore I propose to add the pair of properties listed in (12) and (13) to the lists provided by Dowty in order to account for the telic/atelic distinction:

(12) the participant has the most prominent thematic role in a first subevent (=proto-Agent property)

(13) the participant has the most prominent thematic role in a second subevent (=proto-Patient property)

These two properties are based on the idea that a telic event is composed of at least two subevents, one that precedes the final state or event, and the final state or event itself. In this way, it is possible to already establish a distinction between the *preocupar* verbs and the *temer* verbs in their thematic structures, as the theme of a verb like *preocupar* plays the prominent role in the "triggering" event, and the experiencer in the resulting state.

Thus, if we add these two properties to Dowty's lists, we can rewrite them as in (14) and (15):

(14) Proto-agent properties:
   a. the participant is involved volitionally in the event
   b. the participant has the most prominent thematic role in a first subevent
   c. the participant feels or perceives something
   d. the participant contains or possesses something

(15) Proto-patient properties:
   a. the participant undergoes a change of state
   b. the participant has the most prominent thematic role in a second subevent
   c. the participant is the object of a feeling or a perception
   d. the participant is contained in or enters something else, or is or comes into the possession of another participant

The comparison of this list of properties with Dowty's shows that for the most part they cover the same data. I retained the two first agentive properties (volitional involvement and sentience) and the first patient property ("undergoes change of state"). The two properties concerning the causative character of the sentence and the property of being an "incremental theme" are collapsed into the aspectual distinction.[6] I added the second patient property to ensure symmetry between the two lists. But there are also several important differences between both lists of properties. More particularly, I did not retain the two last properties of Dowty's lists. However, most of the examples proposed by Dowty (1991: 573) to exemplify these properties can be subsumed in the part-whole and possessor-possessed distinction (property d) and in the aspectual distinction (property b), as I demonstrate in Vanhoe (2002).

2.2 A hierarchy of properties

At the same time, and for reasons that soon will become clear, it is necessary to establish a hierarchy between the properties in these two lists: following a suggestion of Alsina (1996: 41), I consider the first two properties of each list to be "primary" properties, the last two properties are "secondary" properties.[7] This hierarchy of properties captures the intuition which also is at the basis of the standard hierarchy of thematic roles, in which agents (property a) or causers (property b) are ranked higher than experiencers (property c). The parallelism between properties (c) en (d) is motivated by the observation that in Spanish, sentences denoting a part-whole relationship often display the same characteristics as experiencer verbs of the *gustar*-class (Vanhoe 2002: 236). Thus, if we analyze the three examples listed in (1)

[6] According to Ackerman and Moore (1999), an incremental theme does not necessarily imply telicity, which is in contradiction with my proposal to collapse both characteristics into one property. According to these authors, in an example like "Kim drank water", although this sentence does not refer to a telic event, the object denotes an incremental theme. I don't have a definitive answer to this problem. However, Ackerman and Moore (1999) characterize an incremental theme as a participant of a predicate which preserves the part-of relation. As *water* in "Kim drank water" does not denote a precise amount of water, it seems difficult to distinguish a part-of relationship in this kind of sentences.

[7] I suspect the effect of this hierarchy could also be reached with an optimality theoretic account, but I haven't fully explored this possibility.

to (3), we can thematically characterize their participants as follows:

- in (1), the subject ("Juan") is a secondary agent (14c) and the object ("el fuego") is a secondary patient (15c)
- in (2), the subject ("el fuego") is simultaneously a primary agent (14b) and a secondary patient (15c), while the object is simultaneously a primary patient (15b) and a secondary agent (14c)
- in (3), the subject ("el teatro") is a secondary patient (15c) and the indirect object a secondary agent (14c)

With the *temer*-verbs, there is no conflict between primary and secondary properties and these verbs will be realized as regular transitive verbs. As primary properties take precedence over secondary properties, the *preocupar*-verbs also will be realized as normal transitive verbs. Although in other contexts the secondary properties of the *preocupar*-verbs seem to play a syntactic role (for instance, in the formation of the middle construction, as shown in Vanhoe 2002), they do not in the mapping of the participants to the different grammatical relations. If we consider agents to map to [−o] arguments, and patients to [−r] arguments, standard mapping theory will do the rest of the job. The idea that the *preocupar*-verbs are normal transitive verbs is in contradiction with most other analyses of these verbs (and not only the transformational ones), as they generally consider them as displaying special properties in their mapping of thematic roles to grammatical functions. However, in my thesis I show that the analysis presented here is also empirically justified, at least in Spanish, while other authors, like Bouchard (1995) and Ruwet (1972) suggest the same for French and English.[8]

3 The *gustar* verbs

3.1 Romance indirect objects are OBJ

The analysis of the verbs of the class of *gustar* is, at least formally, more complex, as the correct grammatical characterization of Romance indirect objects is not immediately evident. In principle, it seems most natural to consider them as obliques.

[8] Zaenen (1993: 145) also considers these verbs as "simple transitive verbs".

However, Alsina (1996: 150ss) gives several arguments against this analysis. His arguments are based on Catalan data, but can be applied directly to Spanish data (as both are closely related Romance languages). They could also be analyzed as thematic or secondary objects (OBJ_θ) but they don't seem to behave like typical thematic or secondary objects in other languages either; contrary to secondary objects, they don't have to be *secondary* (they don't have to be used together with another object) and they are *always* realized with a preposition (*a*):

(16) Juan le dio el libro a María.
John 3SG/DAT gave the book to Mary.
'John gave the book to Mary.'

(3) El teatro le gusta a Juan.
The theater 3SG/DAT pleases DAT John
'John likes the theater.'

(17) Juan le ha mentido a su jefe.
John 3SG/DAT has lied to his director.
'John lied to his director.'

Thus it seems worthwhile to follow a suggestion made by Alsina (1996) and to consider both direct and indirect objects as morphologically distinct instances of the same grammatical function "object". Indeed, many grammatical phenomena in Spanish suggest the similarity of both types of objects. It is true that this analysis implies some important modifications of standard rules, but Alsina (1996) shows that they can be accounted for satisfactorily. Under this hypothesis, we can easily analyse the *gustar* verbs, if we add the typically Spanish, maybe even typically Romance, rules, listed in (18) and (19), to the mapping theory:

(18) **θ-structure to a-structure mapping**
Secondary agents correspond to [+o] arguments in the marked option.

(19) **a-structure to f-structure mapping**
A [+o] argument corresponds to $\begin{pmatrix} \text{OBJ} \\ (\uparrow \text{DAT}) = + \end{pmatrix}$

These rules fulfill a function similar to the one introduced for instance in Bresnan (2001: 309), in order to analyze "secondary patientlike roles". With these rules,

we can derive the grammatical functions of a verb like *gustar* as in (20):[9]

(20) *gustar*:

θ-structure	< P-A^{②},	P-P >
	|	|
a-structure	[+o]	[−r]
	|	|
f-structure	OBJ	SUBJ
	(↑ DAT) = +	

We already know that the experiencer of *gustar* is a secondary agent. Thus, this argument can be mapped to a [+o] argument, and consequently, through (19), to an object marked with dative case. However, the fact that a verb follows rule (18) has to be specified lexically: other verbs, such as *temer*, don't follow this option. This analysis of the verbs of the *gustar* class has several advantages. Most noticeably, it very naturally accounts for the unaccusative characteristics of the *gustar*-like verbs. Just as with normal, intransitive unaccusative verbs, the subject of *gustar* is characterized, in its a-structure, as a [−r] argument.

In addition, this analysis allows us to introduce some regularity in the apparently idiosyncratic behavior of the verbs of the *gustar* class. The mapping exemplified by these verbs can only be obtained with predicates that contain a secondary agent; for that reason it is only possible with verbs that have an experiencer or a "container" or possessor as one of their participants. This is in accordance with what one can find in Spanish. Not only this type of indirect constructions can be found with experiencer verbs but also with verbs expressing a part-whole relation, or a relation of possession (or their negation):

(21) A Juan le falta confianza.
DAT John(IO) 3SG/DAT lacks confidence(SUBJ)
'John lacks confidence.'

(22) A Juan le basta tu palabra.
DAT John(IO) 3SG/DAT suffices your word(SUB)
'Your word is enough for John.'

[9] P-A stands for "proto-agent", P-P for "proto-patient" and P-A^{②} for "secondary agent".

The approach presented here has several advantages over other mapping theories, such as Lexical Decomposition Grammar (Joppen and Wunderlich 1995, Stiebels 2000), which consider this kind of verbs as examples of a purely lexical idiosyncrasy. As such, these approaches can't explain why this idiosyncrasy is limited to particular semantic types of verbs.

3.2 *Leísmo*

This analysis also makes it possible to explain some typically Spanish data in a principled way. As is well known in the Spanish grammatical tradition, Spanish is characterized by what has come to be known as the phenomenon of "leísmo", by which an object, traditionally analyzed as a DO, can be marked, under certain circumstances, with dative morphology (that is, the morphology characteristic of an IO). Thus, in (23), the DO is marked with "normal" accusative case, while in (24), it is marked with dative case:

(23) Juan lo ha visto.
John 3SG/ACC has seen
'John has seen him'

(24) Juan le ha visto.
John 3SG/DAT has seen
'John has seen him'

In most cases *leísmo* is optional, and seems to be triggered by contextual and/or pragmatic factors that are more or less difficult to circumscribe. In addition, *leísmo* seems to be most frequent in European Spanish, more particularly in Northern and Central dialects of Spain. However, with certain verbs, a DO is obligatorily marked with dative morphology, not only in European Spanish, but also in Latin American Spanish. The prototypical example of this is the verb *interesar*, as shown in (25) and (26):

(25) Este libro le interesa (a Juan).
This book 3SG/DAT interests (ACC? John)
'This book interests him (John).'

(26) * Este libro lo interesa (a Juan).
This book 3SG/ACC interests (ACC John)

‘This book interests him (John).’

However, in its “causative” variant, *leísmo* is again optional, as can be seen in (27):

(27) María lo / le ha interesado (a Juan) en el negocio.
María 3SG/ACC 3SG/DAT has interested (ACC John) in the business.
‘Mary has interested John in the business.’

One possible analysis is to say that the object of *interesar* (in its non-causative version) is not a direct but an indirect object: *le* is the normal dative pronoun, and both DO and IO can be marked with the preposition *a*. However, from various points of view, this element behaves more like a direct than an indirect object:

- *interesar* is perfectly possible in an adjectival passive construction, just like the other verbs of the *preocupar*-class, as can be seen in examples (28) to (30)

  (28) Juan está preocupado por el discurso.
  ‘John is worried about the speech.’

  (29) Juan está interesado por el discurso.
  ‘John is interested in the speech.’

  (30) * Juan está gustado por el discurso.
  ‘John is pleased with the speech.’

- this verb is possible in an “absolute construction”, just like the *preocupar*-verbs, and contrary to the *gustar*-verbs, as shown in (31) to (33)

  (31) Preocupado Juan por el incidente,…
  Worried John by the incident
  ‘As John got worried about the incident,…

  (32) Interesado Gustavo repentinamente por los
  Interested Gustavo suddenly by the
  ordenadores,… (Miguel 1992: 244-245)
  computers
  ‘As Gustavo was suddenly interested in computers,…

(33) * Gustado Juan con el café,...
Pleased John with the coffee
'As John was pleased with the coffee,
... se fue sin explicaciones.
... he went away without explanation.'

- generally, an indirect object, contrary to a direct object, is "announced" by a so-called expletive pronoun, as in example (3).[10] With *interesar*, both versions, with and without an expletive pronoun, are possible:

(34) El teatro *?(le) gusta a Juan.

(35) El teatro (le) interesa a Juan.
'The theater interests John.'

In the present analysis, nothing prevents a DO from being marked with dative morphology as on a "deeper" level, DO and IO are instances of the same grammatical function OBJ. This can be expressed very easily in the lexical entry of the verb *interesar*, as represented in (36):

(36) *interesar* V $(\uparrow \text{PRED}) = \text{'interesar}\langle(\uparrow [-\text{o}])(\uparrow [-\text{r}])\rangle\text{'}$
$(\uparrow \text{OBJ DAT}) = +$

Although *interesar* selects a regular direct object (a $[-\text{r}]$-argument), its DAT attribute receives obligatorily (and idiosyncratically) a positive value. In its causative version, this restriction disappears and the object can be a normal direct object, not marked with dative case (27). As a hypothesis, we can state that the presence or absence of *leísmo* is triggered by the value of the DAT attribute, not by the grammatical function itself. It is important to observe that although both with *gustar* and with *interesar* the object is necessarily marked with dative case, on a more abstract level, these verbs have radically different structures: their objects are assigned dative morphology through very different processes; more specifically, *gustar* is an

[10] However, as is observed by Roegiest (1999: 71), this "rule" can be overridden easily in certain contexts (whose exact nature still has to be determined).

unaccusative verb, while *interesar* is, at least at the level of its a-structure, a regular transitive verb.[11]

4 Conclusion

In conclusion, in this paper I tried to show that it is possible to account for data that are traditionally considered to be problematic in the Spanish grammatical tradition, with the formal methods of Lexical Functional Grammar and more particularly of Lexical Mapping Theory. In order to achieve this goal, only some relatively minor modifications of the standard theories were needed. With the hypotheses presented here, it is possible to account for the different kinds of mapping exemplified by three classes of Spanish psychological verbs. In addition we can account very naturally for the unaccusative characteristics of the verbs of the *gustar* class, and explain the presence of dative morphology with certain transitive verbs like *interesar*. In Vanhoe (2002), I show that these hypotheses also make it possible to shed a new light on several other constructions involving psychological verbs, such as their presence in middle constructions, as well as their behavior in constructions involving a binding relation.

References

Ackerman, F. and Moore, J. (1999). 'Telic entity' as a proto-property of lexical predicates. In Butt, M. and Holloway King, T., editors, *Proceedings of the LFG99 Conference, university of Manchester*, `http://csli-publications.stanford.edu/LFG/4/lfg99.html`. World Wide Web. Internet document consulted on 28/9/2001.

Ackerman, F. and Moore, J. (2001). *Proto-properties and Argument Encoding: a Correspondence Theory of Argument Selection.* CSLI Publications, Stanford, California.

[11] The indirect objects of verbs like *mentir* or *hablar*, which are not unaccusative either, are also distinct from the object of *interesar*, as they are [+o]-arguments, not [−r]-arguments.

Alsina, A. (1996). *The role of argument structure in grammar. Evidence from Romance*. CSLI Publications, Stanford, California.

Belletti, A. and Rizzi, L. (1988). Psych-verbs and theta-theory. *Natural language and linguistic theory*, 6:291–352.

Bouchard, D. (1995). *The semantics of syntax.* Univ. of Chicago Press, Chicago.

Bresnan, J. (2001). *Lexical-Functional Syntax.* Blackwell, Malden-Oxford.

Bresnan, J. and Kanerva, J. (1989). Locative inversion in Chicheŵa: A case study of factorization in grammar. *Linguistic Inquiry*, 20(1):1–50.

Butt, M. and Holloway King, T. (2000). Introduction. In Butt, M. and Holloway King, T., editors, *Argument Realization,* pages 1–14. CSLI Publications, Stanford, California.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67:547–619.

Grimshaw, J. (1994). *Argument Structure*. MIT Press, Cambridge. [1990].

Joppen, S. and Wunderlich, D. (1995). Argument linking in Basque. *Lingua*, 97:123–169.

Kelling, C. (2002). Argument realization: French psych verb nominalizations. Handout of a poster presented at the 2002 International Lexical Functional Grammar Conference, National Technical University of Athens, 3-5/07/2002.

Miguel, E. d. (1992). *El aspecto en la sintaxis del español: Perfectividad e impersonalidad.* Ediciones de la Universidad Autónoma de Madrid, Madrid.

Pesetsky, D. (1995). *Zero syntax: Experiencers and cascades.* MIT Press, Cambridge.

Postal, P. M. (1971). *Cross-Over Phenomena.* Holt, Rinehart & Winston, New York.

Roegiest, E. (1999). Objet direct prépositionnel ou objet indirect en espagnol. *Verbum*, XXI(1):67–80.

Ruwet, N. (1972). *Théorie Syntaxique et Syntaxe du Français.* Editions du Seuil, Paris.

Stiebels, B. (2000). Linker inventories, linking splits and lexical economy. In Stiebels, B. and Wunderlich, D., editors, *Lexicon in focus*, pages 213–247. Akademie Verlag, Berlin.

Tenny, C. (1994). *Aspectual Roles and the Syntax-Semantics Interface.* Kluwer, Dordrecht.

Vanhoe, H. (2002). *Aspectos de la sintaxis de los verbos psicológicos en español. Un análisis léxico funcional.* PhD thesis, Universiteit Gent.

Vendler, Z. (1967). *Linguistics in Philosophy.* Cornell UP, Ithaca.

Wechsler, S. (1995). *The Semantic Basis of Argument Structure*. CSLI Publications, Stanford, California.

Zaenen, A. (1993). Unaccusativity in dutch: integrating syntax and lexical semantics. In Pustejovsky, J., editor, *Semantics and the lexicon*, pages 129–161. Kluwer, Dordrecht.

French Causatives : a Biclausal Account in LFG

Nicholas Yates
(Université Libre de Bruxelles)

Proceedings of the LFG02 Conference
National Technical University of Athens, Athens
Miriam Butt and Tracy Holloway King (Editors)
2002
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

The aim of this paper is to question the well-established view in constraint based grammars on Romance causatives, i.e. that they are functionally monoclausal, as well as its most important corollary, that their lexical representation is underspecified with regard to argument structure (Alsina 1996, Abeillé, Godard & Sag 1998) that gets fully specified by merging with another predicate. This is the generally accepted view, even though it contradicts initial claims on lexical integrity. The discussion will rest upon a few problematic sentences of which the monoclausal theories give an unsatisfactory account. I argue that the monoclausality of causatives is only a surface (i.e. a c-structure) monoclausality. The 'merging effect' observed at the level of surface realisation, and the apparent extra arguments of the causative predicate will be accounted for by some other means. The idea is to resort to the previous conception of a biclausal f-structure, but with a more 'complex' use of structure sharing. I will mainly be concerned with the implementation of this concept in Lexical Functional Grammar (LFG) in this paper, but a short comparison with a similar Head-Driven Phrase Structure Grammar (HPSG) approach will support the idea that the proposal has some empirical relevance.

1 Introduction [1]

Romance causatives[2] have been extensively studied in the field of generative grammars for the last 30 years and, more specifically, during the last decade in constraint-based grammars. In the latter framework, there has been a rather wide agreement that Romance causative constructions are formed through some sort of merging of two lexical subcategorisation frames. This generally implies that the lexical entry of the causative verb is underspecified with regard to argument structure (Alsina 1996 for LFG, Abeillé et al. 1998 for HPSG) which will have to be syntactically fully specified by the addition of elements from another predicate's argument structure. Although this is an important departure on conceptual grounds from initial assumptions such as the 'Lexical Integrity Principle' (see Ackerman & Webelhuth 1997 for discussion), it is the view generally adopted in constraint based frameworks.

Let's review some of the basic facts about causatives by examining the following sentences:[3]

(1) Pierre a fait courir **Paul**.
Peter made Paul run.

(2) Pierre a fait construire un bateau **à Paul**.
Peter made Paul build a boat.

(3) Pierre a fait construire un bateau **par Paul**.
Peter had a boat built by Paul.

(4) Pierre **lui** a fait écrire une lettre.
Peter made him write a letter. / Peter had a letter written to him.

(5) Pierre **lui** a fait écrire une lettre **par Marie**.
Peter had a letter written to him by Mary.

[1] the author wishes to thank Marc Dominicy, Fabienne Martin, Philip Miller and at least two anonymous reviewers for insightful comments.

[2] In this paper, I will only discuss French causatives. Many of the general assumptions made here hold for other Romance languages, although there are some notable differences (cf. Frank 1996)

[3] There are two types of constructions for causatives in Romance languages (cf Abeillé et al. 1998) The first one seems to be subject to some kind of merging of two subcategorisation frames ('Il **lui** a fait tuer un homme'). The second one, subject to variation in acceptability, is a classical control construction ('?Il **l'**a fait tuer un homme'). We will only concern ourselves with the first kind in this paper.

(6) Pierre **lui** a fait écrire une lettre **à Paul**.
Peter made him write a letter to Paul.

(7) *Pierre **lui** a fait téléphoner **Marie**.
Peter made Mary call him.

(8) Pierre a fait téléphoner **Marie à Paul**.
Peter made Mary call Paul.

The first three sentences show us the very basic facts about causatives in French. In simple terms, the causative predicate 'faire' (to do, to make) takes an infinitive verb as complement, and apparently shares its arguments with the embedded predicate's arguments. If the embedded verb is intransitive, its subject will be expressed as the causative's direct object. If the infinitive verb is transitive, its subject will be expressed either as an indirect à-object or as a par-phrase (the typical agent complement phrase in French). These are the possible functions that can be occupied by what has been called the 'causee' role.[4] All other arguments keep their grammatical function but seem to be complements to the causative verb. This is supported by observed behaviour with regard to tough-movement, NP-ordering, extraction and upstairs cliticisation. But we will see that the constraints that appear after a closer look at sentences (4) through (8) reveal that the story is somewhat more complicated. But we will postpone comments on these until later on.

I will not dwell on the numerous facts that argue in favour of some sort of merging or dependence of the embedded infinitive verb (see, among others, Kayne 1975, Ruwet 1972, Tasmowski 1984, Burzio 1988). But it should be stressed that many of the claims for monoclausal structure are based on phrase-structural constraints like heavy-NP shift,... Since I keep the assumption that the surface realisation of French causatives (in our case, the c-structure) is monoclausal in nature, I assume that they will be accounted for in the same way here as in other theories. There are some specific behaviours of causatives, though, that do not seem to be explainable through phrase-structural configuration, but rather be dependant upon some internal lexical-structural constraint. These have generally been seen as the main evidence for the monoclausal representation in the constraint-based literature. This includes mainly upstairs cliticisation (Miller & Sag 1997) and functional relation changes of the embedded verb's arguments, from SUBJ to OBJ or OBJ_{dat} (Alsina 1996, Dalrymple & Zaenen 1996,...) I will argue that these phenomena are actually not uniformly pleading in favor of a monoclausal structure.

[4]There has been a lot of discussion regarding this notion of a 'causee' role, but I believe this to be irrelevant to the present discussion.

In derivational frameworks, they are easily accounted for. In Government & Binding (GB), the embedded VP is partially or completely destroyed in the process (Burzio, 1988). In the Relational Grammar (RG) framework, two phrases are merged in 'clausal union' (Fauconnier 1983). The derivational process will impose different constraints on NPs according to their initial relation to the embedded infinitive, whether they are 'deep' subjects or 'deep' dative objects.

This apparently leaves us with a straight alternative. Either we adopt the derivational view (which, of course, we don't want) and there is a structural difference between the initial subject or indirect object. The above examples then follow from general principles (θ-criterion,...). If, on the other hand we adopt the constraint-based view (and we will of course) and the monoclausal structure that comes with it. This, I consider to be an unsatisfactory alternative. I will try to show that it is possible to give an account that stays faithful to the basic assumptions of constraint-based grammars (maybe even more faithful than previous accounts) while having the same empirical coverage that the derivational theories can give.

2 A biclausal structure

In the present paper I challenge the general view by assuming that the internal structure of causatives in French is actually biclausal. In the LFG formalism, the obvious way of expressing this is by having a biclausal f-structure. The specificity of this kind of construction, namely complex predicates, will be that they make a much more 'complex' use of structure sharing. To express this in one 'lexical item', I give in figure 1 a very general template of what a lexical entry for 'faire' should look like.

Faire, V : Pred = 'faire'<SUBJ, VCOMP, OBJ, OBJ$_{[dat]}$>
$(\uparrow X = \uparrow VCOMP\ X')$
$(\uparrow Y = \uparrow VCOMP\ Y')$
...

Figure 1: 'faire'

As hinted above, I will argue that the main pro-monoclausal arguments are actually not uniformly pleading in favour of a monoclausal structure. Let's have a closer look at the constraints on upstairs cliticisation and on the presence of two

dative NPs that the examples (4) to (8) reveal. For Alsina (1992) & (1996) or Frank (1996), the best way to explain clitic climbing is by positing that the two predicate undergo some sort of predicate composition (merging or partial-merging) whereby the arguments of the embedded infinitive verb are added to the a-structure of the causative 'faire'. The resulting complex a-structure is then mapped onto a flat f-structure. I will not review the mechanisms for merging a-structures here since they are rather technical innovations of the standard LFG formalism and my aim is to argue that they are unnecessary. Indeed, it is difficult to account for the fact that there can sometimes be two $\text{OBJ}_{[dat]}$ [5], as in example (6). This would appear to be in direct contradiction with the Coherence Principle (Bresnan 2001), unless one admits that they are different functions. For instance, on could stipulate that the first is an OBJ_{causee} while the second would be an $\text{OBJ}_{recipient/goal}$. But this would ignore a number of general facts about dative objects.
First of all, they can both be cliticised by 'lui', though there are certain restrictions here that we will consider below, and 'lui' is a strictly dative pronoun. So we can't escape the fact that there are two dative NPs subcategorised for by the same predicate if we decide for the argument composition view. And this would be a the unique and strictly construction-specific case in the French language.

What more, there seems to be some semantic correspondence between the subcategorisation frames of the non-causative and the causative versions of faire. For the sake of semantic generalisation, one would want to posit some relation between both kind of predicates by saying, for instance, that the dative object of the causative 'faire' is in fact its recipient/goal θ-role :

(9) Il a fait un gâteau **à Paul**. → Il **lui** a fait un gâteau.

These observations lead me to consider the option of admitting two different θ-roles that happen to be realised by the same preposition, but that can nonetheless co-occur in the subcategorisation frame of one (complex) predicate, as a rather *ad hoc* solution.

I believe the strongest argument in favour of a biclausal approach, comes from the observation that both dative objects are subject to different constraints regarding cliticisation. Some of these are, to my knowledge, unaccounted for in the earlier literature on causatives in constraint-based frameworks, though they are an important part of the discussion on the case in other frameworks like GB or RG.

[5] This would seem to be restricted to cases like example (6) above, where one of the à-objects is cliticised. But, with appropriate stress, sentences with two prepositional dative phrases appear to be acceptable : "Il a fait envoyer une lettre au président$_{recipient/goal}$ à tous les enfants de sa classe$_{causee}$"

Let's return to our examples (4) to (8) above and start by examining sentence (4). When no prepositional phrase is present, the clitic 'lui' can stand either for the causee or the recipient/goal of 'écrire'. The sentence becomes unambiguous, of course, once one adds a prepositional phrase that stands for the causee as in (5) or (6). It is interesting to note that in (6), though both causee and recipient/goal are realised as dative NPs, the sentence is unambiguous. The clitic can only stand for the causee and the prepositional dative phrase can only stand for the recipient/goal of 'écrire'. In (7), the classical example by Kayne (1975), since the direct object 'Marie' would have to be the causee ('Téléphoner' is intransitive, so only the active causative 'faire$_{1b}$' is applicable), 'lui' could only be the recipient/goal, but strangely this sentence is very clearly ungrammatical although the non-clitic version (8) is perfectly acceptable.

These facts are generally not mentioned or not accounted for in the constraint based grammar literature. For instance, in the HPSG analysis of Abeillé et al. (1998), (4) would be an accepted sentence[6] and they would have to resort to pragmatic factors to rule it out. In Alsina (1996), a footnote mentions a complex series of constraints on dative objects, but without going further into the matter.

The solution I am going to propose here draws on several earlier observations about French causatives. In their foundational work, Hyman & Zimmer (1976) noted that there are two basic semantic types of causatives :

(10) Pierre a fait nettoyer les toilettes **au général**.
Peter made the general clean the toilets.

(11) Pierre a fait nettoyer les toilettes **par le général**.
Peter had the toilets cleaned by the general.

In the first one (10), the causee (dative object) is considered to be the main volitional focus of the agent 'Pierre'. His purpose here is that the general, and no one else, would clean the toilets. In the second example (11), his aim seems to be mainly to get the toilets cleaned, be it by the general or anyone else. Alsina (1992) has shown this semantic distinction to be a very general cross-linguistic one. The two types of causatives have often been called the active and passive causative, respectively, because of the fact that the subcategorisation of the complex predicate seems to match that of the active or passive versions of the embedded infinitive

[6]See discussion about causatives in HPSG below.

verb. In some languages, like English, the embedded verb even bears the active or passive morphology alternatively, as can be seen in the translations of sentences (10) and (11).
Now, if we get back to our examples (4) to (8) and classify them according to this basic distinction we get the following results : (6) to (8) are active causatives, (5) is a passive causative and (4) is either. Strikingly, this classification matches exactly that of the sentences allowing cliticisation of the recipient/goal NP; when the causative is passive, the recipient/goal NP of the embedded infinitive verb can appear as a dative clitic on the causative verb.

How can a bi-clausal approach account for this. First, we want to have different entries for 'faire' in the lexicon. We will consider two basic distinctions between different types. On the one hand, we separate active from passive causatives. On the second hand, we want to keep the well known distinction between transitive-subcategorising and intransitive-subcategorising causative verbs. The latter can be considered to be a sub-distinction between active forms only, since, as mentioned above, the transitive-subcategorising form will realise the causee as a direct object and cannot, therefore, realise it as a prepositional par-phrase.

Second, since we want a bi-clausal f-structure, this means that we will have two local subcategorisation domains, one for the causative predicate and one for the embedded infinitive verb. To account for the merging effect, we have to use argument sharing. The crucial innovation I would like to propose here, is for the causative predicate only to have a limited number of slots for linking the embedded verb's arguments, i.e. at the most an OBJ and an OBJ_{dat}. As a result, some arguments of the infinitive verb will not be structure-shared. If we accept as a general principle of well-formedness (of sets of lexical entries) for causative predicates that they will link their objective grammatical functions (OBJ and OBJ_{dat}) to the highest available functions of the embedded lexical items, we obtain the three partial entries in figure 2. These are only a first approximation. A full coverage will probably require a more elaborate set of lexical entries.

I believe a short digression is in order here. The careful reader will have noticed that the entry for 'faire_2', i.e. the so-called 'passive' version of the causative, does not control any of its objects to the embedded subject as would be expected if it were a passive predicate. The functional equation refers to the embedded direct object. This is due to the fact that I consider the embedded VCOMP to be headed by an infinitive verb without subject, a sort of verb form that is only half way through

Faire$_{1a}$, V : Pred = 'faire'<SUBJ, VCOMP, OBJ, OBJ$_{[dat]}$>
$(\uparrow OBJ = \uparrow VCOMP\ \ OBJ)$
$(\uparrow OBJ_{[dat]} = \uparrow VCOMP\ \ SUBJ)$

Faire$_{1b}$, V : Pred = 'faire' <SUBJ, VCOMP, OBJ>
$(\uparrow OBJ = \uparrow VCOMP\ \ SUBJ)$

Faire$_{2}$, V : Pred = 'faire' <SUBJ, VCOMP, OBJ, OBJ$_{[dat]}$>
$(\uparrow OBJ = \uparrow VCOMP\ \ OBJ)$
$(\uparrow OBJ_{[dat]} = \uparrow VCOMP\ \ OBJ_{[dat]})$

Figure 2: lexical entries

its transformation into a passive form.[7] The subject has been removed, allowing the agent role to be optionally mapped onto an oblique par-phrase, but the object has not been raised to subject position. This of course is a rather bold statement since it directly contradicts the subject-condition of LFG (Bresnan 2001), but I believe a number of facts support the idea. Some GB authors (Zubizaretta 1985, Burzio 1986) consider the embedded infinitive verbs of French causative constructions not to be fully-formed words. One of the reasons for positing this is that the embedded so-called passive verb cannot have passive morphology, as is the case in its English counterpart, while standard VCOMP constructions would accept it :

(12) *Il a fait être envoyé une lettre par Marie.
He had a letter sent by Mary.

(13) Il a voulu être envoyé à Paris pour ses études.
He wanted to be sent to Paris for his studies.

Another point of interest is the fact that the infinitive can, in some cases, appear in a context where there is no possible binder for its subject. This is also a unique feature in infinitive constructions. This is the case for the second reading of the following sentence :

(14) Pierre lui a fait téléphoner.
Peter made him phone (someone). / Peter had (someone) phone him.

[7] I am speaking in 'derivational' terms for the sake of clarity, but the above has to be understood as referring to some declarative system of constraints organising the lexicon. I leave this matter open here.

(15) *J'ai convaincu de lui téléphoner.
I convinced (someone) to phone him.

Considering the embedded verb to be some sort of not fully-formed word in which the patient role has not been promoted to subject, I believe, is quite satisfactory from a purely intuitive point of view. It seems to be coherent with the fact that it can't realise its own clitics. Unfortunately, I won't pursue the subject any deeper here, for lack of space, but it should be noticed that the analysis sketched above does not crucially rely on this assumption. We could replace the entry for 'faire$_2$' by the entry given in figure 3 and make the necessary adjustments to the other entries regarding the constraint on passive, while keeping most of the assumptions argued for in this paper.

Faire$_{2'}$, V : Pred = 'faire' <SUBJ, VCOMP, OBJ, OBJ$_{[dat]}$>
$(\uparrow OBJ = \uparrow VCOMP \ \ SUBJ)$
$(\uparrow A\text{-}OBJ = \uparrow VCOMP \ \ OBJ_{[dat]})$
$(\uparrow VCOMP \ \ PASSIVE = \ +)$

Figure 3: 'faire$_{2'}$'

We do, however, loose a convenient way of explaining the following sentence from Tasmowski (1984):

(16) Il lui a fait parvenir la lettre.
He made the letter reach him.

Compare the above sentence to example (7). In line with what has been said above, I consider that unaccusative verbs can be introduced in causative constructions as subcategorising for a direct object only. This would explain why they can be subcategorised for by 'faire$_2$'.

I give in figures 4 and 5, two f-structures to illustrate how our lexical entries interact with their embedded predicates in accordance with for the facts observed in sentences (4) to (8).

As we see, not all of the embedded verb's arguments are controlled by the causative predicate's arguments. Typically, a par-phrase will only be subcategorised for by the embedded verb, as well as the dative objects of ditransitive verbs

(6') 'Pierre lui fait écrire une lettre à Paul.'

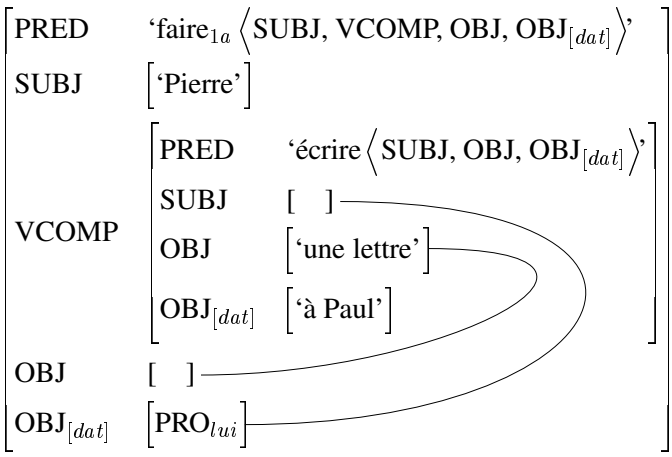


Figure 4: f-structure of an active causative

(5') 'Pierre lui fait écrire une lettre par Marie.'

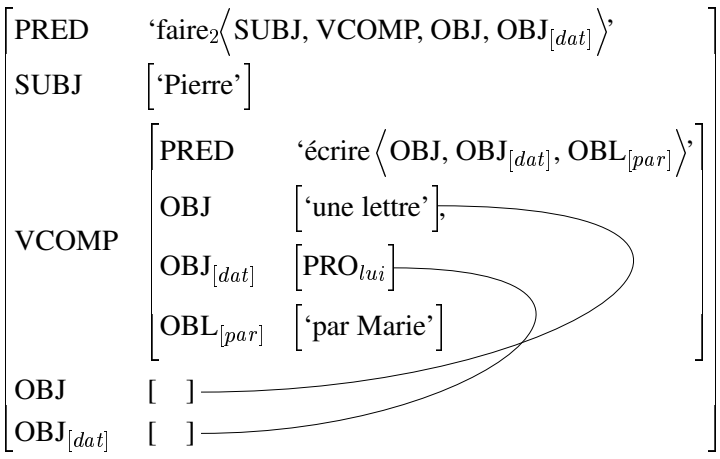


Figure 5: f-structure of a passive causative

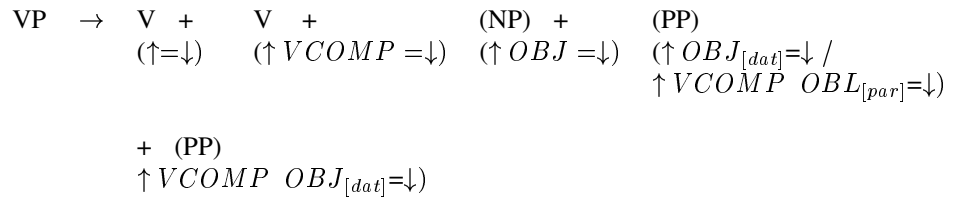


Figure 6: c-structure rule

appearing as complements of faire$_{1a}$. We can now explain the cliticisation facts in a very straightforward manner : Only the functional relations directly subcategorised for by the causative predicate can appear as clitics. This seems trivial, we simply use general rules about clitics in French, nothing else. But it does actually cover all the constraints observed in the previous sentences as the reader can check for him- or herself.

But then, there is still the so-called merging effect observed at the surface level. The fact that in surface realisations, both predicates really appear to have merged their argument structures. The phrase-structural constraints appear not to discriminate between arguments depending on their level of subcategorisation. For instance, the linear order of complements seems shows that they are all part of one single syntactic domain. Although there is a rather strict order for complements in French, this order can be modified in order to emphasize some argument :

(17) Pierre a fait nettoyer par le général toutes les toilettes des casernes.
Peter had the general clean all the toilets in the barracks.

Examples like this one clearly show that there is, at some point, a structural mismatch between an embedded and a flat representation. Since I have argued that this mismatch should not be between a- and f-structure, it has to be between f- and c-structure. So we would probably want to have a complex predicate specific phrase structure rule or set of rules to allow for the correct constructions. This would be something like figure 6.

In definitive, many other phenomena observed about causatives will find the same resolution as in the monoclausal accounts because either they rely on phrase-structural constraints and I have not brought any fundamental change in this area, or they rely on functional relations, but it turns out that they generally affect precisely

the functions that are shared by both arguments. This is, for instance, the case with tough-extraction :

(18) Ce poisson est difficile à faire manger à Paul.
≈ This fish is tough to get Paul to eat.

One apparently problematic case, I believe, rests on a misjudgement. It has sometimes been claimed that the par-phrase could be cliticised by 'lui', in support of the monoclausal theory (Kinouchi 1999):

(19) Il a fait détruire la ville par/?à Jean.
He had the city destroyed by Jean/He made Jean destroy the city.

(20) Il lui a fait détruire la ville.

The active version of (19) is a little awkward because it seems to imply that the important information conveyed by the sentence is that something was done to 'Jean', rather than a city having been destroyed. Kinouchi believes that the absence of awkwardness in the cliticised version implies that 'lui' stands for 'par Jean'. But a closer look at the active version of (19) shows us that it is not that awkward once we understand that it conveys a pragmatically enriched meaning, something like 'He got Jean so mad that he went out and destroyed the whole city!' Strikingly, the cliticised version with 'lui' can only be read with this meaning. For some reason, the presence of the clitic pronoun simply renders it directly accessible. I believe, therefore, we can safely discard the idea that in one very specific case in French (i.e. only causative constructions) a par-phrase can be replaced by a dative clitic.

3 A Short Comparison with HPSG

To support the idea that the previous proposal stems out of empirical rather than purely formal considerations, I will show how the same general ideas can be applied in another constraint-based framework, namely Head-Driven Phrase Structure Grammar (Pollard & Sag, 1994) to produce the same effect.

It is quite difficult to state one-to-one correspondences between different levels of representation in separate grammatical frameworks, because the information is not distributed uniformly at each level. In the case of interest, though, I believe the features containing the relevant information in HPSG are the argument structure list (ARG-ST) and the valence lists (SUBJ and COMPS). The valence lists contain

a number of signs (words or phrases) that get removed from the lists as they are syntactically realised through phrase structure rule. For instance, the COMPS feature of the mother in the HEAD-COMPLEMENT-RULE will have one less element on its COMPS list than its head-daughter. This is how the grammar checks that all and only the right complements are present in the syntax. So the COMPS feature contains information that is directly relevant to surface realisation (something like the equivalent of LFG's functional annotations to c-structure rules.)

ARG-ST, on the other hand is a strictly lexical feature reflecting the internal valence of the predicate. In earlier versions of HPSG, ARG-ST was simply the result of the concatenation of SUBJ and COMPS. The strict equivalence between ARG-ST and valence lists, known as the Argument Realisation Principle (Sag & Wasow 1999) has been rather widely abandoned afterwards. Recent studies (Manning & Sag 1998, Davis & Koenig 2000, Miller & Sag 1997,...) have gone in the way of loosening the tie because some phenomena seem to imply that the isomorphism is not that strong. It is argued that some elements can appear on the ARG-ST list while being absent from the valence lists (pro-drop arguments, clitics,...) and therefore, from phrase structure realisation.

Actually, the information carried by HPSG's ARG-ST is closer to that of LFG's f-structure than to it's a-structure. It is sort of the equivalent of the PRED value and the information carried by the grammatical function features (SUBJ, OBJ,...) It is no surprise then, that the dominant view on causatives in HPSG considers their ARG-ST to be flat and to contain all the arguments of both the causative and embedded predicate. This leads to the same problem as we have observed in the case of LFG approaches.

For Abeillé, Godard, & Sag (1998), causatives are complex predicate obtained through the process of argument composition. They distinguish two types of argument composition : a-composition (where the main predicate adds the embedded predicate's ARG-ST list to its own, as is the case with tense auxiliaries) and c-composition (where only the COMPS list of the embedded predicate is added, as is the case with causatives.[8]) They assume, following Miller & Sag 1997, clitics to be lexical affixes on verbs which are present on the ARG-ST list but not on the COMPS list. Since the dative pronoun in (7) is an element of the embedded predi-

[8]In this approach, the embedded subject only shares its index with the dative NP of active transitive causative constructions. This is therefore the only assumed relational change.

cate's COMPS list,[9] it will be merged onto the ARG-ST of the causative predicate, therefore allowing it to be cliticised. Following the same pattern, since both dative NPs in (6) are present on the ARG-ST of the causative verb, both are predicted to be cliticisable. One can easily see that to account for this we need to be able to distinguish between dative NPs originating on the upper verb's or on the lower verb' ARG-ST. But since the resulting ARG-ST is flat, this is no longer possible.

Now let's apply the same ideas as above to the present framework. We want a biclausal representation at the lexical level to allow us to distinguish between arguments of the causative verb, arguments of the embedded verb and shared arguments. The obvious way of doing this is by simply eliminating the argument composition. Each predicate will have its own ARG-ST fully determined in the lexicon. We also want to allow ourselves the possibility of linking the accusative and dative arguments of the causative to arguments of the embedded verb. This will be done by linking constraints specified in the lexical entries, just as we did in LFG. And finally, we want the complements of both predicate to 'mix' at phrase structure level to render the merging effect. As mentioned above, a good place to state phrase-structural constraints in HPSG is the valence lists. So I will propose that merging takes place in the COMPS list, namely, the causative predicate's COMPS list will be a concatenation of its own complements (i.e. the non-cliticised elements of its ARG-ST excepting the subject) and the COMPS list of the embedded verb. This takes the 'loosening' of the ARP a step further, giving the surface level of representation even more freedom from the lexical level.

An important difference, due to the general architecture of both frameworks, is that in this case, the so-called phrase-structural merging will have to be stated in the lexicon (albeit as an underspecified COMPS list). This is due to the will of classical HPSGers to maintain the number of syntactic rules to a strict minimum, but it is in no way a formal necessity. One could also state construction-specific syntactic rules to account for the COMPS merging. I give the corresponding lexical entries for 'faire$_{1b}$' and 'faire$_2$' in HPSG in figures 7 and 8 and leave it to the reader to verify that they will have the same behaviour as their LFG counterparts.

[9] Abeillé et al. state as a constraint that embedded predicates have to be non reduced. This is necessary to block downstairs cliticisation in c-composition constructions ('*Il a fait le prendre à Paul.'). I assume the same constraint (see Abeillé et al. 1998 for technical details.)

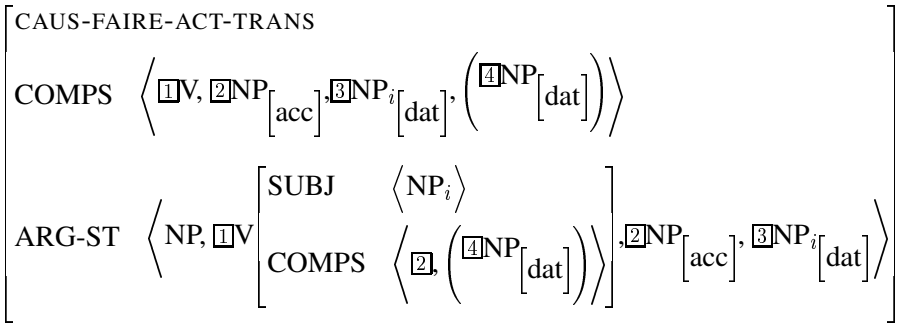


Figure 7: 'Faire$_{1a}$' - HPSG

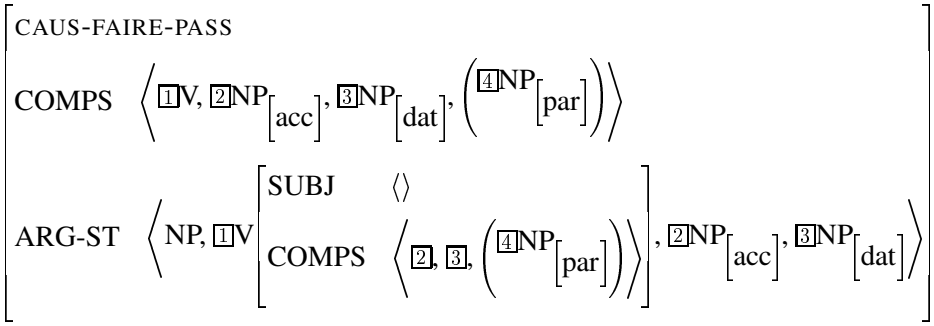


Figure 8: 'Faire$_2$' - HPSG

4 Conclusion

In this paper, I have argued that the generally accepted monoclausal conception of causatives is not well suited to account for the series of constraints revealed in sentences (4) to (8) above. I have shown that these can be easily solved if one admits that both the causative 'faire' and the embedded infinitive have their own local subcategorisation frames and that some but – crucially – not all of the subcategorised functions are shared. The present account arguably requires less technical machinery and formal innovations than its predecessors. It can be summarised in a few general principles:

- The causative predicate 'faire' takes as arguments, a subject, a infinitive verb phrase and at the most two objects (a direct and an indirect object).

- It's objective arguments are structure-shared with the most highly ranked functional relations of the embedded infinitive verb.

- The complements of both predicate are joined in a single domain at phrase structure level (c-sructure). This amounts to the arguments of 'faire', plus the possible extra arguments of the embedded verb that are not controlled by any of the arguments of the causative predicate.

- Only the local grammatical functions of 'faire' can cliticise. The remaining grammatical functions on the infinitive verb ($\text{OBJ}_{[dat]}$ and $\text{OBL}_{[par]}$) will still surface in a flat c-structure as complements of the complex predicate.

To support the empirical argument behind this claim, I have compared the LFG treatment with a similar one in Head-driven Phrase Structure Grammar, where the same pre-formal considerations led me to switch the 'merging point' from the argument structure (ARG-ST) as in Abeillé et al. (1998), to a level of information content that has a closer tie to phrase-structural constraints, namely the COMPS feature.

This paper presents a rather general approach, which needs to be refined by the design of a much more elaborate set of lexical entries. It does not cover more complex case, where unspecified objects have been removed for instance. Constrasts in behaviour between unaccusative and unergative verbs also need more indepth analysis. But I believe the paper makes one point at least, that is that the biclausal approach to French causatives may have been somewhat underestimated and hastily discarded.

References

ABEILLE, A., D. GODARD, I. SAG (1998) Two kinds of compostion in French complex predicates. In E. Hinrichs, T. Nakazawa & A. Kathol, eds. *Complex Predicates in Nonderivational Syntax*. New York:Academic Press. 1-41.

ACKERMAN, F. & G. WEBELHUTH (1997) The Composition of Discontinuous Predicates: Lexical or Syntactic.

ALSINA, A. (1992). On the Argument structure of causatives. *Linguistic Inquiry*. 23(4). 517-555.

ALSINA, A. (1996) *The role of argument structure in syntax: Evidence from Romance*, Stanford:CSLI publications.

BRESNAN, J (2001) *Lexical Functional Syntax*. Oxford: Blackwell.

BURZIO, L. (1986) *Italian Syntax*. Dordrecht: Reidel.

DAVIS, A. & P.-J. KOENIG (2000) Linking as constraints on word classes in a hierarchichal lexicon. *Language*. 76. 56-91.

FAUCONIER, G. (1983) Generalized Union. In L. Tasmowski & D. Willems, eds. *Problems in Syntax*. New York: Plenum. 159-229.

FRANK, A. (1996) A note on complex predicate formation. In M. Butt & T. H. King, eds. *On-line proceedings of the first LFG conference*. http://www-csli.stanford.edu/plublications/LFG1/

HYMAN, L. & K. ZIMMER (1976) Embedded Topic. In C. N. Li, ed.*Subject and Topic*. New York: Academic Press. 189-211.

KAYNE, R. (1975) *French Syntax: the transformational cycle*. Cambridge, MA: The MIT Press.

KINOUCHI, Y. (1999) *Cas syntaxique et cas sémantique en français et en japonais: Quelques remarques sur la Grammaire Relationnelle*. Ph. D. Dissertation, Université de Paris VIII.

MANNING, C. & SAG, I. (1995). Dissociation between Argument Structure and Grammatical Relations. In G. Webelhuth, J.P. Koenig & A. Kathol, eds. *Lexical and Constructional Aspects of Linguistic Explanation*. Standford: CSLI Publications. 63-78.

MILLER, P. & I. SAG (1997) French clitic movement without clitics or movement. *Natural Language and Linguistic Theory* 15, 573-639.

POLLARD, C. & SAG, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press and Stanford: CSLI Publications.

RUWET, N. (1972) *Théories syntaxiques et syntaxe du français*. Paris: Ed. du Seuil.

SAG, I. & WASOW, T. (1999). Syntactic Theory: A Formal Introduction. Stanford: CSLI Publications. TASMOWSKI, L. (1984) ?*Lui faire téléphoner quelqu'un d'autre: une stratégie?. *LinguisticæInvestigations* 8.2. 403-427.

ZAENEN, A. & M. DALRYMPLE (1996) Les Verbes causatifs "polymorphiques": les prédicats complexes en français. *Langages* 30. 79-95.

ZUBIZARETTA, M. L. (1985). The relation between Morphophonology and Morphosyntax: The Case of Romance Causatives. *Linguistic Inquiry* 16.2. 247-289.

Subsumption and Equality:
German Partial Fronting in LFG

Annie Zaenen and Ronald M. Kaplan

Palo Alto Research Center

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002

CSLI Publications

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

Subsumption and Equality: German Partial Fronting in LFG*

Annie Zaenen and Ronald M. Kaplan
Palo Alto Research Center
{Zaenen|Kaplan}@parc.com

1. Introduction

In a previous paper (Zaenen and Kaplan, 1995; henceforth ZK) we developed a general LFG account of West Germanic sentence structure, concentrating on the order of nominal arguments in the Vorfeld and Mittelfeld. The account was based on the interactions between functional uncertainty equations, functional precedence constraints, and phrase structure rules. In Kaplan and Zaenen (forthcoming) we develop our account of the verbal complex. In the present paper we concentrate on another specific problem, partial VP topicalization in German. We extend our previous accounts by showing how subsumption relations rather than equality can model some important properties of partial VP fronting and also the differences in behavior of equi and raising complements in fronted and extraposed environments. We observe more generally that the subsumption relation in LFG and perhaps also in other constraint-based frameworks can provide insightful characterizations of asymmetrical phenomena in natural language that are otherwise difficult to describe.

2. Partial fronting phenomena

As has often been observed, German sentences like the following are grammatical:[1]

(1) a. Das Buch zu geben schien Hans dem Mädchen.
The book to give seemed Hans the girl.

b. Dem Mädchen zu geben schien Hans das Buch.
The girl to give seemed Hans the book.

c. Zu geben schien Hans dem Mädchen das Buch.
to give seemed Hans the girl the book.

* We gratefully acknowledge Stefan Riezler and Anette Frank for providing some of the data we discuss in this paper. We also thank Miriam Butt, Mary Dalrymple, and Detmar Meurers for helpful comments on an earlier draft. The usual disclaimers apply.

[1] The acceptability of this type of sentence varies and depends heavily on discourse factors, which we ignore here.

d. Dem Mädchen das Buch zu geben schien Hans.
The girl the book to give seemed Hans.

'Hans seemed to give the girl the book.'

In (1a) the verb is topicalized together with one of its non-subject arguments, *Das Buch*, in (1b) the other argument is chosen. (1c) gives a version in which only the verb is topicalized. The fourth version exemplifies the fronting of the complete VP. Following the literature, we will refer to versions (1a) to (1c) as instances of partial VP fronting (henceforth PVPF).

Early discussions of PVPF focused mainly on examples in which the verb was fronted together with one or more of its non-nominative dependents, as in (1). But it has long been observed (Uszkoreit, 1987) that the subjects of unaccusative verbs can also be fronted. More recent literature starting with Haider (1990) has also drawn attention to PVPF sentences with unergative subjects. Both cases are exemplified in (2):

(2) a. Ein Fehler unterlaufen ist ihr noch nie.
An error happened-to is her still never
'Until now she has never made a mistake.'

b. Ein Aussenseiter gewonnen hat hier noch nie.
An outsider won has here still never.
'Until now no outsider has won here.'

PVPF has mainly attracted attention because the Vorfeld is occupied by material that cannot always be a constituent in the Mittelfeld. In some theories, for instance, in transformational ones, this presents a problem. In the next section (Section 3) we will extend to German the traditional LFG approach to topicalization and show that that particular problem does not arise. In HPSG, the phenomenon has attracted attention because it bears on the way the Mittelfeld is structured (see e.g. Nerbonne, 1995) and the way argument saturation works (see e.g. Müller, 1999). The assumptions we have made in earlier papers about West Germanic word order have as a consequence that the structure of the Mittelfeld is irrelevant for our analysis, but the issue of how to state the constraints on argument saturation do arise. In Section 4 we lay out a subsumption-based analysis of partial verb phrase fronting and show how it improves on a traditional LFG account. In Section 5 we discuss the interaction between PVPF and raising and control constructions. We conclude with a short discussion of the different roles of subsumption and equality in linguistic modeling.

3. Topicalization in LFG

Topicalization, like other long distance phenomena, is modeled in LFG through functional uncertainty (Kaplan & Zaenen, 1989). Functional uncertainty is a straightforward extension to the basic mechanism for describing simple functional relationships in LFG. A basic equation such as (↑ XCOMP) = □ appearing in a phrase-structure rule is satisfied just in case the f-structure corresponding to the mother node of the c-structure expansion (the f-structure denoted by ↑) has

an XCOMP attribute whose value is the f-structure corresponding to the daughter node of the expansion (the □ f-structure).

The problem with long distance dependencies is that the relationship between two f-structures is not determined uniquely by the positions of the phrasal constituents to which they correspond. Consider the topicalized sentences in (3):

(3) a. Mary John likes.
b. Mary John says that Bill likes.
c. Mary John says that Bill believes that Henry likes.
d. Mary John says that …

In the first one *Mary* is understood both as the TOP(ic) of the sentence and also as the OBJ of *likes*. An equation (↑ OBJ) = □ associated with the fronted *Mary* NP would properly characterize this within-clause relationship. In the second sentence *Mary* is still understood as the object of *likes*, but *likes* is now the predicate of a complement of the higher verb *says*, and the appropriate annotation for defining *Mary*'s within-clause function would be (↑ COMP OBJ) = □. For the third sentence the equation would be (↑ COMP COMP OBJ) = □, and in general for every additional level of embedding that might happen to be in the main clause, the path of functions appropriate for *Mary* would be lengthened with an additional COMP. The uncertainty in how to annotate the fronted NP comes from the fact that there is no information available at its surface position to determine exactly which of these possible equations correctly captures its functional relationship to the embedded clause.

Functional uncertainty provides a simple way of defining a family of equations while still leaving open the choice of exactly which member of the family will turn out to be consistent with an embedded f-structure. For this particular construction, the equations in the family all have functional paths that belong to the regular language COMP* OBJ, and the infinite family of appropriate equations can be specified in the single equation (↑ COMP* OBJ) = □. In the general case, suppose that *f* and *g* are f-structures and that □ is an expression denoting a regular language of functional paths. Then we assert that

(4) $(f\ \square) = g$ holds if and only if $(f\ x) = g$ holds for some string x in the language □.

Kaplan and Zaenen (1989) give a somewhat more precise definition and discuss an initial set of linguistic applications for this device; Kaplan and Maxwell (1988) show that it has attractive mathematical and computational properties.

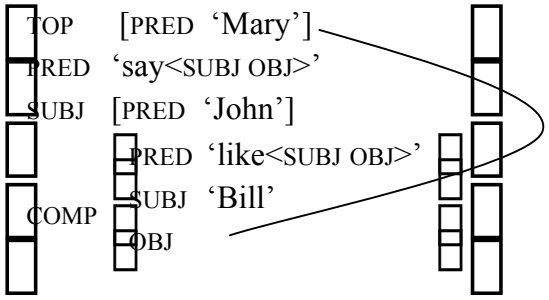
Within this framework we see that the English topicalization patterns in (3) can all be derived by means of the c-structure expansion and the uncertainty equation in the following rule:[2]

[2] It is customary in LFG rules not to specify the ↑ = □ equations that identify the heads and coheads of a construction, letting those be the unmarked relations between mother and daughter f-structures. In this paper we do not follow that abbreviatory convention and instead explicitly mark all f-structure connections so that the flow of information is easier to discern.

(5) S' → NP S
(↑ TOP) = ↓ ↑ = ↓
(↑ COMPS* NGF) = ↓

Here we use COMPS as an abbreviatory symbol that ranges over COMP and XCOMP, and NGF ranges over the usual set of nominal grammatical functions (SUBJ, OBJ, OBJ2,…). This rule provides the appropriate f-structure for sentence (3b), for example, as shown in (6):

(6)



The linking line in this diagram indicates that the values of the TOP and COMP OBJ functions have exactly the same attributes and values, including the same instantiations of the semantic-form PRED values (Kaplan & Bresnan, 1982).

The S' rule for German topicalization is an elaboration of the rule in (5) and also of the S' rule we previously proposed for Dutch, rule (29) of Zaenen and Kaplan (1995). There is little basis in German for making a categorial distinction between S and VP, since in German nominative subjects can appear interspersed with non-nominative NP's. For German we therefore conflate S and VP into a single category which, for want of a better name, we will call S|VP. S|VP has expansions that permit all possible grammatical functions. Our German rule (7) specifies this category for the Mittelfeld position and also includes this as one of the possible expansions in the fronted Vorfeld position. Also, the Vorfeld and Mittelfeld are separated by a tensed verb in second position.

(7) S' → XP V (S|VP)
↑ = ↓ ↑ = ↓
(↑ TENSE)

where XP = { NP | S|VP | …}
(↑ TOP) =↓ (↑ TOP) =↓
(↑ COMPS* NGF) = ↓ (↑ XCOMP* XCOMP) = ↓

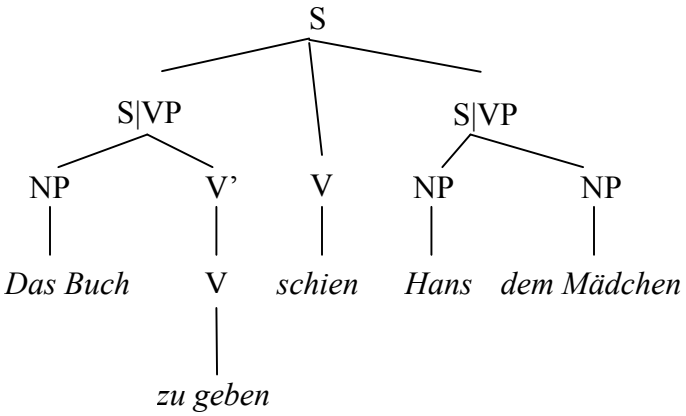
The S|VP realization of the XP is of course what allows for the topicalization of (S|)VP's in the Germanic languages. The partial fronting found in (1) and (2) arises from the optionality of the constituents in the S|VP, as we have also discussed in previous papers. In those papers the phrase structure rules we proposed were for Dutch; we repeat them here with some trivial adaptations for German. Rule (8a) corresponds to (27) in Zaenen and Kaplan (1995) and (8b) is a German variant of their (13).

(8) a. S|VP → NP* (V') (S|VP)
(↑ COMPS* NGF) = ↓ ↑ = ↓ (↑ XCOMP* COMP) = ↓

b. V' → (V') (V)
(↑ XCOMP) = ↓ ↑ = ↓
(↑ XCOMP^{+} NGF) $\prec_f$ (↑ NGF)

Rule (8a) allows for any number (including zero) of NP's coming before the optional V' cluster of verbs, and it permits an optional post-verbal complement. The functional role of individual NP's is expressed by the associated uncertainty equation; this allows each NP to be a nominal function of a verb at any depth of COMPS (XCOMP or COMP) complement embedding. Arguments for a verb cluster in German are given in Hinrichs and Nakazawa (1995). The verb cluster is optional in this rule because the f-structure head may also be realized as the tensed second-position verb of rule (7).[3] Rule (8b) provides a left-branching expansion for German verb clusters, with each verb serving as the XCOMP head of the verb immediately to its right.[4] The c-structure that these rules provide for sentence (1a) is sketched in (9), and the associated f-structure is shown in(10):

(9)

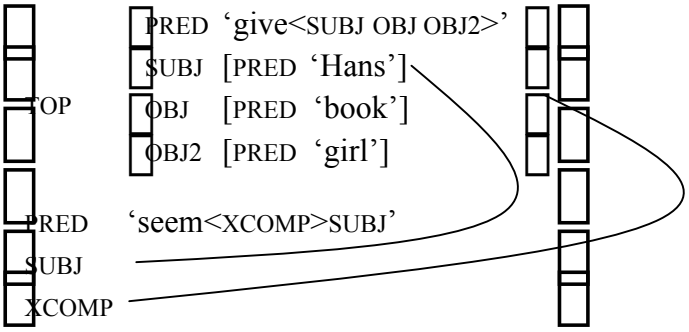


[3] Note that even though all the constituents in (8a) are optional, the LFG prohibition against any empty nodes (Kaplan & Zaenen, 1989) means that at least one daughter must appear in any expansion of S|VP. Thus the optionality of the Mittelfeld S|VP in (7) is necessary to allow for sentences with no post-verbal constituents, such as (i):

(i) Hans läuft.
Hans runs.

[4] The f-precedence constraint in (8b) imposes appropriate ordering constraints on the NP arguments for all the verbs in an XCOMP hierarchy. Nominal word order and order within the verb cluster were the major foci of earlier papers (Zaenen & Kaplan (1995); Kaplan & Zaenen (in press), but they are not relevant to the present discussion. We do not comment on word-order constraints in the remainder of the present paper.

(10)



For this sentence the uncertainty path chosen for the Vorfeld in the S' rule (7) consists of a single XCOMP, and the outer linking-line indicates that the TOP and XCOMP have identical internal functions and features. The uncertainties for the NP's in the fronted S|NP expansion resolve to the singleton paths OBJ and OBJ2, and the uncertainty for *Hans* in the Mittelfeld S|VP resolves to SUBJ. The inner linking-line marks the fact that *scheinen* 'to seem' is a raising verb, and thus its lexical entry includes a standard functional-control equation that identifies its SUBJ with the SUBJ of its XCOMP, as shown in (11):

(11) scheinen V (↑ PRED) = 'seem<(↑ XCOMP)>(↑ SUBJ)'[5]
(↑ SUBJ) = (↑ XCOMP SUBJ)

Because of the equality relations on the topicalized S|VP, the functional information in the fronted constituent combines with the information in the Mittelfeld S|VP and the lexical equation of functional control, and the result is an f-structure that satisfies the Completeness and Coherence requirements of *geben*.

These are relatively simple rules, but they account for a surprising amount of the syntactic data. The subcategorization requirements for all the sentences in (1) are satisfied, even though those sentences receive quite different c-structures. And because they do not incorporate a distinguished and obligatory position for the German subject, allowing the subject to appear among any of the nominals in the S|VP, these rules also provide appropriate analyses for the examples in (2). The only kind of partially fronted VP's that these rules systematically exclude are examples that are in fact ungrammatical. One such example is shown in (12).

(12) *Müssen wird er ihr ein Märchen erzählen.
must will he her a story tell.
'He will have to tell her a story.'

In this example *erzählen* must be assigned as the head of the XCOMP of the matrix verb *wird*, because the V' cluster has no internal uncertainty equations. Resolving the uncertainty COMPS* to the empty sequence causes *müssen* also to be assigned as the XCOMP head, and the result is an

[5] Recall that, according to standard LFG conventions, the angled brackets around the XCOMP indicate that it is a semantic argument whereas the fact that the SUBJ is outside the brackets means that it is a non-semantic grammatical function. The Semantic Completeness condition (Kaplan & Bresnan, 1982) requires the f-structure of a semantic argument to have its own semantic-form PRED value.

inconsistency with *erzählen*. But if the uncertainty is resolved to any longer sequence of XCOMP's or COMP's the resulting f-structure will be incoherent, since *erzählen* does not subcategorize for either of those functions. Other ungrammatical sentences are illustrated in (13) and (14). These are disallowed because they violate the Coherence and Completeness conditions respectively.

(13) * Dem Mädchen ein Märchen wird er ihr erzählen.
The girl a story will he her tell
'He will tell the girl her a story.'

(14) * Dem Mädchen gegeben hat er.
The girl (Dat) given has he
'He has given the girl.'

We note that under our analysis there are no problems with differences in the content of constituents in the Vorfeld and the Mittelfeld, since all S|VP elements are optional. Our proposal, however, does have two drawbacks. First, it allows for sentences such as (15), where two arguments of a verb have been fronted without their verb. This is allowed because the same c-structure expansions are possible for S|VP in both the Vorfeld and the Mittelfeld, and thus the V', and hence the V, is not required in the fronted constituent.

(15) * Ihr ein Märchen wird er erzählen.
Her a story will he tell
'He will tell her a story.'

Second, the proposal does not record at the f-structure level which parts of the S|VP are topicalized and which ones are not.[6] Whether this is important or not depends on one's view of the interaction between the f-structure and other modules of linguistic information: for argument structure and purely syntactic wellformedness conditions, this information is not important. But if we assume that there is a discourse-structural difference between the various versions of (1) and that the discourse structure is read off the f-structure without separate input from the c-structure (and even without covert c-structure information via inverse correspondence relations), the account given is inadequate. In the next section we revise our account so that it does distinguish the topicalized from the untopicalized grammatical functions. This revision also solves the verb-less fronting problem illustrated in sentence (15).

4. A Subsumption Analysis of Topicalization

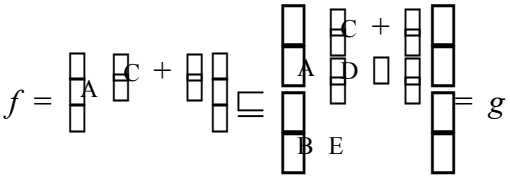
Most current constraint-based theories of syntax use an equality relation (or its unification equivalent) as the main device for combining information from various surface locations via path specifications. Equality is a symmetric relation and thus can provide no account for the often remarked-upon asymmetry, or even anti-symmetry, of syntactic relations. Our theoretical framework makes available a subsumption relation in addition to equality, and subsumption permits us to model asymmetric syntactic dependencies. In this section we review the difference

[6] Other cases of the same and related problems of representation are discussed by King (1997).

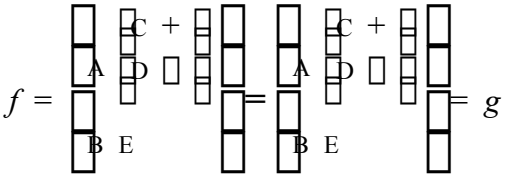
between subsumption and equality and then show how subsumption constraints can be used to solve the two residual problems we have just described.

Informally, subsumption establishes an ordering relation between two units of information, stating that the one subsuming the other contains less information (or is less specific or more general) than the one that is subsumed. A formal definition of the subsumption relation between two functional entitites *f* and *g*, notated as $f \sqsubseteq g$, is given in (16), along with an illustration of two f-structures that satisfy the relation. For comparison we show in (17) a parallel formal definition of equality. Dom(*f*) in these definitions denotes the domain of the f-structure *f*, the set of all its attribute-symbols.

(16) Definition of Subsumption: $f \sqsubseteq g$ iff
f and *g* are the same symbol or semantic form, or
f and *g* are both f-structures, Dom(*f*) $\subseteq$ Dom(*g*), and $(f\ a) \sqsubseteq (g\ a)$ for all $a \in$ Dom(*f*), or
f and *g* are both sets and every element of *f* $\sqsubseteq$ some element of *g*



(17) Definition of Equality: $f = g$ iff
f and *g* are the same symbol or semantic form, or
f and *g* are both f-structures, Dom(*f*) = Dom(*g*), and $(f\ a) = (g\ a)$ for all $a \in$ Dom(*f*), or
f and *g* are both sets, every element of *f* = some element of *g* and every element of *g* = some element of *f*.



We note in passing that subsumption is the more primitive relation, since equality can also be defined as the symmetric combination of a subsumption with its inverse:

(18) $f = g$ iff $f \sqsubseteq g$ and $g \sqsubseteq f$ (symmetry)

Returning to the analysis of partial VP fronting, let us examine the rules and simplified representations for the variant given in (1a), repeated here for convenience:

(1)	a.	Das	Buch	zu	geben	schien	Hans	dem	Mädchen.
		The	book	to	give	seemed	Hans	the	girl.

'Hans seemed to give the girl the book.'

We revise the fronted S|VP expansion of the S' rule given in (7) by replacing its uncertainty equation with a subsumption relation, as shown in (19). For the sake of concreteness, we give in (20a) the particular instance of this uncertainty necessary for example (1a). In (20b) we show the expansion of rule (8a) that derives the Vorfeld S|VP, and (20c) derives the Mittelfeld S|VP for this example.

(19) S' → S|VP V S|VP
(↑ TOP) = ↓ ↑ = ↓ ↑ = ↓
↓ ⊑ (↑ XCOMP* XCOMP) (↑ TENSE)

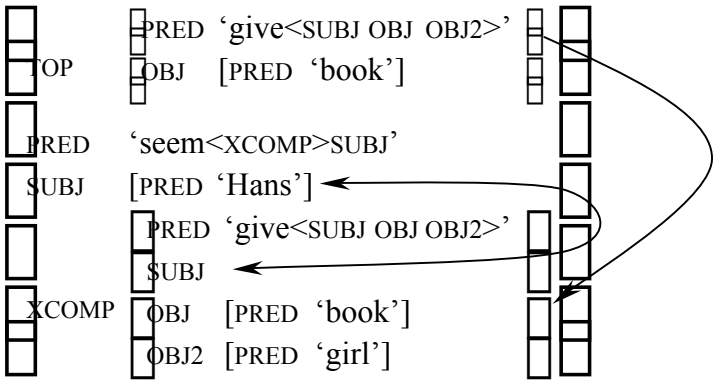
(20) a. S' → S|VP V S|VP
(↑ TOP) = ↓ ↑ = ↓ ↑ = ↓
↓ ⊑ (↑ XCOMP) (↑ TENSE)

b. S|VP → NP V'
(↑ OBJ) = ↓ ↑ = ↓

c. S|VP → NP NP
(↑ SUBJ) = ↓ (↑ OBJ2) = ↓

With the subsumption constraint instead of equality in rule (19), the f-structure in (21) instead of the one in (10) is assigned to sentence (1a).

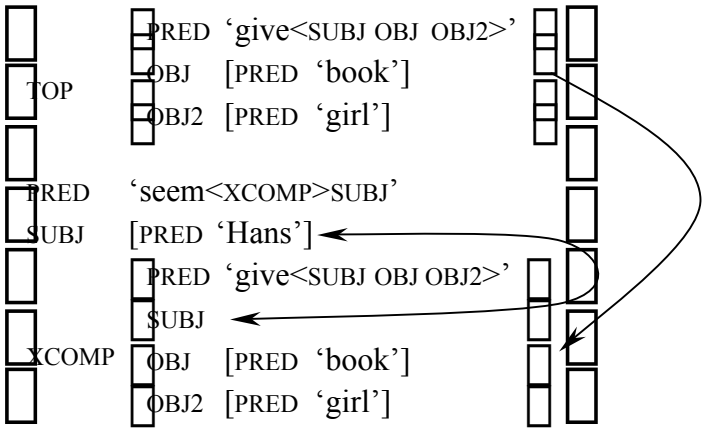
(21)



Here we have placed an arrow on the linking line between the TOP and XCOMP f-structures to indicate the asymmetric flow of information specified by the subsumption constraint: all properties of the TOP f-structure (including the particular instantiation of the semantic form) follow the arrow and appear in the XCOMP, but not vice versa. In constrast, the two arrows on the line linking the SUBJ and XCOMP SUBJ represent the symmetric flow of information between these equated f-structures, and in this case we continue the abbreviatory convention of not displaying the identical properties of the two f-structures (Kaplan & Bresnan, 1982). As can be seen from this diagram, the subsumption relation insures that the information from the topicalized position combines with additional Mittelfeld information to form the XCOMP f-structure, just as in our initial equality-based proposal. But now the TOP and XCOMP values are kept distinct, and f-structures produced under the subsumption analysis thus clearly show which properties of the

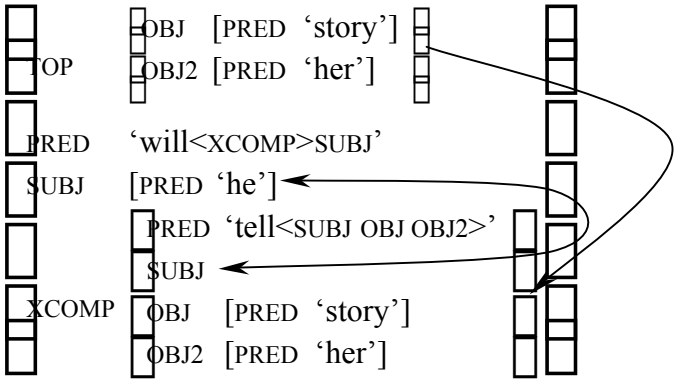
XCOMP have been topicalized. This can be seen by comparing (21) with (22), the f-structure for (1d):

(22)



The subsumption relation also has the effect of ruling out sentences such as (15), where two nominals appear without a verb in the fronted S|VP position, as shown in (23):

(23)



Without the verb the TOP f-structure for this example contains two governable but ungoverned functions and the Coherence condition is therefore not satisfied.

Subsumption does have one undesirable consequence, however: the TOP f-structures for grammatical PVPF sentences such as those in (1) now do not contain all the functions required by their PREDs and thus would be incomplete. This is a technical difficulty that we remedy by extending the definition of Completeness of Kaplan and Bresnan (1982) to one that it is sensitive to subsumption relations:

(24) An f-structure *g* is complete if and only if each of its subsidiary f-structures is either locally complete *or subsumes a subsidiary f-structure of g that is locally complete*.

As specified by Kaplan and Bresnan, an f-structure is locally complete if it contains all the governable functions that its predicate governs.

5. The interaction between PVPF and raising and equi

Whereas the account above replaces equality with subsumption to model the basic patterns of partial verb phrase fronting, it is interesting to look at the way PVPF interacts with equi and raising, other phenomena that are also traditionally modeled with equality relations. Meurers and de Kuthy (2001) discuss the following contrast:

(25) a. * Ein Aussenseiter zu gewinnen versuchte hier noch nie.
An outsider to win tried here still never
'An outsider never tried to win here.'

b. Ein Aussenseiter zu gewinnen schien hier eigentlich nie.
an outsider to win seemed here actually never
'An outsider never actually seemed to win here.'

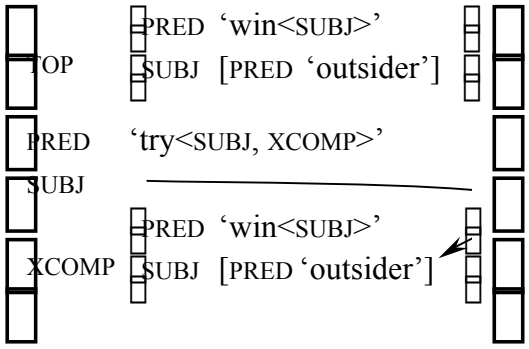
Meurers and de Kuthy attribute this contrast to the difference between equi (25a) and raising (25b) constructions.

In traditional LFG accounts the lexical entries of both equi and raising predicates contain an equation of functional control (such as (↑ SUBJ) = (↑ XCOMP SUBJ) in (11)) that identifies the controller's matrix grammatical function (SUBJ in the case of subject-raising or subject-equi verbs such as *scheinen* and *versuchen* 'to try') with the subject of the complement. The main difference between equi and raising predicates is that for an equi verb the matrix controller is a semantic argument, indicated by the appearance of its grammatical function inside the brackets of the semantic form, while the controller of a raising verb is non-semantic, indicated by its appearance outside the brackets (cf. (11). This difference is not enough to explain the contrast in (25). But we can account for this contrast quite easily by using subsumption instead of equality for the control relation of equi but not raising verbs, as shown in the following lexical entry for *versuchen*:

(26) versuchen V (↑ PRED) = 'try <(↑ SUBJ) (↑ XCOMP)>'
(↑ SUBJ) $\sqsubseteq$ (↑ XCOMP SUBJ)

For sentence (25a) this gives rise to the information dependencies diagrammed in (27):

(27)



This shows that any information defined by the matrix subj will also appear in the xcomp subj, in accordance with the subsumption constraint. But information does not flow in the opposite direction, so for this sentence the matrix subj in fact has no information at all. The top-level f-structure is therefore incomplete, and the sentence is ungrammatical. On the other hand a sentence like (28), in which the subj is fronted as an NP and not part of an xcomp, will receive the coherent and complete f-structure (29) because the properties of the matrix subj do flow down to the xcomp.7

---

[7] Note that the use of subsumption does not solve the well-known problem with case agreement in equi constructions exemplified in (i) (adapted from Berman, 1999):

(i) Ich habe den Burschen geraten, einer nach dem anderen zu kündigen.
I have the boys(D) advised one(N) after another to quit.
'I have advised the boys that they one after the other quit.'

This is an example of a second-object equi construction, and we see that although the controller of the embedded subject is in the dative, the adverbial phrase *einer nach dem anderen* that presumably agrees with the embedded subject is in the nominative. The controller and the embedded subject thus do not share their case values in equi constructions. The proper behavior is characterized by the following lexical entry:

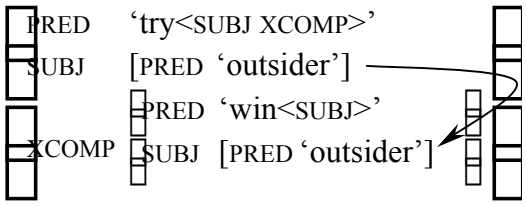
(ii) raten V (↑ PRED) = 'advise< (↑ SUBJ) (↑ OBJ2) (↑ XCOMP)>'
(↑ OBJ2)/CASE ⊑ (↑ XCOMP SUBJ)/CASE
NOM ∈ (↑ XCOMP SUBJ CASE)

Here we have used the restriction operator of Kaplan and Wedekind (1993) to amend the subsumption relation. The effect is that the CASE of the controller, unlike all other features and functions, does not flow down to the embedded subject. Instead the embedded subject is specified explicitly as being nominative by the last constraint; we treat CASE as a set-valued feature in accordance with the Dalrymple and Kaplan (2000) account of feature indeterminacy. This constraint most likely follows from a general convention that identifies nominative as an unmarked or default value for German case.

An alternative solution is to apply the subsumption relation to (↑ OBJ2 PRED) and (↑ XCOMP SUBJ PRED) and thereby enforce sharing only of the semantic form. This would also eliminate the matching requirement on the case feature, as well as the sharing of all other features and values. Presumably the instantiation of the semantic form provides sufficient information for proper semantic interpretation. Which solution is most plausible therefore depends in part on on the treatment of contrasts such as the following:

(iii) Ich versuche mich/*sich zu waschen.
I tried myself/*oneself to wash
'I tried to wash myself.'

(28) Ein Aussenseiter versuchte hier zu gewinnen.
An outsider tried here to win.
'An outsider tried to win here.'

(29) 

This solution locates the ungrammaticality of (25a) precisely in the relation for equi between the SUBJ and XCOMP SUBJ, without appealing directly to any special characteristics of the PVPF construction. The ungrammaticality follows because fronting the subject along with the verb puts it unmistakably in a c-structure position within the complement clause. Its within-clause grammatical function is assigned by virtue of its position in the embedded clause, but this does not establish any connection to the matrix predicate. This solution accounts for a wider range of data, as shown in (30). These examples involve the two-place complement verb *gefallen* 'to please' so that we can observe the difference in behavior between subject and non-subject complement functions.

(30) a. Ein Student versuchte dem Professor noch nie zu gefallen.
A (N) student tried the (D) professor still never to please

b. Dem Professor versuchte ein Student noch nie zu gefallen.
The (D) professor tried a (N) student still never to please.

c. * Ein Student zu gefallen versuchte dem Professor noch nie.
A (N) student to please tried the (D) professor still never

The PRED-subsumption solution will work if we consider the person agreement here to be semantic; if we see it as syntactic, the restriction solution seems more appropriate.

An anaphoric-control account of equi, as proposed by Andrews (1982) for Icelandic and commonly used in other LFG analyses, is also a way of avoiding the case mismatch. We can see this as similar to the PRED-subsumption solution except that the equi verb provides a 'PRO' value as the PRED of the embedded subject rather than the controller's semantic form. The instantiation of the explicitly specified 'PRO' rules out sentences such as (iv), but additional principles are necessary to insure that this particular 'PRO' is anaphorically linked to the matrix controller (cf. Chapter 12 of Dalrymple, 2001).

(iv) * Ich habe den Burschen geraten, sie zu kündigen.
I have the boys advised they to quit.
'I advised the boys to quit.'

Berman (2001) and Dalrymple and Kaplan (2000) suggest that the nominative case of the adverbial in (i) is due to the strong correlation in German between case and grammatical function. On this view the adverbial does not agree with the case value in the subject's f-structure; instead, it appears as nominative because that is the case associated with the grammatical function (SUBJ) that it modifies. This solution will not work for instances of quirky case in Icelandic, and it also fails in German raising examples, as discussed below in footnote 10.

d. Dem Professor zu gefallen versuchte ein Student noch nie.
The (A) professor to please tried a (N) student still never

e. * Noch nie hat dem Professor versucht, ein Student zu gefallen.
Still never has the (D) professor tried a (N) student to please

f. Noch nie hat ein Student versucht, dem Professor zu gefallen.
Still never has a (N) student tried the (D) professor to please

'Until now a student never tried to please the professor.

Our subsumption solution does not predict ungrammaticality when either subjects or non-subjects are fronted as NP constituents unaccompanied by the verb, as seen in (30a-b) (and also (28) above). These are accounted for by the NP realization of the XP in rule (7), where the uncertainty there is resolved to either SUBJ (30a) or XCOMP OBJ (30b). Sentence (30c) resembles (25a) in that the complement subject is fronted along with the complement verb, and it is also ungrammatical. In contrast, when the object and verb are fronted together, as in (30d), the sentence is quite acceptable. This is because the complement object does not need to bear any particular relation to the matrix verb. Sentences (30e-f) are instances of extraposition, not topicalization, but they show a similar pattern. Here also the postposed NP's belong to S|VP of the embedded clause and their overt position assigns them the within-clause function. For the postposed subject the subsumption relation does not allow satisfaction the Completeness condition for the matrix, but there is no violation for postposted nonsubjects.

We now turn to the grammatical raising example (25b). In raising constructions the subject of the complement clause bears no semantic relation to the matrix verb, but an array of standard arguments shows that it does bear a syntactic relationship to the matrix. For example, one well-known argument is based on subject-verb agreement facts as illustrated in the contrast between (2a-b), repeated here, and (31a-b).[8] In spite of the fact that the subject appears overtly with the verb in the embedded clause, the higher verb agrees with it. This shows that it functions also as the matrix subject.

(2) a. Ein Fehler unterlaufen ist ihr noch nie.
An error-sg happened-to is-sg her still never
'Until now she has never made a mistake.'

b. Ein Aussenseiter gewonnen hat hier noch nie.
An outsider-sg won has-sg here still never.
'Until now no outsider has won here.'

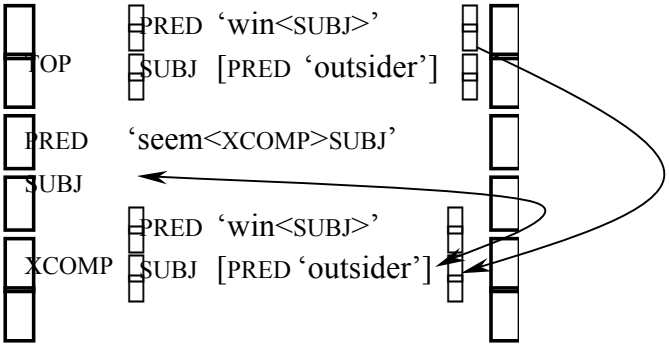
(31) a. * Manche Fehler unterlaufen ist ihr noch nie.
Many error-pl happened-to is-sg her still never
'Until now she has never made many mistakes.'

[8] We assume here that *sein* 'to be' and *haben* 'to have' are raising verbs.

b. * Ein Aussenseiter gewonnen haben hier noch nie.
An outsider-sg won have-pl here still never.
'Until now no outsider have won here.'

The way the raising construction works in traditional LFG analyses was sketched above in the lexical entry, rules and diagrams (11) to (21). Under this analysis the relation between a raised subject and an XCOMP subject is one of equality. This equality insures that *ein Aussenseiter* 'an outsider' is interpreted as the subject of *scheinen* 'to seem' as well as that of *gewinnen* 'to win', as shown in (32):

(32)



Unlike our formalization of equi, in our account of raising we retain the equality in the functional control relation. Equality predicts the grammaticality of (25b) and the sentences in (2), and it predicts the ungrammaticality of the sentences in (31).

Our analysis allows a raising subject to occur overtly within the c-structure constituent that is annotated as an XCOMP and thus not to be raised at all in the c-structure. The raising effect is entirely due to the equality relation (see Zaenen, 1989, for an early version of this analysis for Dutch). This makes the prediction that with raising verbs we can find overt subjects in extraposed complements, as illustrated by the example in (33) (from Meurers and De Kuthy, 2001):[9]

(33) Obwohl damals anfing, der Mond zu scheinen.
Even though back then begun the moon (N) to shine.
'Even though the moon had begun to shine back then.'

This example contrasts with the ungrammatical equi sentence (30e).

[9] Another potential example, taken from the web, is the following.

(i) Es scheint sich aber allgemein die Form Bergfried durchgesetzt zu haben.
It seems refl however generally the form Bergfried imposed to have.
'The form Bergfried seems, however, to have imposed itself.'

Here the analysis depends on exactly how the conditions on so-called there-insertion (*es* in German) are stated for German. We will not go into this.

Our account of raising also extends properly to the cases of German object raising, the so-called AcI constructions. As has been discussed recently in Meurers and De Kuthy (2001), in AcI constructions the subject of the complement verb appears as an accusative, as illustrated with the PVPF example in (34):

(34) Den Kanzler tanzen sah der Oskar.
The (A) chancellor dance saw the Oskar.
'Oskar saw the chancellor dance.'

This kind of sentence is accounted for straightforwardly by our proposal under the assumptions that the raising verb *sehen* 'to see' takes an accusative object, that its object and the subject of the embedded verb are related by equality, and that nominative case is not obligatorily assigned in the infinitive clause.[10]

6. Conclusions

In this paper we propose a new analysis of German partial VP fronting and show how it interacts with raising and equi. The main new ingredient is the use of subsumption in addition to equality in modeling the flow of information. Subsumption combines with the optionality of most c-structure constituents within the S|VP and previously proposed uncertainty equations to give the right results. Unlike the previous account, it explicitly records in the f-structure which parts of the complement clause appear in topic position. We also use subsumption to model the relation between the subject of an equi-verb complement and the grammatical function it serves in the matrix.

For the interaction between raising and fronting we crucially rely on the fact that equality relations allow information to be realized in either of the c-structure positions between which a functional equality holds. This allows us to have (f-structure) raising without (c-structure) raising as was first pointed out in Zaenen (1989). Our account of the interaction of VP fronting with raising is in this respect similar to the one proposed by Meurers and De Kuthy (2001) who rediscovered the 'raising without raising' solution in an HPSG framework.

The German raising and equi facts handled here are similar but not identical to those found in French Stylistic Inversion. As discussed in Zaenen and Kaplan (2002), in French both subject to subject raising and subject-controlled equi are best handled with equality relations whereas both object raising and object-controlled equi require subsumption.

[10] Here the understood subject also imposes accusative case on the adjuncts that agree with it, as shown by the following example (from Müller, 1999).

(i) Der Wächter sah die Männer einen nach dem anderen weglaufen
The guard saw the men (A) one (A) after the other run-away

This is to be expected: because of the equality relation, the case of the object is also the case of the embedded subject, and the adjunct agrees with it.

The use of subsumption addresses fundamental questions about the nature of information flow in syntax. Most transformational theories promote an asymmetric or even anti-symmetric view. Constraint-based formalisms have tended to stress the non-directionality of information flow. The subsumption relation permits a characterization of asymmetric syntactic dependencies that cannot be easily encoded in phrase structure constraints and thus allows for simple models of phenomena that otherwise are difficult to describe.

References

Berman, J., 1999. Does German Satisfy the Subject Condition? *Proceedings of the LFG99 Conference*. Stanford: CSLI Online Publications, http://www-csli-stanford.edu/publications.

Dalrymple, M., 2001. *Lexical functional grammar*, Syntax and Semantics, vol 34. New York: Academic Press.

Dalrymple, M. and R. M. Kaplan, 2000. Feature indeterminacy and feature resolution. *Language* 76, 759-798.

Haider, H., 1990. Topicalization and other puzzles of German syntax. In G.Grewendorf and W. Sternefeld (eds.), *Scrambling and barriers*. Amsterdam: John Benjamins, 93-112.

Hinrichs, E. and T. Nakazawa, 1995. Linearizing AUXs in German verbal complexes. In J. Nerbonne, K. Netter and C. Pollard (eds.), *German in Head-Driven Phrase Structure Grammar*. Stanford: CSLI Publications, 11-38.

Kaplan, R. M. and J. Bresnan, 1982. Lexical Functional Grammar: A formal system for grammatical representation. In J. Bresnan (ed.), *The Mental representation of grammatical relations*. Cambridge: The M.I.T. Press, 173-281.

Kaplan, R. M. and J. T. Maxwell III, 1988. An algorithm for functional uncertainty. In *Proceedings of COLING-88* pps. 297-302.

Kaplan, R. M. and A. Zaenen, 1989. Long-distance dependencies, constituent structure, and functional uncertainty. In M. Baltin and A. Kroch, eds., *Alternative conceptions of phrase structure*. Chicago: The University of Chicago Press,17-42.

Kaplan, R. & A. Zaenen, in press. West Germanic verb clusters in LFG. In G. Kempen & P. Seuren (eds.), *Germanic verb clusters*. Amsterdam: John Benjamins.

Kaplan, R. M. and J. Wedekind, 1993. Restriction and correspondence-based translation. *Proceedings of the 6th Conference of the Association for Computational Linguistics European Chapter*, Utrecht University, 193-202.

King, T. H., 1997. Focus domains and information-structure. *Proceedings of the LFG97 Conference*. Stanford: CSLI Online Publications, http://www-csli-stanford.edu/publications.

Meurers, D. and K. De Kuthy, 2001. Case assignment in partially fronted constituents. In C. Rohrer, A. Rossdeutscher, and H. Kamp (eds.), *Linguistic form and its computation*. Stanford: CSLI Publications, 29-64.

Müller, S. 1999. *Deutsche Syntax deklarativ*. Tübingen: Niemeyer.

Nerbonne, J., 1995. Partial verb phrases and spurious ambiguities. In J. Nerbonne, K. Netter and C. Pollard (eds.), *German in Head-Driven Phrase Structure Grammar*. Stanford: CSLI Publications, 109-150.

Uszkoreit, H., 1987. Complex fronting in German. In G. J. Huck and A. Ojeda (eds.), *Discontinuous constituency*, Syntax and Semantics, vol. 20. New York: Academic Press, 405-425.

Zaenen, A., 1989. The place of bevallen (please) in the syntax of Dutch. Report SSL-89-17, Xerox Palo Alto Research Center, Palo Alto.

Zaenen, A., and R. M. Kaplan, R. M., in press. Subject inversion in French: Equality and inequality in LFG. In C. Beyssade, O. Bonami, P. C. Hofherr, & F. Corblin (eds*.), Empirical issues in formal syntax and semantics 4*. Paris: Presses Universitaires de Paris-Sorbonne.

Zaenen, A. and R. M. Kaplan, 1995. Formal Devices for linguistic generalizations: West Germanic word order in LFG. In Dalrymple, M., R. Kaplan, J. Maxwell, and A. Zaenen (eds.), *Formal issues in Lexical-Functional Grammar*. Stanford: CSLI Publications, 215-239.

TIGER TRANSFER
Utilizing LFG Parses for Treebank Annotation

Heike Zinsmeister

University of Stuttgart
Institute for Natural Language Processing
zinsmeis@ims.uni-stuttgart.de

Jonas Kuhn

Stanford University
Department of Linguistics[1]
jonask@mail.utexas.edu

Stefanie Dipper

University of Stuttgart
Institute for Natural Language Processing
dipper@ims.uni-stuttgart.de

Proceedings of the LFG02 Conference

National Technical University of Athens, Athens

Miriam Butt and Tracy Holloway King (Editors)

2002
CSLI Publications
http://csli-publications.stanford.edu/

[1]Jonas Kuhn was at the University of Stuttgart when the main work reported in this paper was performed.

Abstract

Creation of high-quality treebanks requires expert knowledge and is extremely time consuming. Hence applying an already existing grammar in treebanking is an interesting alternative. This approach has been pursued in the syntactic annotation of German newspaper text in the TIGER project. We utilized the large-scale German LFG grammar of the PARGRAM project for semi-automatic creation of TIGER treebank annotations. The symbolic LFG grammar is used for full parsing, followed by semi-automatic disambiguation and automatic transfer into the treebank format. The treebank annotation format is a 'hybrid' representation structure which combines constituent analysis and functional dependencies. Both types of information are provided by the LFG analyses.

Although the grammar and the treebank representations coincide in core aspects, e.g. the encoding of grammatical functions, there are mismatches in analysis details that are comparable to translation mismatches in natural language translation. This motivates the use of transfer technology from machine translation.

The German LFG grammar analyzes on average 50% of the sentences, roughly 70% thereof are assigned a correct parse; after OT-filtering, a sentence gets 16.5 analyses on average (median: 2). We argue that despite the limits in corpus coverage the applications of the grammar in treebanking is useful especially for reasons of consistency. Finally, we sketch future extensions and applications of this approach, which include partial analyses, coverage extension, annotation of morphology, and consistency checks.

1 Introduction

This paper reports on work done in the context of the TIGER project.[2] The project aims at creating a large German treebank, the TIGER treebank (Brants et al. 2002), and at developing search tools (TIGERSearch, Lezius (2002)) for exploiting the information encoded in the treebank. The annotation is very detailed in that it encodes information about part-of-speech: e.g. NN, VMFIN (common noun, finite modal verb); morphology: e.g. Masc.Nom.Sg; syntactic category: e.g. NP, PP (noun/prepositional phrase); and grammatical function: e.g. SB, OP (subject, prepositional object). In order to represent the functional dependency relations in a compact graph format with the sequential word string at the terminal nodes, a generalized tree format was adopted which includes crossing branches (cf. the NEGRA project, Skut et al. 1997; for an example, see Figure 5 below). In this format, for instance, a nominal phrase may form a discontinuous constituent with an extraposed relative clause that modifies it. (A more familiar phrase structure format involving traces can be obtained with a conversion routine.)

Two different annotation methods are used in the TIGER project: (i) an interactive combination of a cascaded probabilistic parser (Brants 1999) and manual annotation with the ANNOTATE tool (Plaehn and Brants 2000); (ii) parsing by a symbolic LFG grammar, followed by manual disambiguation and automatic transfer into the TIGER format (Dipper 2000). Technique (i) is the main line in the annotation process; (ii) has a more experimental status. A motivation for the use of (ii) was to explore to what extent a preexisting broad-coverage unification grammar of German can be exploited for an annotation project. Since the corpus is supposed to satisfy high standards of quality (in particular consistency), each sentence is generally annotated independently by two annotators. In cases of mismatch the annotators have to go over the sentence again in a discussion session. The use of two entirely independent methods

[2]The TIGER project (URL: `http://www.ims.uni-stuttgart.de/projekte/TIGER/`) is funded by the *Deutsche Forschungsgemeinschaft* (DFG).

We would like to thank Anette Frank (DFKI Saarbrücken, formerly at XEROX Grenoble) for her input and great help with the transfer component. Thanks also to Bettina Schrader who was not only responsible for the disambiguation of the LFG analyses but also implemented parts of TIGER transfer. Credit for the original idea of exploiting an existing transfer module goes to Martin Emele. Besides these three people, we would like to thank Stefan Evert, Jan Anderssen, Hannah Kermes, and the audiences at the Workshop on "Syntactic Annotation of Electronic Corpora" (Tübingen, 2000) and at the "Third Workshop on Linguistically Interpreted Corpora" (LINC – Leuven, 2001), for discussion and comments on previous versions of this paper.

is an additional way of ensuring consistency. With the more mechanical grammar-based approach, low-level mistakes resulting from carelessness are less likely to appear. To a certain degree this outweighs the problem that the grammar of course does not cover all the constructions appearing in a newspaper corpus.

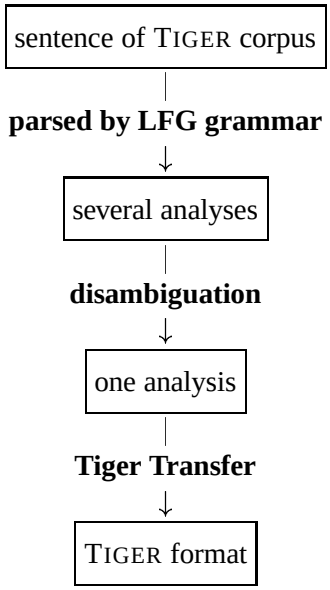


Figure 1: Scenario of annotation by LFG

This report focuses on the grammar-based annotation approach (ii), which is depicted in Figure 1, and in particular on its transfer component: TIGER Transfer. The paper is organized as follows: Section 2 gives a short introduction to the German LFG grammar that is used in the annotation and also addresses the disambiguation task. Section 3 presents the transformations that an LFG analysis of a sentence undergoes on its way to the TIGER representation. Section 4, then, gives some statistics. Finally, Section 5 concludes the paper with an outlook on future work.

2 The LFG Analysis

2.1 German LFG Grammar

The German LFG grammar (Dipper to appear) was developed in the PARGRAM project,[3] using the Xerox Linguistic Environment (XLE). Analyzing a given sentence with the LFG grammar yields two representations, the constituent structure (c-structure) and the functional structure (f-structure). C-structure encodes information about morphology, constituency, and linear ordering. F-structure represents information about predicate argument structure, about modification, and about tense, mood, etc.

Figure 2 shows the LFG c-structure for a simple sentence from the TIGER corpus: *Hier herrscht Demokratie* 'Democracy rules here'. The example demonstrates both, familiar aspects of German syntax and more technically motivated specialities of this particular grammar implementation.[4] The latter in-

[3]URL: `http://www.parc.com/istl/groups/nltt/pargram`

[4]The German LFG grammar encodes a generalized CP-analysis of German: The finite verb *herrscht* thus occupies the C-position, preceded by the adverb phrase *hier* in the specifier position of CP. The NP *Demokratie* is immediately dominated by Cbar. For processing reasons, there is no VP-projection covering the 'Mittelfeld' and hence dominating the NP. Note finally, the root node does not only dominate the clausal projection of the sentence but also its identification and its final punctuation

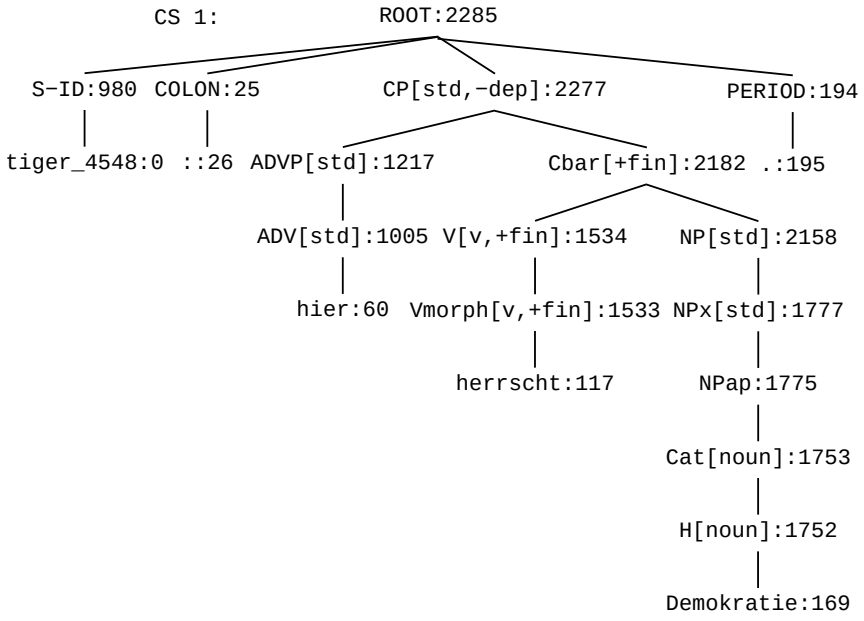


Figure 2: **C-structure** of tiger_4548: *Hier herrscht Demokratie.*

clude a fine-grained differentiation of category symbols, among others 'complex' category symbols with category-level features in squared brackets, like `V[v,+fin]`[5]. Figure 3 shows the f-structure of *Hier herrscht Demokratie* 'Democracy rules here'. The leftmost bracket opens the feature structure of the main predicate, the verbal predicate *herrscht*. This f-structure includes several grammatical functions, which have embedded f-structures as values: ADJUNCT points to the adverbial predicate *hier* (embedded in a set, since there can be several adjuncts), SUBJ(ect) points to the predicate *Demokratie*. In addition the f-structures contain morphosyntactic information like TENSE, MOOD, CASE, and GENDER. In LFG, the level of c-structure is related to the f-structure by the function ϕ which maps each c-structure node to a feature structure (a many-to-one mapping). The mapping relations between c-structure nodes and f-structures are indicated by indices. In Figure 2, for instance, the c-structure nodes representing the projections of the adverb *hier* are indexed with 60, 1005, and 1217, respectively. The function ϕ maps the nodes to the feature structure that is the value of the (set-valued) feature ADJUNCT on f-structure, cf. Figure 3.

2.2 Disambiguation

Almost every sentence of a newspaper corpus is syntactically ambiguous. There are structural ambiguities such as different attachment sites of adjuncts and word-level ambiguities due to ambiguous inflectional marking, homographic word forms or alternatives in subcategorization. Hence the grammar output has to be disambiguated, which normally means that a human annotator has to select the correct analysis.[6]

mark (S-ID and COLON).

[5] `V[v,+fin]` denotes a verbal category of the subtype 'finite full verb'.

[6] Shallow parsing approaches typically employ a more deterministic strategy, i.e., they combine parsing and disambiguation (from corpus-based training), producing just a single analysis. Quite obviously in the scenario of creating a high-quality treebank annotation, manual control is indispensable. In the approach using the ANNOTATE tool, this is ensured through the interactive cascaded procedure involving the human annotator at all levels. In our case, the entire analyses are presented to the

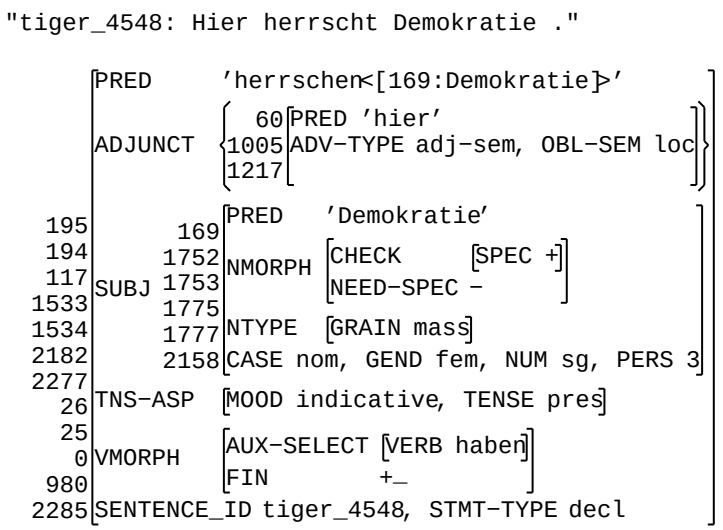


Figure 3: **F-structure** of tiger_4548: *Hier herrscht Demokratie.*

XLE supports this disambiguation in so far as it packs all different readings into one complex representation that can easily be browsed by the human annotator. On average, however, a sentence of the TIGER corpus receives several thousands of LFG analyses. Obviously it is impossible to disambiguate those analyses manually. For that reason, XLE provides a (non-statistical) mechanism for suppressing certain ambiguities automatically. We illustrate both kinds of disambiguation in the following paragraphs.

Manual Disambiguation The parses of one sentence are represented in a packed feature structure chart, cf. Maxwell and Kaplan (1989): the features common to all readings of the sentence are represented a single time; feature constraints that do not hold in all readings are marked by context variables. The result is an f-structure that is annotated with variables to show where alternatives are possible. After some training, this representation is easily readable for the annotator. Manual selection of the correct analysis is done either by picking the corresponding c-structure tree or by clicking on the respective variables in the f-structure. XLE moreover supports manual disambiguation by various other browsing tools applied to c-structure as well as to f-structure (King et al. to appear describe these tools in detail).

In Example (1), ambiguity in case marking gives rise to two different predicate argument structures (and to two different constituent structures). *Der Stiftung* 'the foundation' can either be dative or genitive, i.e., it can either function as indirect object to the ditransitive verb *verkaufen* 'sell' or as genitive attribute to *Haus* 'house' (with a transitive version of *verkaufen*, which is likewise possible), cf. the readings in (2). In Figure 2.2, the f-structure alternatives that are restricted to the transitive reading are annotated with the variable `a:1`, the ditransitive ones with `a:2`, respectively. The resolution of this ambiguity requires context knowledge and has to be done manually.

(1)	Die	Stadt	verkaufte	das	Haus	der	Stiftung.
	The	town	sold	the	house	the	foundation

human annotator grammar includes many more explicit grammatical constraints than the grammars that are used in a shallow parser) .

(2) a. [Die Stadt]$_{NOM}$ verkaufte [das Haus]$_{ACC}$ [der Stiftung]$_{DAT}$.
'The town sold the house to the foundation.'

b. [Die Stadt]$_{NOM}$ verkaufte [das Haus [der Stiftung]$_{GEN}$]$_{ACC}$.
'The town sold the house of the foundation.'

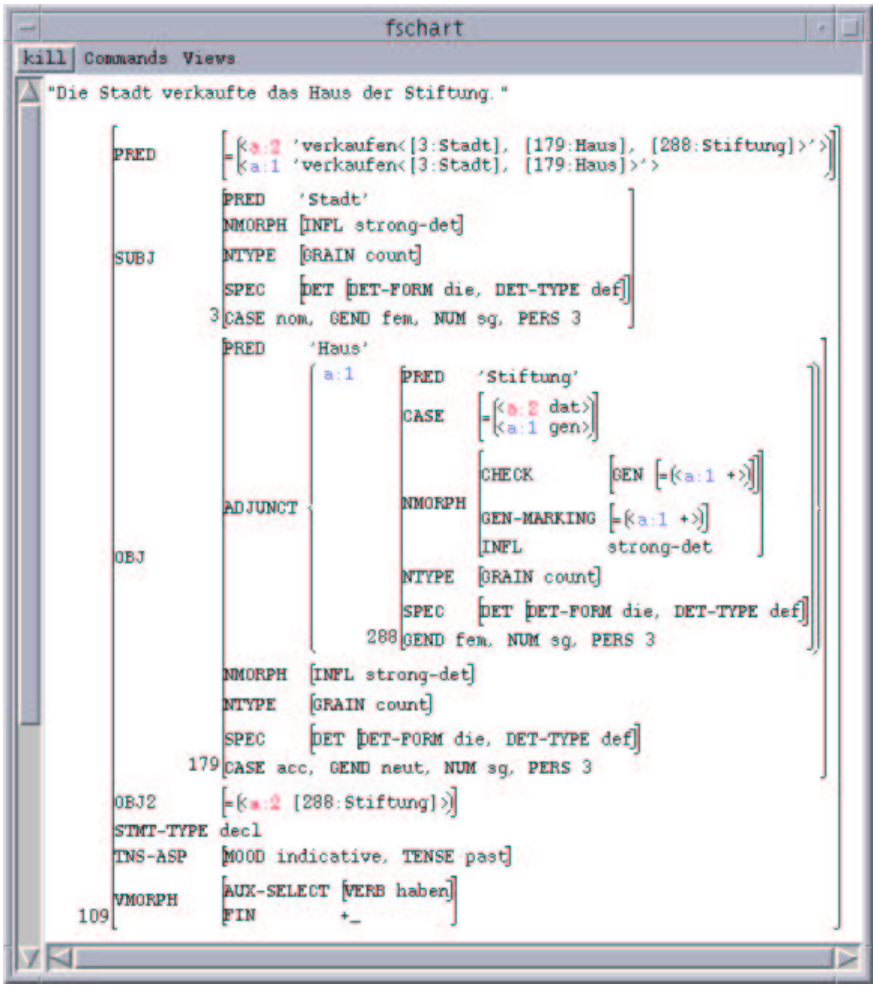


Figure 4: **Packed f-structure-chart** of *Die Stadt verkaufte das Haus der Stiftung.*

Automatic Disambiguation Example 1 is even more ambiguous. *Das Haus* and *die Stadt* can either be nominative or accusative, i.e. subject or direct object of the clause, cf. the readings in given in (3). In this case, one reading is rather improbable, namely that with the object occupying the first position (the 'Vorfeld'). When like in this example, the case marking is ambiguous (but not when the accusative is overtly marked), then there is a strong dispreference against the reading with the object in the Vorfeld position. This phenomenon is sometimes called the *word order freezing effect* (compare (Kuhn 2001, sec. 4.2), Lee (2001)). This dispreference, as well as other biases, is exploited by a (non-statistical) mechanism that XLE provides for suppressing ambiguities automatically. The mechanism consists of a constraint ranking scheme inspired by Optimality Theory (OT), see Frank et al. (2001). Each rule and each lexicon entry may be marked by 'OT marks'. When a sentence is parsed, each analysis is annotated

by a multi-set of OT marks, thereby keeping a record of all OT-marked rules and lexicon entries that were used for the respective parse. The grammar contains a ranked list of all OT marks. When an ambiguous sentence is parsed, the OT mark multi-sets of all readings compete with each other. A multi-set containing a higher ranked OT mark than another multi-set is filtered out, thus suppressing highly improbable or marked readings, and reducing the number of ambiguities the human disambiguator has to deal with. (Note that no suppressing of parses happens in the absence of ambiguity.)

(3) a. [Die Stadt]$_{ACC}$ verkaufte [das Haus]$_{NOM}$ [der Stiftung]$_{DAT}$. [*improbable*]
'The house sold the town to the foundation.'

b. [Die Stadt]$_{ACC}$ verkaufte [das Haus [der Stiftung]$_{GEN}$]$_{NOM}$. [*improbable*]
'The house of the foundation sold the town.'

A mark 'ObjInVorfeld', for example, forces to disprefer a direct or indirect object in the 'Vorfeld' (rather than subject or adjunct). Thus the improbable readings in (3) are suppressed. For Sentence (1), the German LFG grammar reduces the eight possible readings[7] to two by means of the OT filter mechanism. A further option would be to integrate a probabilistic disambiguation of sentence analyses (cf. Riezler et al. 2000).

3 From LFG to TIGER

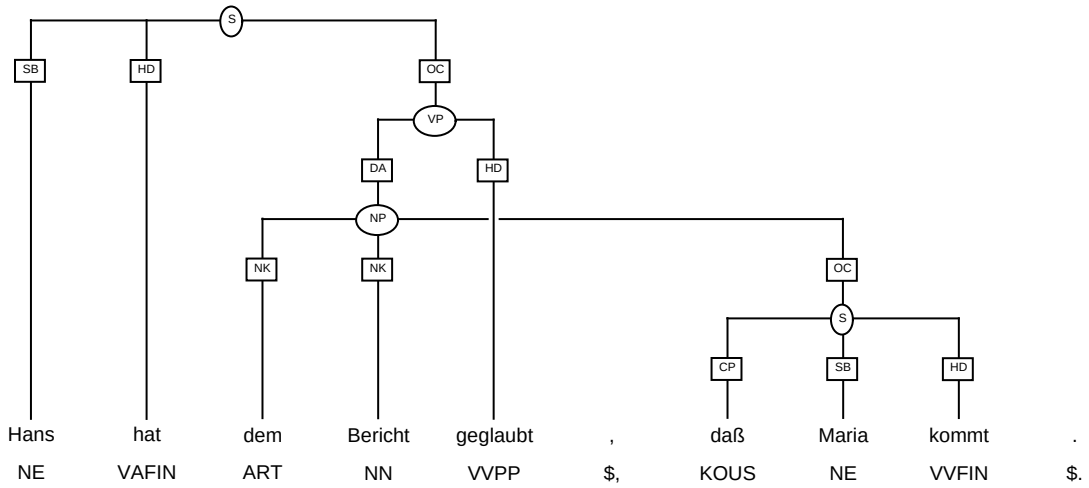
This section deals with the mapping of the LFG grammar output to the TIGER format. Section 3.1 illustrates the close correspondence between LFG f-structure and TIGER graphs. Section 3.2 gives criteria for implementing the conversion procedure and motivates the particular way of splitting up the task in subtasks. Section 3.3 presents preprocessing steps that make the LFG output suited for the actual transfer component. Sections 3.4 and 3.5 deal with the actual transfer. Finally, Section 3.6 concludes the section with the postprocessing steps that complete the mapping.

3.1 LFG F-structure and Tiger Graphs

The initial motivation for adopting the LFG-based annotation method was that despite many differences in details, the syntactic analyses of the LFG grammar and the TIGER graph representation are very similar at the level of functional/dependency structure: Both, LFG's f-structure and the TIGER graph representation, model a dependency structure at a comparable degree of granularity. A further similarity is that in both representations, the dependency structure is explicitly related to the word string. In the case of LFG, this relation is mediated through the level of c-structure; in TIGER it is coded directly into the dependency graph. The TIGER format gives dependency structure priority over phrase structural constituency. It makes use of a generalized tree graph notation which allows for crossing branches. In both schemes, the situation can arise that a single f-structure/dependency-structural constituent corresponds to a set of non-adjacent terminal nodes, i.e., nodes with some intervening word material belonging to a higher-level f-structure/constituent. For example, in *Hans hat dem Bericht geglaubt, daß Maria kommt* 'Hans believed the report that Maria will come', the embedded clause is a complement of the non-adjacent nominal *Bericht*, cf. the representations in Figure 5. The dependency which yields the crossing branch in the TIGER graph is encoded in the LFG f-structure as the complement clause embedded under a COMP feature of the nominal predicate.

Along with these high-level similarities there are a number of differences in the representation conventions of TIGER vs. the German LFG grammar. One main difference is due to the fact that redundant

[7] The discussed four readings are doubled by an additional ambiguity of the verb. It also allows a (dispreferred) subjunctive reading.



"Hans hat dem Bericht geglaubt, daß Maria kommt."

$$\begin{bmatrix} \text{PRED} & \text{'glauben<[1:Hans], [52:Bericht]>'} \\ \text{SUBJ} & 1\begin{bmatrix}\text{PRED 'Hans'}\end{bmatrix} \\ \text{OBJ2} & 52\begin{bmatrix} \text{PRED} & \text{'Bericht<[163:kommen]>'} \\ \text{COMP} & 163\begin{bmatrix}\text{PRED} & \text{'kommen<[186:Maria]>'} \\ \text{SUBJ} & 186\begin{bmatrix}\text{PRED 'Maria'}\end{bmatrix}\end{bmatrix}\end{bmatrix} \\ 32 \end{bmatrix}$$

Figure 5: Long distance dependencies in TIGER graph and LFG f-structure representation

information is not encoded in the TIGER format (which we will deal with in Section 3.6). For example, phrasal categories are always given in LFG c-structure, independent of the complexity of the phrase, whereas in TIGER only complex constituents are dominated by a phrasal node (e.g. VP), compare the representation of the intransitive *kommt* 'comes' vs. the transitive (i.e. complex) *geglaubt* 'believed' in the TIGER graph in Figure 5. Other discrepancies exist in the representational conventions for particular phenomena (dealt with in Section 3.5). For example, in the TIGER representation, the verbal phrase *dem Bericht geglaubt* 'believed the report' (containing the full verb participle *geglaubt*) is embedded as a clausal object (OC) under the auxiliary *hat*. The LFG grammar does not employ a nesting analysis on the functional level. The full verb and the auxiliary occur at the same f-structure level.

In summary, differences between the two representation schemes fall in two distinct categories: formalism-inherent differences and differences in representational convention or linguistic analysis. The latter could in principle be overcome *within* either of the two formal frameworks, e.g., by using the LFG formalism for writing a new grammar that uses category symbols according to the exact specifications of the TIGER annotation scheme (although this is certainly not a practical option, since the grammar has its independent motivation the way it is).

3.2 Criteria for implementing the conversion procedure

The observed systematic relation between the source and the target format of the required conversion makes it realistic to implement a mechanical routine for this conversion. At the same time, subtleties in the differences of the second kind (differences in convention or linguistic analysis) have to be approached with care. In particular, one has to be aware that neither the source nor the target format conventions are specified in all detail; they are not even fixed once and for all, but may undergo occasional changes. (Pre-

sumably they cannot be fixed in principle as long as one keeps applying the grammar and the annotation scheme to new corpus material.)

We put great emphasis on the criterion of flexibility in the specification of the conversion procedure in order to be able to react to modifications in the source and target format conventions. This excluded a monolithic implementation of the conversion step. Rather a design with a declarative specification of the convention-related conversion-steps was vital. As mentioned above, it is possible to make all conversions but the ones concerning formalism-inherent differences *within* either one of the formalisms. The decision we made was to exploit this fact and stay within the formal framework of LFG for most of the conversion procedure[8] – until a final low-level conversion into the notational format of the TIGER treebank (this will be called the *postprocessing* below, see Figure 6). The great advantage of this move was that we could exploit existing systems for modifying grammatical analyses within a linguistic formalism: transfer systems as used in machine translation. Although the present context of application is quite different, the task is very similar. In our case the source and target structures do not originate from grammars of different languages, but from different systems of representation for syntactic data of the same language.

In order to ensure a high-quality conversion, our goal has been to rely on *systematic* differences, which can be converted mechanically, as much as possible. Besides the more or less notational rewriting step that we perform towards the end of the conversion there are other highly systematic differences between the formalisms, concerning the relation between functional/dependency structure and the surface string. Basically, LFG's c-structural information has to be folded into the f-structure (and reduced to a subset of relevant categorial information). It turned out convenient to perform this conversion right at the beginning (as what we call *preprocessing* below). Having separated out these two formalism-inherent aspects of the conversion, the remaining step can focus on the more linguistically involved conversion aspects as part of the transfer proper.

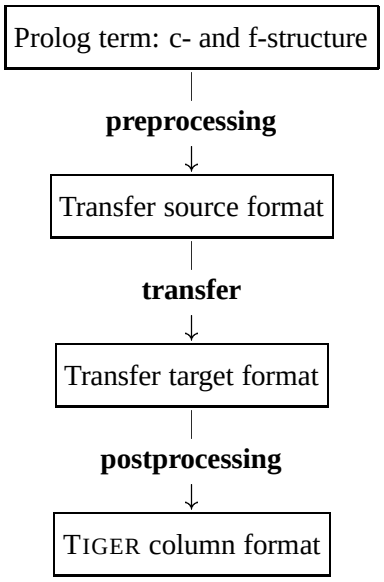


Figure 6: Subprocedures of the conversion routine

The resulting modular design shown in Figure 6 has some obvious advantages: Changes in the grammar or annotation scheme (unless radical) should only affect this 'middle' step. Testing of the conversion

[8]One advantage of this is that the XLE visualization tools for LFG structures can be used even at a stage in which the representation has already been converted quite a long way (cf. Figures 8 and 9 below).

steps can be performed separately, and it becomes an option simply to replace the postprocessing step, for instance, if a different target format (e.g. XML) is desired.

3.3 Preprocessing

Since the transfer step proper proceeds along the f-structures of the LFG representation, it has to be ensured that all the information required to construct the TIGER target structures is accessible from f-structure. This is the task of the preprocessing step. Note that information about linearization of the individual words – although not encoded at the level of f-structure – can be ignored in our context: the word string remains identical, therefore it would be redundant to keep track of word order information during transfer. As long as unique reference to the words is made, the linearization of the string can easily be overlaid over the target TIGER graph – crossing edges result automatically.

Nevertheless, reference to c-structure categories is required in order to construct the correct category labels in the target structure. Thus the preprocessing step implements a general format conversion of the LFG structures, folding the c-structural information into the f-structure. (Not all the c-structural information is required, but the transfer can easily eliminate irrelevant information.)

The LFG grammar development and parsing system XLE provides an export format for the syntactic analyses: a Prolog term, containing flat lists of f-structure and c-structure descriptions, cf. Figure 7. The preprocessing step was implemented as a Prolog program taking this format as input and producing a similar term with an 'enriched' f-structure as output. Following the modular philosophy argued for in the previous section, the preprocessing program performs only highly systematic, canonical modifications, leaving phenomenon-specific decisions to the transfer step.

What the program effectively does is traverse the c-structure from the root node, keeping track of the c-structure/f-structure correspondence. This leads to a set of partial subtrees consisting only of connected co-projecting c-structure nodes. A feature representation of such subtrees is then added to the respective f-structure under a special feature `TT_TREE`. Some special care has to be taken since a single f-structure may have several corresponding connected subtrees; furthermore, it has to be made sure that the connection to the words in the string remains recoverable, using a special feature `TT_TERM-CAT`.[9]

A detail of an enriched f-structure is shown in Figure 8 (here, not the Prolog term itself is shown, but the XLE display of it – the modified export format can be read in and displayed again). Note how the subtree projected by the adverb *hier* is merged into the feature structure of the corresponding predicate. The pointer feature `TT_PHI` (somewhat redundantly) has the feature structure of *hier* as its value. This configuration allows to test for features and values even in more complex structures without moving through the recursive tree structure. The integrated c-structure information does not only include information about the syntactic categories (e.g. `ADV[std]`, `AVDP[std]`) but also sublexical information like part of speech and lemma (e.g. `+Adv+Common`, `hier`), and the pointer to the surface token (`TT_SFF_ID`).

The preprocessing routine, in addition, splits parametrized category symbols like `ADVP[std]` into functor-argument lists. This is relevant for generalizations over parametrized features since the transfer system does not allow to use regular expressions on label names.

3.4 The Transfer Grammar

We made use of the transfer system of the XEROX Translation Environment by Martin Kay (XTE) which is part of the XLE development platform. The transfer component is a rule rewriting system based on Prolog. As mentioned in Section 3.1, differences in representation or linguistic analysis lead to structural

[9] This part is in fact non-trivial since the PARGRAM LFG grammar does not operate on a string of fullform words, but on the output of a morphological analyzer which adds branching to the c-structure that would not be recoverable in the target structure, given just the word string.

```
fstructure('tiger_4548:  Hier herrscht Demokratie .',
      [...]
      % Constraints:
      [
      cf(1,eq(attr(var(0),'PRED'),semform('herrschen',3,[var(1)],[]))),
      cf(1,eq(attr(var(0),'ADJUNCT'),var(2))),
      cf(1,eq(attr(var(0),'SENTENCE_ID'),'tiger_4548')),
      cf(1,eq(attr(var(0),'STMT-TYPE'),'decl')),
      cf(1,eq(attr(var(0),'SUBJ'),var(1))),
      cf(1,eq(attr(var(0),'TNS-ASP'),var(3))),
      cf(1,in_set(var(5),var(2))),
      cf(1,eq(attr(var(5),'PRED'),semform('hier',1,[],[]))),
      cf(1,eq(attr(var(5),'ADV-TYPE'),'adj-sem')),
      cf(1,eq(attr(var(1),'PRED'),semform('Demokratie',7,[],[]))),
      cf(1,eq(attr(var(1),'CASE'),'nom')),
      cf(1,eq(attr(var(1),'GEND'),'fem')),
      cf(1,eq(attr(var(1),'NUM'),'sg')),
      cf(1,eq(attr(var(1),'PERS'),'3')),
      cf(1,eq(attr(var(3),'MOOD'),'indicative')),
      cf(1,eq(attr(var(3),'TENSE'),'pres')),
      ],
      % C-Structure:
      [
      [...]
      cf(1,subtree(2273,'CP[std,-dep]',-,1217)),
      cf(1,phi(2273,var(0))),
      cf(1,subtree(1217,'ADVP[std]',-,1005)),
      cf(1,phi(1217,var(5))),
      cf(1,cproj(1217,var(10))),
      cf(1,subtree(1005,'ADV[std]',1004,111)),
      cf(1,phi(1005,var(5))),
      cf(1,subtree(1004,'ADV[std]',-,61)),
      cf(1,phi(1004,var(5))),
      cf(1,terminal(62,'hier',60)),
      cf(1,phi(62,var(5))),
      cf(1,terminal(112,'+Adv+Common',60)),
      cf(1,phi(112,var(5))),
      [...]
      cf(1,surfaceform(0,'tiger_4548',0,3)),
      cf(1,surfaceform(26,':',3,16)),
      cf(1,surfaceform(60,'hier',16,21)),
      cf(1,surfaceform(117,'herrscht',21,30)),
      cf(1,surfaceform(169,'Demokratie',30,36)),
      cf(1,surfaceform(195,'.',36,49))
      ]).
```

Figure 7: tiger_4548: Prolog term: c- and f-structure (details)

mismatches that are comparable to transfer ambiguities in natural language translation. It was reasonable, therefore, to develop the mapping in an actual transfer environment. Using a tested and integrated system had the additional advantage that neither a specification language nor the processing routines had to be developed. The main focus lay on the interface routines and the specification of the mapping.

The preprocessed LFG parse, cf. Figure 8, functions as transfer source format. It is a flat list of predicate-value pairs in which, for example, the LFG function SUBJ(ect) is represented as a two-place predicate that takes two f-structure indices as arguments: `subj(X,Y)`[10]. The predicate-value list is transformed step by step into the TIGER target format. The Transfer rewriting rules apply in an ordered way to the gradually changing set of predicates, which means that the output of a given rule is the input of the subsequently following rule. The architecture of the grammar file therefore mirrors the order of

[10] `Subj(X,Y)` reads as 'the predicate of f-structure Y is the subject of the predicate of f-structure X'. As common in Prolog, constants begin with lower case, variables begin with upper case or an underscore.

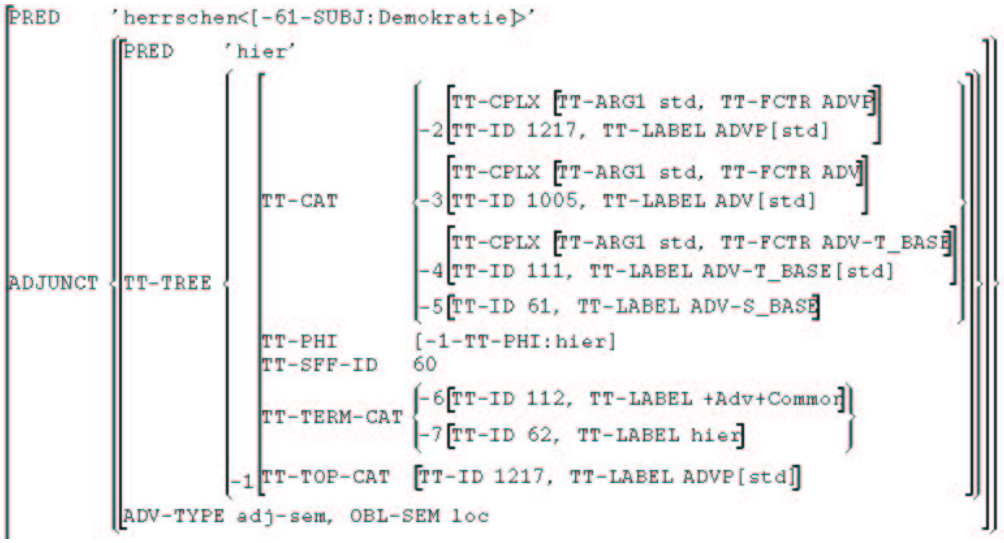


Figure 8: tiger_4548: Detail of the enriched f-structure (Transfer target format)

potential rule application.

Rules Input predicates are on the left-hand side of a rule; output predicates are on its right-hand side. Input and output predicates are separated by a rewriting symbol, the operator '`==>`'. The most basic rules simply rewrite the name of the predicate and pass on the values of the arguments unchanged. For example, the LFG function SUBJ is mapped to the TIGER function SB, the function OBLAGT (the optional agent in a passive clause) to SBP, respectively. The argument slots are not manipulated, i.e. the dependency structure is passed on unaltered.

(4)
```
subj(X,Y)    ==>  sb(X,Y).   %  subject
oblagt(X,Y)  ==>  sbp(X,Y).  %  agent in passives ("von"-phrase)
```

A predicate on the input side of a rule is deleted from the input set of predicates, and a predicate on the output side of a rule is added to the output set of predicates. Rules may be contextually restricted by positive and negative tests. The operator '+' preceding a predicate indicates that the predicate is required in the input set for the rule to apply although the rule does not affect the predicate itself. A preceding '−' triggers a negative test: the rule is applied only if the predicate is absent. For example, in (5), the LFG function `XCOMP` is mapped to the TIGER function `PD` (predicative) only if there is a coindexed predicate `XCOMP-TYPE(X,'copula')`. Otherwise, `XCOMP` is rewritten as the TIGER function `OC` (clausal object).

(5)
```
+xcomp_type(Y,'copula'),  xcomp(X,Y)  ==>  pd(X,Y).  %  predicative
                          xcomp(X,Y)  ==>  oc(X,Y).  %  default
```

A zero on the right-hand side encodes the empty set, see e.g. (6). In this case, all predicates on the left-hand side of the rule are deleted from the set of predicates without replacement.[11]

(6)
```
tt_tree(_,_)  ==>  0.  %  deletion
```

[11]The underscore is the common symbol for the anonymous variable in Prolog.

Macros and Templates The system allows for the definition of macros and templates—short-hand notations of sets of predicates and transfer rules, respectively. They do not only facilitate rule development but also the adaptation to changes in either the source or target format. A format change, then, only requires to adapt a given macro or template but not all occurrences of it in the grammar.

The macro TT_VERB_MORPH in (7), for example, is an abbreviation of the predicates encoding the information relevant for the mapping of verbal morphology tags. The predicates include TENSE and MOOD, which are embedded in the verbal TNS-ASP feature; and the corresponding person and number information which LFG encodes as PERS and NUM embedded in the subject feature (SB).

(7)
```
tt_verb_morph(V,Person,Number,Tense,Mood):=

   +tns_asp(V,V1),
      +tense(V1,Tense),
      +mood(V1,Mood),

   +sb(V,V2),
      +pers(V2,Person),
      +num(V2,Number).
```

Just like macros are short-hand forms of predicates, templates are short-hand forms of rules. In (8), the template LABEL2HEAD expands to a rule that maps a c-structure label, FCTR, to a TIGER head relation, HEAD. The rule passes on the form and index of the surface token. In addition, it introduces a part-of-speech tag. The input predicates share their first argument, i.e., they are all connected to a specific feature structure in the input representation. TT-PHI is a pointer that relates c-structural information to the corresponding f-structure index. It allows one to link the target predicates directly to the dependency structure.

(8)
```
label2head(Fctr,Arg1,Head,Pos) ::

   cat_label(V,Fctr,Arg1),            %  macro for category label
   tt_phi(V,V0),                      %  pointer to f-structure
   sff_in(V,Id,Form)                  %  macro for linking of sur-
                                      %  face form and surface id
   ==>
   ti_terminal(V0,Head,Pos,Id,Form).  %  macro for TIGER heads
```

The actual transfer rules instantiate templates in that all argument slots are filled with constants, see (9)[12]. Templates can also encode a sequence of rules. In this case, the grammar compiler expands the instantiated rule accordingly.

(9)
```
label2head('ADV','std',hd,'ADV').      %  standard adverbs

fctr2head('ADV',hd,'PWAV').            %  interrogative and
                                       %  relative adverbs
label2head('PAdv','std',hd,'PROAV').   %  pronominal adverbs

fctr2head('PAdv',hd,'PWAV').           %  interrogative and
                                       %  relative pronominal
                                       %  adverbs
```

[12] LABEL2HEAD and FCTR2HEAD differ only with respect to the category label. FCTR2HEAD generalizes over the value of Arg1. LABEL2HEAD can express FCTR2HEAD if its second argument is instantiated with the anonymous variable. In this case any value of Arg1 matches the input requirements. Since it turned out to be less efficiently processed, the grammar rules make no use of this encoding option.

Structure of the Grammar The transfer component applies the rules in the order of specification, i.e. the transfer grammar mirrors the mapping process. Each compiled rule is called only once in the mapping process and is then applied to all predicates that match the rule input requirements. The grammar is organized in seven main parts:

(i) Redundant predicates are deleted. E.g. the subject feature of adjectives in LFG has no correspondence in the TIGER format[13]. It is therefore deleted and will not interfere with the mapping process.

(ii) For more efficient processing, all set-valued features are rewritten as relational predicates. The rule in (10) introduces `X_ADJUNCT` which is a temporary predicate in the sense that it is neither present in the transfer input nor in the transfer output. It is both, introduced and subsequently deleted in the process of transfer.[14]

(10)
```
+adjunct(V,V1), in_set(V2,V1)    %  set-valued
==> x_adjunct(V,V2).             %  relational

adjunct(_,_) ==> 0.              %  deletion
```

(iii) Grammatical functions that are encoded in the LFG f-structure are mapped to target predicates. In many cases the predicate names are just rewritten and the functional structure is passed on unchanged, e.g. `subj(V,V1) ==> sb(V,V1).`

(iv) Syntactic categories and syntactic heads are mapped in combination with part-of-speech and morphology tags. More specific rules thereby precede more general rules. If not all functional structure is part of the input, the mapping inserts structure, see Section 3.5 for a more detailed discussion of this.

(v) A repair section follows after the mapping proper. It includes rules that map temporary predicates on target predicates. For instance, the temporary head of finite auxiliaries `HD_AUX` is mapped on the (more general) target head function `HD`. The section also includes rules that 'repair' target format notation. For example, the TIGER format distinguishes between prepositional modifiers of nouns and other noun modifiers. The former are labelled 'modifier of the noun to the right' (`MNR`) whereas the rest is uniformly labelled 'noun kernel element' (`NK`). It is much simpler to map all noun modifiers alike and only to check for the specific case at the end of the mapping process, cf. (11).

(11)
```
nk(V,V1),+ti_cat(V1,'PP') ==> mnr(V,V1).  %  prepositional
                                          %  modifier of noun
```

The mapping rules are followed by two further rule sections. (vi) Robustness rules check for all functional labels and terminals whether they are integrated in the dependency structure – which is a necessary prerequisite for the canonical postprocessing convertion to the TIGER column format. If necessary a fragment relation is inserted. (vii) Finally, all non-target predicates are deleted.

3.5 Transfer Phenomena

In contrast to natural language transfer, TIGER Transfer leaves the surface string unchanged. The task is to map a limited set of grammatical features into another limited set of grammatical features. Although there are many trivial cases, the format conversion is more complex than a simple mapping of two feature sets. Due to differences in representation chosen for particular linguistic phenomena, there are mismatches that are comparable to 'transfer ambiguities' in natural language translation (for the latter see e.g. Kameyama et al. 1991, Emele et al. 2000).

[13] In the case of attributive adjectives, for instance, the subject points to the modified head noun.

[14] Doug Arnolds (p.c.) made us aware of the drawbacks temporary predicates might have if they are used in a large grammar.

Ambiguous Predicates A simple example was introduced in Section 3.4, the mapping of XCOMP to PD or OC, depending on the value of the feature XCOMP. A more complex case is the translation of the predicate ADJUNCT. It is a very general function in the German LFG and corresponds to three different TIGER functions, i.e., it is three-fold ambiguous. The conditions for resolving the ambiguity are not encoded in a specific feature, but have to be found independently. In (12), ADJUNCT is mapped to the function 'genitive attribute' (AG) if the embedding predicate is a noun (marked by NTYPE) and the embedded predicate has a case feature with the value gen(itive) – unless it is an attributive adjective (ATYPE attributive). Other ADJUNCTs within nouns are mapped to 'noun kernel element' (NK). The default mapping of ADJUNCT is to the function 'modifier' (MO).[15]

(12)
```
+ntype(V,_), +case(V1,'gen'),     %  required context
-atype(V1,'attributive'),         %  negative condition
x_adjunct(V,V1),
==>
ag(V,V1).                         %  'genitive attribute'

+ntype(V,_), x_adjunct(V,V1)      %  other noun modifiers
==>
nk(V,V1).                         %  'noun kernel element'

x_adjunct(V,V1)                   %  default:  modifiers of
==>                               %  verbs and adjectives
mo(V,V1).                         %  'modifier'
```

Some ambiguities cannot be resolved by the transfer component. For example, TIGER distinguishes two potential functions of prepositional phrases in predicative constructions, see (13). PPs with an abstract meaning, i.e. idiomatic chunks, are analyzed as predicatives (PD), all other PPs as modifiers (MO). Transfer provides only the function MO here. The adaptation to the specific TIGER edge label has to be done manually after transfer, e.g. with the (semi-automatic) ANNOTATE tool.

(13) a. PP functions as predicative:
Sie ist auf der Hut ('She is on her guard')

b. PP functions as modifier:
Sie ist im Garten ('Sie is in the garden')

Structural Changes There are 'head switch' transformations in natural language translation in which the head of the source structure becomes a dependent element in the target structure and a former dependent element becomes the head of the constituent. In Example (14), *like* is the matrix predicate which subcategorizes the infinitive *come*. In the corresponding German example in (15), *komme* is the main predicate which corresponds in meaning to the English embedded infinitive, *come*. The meaning of *like* is expressed by the adverb *gern* 'gladly' that modifies *komme*.

(14) I like to come

(15)

Ich	komme	gerne
I	come	gladly

In the mapping of LFG to TIGER representations, there are constellations that resemble head switch. Figure 5 in Section 3.1 shows how the two systems represent the concept 'head of a clause' differently. In the German LFG, on the one hand, the main verb is always the main predicate of the clause. Since

[15] ADJUNCT is mapped to the temporary predicate X_ADJUNCT, cf. Section 3.4

temporal (or passive) auxiliaries do not have lexical meaning on their own, they co-project with the main verb on f-structure. In TIGER, on the other hand, the finite verb is analyzed as the head of the clause, independent of whether the verb is a main verb or an auxiliary. Accordingly, in analytic tenses, like perfect, the finite auxiliary is the clausal head and the main verb becomes the head of an embedded function `OC` ('clausal object'). The mapping has to reorganize the verbal heads and insert additional structure. The restructuring is guided by c-structural information.

The natural language mapping of Example (14) to Example (15) does not only affect the hierarchical structure but also the distribution of the arguments: The subject of the predicate *like* in English becomes the subject of the predicate *komme* in German. A more explicit case of argument rearrangement is given in the mapping of Example (16) to Example (17). The German verb *kennenlernen* is translated compositionally into the expression *get to know* in English. The mapping splits the arguments of *kennenlernen* and relates the subject (*sie/they*) to the structurally higher verb *get* and the object (*sich/each other*) to the structurally embedded verb *know*.

(16)	(Er	glaubt,)	dass	sie	sich	kennenlernten
	he	thinks	that	they	REFL	got_to_know

(17) (He thinks) that they got to know each other

Sometimes, arguments have to be rearranged in grammar mapping, as well. If you go back to Figure 5, again, you see that the German LFG treats all arguments the same: both the subject (SUBJ) and the indirect object (OBJ2) belong to the main predicate on f-structure. TIGER, in contrast, distinguishes between subjects and all other arguments. Only the subject is always dependent of the finite head. The other arguments are dependent on the main predicate; if the main predicate is embedded, the arguments are embedded as well, cf. the main predicate *geglaubt* and the indirect object (DA). Argument rearrangement is encoded in the repair section of the transfer grammar.

3.6 Postprocessing Steps

After the renaming and restructuring procedures illustrated in the preceding sections, further small modifications complete the mapping LFG – TIGER.

Morphology Tags Currently the TIGER treebank format supports only one slot for morphology tags. In LFG, the morphological features are distributed on different predicates, e.g. `gend(V,'Masc'), case(V,'Nom')`. They are collected during the transfer process. A Perl script subsequently joins all morphological features belonging to one token into one string. The morphology mapping could have been integrated in the transfer grammar, as well. But it would have slowed down the transfer procedure considerably.[16]

TIGER Column Format The output of the transfer proper is a Prolog file. A canonical postprocessing step (implemented in Prolog) converts the Prolog file, cf. Figure 9, into the TIGER column format, cf. Figure 10.

Tokenizer Modifications Furthermore, certain string manipulations of the tokenizer used in LFG parsing have to be undone. These string manipulations involve certain punctuation marks (commas surrounding embedded clauses), upper and lower case (sentence-initial), hyphenated compounds, and quotation marks. Hyphenated compounds such as *CDU-Frauen* 'CDU women' are split into two tokens by the LFG

[16] Because the system does not allow for variables ranging over predicates.

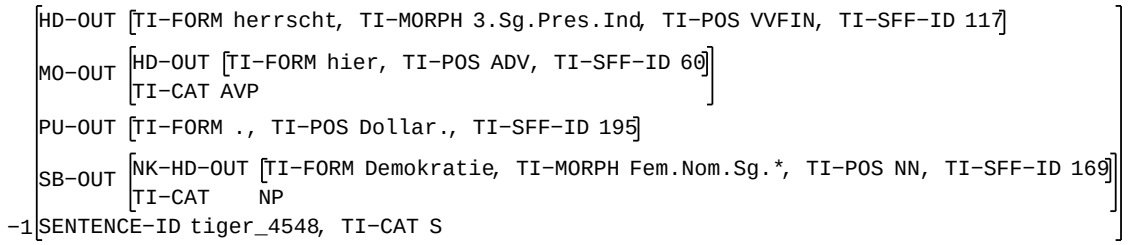


Figure 9: Transfer target format of tiger_4548

```
#FORMAT 3
#BOS 4548 102 947689949 1%% LFG 4548
hier            ADV      -               HD       502
herrscht        VVFIN    3.Sg.Pres.Ind   HD       500
Demokratie      NN       Fem.Nom.Sg.*    NK       501
.               $.       -               -        0
#500            S        -               -        0
#501            NP       -               SB       500
#502            AVP      -               MO       500
#EOS 4548
```

Figure 10: TIGER column format of tiger_4548

tokenizer, and analyzed accordingly by both the LFG morphology component and the transfer algorithm. In contrast, the TIGER annotation scheme treats hyphenated compounds as one token, therefore requiring adjustments after transfer. Concerning punctuation, the tokenizer problems are more difficult to solve. Typically, commas surround embedded clauses, but the second comma is omitted if the embedded clause immediately precedes the sentence final punctuation. However, in order to avoid different rule versions of embedded clauses (with and without second comma), the LFG tokenizer provides for additional commas in front of the sentence final punctuation mark. These additional commas have to be deleted after transfer. Quotation marks in German may enclose phrases and chunks as well as arbitrary sequences of words (non-constituents). They can even cross sentence boundaries. Hence stating a rule that deals with at least the most common types of quotation is nearly impossible. Therefore the LFG tokenizer deletes quotation marks and the postprocessing procedure has to recover them. All those modifications are dealt with mainly by a string comparison done directly after transfer. It compares transfer output strings to the original input sentences before LFG parsing. In case of faulty output, missing elements are inserted or superfluous ones deleted as appropriate. The repaired structures are then checked and, if necessary, completed by the human annotator using the semi-automatic ANNOTATE tool.

Redundant Information As already mentioned in Section 3.1, certain non-branching category projections are not represented in the TIGER format. This points to another specific feature of the TIGER annotation format, namely the tendency to avoid redundancy. To illustrate this feature, see the graphs of the example tiger_4548 shown in two different versions in Figure 11. The first tree is the current output version of TIGER Transfer as presented so far. The second tree shows the actual TIGER version: all nodes dominating just one daughter node have been omitted. This is motivated by the fact that the missing information is redundant in the sense that it can be inserted automatically. Furthermore, the annotation task becomes easier if the structure is less complex and less nodes have to be checked and corrected by the human annotator. Finally, since the structure is flatter, bigger parts of a tree can be displayed on the screen. In TIGER Transfer the respective projections are deleted in the course of the postprocessing.

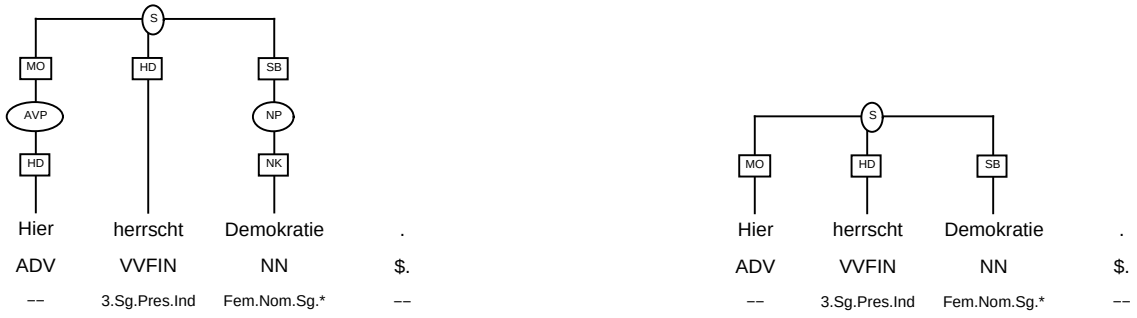


Figure 11: TIGER graph with and without non-branching nodes

4 Results

The German LFG Grammar analyzes on average 50.1% of the sentences, roughly 70% thereof are assigned a correct parse[17]. The sentences that are parsed contain 16.0 words on average and the average parsing time is 8.29 sec/sentence. Without using the OT filter mechanism a sentence gets 35,577 analyses on average (median: 20). After OT filtering, the average number of analyses drops to 16.5 (median: 2), cf. Dipper (2000). About 2,000 sentences of the TIGER corpus have been annotated by LFG parsing.

Although there are limits to the corpus coverage, the application of the grammar in annotation seems useful, in particular when combined with another annotation method. High quality treebanking requires at least two annotation passes for each sentence, and a comparison step. If one of the passes is classical manual annotation and the other pass is based on a full-depth grammar, the strength of the two techniques are combined, leading to a more consistent overall result.

5 Outlook

In the last section, we will discuss potential extensions to the LFG-based annotation: (i) partial analyses ('fragments'); (ii) morphology and lemma annotation; (iii) consistency checks. All three application tasks would benefit from a coverage extension of the grammar, an outline of which we will discuss at the very end of this paper.

Partial analyses For the annotation of new sentences a feature of the XLE system can be exploited, the 'Fragment mechanism'. This mechanism allows for every sentence to get at least a partial analysis. Hence the LFG-based annotation is no longer restricted to 35% of the corpus.

The fragment mechanism works as follows. A sentence is first parsed as usually; only if it does not get an analysis, XLE reparses the sentence. In this second parse XLE only tries to construct certain constituents, specified in advance by the grammar writer, e.g. `NP[std]`, `PP[std]` (ordinary NPs and PPs). Typically those constituents are maximal projections as defined in the grammar. Hence they are not chunks but may be complex and even recursive; an adjacent relative clause, e.g., is always part of the respective NP-fragment. Other constituents that are suited for fragments are subordinate clauses (adverbial and subcategorized), since their left and right boundaries are marked clearly.

First experiments with fragments were very promising. For instance, noun chunks were analyzed with a precision of 89% and a recall of 67%; prepositional chunks were found with a precision of 96%

[17] 10% of the sentences failed because of gaps in the morphological analyzer; 6% failed because of storage overflow or timeouts (with limits set to 100 MB storage and 100 seconds parsing time). 10% of the parsed sentences were not evaluated wrt. the correct analysis because they received more than 30 analysis after OT-filtering.

(ignoring PP-attachment for the evaluation) and a recall of 79%, cf. Schrader (2001).

The partial analyses yielded by the fragment mechanism would have to be mapped into partial analyses in the TIGER format and subsequently completed by human annotators supported by the tool ANNOTATE. In this scenario, TIGER Transfer has to be modified to deal with partial input and output.

Annotation of morphology It is planned in the TIGER project to add an annotation of morphology and lemma information to the sentences annotated syntactically so far; new sentences will be annotated with all information types. For this task the LFG grammar can be exploited easily. Each LFG analysis of a sentence automatically contains morphology and lemma information. The transfer already provides for a mapping of the respective tags.
There is is even a straightforward way to supply additional morphology and lemma information for sentences already annotated with syntactic structure, without having to disambiguate the parses again manually. Parts of the annotation of a sentence (e.g. predicate-argument structure, adjunct attachment) are transformed into Prolog terms. The sentence is parsed as usual and all analyses are stored in the Prolog export format. A test routine then picks out the analysis corresponding to the Prolog terms derived from the annotation. Thus the already existing annotation replaces the manual disambiguation step. Now the morphology and lemma information of the selected analysis is converted to the TIGER format. This way, large parts of the already annotated corpus can be automatically enriched by the LFG grammar with new information.

Consistency checks The method sketched in the preceding paragraph can also be used to perform consistency checks. Especially in the domain of part-of-speech tags, human annotators easily overlook errors. However, for a high quality corpus such as the TIGER treebank, correct part-of-speech tags are as important as correct structures. Likewise for application such as TIGERSearch (a query tool for the TIGER corpus), it is often necessary to rely on part-of-speech information, e.g. when looking for postnominal adverbs as in *Hans selbst* 'Hans himself'. The flat annotation style applied in TIGER makes this point even more important. In TIGER there is no structural property (node, label) in an NP differentiating between, e.g. a determiner, an attributive adjective, and the head noun. The only difference is their part-of-speech tags `ART`, `ADJA`, `NN/NE`, respectively.

As sketched above, information such as predicate-argument-structure of the existing annotation will be mapped to Prolog terms thus abstracting from part-of-speech tags. The sentence is parsed and disambiguated automatically, and the transfer generates the TIGER representation format. This way, part-of-speech tag errors will be detected automatically.

Coverage extension Furthermore the annotated corpus could be exploited for extending the coverage of the grammar – this would of course improve the results of the mentioned tasks performed with the grammar. With these applications in mind, it would be justified to aim particularly at systematic extension of coverage with respect to the given corpus.

There are two ways in which it is realistic to expect a possible coverage extension, now that the annotated TIGER corpus is available. The first strategy relies on 'classical' grammar writing, i.e., involving a linguist who identifies missing rule parts or lexicon entries. When this technique is applied, trying to extend a grammar that already has a relatively broad coverage, a predominant problem is the detection of unintended interactions. How can one exclude that a rule modification required for a given sentence cause the grammar to break down on a number of other sentences? Only if a sufficient amount of realistic sentences is used for a comparative test can modifications be accepted with reasonable confidence.

Here, the treebank can be used as a test suite in regression testing. While any collection of sentences can be used to compare the grammar behavior before and after a modification, only the annotation of

the correct analysis will guarantee the desired grammar behavior. (In very short, unambiguous sentences this is less of a problem, but when realistic sentences change from eighty to sixty readings, inspection of the solutions would be required to make sure that the twenty readings lost are irrelevant.) Technically, the TIGER-derived test suite used in grammar development will be a list of strings annotated with target structures in a similar way as proposed in Kuhn (1998) (we use a slightly different scheme now, in which XLE's export representation is compared with the stored target representation). In essence, the annotation structures will specify predicate-argument structure and modifier attachment (as mentioned above), but leave further details open to the grammar.

Let us now turn to the second strategy we would like to experiment with for extending coverage with respect to the TIGER corpus. The basic idea turns on the fact that a high-quality symbolic grammar has to be highly restricted when it is applied to unseen text, in order to avoid that overgeneration leads to a proliferation of readings per sentence. For instance, the subcategorization frames listed in the verb lexicon are kept very restricted, although parsing failures are often due to missing frames for a known verb. But the alternative of relaxing such restriction would require some external control of the additional readings that would arise.

With the target annotation present in the TIGER corpus, this external control is actually in place. So, when the grammar is used to reparse the corpus, it makes sense to relax some of the strict conditions encoded in the lexicon and rules. The larger set of readings arising for each sentence will be cut down immediately by allowing only parses that meet the annotated predicate-argument structure (again using the test suite). Thus, not only sentences that are covered by the 'classical' grammars could be reparsed successfully, but hopefully also some significant proportion of previously uncovered sentences.

Ultimately, one could run training experiments with the statistical model of Riezler et al. (2000) over the relaxed grammar. The TIGER corpus would serve as 'complete data', i.e. supervised training material, so the relaxed grammar could then also be applied on unseen sentences with external control over the readings (in this case the external control is exerted by the statistical model). Evaluating such an experiment can be expected to be highly revealing about the information sources required for a large-scale unification grammar.

References

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Teebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Brants, Thorsten. 1999. *Tagging and Parsing with Cascaded Markov Models – Automation of Corpus Annotation*, volume 6 of *Saarbrücken Dissertations in Computational Linguistics an d Language Technology*. Saarbrücken, Germany.

Dipper, Stefanie. 2000. Grammar-based Corpus Annotation. In *Proceedings of LINC-2000*, pp. 56–64, Luxembourg.

Dipper, Stefanie. to appear. *Implementing and Documenting Large-scale Grammars – German LFG (working title)*. PhD thesis, IMS, University of Stuttgart.

Emele, Martin, Michael Dorna, Anke Lüdeling, Heike Zinsmeister, and Christian Rohrer. 2000. Semantic-based Transfer. In *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 359–376.

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.

Kameyama, Megumi, Ryo Ochitani, and Stanley Peters. 1991. Resolving Translation Mismatches with Information Flow. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*, Berkeley, CA.

King, Tracy Holloway, Stefanie Dipper, Annette Frank, Jonas Kuhn, and John Maxwell. to appear. Ambiguity Management in Grammar Writing. *Journal of Language and Computation*.

Kuhn, Jonas. 1998. Towards Data-intensive Testing of a Broad-coverage LFG Grammar. In Bernhard Schröder, Winfried Lenders, Wolfgang Hess, and Thomas Portele (eds.), *Computers, Linguistics, and Phonetics between Language and Speech, Proceedings of the 4th Conference on Natural Language Processing – KONVENS-98*, pp. 43–56, Bonn. Peter Lang.

Kuhn, Jonas. 2001. Generation and parsing in Optimality Theoretic syntax – issues in the formalization of OT-LFG. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality-theoretic Syntax*, pp. 313–366. Stanford: CSLI Publications.

Lee, Hanjung. 2001. Markedness and Word Order Freezing. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality-theoretic Syntax*, pp. 63–128. Stanford: CSLI Publications.

Lezius, Wolfgang. 2002. *Ein Werkzeug zur Suche auf syntaktisch annotierten Textkorpora*. PhD thesis, IMS, University of Stuttgart.

Maxwell, John, and Ron Kaplan. 1989. An Overview of Disjunctive Constraint Satisfaction. In *Proceedings of the International Workshop on Parsing Technologies*, Pittsburgh, PA.

Plaehn, Oliver, and Thorsten Brants. 2000. Annotate – An Efficient Interactive Annotation Tool. In *Proceedings of ANLP-2000*, Seattle, WA.

Riezler, Stefan, Detlef Prescher, Jonas Kuhn, and Mark Johnson. 2000. Lexicalized Stochastic Modeling of Constraint-based Grammars using Log-linear Measures and EM Training. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00), Hong Kong*, pp. 480–487.

Schrader, Bettina, 2001. Modifikation einer deutschen LFG-Grammatik für Partial Parsing. Studienarbeit.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. In *Proceedings of ANLP-97*.

Heike Zinsmeister
zinsmeis@ims.uni-stuttgart.de
IMS
University of Stuttgart
Azenbergstr.12
Postfach 10 60 37
D-70049 Stuttgart
Germany

Stefanie Dipper
dipper@ims.uni-stuttgart.de
IMS
University of Stuttgart
Azenbergstr.12
Postfach 10 60 37
D-70049 Stuttgart
Germany

Jonas Kuhn
jonask@mail.utexas.edu
Department of Linguistics
1 University Station, B5100
University of Texas at Austin
Austin, TX 78712-1196
USA