Proceedings of LFG07

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Contents

1 Editor's Note

The program committee for LFG'07 were Kersti Börjars and Aoife Cahill. We would like to thank them again for putting together the program that gave rise to this collection of papers. Thanks also go to the executive committee and the reviewers, without whom the conference would not have been possible. Again this year one of the editors also played a part in the local organizing committee, which consisted of Adams Bodomo, Joan Bresnan and Tracy Holloway King, who worked together to put on yet another successful conference. We would like to thank the Department of Linguistics at Stanford and the LSA Institute for assistance and logistics in putting on the conference. We express our appreciation to Powerset for their sponsorship and would like to thank Stefan Müller for providing us with pdflatex tools that helped compile the proceedings.

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

DIFFERENTIAL OBJECT MARKING AND TOPICALITY IN TIGRINYA

Nazareth Amlesom Kifle
University of Bergen

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

Various researchers have used coding strategies such word order, case and pronominal marking to predict asymmetries between different object functions and/or information structure roles such as topics and foci. Moreover, some studies have also suggested that there exists a correlation between different grammatical functions and information structure roles. This paper analyzes object marking in double object construction in Tigrinya. Tigrinya employs word order, case and pronominal marking for coding grammatical functions and information structure roles. Differential marking of objects depends on definiteness/specificity which simultaneously triggers case and pronominal marking. In Tigrinya this double marking strategy of definite objects implies two interdependent motivations for differential object marking. Case marking is employed to contrast definite object functions with subjects, or in other words, to create a resemblance between different object functions. Whereas pronominal marking is employed to create similarity in information structure roles between topical objects and topical subjects. Moreover, based on the pattern that applicative constructions in Tigrinya reveal, this paper argues that there is no correlation between the primary object (*OBJ*) and secondary object (*OBJ*$_\theta$), i.e. the core object functions attested in *LFG* (Lexical Functional Grammar), and the *topic* and *focus* information structure roles. Since languages vary as to which object: the base or the applied object, reveals more primary object properties, accordingly, this variation is reflected by which object associates with which information structure role.

1 Introduction

It is a widely attested phenomenon that languages code their object functions variedly (e.g. Comrie, 1979; Khan, 1984; Bessong, 1985; Croft, 1988; Aissen, 2003a, among others). Bessong (1985) designated this phenomenon as *differential object marking* (*DOM*). In some languages purely semantic factors such as *animacy* and *definiteness*, and in others information structure roles alone, i.e. *topic* and *focus*, or both trigger variation in object marking. For example, in Romanian animate-referring pronouns and proper nouns (Farkas, 1978) and

[†] I would like to thank the audience at the LFG07 conference for their useful comments and criticisms. I wish specially to acknowledge the comments and suggestions of Mary Dalrymple, Yehuda Falk and Sam Mchombo. I am also grateful to Miriam Butt who during her visits to Bergen and our ParGram (Parallel Grammar Project) meetings endowed me with valuable insights. I also wish to thank the editors of the 'Proceedings of the LFG07 Conference', whose comments and editing have influenced the quality of this paper. My special thanks also goes to my supervisor, Helge Dyvik, whose support and guidance have being crucial throughout the development and maturation of my research.

in Hebrew definite objects (Givón, 1978) are case marked. In Bantu languages animacy and definiteness/specificity determine pronominal marking of objects (Morimoto, 2002). In Semitic languages such as Amharic and Syriac definiteness as well as discourse prominence triggers case and pronominal marking in direct objects (Khan, 1984). However, in some languages verbal affixes do not always correspond with argument functions. For example, in Maithili (Indo-Aryan) the controllers of the verbal affixes can be objects with various semantic roles, obliques, possessors etc., as long as they are salient in the discourse context (Dalrymple and Nikolaeva, 2007). Aissen (2003a) investigated languages in which *DOM* depends on semantic factors to trigger dependent marking (case), and she proposed a unified generalization of the phenomena that predicts the relative markedness of objects based on the degree of prominence on the dimensions of animacy and definiteness (1). These scales indicate that the higher a direct object occurs in the hierarchy, the more likely it is to be case marked.

(1) a. Animacy Scale (Aissen, 2003a, 442)
Human > Animate > Inanimate

b. Definiteness Scale (Aissen, 2003a, 444)
Pronoun > Name > Definite > Indefinite Specific > NonSpecific

In a recent study, Dalrymple and Nikolaeva (2007) proposed a new theory of differential object marking which accounts for the information structure role of *'secondary topic'*. In their study the designation *'secondary topic'* refers to the object argument which assumes the highest discourse function after the *'primary topic'*, a discourse function that corresponds with the subject argument. Let us first give a working definition for the terms *topic* and*focus*. According to Lambrecht (1998, 118) *topic* refers to the entity that the proposition expressed in an utterance is ABOUT, and *focus* refers to the new information or pragmatic assertion added on to the pragramtic presupposition (old information). Dalrymple and Nikolaeva (2007) assert that languages treat secondary topics distinctively by coding them morphologically, either through verbal affixes or case marking, and by assigning them to a particular grammatical function, or both. Their observation goes inline with the *'theory of agreement'* proposed by Bresnan and Mchombo (1987) and the study of *'object asymmetries'* (Bresnan and Moshi, 1993; Alsina and Mchombo, 1993; Alsina, 1996) developed within the LFG (Lexical Functional Grammar) framework. Bresnan and Mchombo (1987) analyze the subject marker as an ambiguous marker between grammatical and anaphoric agreement, and the object marker as an unambiguous anaphoric/topic agreement marker. Moreover, Bresnan and Moshi (1993) use restrictions on word order and pronominal marking to predict syntactic properties of objects in constructions such as the dative shift and the applicative. They classify Bantu languages into *symmetric* and *asymmetric* languages with regard to the syntactic behaviors of their objects. In symmetric applicatives both the *verbal object* (VO), an object that a verb is initially subcategorized for as its basic argument, and *applied object* (AO), an object that a verb is subcategorized for by virtue of being marked with an applicative morpheme, reflect primary object properties. On the other hand, in asymmetric applicatives only the AO reflects primary object properties. The primary object properties are properties that a single object of a mono-transitive verb reveals by occupying the immediate post verbal position, controlling pronominal agreement, and assuming the subject function in passivization. These properties are represented by the feature [-r] which indicates the non-restricted nature of the object that acquires them. In LFG this object receives the designation *OBJ*. On the other hand, the object that does not possess such properties is assigned a [+r] feature, and is designated as secondary object or OBJ_θ. OBJ$_\theta$ is restricted to specific semantic roles such as theme, instrumental, locative, etc. (depending on individual languages), and the subscrip 'θ' is a variable that represents the class of semantic roles that OBJ$_\theta$ can be associated with (Bresnan and Kanerva, 1986; Bresnan, 2001; Dalrymple, 2001). Dalrymple and Nikolaeva (2007) maintain that there exists an obligatory linkage between grammatical functions and information structure roles. Based on their observation of data from Ostyak and Chatino, they argue that secondary topics correspond to primary objects (*OBJ*) and the non-topical (focus)/unmarked objects to secondary objects (OBJ$_\theta$).

This paper aims to investigate the conditions that instigate DOM on the one hand, and to describe the functions of the different grammatical strategies involved in marking grammatical functions and discourse functions in Tigrinya double object constructions on the other hand. This paper will be organized in the following way. First, object marking strategies in mono-transitive, distransitive and applicative constructions will be presented. Second, syntactic object properties will be described in order to distinguish between the types of objects that occur in double object constructions. Third, the function of pronominal marking in information structure roles will be analyzed. Fourth, the correlation of information structure roles to grammatical functions will be demonstrated. Finally, concluding remarks will be forward.

2 Object marking in Tigrinya

Tigrinya employs a SOV order in its syntax (Raz, 1980; Tesfay, 2002; Girma, 2003; Weldu, 2004). However, this order is not strictly followed when nominal constituents are either head-marked or/and dependent-marked since under these conditions the arguments can be reordered in various combinations for pragmatic reasons. Subjects are unmarked for case, but are obligatory marked with pronominal suffix. It is a cross-linguistically attested phenomenon that caselessness triggers agreement, but not the other way round (Falk, 2006, 101). Moreover, an indefinite object is neither case nor pronominally marked. Only definite and discourse prominent specific objects trigger case and pronominal marking. The subject and the object pronominal suffixes code the gender, number and person agreement values. This is illustrated in (2).[1]

(2) a. ላም ብዕራይ ርእያ ።
lamɨ bɨʕɨrayɨ rɨʔɨy-a.
cow.FSg bull.MSg Perf.see-SM.3FSg
'a cow saw a bull.'

b. *ብዕራይ ላም ርእያ ።
bɨʕɨrayɨ lamɨ rɨʔɨy-a.
bull.MSg cow.FSg Perf.see-SM.3FSg

Example (2a) shows the unmarked order where the verb carries only a subject pronominal suffix. If we switch the order of the subject and the object as in (3b), the sentence becomes ungrammatical which is evidenced by the agreement mismatch: the verb codes a feminine subject, but the nominal in the subject position shows a masculine gender value. When a definite object is marked with case and verbal suffix, the word order becomes unbounded as in (3).

(3) a. እታ ላም ነቲ ብዕራይ
ʔɨt-a lamɨ n-ät-i bɨʕɨrayɨ
Det-3FSg cow.FSg Obj-Det-3MSg bull.MSg
ርእያቶ ።
rɨʔɨy-a-to.
Perf.see-SM.3FSg-OM$_1$.3FSg
'The cow saw the bull.'

b. ነቲ ብዕራይ እታ ላም
nät-i bɨʕɨrayɨ ʔɨta lamɨ
Obj-Det.3MSg bull.MSg Det.3FSg cow.FSg
ርእያቶ ።
rɨʔɨy-a-to.
Perf.see-SM.3FSg-OM$_1$.3FSg
'The bull, the cow saw it.'

[1] Glossing abreviations: Appl: applicative, Def: definite, Det: determiner, F: feminine, Imperf: imperfective, Indef: indefinite, Infin: infinitive, M: masculine, O: object, Obj: objective case, OM$_1$: OBJ marker, OM$_2$: OBJ$_\theta$ marker, Pass: passive, Past: past tense, Perf: perfective, Pl: plural, Rel : relative, Pres: present tense, Poss: possessive, SM: subject marker, Sg: singular, Su: Subject, TOPIC1: primary topic, TOPIC2: secondary topic , TOPIC3: tertiary topic and VN: verbal noun

c.	*እታ	ላም	ነቲ	ብዕራይ	ርእያ ።
	ʔɨt-a	lamɨ	n-äti	bɨʕɨrayɨ	rɨʔɨy-a.
	Det-3FSg	cow.FSg	Obj-Det.3MSg	bull.MSg	Perf.see-SM.3FSg

In (3a) and (3b) the verb bears a obligatory pronominal suffix for the definite object. Example (3c) shows that a clause becomes ungrammatical if the verb does not code the definite object. In addition, the definite object is obligatorily marked by a prepositional particle '**ን**/nɨ'. This case marker non-distinctly codes definite accusative objects and dative objects regardless of their definiteness status. This marker is referred as *'objective case'* in this paper.

Sometimes specificity can trigger case and pronominal marking. When a specific object argument is understood as being affected by the action/event that the verb denotes, then it can trigger pronominal marking as in (4).

(4)	a.	ንሓደ	ሰብኣይ	ኪሕግዘኒ
		nɨ-ḥadä	säbɨʔayɨ	k-i-ḥɨgɨz-ä-ni
		obj-one.M	man.Sg	Infin-Imperf.3-help-SM.MSg-OM$_1$.1Sg
		ሓቲተዮ ።		
		ḥatit-ä-yo		
		Perf.ask-SM.1Sg-OM$_1$.3MSg		

'I asked a (certain) man to help me.

b.	እቲ	መምህር	ትማሊ	ንሓደ	ተመሃራይ	መጽሓፍ
	ʔɨt-i	mämɨhɨrɨ	tɨmali	nɨ-ḥadä	tämäharayɨ	mäṣɨḥafɨ
	Det-3MSg	teacher.Sg	yesterday	Obj-one.M	student.MSg	book-Sg
	ሂቡዎ ።					
	hib-u-wo.					
	Perf.give-SM.3MSg-OM$_1$.3MSg					

'Yesterday the teacher gave a book a (certain) student.'

In Tigrinya the numeral 'one' is used to mark specificity. As examples (4a) and (4b) show, the specifier 'one' is marked with the objective case '**ን**/nɨ' and the specified argument controls the pronominal suffix.

Tigrinya has two object pronominal forms. One form is associated with VOs. For example, in '**በሊዑዎ**-bäliʕ-u-wo/*eat-SM.3MSg-OM$_1$.3MSg*' the 'wo' suffix codes a theme object, and in '**ሂቡዎ**-hib-u-wo/*give-SM.3MSg-OM$_1$.3MSg*' it can mark either a theme or a recipient object depending on which one is more topical. The second form is composed of the prepositional clitic '**ለ**/lɨ' and pronominal suffixes. For example, the object suffix 'lu' in '**በሊዑሉ**/bäliʕ-u-lu/*eat-SM.3MSg-OM$_2$.3MSg*' is made up of 'l' which denotes a beneficiary, maleficiary, instrumental or locative semantic roles, and 'u' a third person masculine singular agreement values. However, '**ለ**/lɨ' can never be associated with a theme/patient object argument.[2]

[2] There is no one-to-one correspondence between the two object pronominal forms and their meanings. For example, OM$_1$ marks definite object arguments of transitive and ditransitive

Tigrinya is not strictly a head final language. When the verb carries agreement suffixes for both the subject and the object, it can be pre-posed as in (5).

(5)

“**ደው**	**በል**”	**ይብሎ**	**ሓደ**
däwɨ	bäl-ɨ	yɨ-bɨlo	ḥade
still	Imper.be-SM.2MSg	Imperf.SM.3MSg-say-OM$_1$.3MSg	one

ካብቶም	**ቆልዑ**	**ነቲ**	**ሰብኣይ ።**
kabɨtomɨ	qolɨʕu	n-ät-i	säbɨʔayɨ.
of-Det.3MPl	child.Pl	Obj-Det-3MSg	man.Sg.

‘ “Stop!” says, one of the children to the man. / One of the children tells the man to stop.

(Newspaper corpus: Hadas Ertra 2007, Issue 16, no. 236)

In this example, the verb is fronted, and both the subject and object follow it. The subject and the object can also be dropped, and in this way the verb can stand alone as a complete clause.

Therefore, case and pronominal marking of objects in monotransitive clauses is determined by definiteness and specificity. In the following section we will extend this discussion to analyze double object constructions. Verbs in Tigrinya admit only one object pronominal suffix at a time. Since restrictions on pronominal marking have been used to predict object properties in double object constructions (e.g. Bresnan and Kanerva, 1986; Bresnan and Moshi, 1993; Alsina and Mchombo, 1993; Harford, 1993, for Bantu languages), we will investigate the syntactic restriction in double object constructions in order to characterize syntactic properties of object functions.

2.1 Objects in ditransitive clauses

In Tigrinya double object constructions that involve ditransitive verbs employ different syntactic restrictions than applicative constructions. Ditransitive verbs such as '**ወሃበ**/wähabä-*give*', '**ዓደለ**/ʕadälä-*distribute*', '**ነገረ**/nägärä-*tell*' and '**መሃረ**/mäharä-*teach*, etc.' initially subcategorize for two object arguments. These objects are coded with the same form of pronominal affix. Let us first consider a clause with two indefinite objects (6).

verbs, and it also codes affected AO of intransitive verbs, as in **መጺኡዋ**-mäṣi-u-wa/came-SM.3MSg-OM$_1$.3FSg which means *'He/it came/arrived to her'*. Therefore, we adopt the glosses OM$_1$ and OM$_2$ as identifiers of the two morphological forms rather than as designators of meanings.

(6) a. **እቲ መምህር ትማሊ ንተመሃሮ መጽሓፍቲ**
ʔɨt-i mämɨhɨrɨ tɨmali nɨ-tämäharo mäṣɨḥafɨ-ti
Det-3MSg teacher.Sg yesterday Obj-student.Pl book-Pl
ዓዲሉ ።
ʕadil-u.
Perf.distribute-SM.3MSg
‘Yesterday the teacher distributed books to students.’

b. **እቲ መምህር ትማሊ ንተመሃሮ መጽሓፍቲ**
ʔɨt-i mämɨhɨrɨ tɨmali nɨ-tämäharo mäṣɨḥafɨ-ti
Det-3MSg teacher.Sg yesterday Obj-student.Pl book-Pl
ዓዲሉ ።
ʕadil-u.
Perf.distribute-SM.3MSg
‘Yesterday the teacher distributed books to students.’

As the examples in (6) show there is no fixed position to code these objects. They can only be distinguished by their case marking; indefinite theme objects are unmarked, while recipient objects are marked with the objective case '**ን**/nɨ'. When both objects are indefinite, neither of them can control verbal suffix. However, the two clauses express different emphasis, in (6a) emphasis is neutral, but in (6b) the pre-posed theme object is more emphasized. An analogous pattern is attested in **ትግረ** /Tɨgɨrä (Raz, 1980), an Abyssinian Semitic language closely related to Tigrinya.

Similarly, in a ditransitive clause that involves a definite recipient object and an indefinite theme object the word order is unbound as in (7).

(7) a. **እቲ መምህር ነቶም ተመሃሮ መጽሓፍቲ**
ʔɨt-i mämɨhɨrɨ n-ät-omɨ tämäharo mäṣɨḥafɨ-ti
Det-3MSg teacher.Sg Obj-Det-3MPl student.Pl book-Pl
ዓዲሉዎም ።
ʕadil-u-wom.
Perf.distribute-SM.3MSg-OM$_1$.3MPl
‘The teacher distributed books to the students.’

b. **እቲ መምህር መጽሓፍቲ ነቶም ተመሃሮ**
ʔɨt-i mämɨhɨrɨ mäṣɨḥafɨ-ti n-ät-omɨ tämäharo
Det-3MSg teacher.Sg book-Pl Obj-Det-3MPl student.Pl
ዓዲሉዎም ።
ʕadil-u-wom.
Perf.distribute-SM.3MSg-OM$_1$.3MPl
‘The teacher distributed books to the students.’

When objects are different in terms of case marking, they are not ordered in relation to each other. Moreover, only a definite object can control pronominal suffixes, thus in examples (7a and 7b) the recipient object is pronominally marked. However, when the theme object is definite, then word order becomes constrained, and the theme object is pronominally marked as in (8).

(8) a. እቲ መምህር ነቲ መጽሓፍቲ ንተመሃሮ
ʔɨt-i mämɨhɨrɨ nä-t-i mäṣɨḥafɨ-ti tämäharo
Det-3MSg teacher.Sg Obj-Det-3MSg book-Pl Obj-student.Pl
ዓዲሉዎ ።
ʕadil-u-wo.
Perf.distribute-SM.3MSg-OM$_1$.3MSg.

‘The teacher distributed the books to students.’ [3]

b. *እቲ መምህር ንተመሃሮ ነቲ መጽሓፍቲ
ʔɨti mämɨhɨrɨ nɨ-tämäharo n-ät-i mäṣɨḥafɨ-ti
Det.3MSg teacher.Sg Obj-student.Pl OBJ-Det-3MSg book.Pl
ዓዲሉዎ ።
ʕadil-u-wo.
Perf.distribute-SM.3MSg-OM$_1$.3MSg

When the theme object is definite, it obligatorily precedes the recipient object, as in (8a). Since both objects appear similar in terms of their case marking, thus they are coded by their position. As a result, if their order is switched, the sentence becomes ungrammatical, as in (8b). Similarly, when both objects are definite, word order becomes bound, but the verb can bear a pronominal suffix for either object depending on discourse prominence, as in (9).

(9) a. እቲ መምህር ነቲ መጽሓፍቲ ነቶም
ʔɨt-i mämɨhɨrɨ n-ät-i mäṣɨḥafɨ-ti n-ät-omɨ
Det-3MSg teacher.Sg Obj-Det-3MSg book-Pl Obj-Det-3MPl
ተመሃሮ ዓዲሉዎም ።
tämäharo ʕadil-u-wom.
student.Pl Perf.distribute-SM.3MSg-OM$_1$.3MPl

‘The teacher distributed the books to the students.’

b. እቲ መምህር ነቲ መጽሓፍቲ ነቶም
ʔɨt-i mämɨhɨrɨ n-ät-i mäṣɨḥafɨ-ti n-ät-omɨ
Det-3MSg teacher.Sg Obj-Det-3MSg book-Pl Obj-Det-3MPl
ተመሃሮ ዓዲሉዎ ።
tämäharo ʕadil-u-wo.
student.Pl Perf.distribute-SM.3MSg-OM$_1$.3MSg

‘The teacher distributed the books to the students.’

Example (9) shows that a definite theme object must precede a recipient object. Another interesting observation is that in this context either the definite theme object, as in (9b) or the definite recipient object, as in (9a) can be marked with a pronominal suffix depending on the speaker's choice of which referent to highlight. Therefore, definiteness constrains both objects equally. As implied in these examples, conditions on animacy do not have a bearing on object marking in Tigrinya. Had it been relevant, the recipient object would be prioritized over the theme object for pronominal marking.

[3] In Tigrinya a plural form of an inanimate noun (e.g. *'books'* in (8)) has a collective reading. It is determined by a masculine singular article, and the verb agrees with the determiner.

2.2 Objects in double object applicative constructions

A double object applicative construction codes a *VO* and an *AO*. These objects are coded by distinct pronominal forms: OM$_1$ and OM$_2$ respectively. In Tigrinya various semantic roles such as a *beneficiary, maleficiary, instrumental, locative* and *goal* can be expressed applicatively. Applicative constructions involve different syntactic restrictions than ditransitive constructions. Lets first consider an applicative construction that involves a theme vs. beneficiary objects as in (10).

(10) a.

ንስኻ	ንዮናስ	ዓጋዜን	ሃዲን-ካ-ሉ ።
nɨsɨ-k̲a	nɨ-yonasɨ	ʕagazenɨ	hadinɨ-ka-lu
Pro-2MSg	Obj-Yonas.M	deer.Sg	Perf.hunt-SM.2MSg-OM$_2$.3MSg.

'You hunted (for) Yonas a deer.'

b.

ንስኻ	ዓጋዜን	ንዮናስ	ሃዲን-ካ-ሉ ።
nɨsɨ-k̲a	ʕagazenɨ	nɨ-yonasɨ	hadinɨ-ka-lu
Pro-2MSg	deer.Sg	Obj-Yonas.M	Perf.hunt-SM.2MSg-OM$_2$.3MSg.

'You hunted (for) Yonas a dear.'

c.

ንስኻ	ነታ	ዓጋዜን	ንዮናስ	ሃዲን-ካ-ሉ ።
nɨsɨ-k̲a	n-ä-ta	ʕagazenɨ	nɨ-yonasɨ	hadinɨ-ka-lu
Pro-2MSg	Obj-Det-3FSg	deer.Sg	Obj-Yonas.M	Perf.hunt-SM.2MSg-OM$_2$.3MSg.

'You hunted (for) Yonas the dear '

An applicative construction with a theme vs. beneficiary AO is not bound in terms of its word order. As examples (10a) and (10b) show either object can occur in either position. Moreover, the verb always codes the AO regardless of whether the VO is definite or not, as in (10c). This implies that an AO is the most topical object; in fact AOs are always individuated or definite objects. As Donohue states (in Peterson, 2007, 83) *"the essential function of applicative constructions is to indicate that the entity the construction refers to has a greater discourse salience or topic continuity than would otherwise be expected of it"*. Moreover, since an AO acquires its core object status by virtue of the applicative morpheme, if the verb does not bear this morpheme, the construction ceases to be an applicative clause. Since the beneficiary and recipient roles lack distinct prepositions for their oblique expression, they can only be expressed in double object constructions.

Applicative constructions with applied roles such as the instrumental and locative reveal slightly different syntactic restrictions. For example, unlike the objects with beneficiary vs. theme roles, the instrumental/locative vs. theme objects are required to stay in a fixed position, as in (11).

(11) a. እቲ ሰብኣይ ነቲ ፋስ ዕንጨይቲ
?ɨt-i säbɨ?ayɨ n-ät-i fasɨ ʕɨnɨç̣äyɨti
Det-3MSg man.MSg Obl-Det-3MSg ax.Sg wood.Sg
ፈሊጹሉ ።
fäliṣ-u-lu
Perf-chop-SM.3MSg-OM$_2$-3MSg
'The man chopped wood with an ax.'

b. እቲ ሰብኣይ ነቲ ፋስ ነቲ ዕንጨይቲ
?ɨt-i säbɨ?ayɨ n-ät-i fasɨ n-ät-i ʕɨnɨç̣äyɨti
Det-3MSg man.Sg Obl-Det-3MSg ax.Sg Obj-Det-3MSg wood.Sg
ፈሊጹሉ ።
fäliṣ-u-lu
Perf-chop-SM.3MSg-OM$_2$-3MSg
'The man chopped the wood with the ax.'

c. *እቲ ሰብኣይ ዕንጭይቲ ነቲ ፋስ
?ɨt-i säbɨ?ayɨ ʕɨnɨç̣äyɨti n-ät-i fasɨ
Det-3MSg man.MSg wood.Sg Obl-Det-3MSg ax.Sg
ፈሊጹሉ ።
fäliṣ-u-lu
Perf-chop-SM.3MSg-OM$_2$-3MSg

In applicative constructions that involve a theme vs. instrumental/locative object, the AO must precede the VO regardless of whether the VO is definite or not, as in (11a, 11b). If we reverse the order, the construction becomes ungrammatical as in (11c). Moreover, like in a beneficiary vs. theme applicative construction, the verb always codes the applied roles. However, if the VO is topicalized instead of the AO, the instrumental/locative roles are expressed obliquely since they posses distinct prepositions ('ብ/bɨ-' instrumental and 'ኣብ/?abɨ' locative) as in (12).

(12) እቲ ሰብኣይ ነቲ ዕንጭይቲ ብፋስ
?ɨt-i säbɨ?ayɨ n-ät-i ʕɨnɨç̣äyɨti bɨ-fasɨ
Det-3MSg man.Sg Obl-Det-3MSg wood.Sg with-ax.Sg
ፈሊጹዎ ።
fäliṣ-u-wo
Perf-chop-SM.3MSg-OM$_1$-3MSg
'The man chopped the wood with an ax.'

In example (12) the verb codes a definite VO, thus the instrumental role is expressed in an oblique phrase. In terms of word order, the definite theme object must precede the oblique phrase.

To sum up, Tigrinya employs a complex interaction of word order, case and pronominal marking in coding objects. Since unmarked objects are not ordered in relation to each other, verb adjacency cannot be taken as an argument for determining primary object properties. However, restrictions on pronominal marking display asymmetry between the two object. In the following section we will investigate if the restrictions on pronominal marking correlate with the passive typology that characterizes Tigrinya.

3 Primary object properties

A vast body of research in object asymmetries uses the correlation of properties such as pronominal marking and passive typology as a proof for primary objecthood (Bresnan and Moshi, 1993; Alsina and Mchombo, 1993; Alsina, 1996). These studies claim that the underlying properties of a language manifested in passive typology are one and the same as those manifested by the descriptive properties of a language, i.e., restrictions on word order and pronominal marking. In Tigrinya a ditransitive verb can bear a pronominal suffix for either of the two objects. Thus, in this regard both objects may display primary object properties. However, in an applicative construction only the AO controls the pronominal suffix; thus only AOs may display primary object properties with respect to pronominal marking. Bellow we will compare these properties with those reflected in passivization. Let us first consider example (13).

(13) a. እቶም ተመሃሮ መጽሓፍቲ ተዋሂቦም ።
ʔɨt-omɨ tämäharo mäṣɨḥafɨti tä-wahib-omɨ
Det-3MPl student.Pl book.Pl Pass-Perf.give-SM.3MPl
'The students are given books.'

b. እቲ መጽሓፍቲ ንተመሃሮ ተዋሂቡ ።
ʔɨt-i mäṣɨḥafɨti nɨ-tämäharo tä-wahib-uɨ
Det-3MSg teacher Obl-Det-3MPl student.Pl book.Pl
'The books are given to students'

Since the recipient (13a) and the theme (13b) arguments can function as subjects in passivization, both display primary object properties. Another strong piece of evidence for primary objecthood is the ability of the passive verb to admit object suffixes, as in (14). Asymmetric type languages like Chichewa lack this property (Bresnan and Moshi, 1993; Alsina and Mchombo, 1993).

(14) a. እቶም ተመሃሮ ነቲ መጽሓፍቲ
ʔɨt-omɨ tämäharo n-ät-i mäṣɨḥafɨ-ti
Det-3MPl student.Pl Obj-Det-3MSg book-Pl
ተዋሂቦምዎ ።
tä-wahib-om-wo.
Pass-Perf.give-SM.3MPl-OM$_1$.3MSg
'The students are given books.' [4]

[4] This sentence can also have a reflexive reading *'The students gave themeselves to the books.'* since the passive and the reflexive verb forms are marked with the same morphological form.

b. **እቲ መጽሓፍቲ ንተመሃሮ**
ʔɨt-i mäṣɨḥafɨ-ti nɨ-tämäharo
Det-3MSg book-Pl Obl.students.Pl
ተዋሂቡዎም ።
tä-wahib-u-womɨ.
Pass-Perf.distribute-SM.3MSg-OM$_1$.3MPl
‘The books are given to students.’

The passive verb in (14a) bears a subject and an object pronominal suffixes for the recipient and the theme arguments respectively, but example (14b) shows the reverse, here the theme role is expressed as a subject and the recipient as an object. As these examples show Tigrinya displays an alternating passive type in ditransitive constructions. Therefore, both objects exhibit primary object properties with respect to passivization as well. However, in applicative constructions only the theme role can function as a subject in passivization, as in (15).

(15) a. **እቲ መጽሓፍ ንሳባ ተገዚኡላ ።**
ʔɨt-i mäṣɨḥafɨ n-saba tä-gäzi-u-la
Det-3MSg book.Sg Obl-Saba.F Pass-perf.buy-SM.3MSg-OM$_2$.3FSg
'The book was bought (for) Saba.'

b. *****ሳባ መጽሓት ተገዚኣ ።**
saba mäṣɨḥafɨ tä-gäzi-ʔa
Saba.F book.Sg Pass-perf.buy-SM.3FSg

In (15a) the theme role is expressed as a subject, and the beneficiary role as an object. However, applied roles such as beneficiary/locative/intrumental can never be expressed as subject functions, as in (15b).

The type of asymmetry displayed by Tigrinya applicative constructions is different in a crucial way than the asymmetry type found in Bantu languages. In Bantu languages the AO displays primary object properties. While in Tigrinya, with respect to pronominal marking, the AO shows primary object properties, but with respect to passivization only the theme object reflects primary object properties. And thus, passivization and pronominal marking reflect uncorrelated properties. In addition, the passive verb can admit a pronominal suffix for the AO as in (15a). Therefore, Tigrinya has symmetric objects both with the [-r] features classified as OBJs in its ditransitive clauses. In contrast, in its applicative construction it has asymmetric objects, with the AO getting the [+r] feature and thus classified as OBJ$_\theta$ and the VO getting the [-r] feature and classified as OBJ.

4 Object marking and information structure roles

A vast body of research predicts a correlation between grammatical agreement and discourse functions. Among these, Givón's (1976) typological study has been very influential. Givón systematically explained various diachronic data and demonstrated that agreement markers had historically evolved from topic pronouns to clitic pronouns then to redundant agreement markers. Givón claims that agreement and anaphoric marking are the same processes and that they cannot be distinguished either diachronically or synchronically. His proposal regarding the puzzling differences between the pronominal and nominal structure found in the *imperfective* and *perfective* verb conjugation systems in Semitic languages is specially commended in Semitic studies. Tigrinya, like its Semitic peers, has two types of verb conjugation systems, the *imperfective* and the *perfective*. The imperfective verb conjugation is a prefix one which displays partial agreement specification as a prefix and partial specification as a suffix which shows a 'person-stem-(gender, number)' ordering (e.g. Amharic in Baye, 2006, 196). However, in the perfective verb form the subject pronominal marker is a suffix. It is beyond the scope of this paper to outline the historical development in word order and agreement marking in Tigrinya; however, it would suffice to say that the differences between the imperfective (e.g. ይጽሕፍ/yɨ-ṣɨḥɨf-ɨ/SM.3-write-SM.MSg) and the perfective (e.g. ጽሓፈ/ṣɨḥaf-ä/write-SM.3MSg) subject pronominal forms on the one hand, and the perfective and gerundive (e.g. ጽሒፉ/ṣɨḥif-u/write-SM.3MSg) subject pronominal forms on the other hand, reflect different grammaticalization processes. Nevertheless, the different forms function as pronominal subject affixes.

The morphological similarity between independent pronouns, and the subject and object pronominal affixes seem to support the basic claim that the pronominal affixes evolved from topic pronouns to agreement markers. The prefix pronominal system shows little resemblance to the independent pronouns in Tigrinya. However, the gerundive form is quite similar to the endings of independent pronouns as in table (1). [5]

Table 1: *Pronoun and pronominal affixes*

Values	**Subjective**	**Objective**	**Perf.eat-SM-OM**
Pro.3MSg	ንሱ-nɨs-u	ንዓኡ-nɨʕaʔ-u	በሊዑዎ-bäliʕ-u-(w)o
Pro.3FSg	ንሳ-nɨs-a	ንዓኣ-nɨʕaʔ-a	በሊዓታ-bäliʕ-a-(t)a
Pro.3MPl	ንሳቶም-nɨsat-omɨ	ንዓኦም-nɨʕaʔ-omɨ	በሊዖምዎም-bäliʕ-omɨ-(w)omɨ
Pro.3FPl	ንሳተን-nɨsat-enɨ	ንዓአን-nɨʕaʔ-enɨ	በሊዐንአን-bäliʕ-enɨ-(ʔ)änɨ

[5] Vowel sequence is not permitted in Tigrinya syllabic structure. Therefore, epenthetic segments such as 'w' and 't' are inserted between the subject and the object pronominal suffixes for

This table shows that the subject and object suffixal conjugation of the gerundive verb are etymologically related to the personal pronouns in Tigrinya.

The theory of agreement proposed by Bresnan and Mchombo (1987) has influenced a wide body of research in LFG. Bresnan and Mchombo convincingly demonstrated that subject pronominal affixes are ambiguous markers of grammatical and anaphoric agreement; whereas, object pronominal suffixes are only a topic-anaphoric markers. First, the fact that the anaphorically linked arguments and pronominal affixes in a discourse are required to show gender, number and person agreement reflects the anaphoric function of pronominal affixes. Second, in languages like Tigrinya the object pronominal marker is induced by definiteness. Therefore, it can only mark referential, salient and individuated object arguments; and thus it is a topic marker rather than a grammatical agreement marker. On the other hand, the subject marker is obligatory, and it can correspond with non-referential and non-topical subject. For example, Lambrecht (1998, 137) argues that in a context where the whole predicate is focused, the subject is not a topic since the whole proposition is covered by the focus discourse function. The subject marker functions as an anaphoric marker when it corresponds with topical subject NPs in a discourse. We will illustrate this by way of examples from a real discourse context as in (16).[6]

(16) **ኣብ ማዕዶ፡ ሓደ ምትሃት ዚመስል**
ʔabɨ maʕɨdo: ḥadä mɨtɨhatɨ z-i-mäsɨlɨ
at distance: one.MSg ghost Rel-Imperf.SM.3-resemble.SM.MSg
ጻዕዳ ነገር ረኣኹ። ናባይ ምስ ቀረበ ግን፡
ṣaʕɨda nägärɨ räʔa-k̲u ። nab-ayɨ mɨsɨ qäräbä gɨnɨ:
white thing Perf.see.SM.1Sg:: to-1Sg when Perf.near-SM.3MSg but:
ጀለብያ ዘለበሰ ቆልዓ ምኻኑ
ğäläbɨya zɨ-läbäs-ä qolɨʕa mɨ-k̲an-u
jelabia Rel-Perf.wear-SM.3MSg child.Sg VN-be-Poss.3MSg
ተገንዘብኩ።
tä-gänɨzäbɨ-ku ።
Perf.realize-SM.2Sg.

'At a distance, I saw a white thing which resembled a ghost. But when it neared me, I realized its being (it was) a child that wore a Jellabia (robe).'

(Source: Hadas Ertra 2007, Issue 17, no.13)

the 3MSg 'u' and 3FSg 'a' as it is shown in the first and the second rows in this table.

[6] This excerpt is taken from a Tigrinya newspaper 'Hadas Ertra' column series called *'One World'*. The columnist, Amanuel Sahle, is a famous journalist and a linguist. His book 'A Comprehensive Tigrinya Grammar' is one of the most referred to work in Tigrinya studies. He is a member of the 'Medial Language Standardization Committee' in the Eritrean Ministry of Information. Amanuel is believed to be a good writer and a model to other journalists on how to write good/appropriate Tigrinya. Thus, I believe the quality of the text is guaranteed and that the examples employ a standard use of the issue at hand.

In the above discourse, the antecedent of the incorporated subject pronoun (SM.1Sg) that the verbs *'see'* and *'realize'* bear is not realized either as an independent pronoun or as a full NP. The referent can only be recovered from the discourse context. Since the text is a narrative discourse and employs the 'first person narrative' technique, the speaker/writer refers to himself through the incorporated pronoun 'I'. Bresnan and Mchombo (1987) state that in order *'to satisfy the completeness and coherence conditions [such] argument functions (SUBJ, OBJ, etc.) must be expressed syntactically within the phrase structures headed by the predicator, or expressed morphologically on the head itself, or else remain unexpressed'.* They also stress that only the anaphoric agreement relations can be non-local to the agreeing predicator. Under these conditions then, the subject pronominal suffix functions as an anaphoric or a topic marker in this sentence since it agrees with an argument which is not locally present in the same clause. In this sentence, the object argument is new information in this discourse context. The numeral '**ሓደ**-ḥadä/one.M' can function as a marker of specificity or indefiniteness depending on the basic meaning of the verb. In this sentence it introduces an indefinite object, since this object does not control any verbal suffix. Thus the object is required to stay in the same clause as the predicator, and it assumes a focus discourse function. The second sentence consists of a dependent and an independent clause which are demarcated by the sentence adverbial 'but/however'. The dependent and independent clauses denote old and new information respectively. The verb *'near-SM.3MSO'* in the dependent clause contains a subject incorporated pronoun which corresponds to the object antecedent 'ḥadä mɨtɨhatɨ z-i-mäsɨlɨ ṣaʕɨda nägärɨ- *a white thing which resembles a ghost*'. Whereas, the verb in the independent clause 'Perf.realize-SM.2Sg' contains a subject incorporated pronoun which agrees with the subject incorporated pronoun in the previous sentence. Thus, this examples illustrate that the subject and the object pronominal affixes are incorporated pronouns which anaphorically link to topic NPs or even to another incorporated pronoun in a discourse.

Moreover, the subject pronominal affixes can also function as grammatical agreement markers. Constructions which involve psyche verbs in Tigrinya code non-referential subjects, and thus they are non-topical. These constructions are characterized by OSV word order where the topic object is preposed and the non-referential subject is postposed as in (17).

(17)	**ሕጂ፡**	**(ኣነ/ንዓይ)**	**ደኺሙኒ**	**ኣሎ።**
	ḥɨǧi:	(ʔanä/nɨʕayɨ)	däk̠im-u-ni	ʔal-o.
	now:	(I/me)	Perf.tire-SM.3MSg-OM$_1$.1Sg	Pres.exist-SM.3MSg.

'Now, I am tired./ Lit. Now, it has tired me.'

Example (17) shows that the main verb 'däk̠im-u-ni/tired-it-me' codes a non-referential 3MSg subject and a 1sg experiencer object, and the auxiliary 'ʔal-o/exist.it' codes a non-referential 3MSg subject. In such constructions either the nominative or the objective personal pronouns can be used as referents of the object markers. It is a widely observed property in Tigrinya that topical objects get a nominative case, and that makes them comparable to subjects.

5 Information structure roles and grammatical function alignment

One of the key points that Aissen (2003a) makes in her theory of *DOM* is the correlation between grammatical functions and the semantic conditions that induce grammatical marking. Subjects are assumed to be high in prominence and objects are low. She characterizes this type of relationship as *'markedness reversal'* which denotes that the semantic features that are marked for subjects are unmarked for objects and vice versa. The relative markedness of grammatical functions is expressed through the HARMONIC ALIGNMENT of the relational hierarchy (given in example 1) either on the animacy or the definiteness dimension. For example, the harmonic alignment for the definiteness features is schematized in (18).

(18) *Su/Pron >> *Su/Name >> *Su/Def-Spec >> *Su/Non-spec
*Obj/Non-spec >> *Obj/Def-Spec >> *Obj/Name >> *Obj/Pron

This diagram shows that subjects positioned on the left-most adge of the hierarchy are more marked than those at the right-most adge, and the opposite holds for objects. The main point behind such a representation of DOM is to underline the function of grammatical marking. According to Aissen (2003a) grammatical marking is employed in order to differentiate subjects from objects. For example, since definite objects are functionally similar to subjects in terms of prominence, they carry grammatical marking that contrasts them with subjects. However, Dalrymple and Nikolaeva (2007) argue that DOM *"arises from the need to give an overt expression to properties that is* **common** *to objects and subjects"*. In their view case and agreement marking have a *'coding'* function rather than a discriminatory function. They claim that their approach accounts for languages such as Persian and Maithili which assign grammatical marking to secondary topics independently of their syntactic roles.

Tigrinya which involves case and pronominal marking seems to divide the two functions- 'discriminatory' and 'coding' suggested by Aissen and Dalrymple and Nikolaeva (2007) respectively, between these two coding strategies. For example, in monotransitive clauses subjects and indefinite objects are comparable in terms of their case marking: both are unmarked. However, a definite theme object contrast with a subject since the former is marked whereas the latter is unmarked. On the other hand, in double object clauses an indefinite theme object and an object bearing other semantic roles contrast with each other since the former is unmarked and the latter is marked, but they appear comparable when the theme is definite. The discriminatory function is even more pronounced when word order is considered. Whenever the two categories are comparable, word order becomes bound, and when they contrast it becomes unbound. In terms of pronominal marking, subjects are obligatorily marked with pronominal suffixes. But, subject pronominal affixes do not always code topical subjects. Thus, only the anaphoric function of the subject pronominal affixes and object pronominal suffixes underline the similarities between a topical subject and topical object.

Various researchers have indicated that there exists a tendency for a certain grammatical function to link to a centain information structure role. For example, in their comparative study of Hindu/Urdu and Turkish, Butt and King (2000) analyze the weak/nonspecific object which assumes a focus discourse function as a primary object *OBJ*, and the strong/specific objects which are non-focused as a *OBJ*$_\theta$. In contrast, Dalrymple and Nikolaeva (2007), based on the pattern revealed in Ostyak and Chatino, argue that in these languages the secondary topics link to OBJs while the non-topic object to *OBJ*$_\theta$. They maintain that *" it is the marked, topical object rather than the unmarked, non-topical object that displays more properties characteristics of core grammatical functions."* They schematized the alignment of information structure roles to grammatical function as in (19):

(19) Dalrymple and Nikolaeva (2007)

TOPIC	TOPIC2	FOCUS
|	|	|
SUBJ	OBJ	OBJ$_\theta$/ OBL$_\theta$

However, this correlation cannot predict the relative prominence displayed by objects in Tigrinya. In applicative constructions as discussed in section 3, even though applied objects control pronominal marking, and thus are topical, they do not acquire primary object properties. Therefore, there is no correlation between primary object functions and secondary topics. Tigrinya applicative constructions reveal the pattern schematized in (20).

(20) Alignment in Tigrinya applicative constructions

TOPIC1	TOPIC2	TOPIC3	FOCUS
|	|	|	|
SUBJ	OBJ$_\theta$	OBJdef	OBJindef/OBL$_\theta$

The double marking possibility allows two grammatically marked topic objects in Tigrinya. In double object constructions the object that is prioritized for pronominal marking is high in prominence, while a case marked definite object is less prominent when both occur in the same clause. Thus the former is assigned a secondary ranking and the latter a tertiary ranking in topicality. As we can see, this pattern differs from the pattern proposed by Dalrymple and Nikolaeva (2007) in (19). Therefore, this suggests that the correlations between grammatical functions and information structure roles vary from language to language: thus it is language specific.

6 Conclusion

Tigrinya employs word order, case and pronominal affixes in marking grammatical functions and discourse functions. *DOM* is triggered by definiteness and specificity. Definite and discourse prominent specific objects are both head and dependent marked. This double marking strategy implies that there are two motivations for *DOM*. Case marking is employed to contrast definite objects with subjects, or in other words, to create a resemblance between different object functions. Whereas pronominal marking is employed to create similarity in information structure roles between topical objects and topical subjects.

Moreover, Tigrinya makes a formal distinction between ditransitive constructions and applicative constructions. Verbs in ditransitive clauses subcategorize for two VOs and applied verbs subcategorize for a VO and an AO. Tigrinya reveals symmetric properties of objects in its ditransitive constructions, and asymmetric properties in its applicative constructions. However, the type of asymmetry that Tigrinya shows is the reverse version of the asymmetry that languages like Chichewa (Bantu) have. In Tigrinya an AO does not acquire all the syntactic properties of a single object in monotransitive constructions. Even though objects with applied roles control pronominal marking, they cannot assume a subject function in passivization. This challenges the correlation claimed between the passive typology and the restrictions on pronominal marking. The double object data from Tigrinya suggests that the two morphosyntactic operations belong to different grammatical processes.

Therefore, this paper argues that the applicative processes is a topicalization operation in which the AO assumes a more prominent discourse function than the VO, and the applicative morpheme functions as a topic/anaphoric marker in accordance with what is asserted by Bresnan and Mchombo (1987). However, languages vary in the assignment of grammatical function to AO. In some language, for example in Bantu, it assumes the primary object function, and in others, for example in Tigrinya, it assumes the secondary object function. In Tigrinya, the property of being a subject function in passivization is reserved for the VO. The VO assumes a less prominent discourse function *tertiary topic*, i.e. less prominent than the secondary topic in its definite status when both occur in the same clause. A definite VO, even though it does not have precedence for pronominal marking over the AO, is case marked, a property which is also acquired by a definite object of a monotransitive verb. However, an indefinite/unspecific VO cannot control a pronominal and cannot be case marked; thus it assumes a non-topic/focus discourse function. Therefore, since the primary object property displayed by passivization may not correlate with those properties displayed by restrictions on word order and pronominal marking, further research must demonstrate which properties must be taken as basic in order to determine primary objecthood.

References

Aissen, Judith. 2003a. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* 21(3), 435-483.

Alsina, Alex. 1996. Passive Types and The Theory of Object Asymmetries. *Natural Language and Linguistic Theory* 14, 673-723.

Alsina, Alex and Mchombo, A. Sam. 1993. Object Asymmetries and the Chichewa Applicative Construction. In Sam Mchombo (ed.), *Theoretical Aspects of Bantu Grammar*, volume 1, Chapter 1, Stanford, CA: CSLI Publications.

Baye, Yimam. 2006. The Interaction of Tense, Aspect, and Agreement in amharic Syntax. In John Mugane, John P. Hutchingson and Dee A. Worman (eds.), *Selected papers: Proceedings of the 35th Annual Conference on African Linguistics*, pages 193-202, Somerville, MA: Cascadilla Proceedings Project.

Bessong, Georg. 1985. *Empirische Universalienforschung: Differentielle Objektmarkierung in den Neuiranischen Sprachen*. Tübingen: Gunter Narr.

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

Bresnan, Joan and Kanerva, Jonni. 1986. Locative Inversion in Chichewa: A Case Study of Factorization in Grammar. *Linguistic Inquiry* 20(1), 1-50.

Bresnan, Joan and Mchombo, Sam. 1987. Topic, Pronoun, and Agreement in Chichewa. *Language* 63(4), 741-782.

Bresnan, Joan and Moshi, Lioba. 1993. Object Asymmetries in Comparative Bantu Syntax. In Sam Mchombo (ed.), *Theoretical Aspects of Bantu Grammar*, volume 1, Chapter 2, pages 47-87, Stanford, CA: CSLI Publications.

Butt, Miriam and King, Tracy Holloway. 2000. Null Elements in Discouse Structure. In K. V. Subbarao (ed.), *Papers from the NULLS Seminar*, Delhi: Motilal Banarsidass.

Comrie, Bernard. 1979. Definite and Animate direct objects: A natural class.

Linguistica Silesiana III, 13-21.

Croft, William A. 1988. Agreement vs case marking and direct objects. In Michael Barlow and Charles A. Ferguson (eds.), *Agreement in Natural Languages: Approaches, Theories, Descriptions*, pages 159-179, Stanford, CA: CSLI Publications.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. San Diago: USA: Academic Press.

Dalrymple, Mary and Nikolaeva, Irina. 2007. Topicality and nonsubject marking: Agreement, casemarking, and grammatical function. MS, 2007.

Falk, N. Yehuda. 2006. *Subject and Universal Grammar: An Explanatory Theory*. Cambridge Studies in Linguistics 113, New York: Cambridge University Press.

Farkas, Donka. 1978. Direct and Indirect Object Reduplication in Romanian. In *Papers from the Fourteenth Regional Meeting of the Chicago Linguistic Society*, pages 88-97.

Girma, Awgichew Demeke. 2003. *The Clausal Syntax of Ethio-Semitic*. Ph. D. thesis, Department of Linguistics, University of Tromsø, Tromsø.

Givón, Talmy. 1976. Topic, Pronoun, and Grammatical Agreement. In Charles N. Li (ed.), *Subject and Topic*, New York: Academic Press.

Givón, Talmy. 1978. Definiteness and Referentiality. In J. Greenberg (ed.), *Universals of Human Language*, volume 4, pages 291-330, Stanford, CA: Stanford University Press.

Harford, Carolyn. 1993. The Applicative in Chishona and Lexical Mapping Theory. In Sum A. Mchombo (ed.), *Theoretical Aspects of Bantu Grammar*, volume 1, Chapter 3, Stanford, CA: CSLI Publications.

Khan, Geoffrey. 1984. Object Markers and Agreement Pronouns in Semitic Languages. *Bulletin of the School of Oriental and African Studies, University of London* 47(3), 468-500.

Lambrecht, Knud. 1998. *Information structure and sentence form: topic, focus and the mental representation of discourse referents*. Cambridge University Press, second edition.

Morimoto, Yukiko. 2002. Prominence Mismatches and Differential Object marking in Bantu. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG02 Conference*, Stanford, CA: CSLI publications.

Peterson, David. 2007. *Applicative Constructions*. Oxford: Oxford University Press.

Raz, Shlomo. 1980. Tigre Syntax and Semitic Ethiopian. *Bulletin of the School of Oriental and African Studies, University of London* 43(2), 235-250.

Tesfay, Tewolde Yohannes. 2002. *A modern Grammar of Tigrinya*. Via G. Sanvonarola, 1-00195 Roma: U.Detti.

Weldu, Michael Weldeyesus. 2004. Case Marking Systerms in Two Ethiopian Semitic Languages. *Colorado Research in Linguisitcs* 17(1), 1-16.

THE ARCHITECTURE OF I-STRUCTURE

Maia Andréasson
University of Aarhus & Göteborg University

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

In this paper, the inventory and the architecture of a separate i-structure representation in LFG are discussed in relation to Swedish data. It is argued that a discourse function SCENE needs to be distinguished from RHEME and GROUND. It is furthermore proposed that a characteristic that singles out sentence adverbials from other clausal modifiers is their ability to function as focus operators (cf. Rooth 1992) and that FOCUS (as is ACTIVATION) is a discourse feature, separate from the discourse functions. The analysis builds on data from a corpus study of Swedish word order (Andréasson 2007) where the information dynamics of the sentence is found to be the key to explaining much of the possible word order variation.

1 Introduction

Much recent work within LFG deals with word order phenomena in relation to the information structural component of the grammar. Just a few examples are Butt and King 1996, 2000; Choi 1997, 1999; Cook 2001; Cook and Payne 2006; King 1995; 1997; Mycock 2007; O'Connor 2006. Over the years the analyses have shifted from realising discourse function such as TOPIC and FOCUS as Grammatical Discourse Functions in f-structure to proposing a separate and more elaborated representation, mostly called *i-structure*.

In this paper I discuss the architecture of a separate i-structure representation in LFG in relation to Swedish data, mainly concerning different adverbial categories, and their function and placement. In particular, I discuss the discourse function SCENE, and the role of sentence adverbials as FOCUS OPERATORS (cf. Rooth 1992; 1996). The analysis builds on generalisations from the corpus study of Swedish word order in Andréasson (2007) where the information dynamics of the sentence is found to be the key to explaining much of the possible variation.

Following Börjars, Engdahl and Andreasson (2003) and Andréasson (2007), I assume a flat structure in the area following the finite verb in Swedish main clauses (or the subordinating conjunction in subordinate clauses). The c-structure of a main clause where the main verb is non-finite – the sentence in example (1) – is illustrated in figure 1 below.[1]

(1) Därför har Ellis förstås inte gett Síle lammet.
That's-why have-PRS *Ellis of-course* NEG *give Síle lamb*-DEF
'That's why Ellis hasn't given Síle the lamb.'

[†] I thank the audience of LFG07 in Stanford University for helpful comments.

[1] In main clauses where the main verb is finite, the clause does not have a VP, (see Börjars, Engdahl and Andreasson (2003) and Andréasson (2007)).

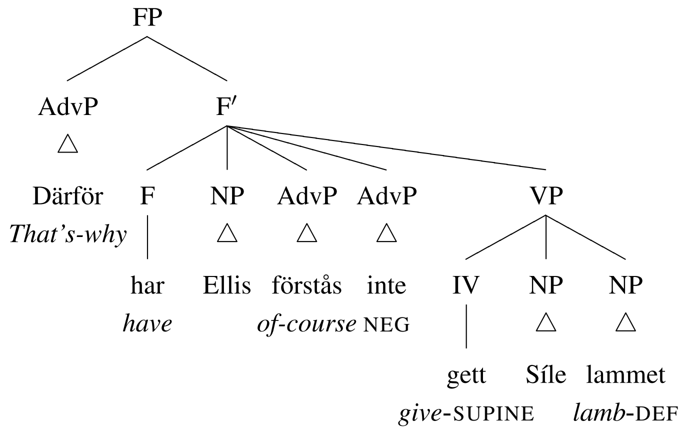


FIGURE 1: *C-structure*

There are substantial possibilities for word order variation in the area of the clause following the finite verb/subordinating conjunction in Swedish, here called the F′ domain.[2] Most of the variation takes place in V2 or V1 clauses with a finite main verb, but there is also the possibility of word order variation between subjects and adverbials in the F′ domain regardless of whether there is a VP or not.

2 The terminology of information structure

A major factor that influences word order in many languages is information structure, or *information dynamics*, which is the term that I use. By information dynamics I understand the relation between on the one hand the speaker's assumptions and intentions and on the other hand the information packaging of the linguistic expression. The term *information dynamics* thus covers more than *information structure* which is sometimes used to denote only the packaging aspect.

The term *information structure* was introduced by Halliday (1967), and since this component of grammar relates to several other components, syntacticians, text linguists, and phoneticians have developed terminologies for this notion that are seemingly similar, but at a closer look are entirely different (for an elaborated discussion, see Vallduví and Engdahl 1996; see also, for example, O'Connor 2006).

When syntacticians use the notion 'topic' in terms like *topicalisation*, this means an element in the beginning of a clause, mostly a constituent with a canonical position elsewhere in the clause being moved to an initial position. For the text linguist the notion may relate information in several separate clauses, as is the case for the term *continuous topic*. The phonetician may use the term *focus* denoting a stress pattern for emphasised elements in a clause, while some grammarians use the

[2] To avoid discussion on whether the functional projection headed by the finite verb/subordinating conjunction should be CP or IP, I employ the dummy F for *Functional*.

term for the very constituent that is emphasised and yet others employ the terms *topic* and *focus* for the partition of a clause in pragmatic relations (or discourse functions). For this reason the notions used in this article are defined explicitly in this section.

I make a distinction between **discourse functions** (DF) like RHEME, GROUND and SCENE, see figure 2 below and section 2.1, 2.2, and 3 below, and **discourse features** like FOCUS (cf. Rooth 1992; 1996) and ACTIVATION, see section 2.3 (Gundel, Hedberg and Zacharski 1993; Lambrecht 1996; O'Connor 2006). All these concepts are formalised in the LFG i-structure, see section 5.

Term	Definition
RHEME	the information in a statement that is intended to increase the listener's knowledge
GROUND	constituents that relate the rheme to questions the speaker assumes are under discussion
SCENE	constituents that relate the proposition to a temporal, spatial or circumstantial context, that is not under discussion

FIGURE 2: *Discourse functions in LFG i-structure*

A brief note is needed on my use of the term FOCUS. I adapt the notion of FOCUS of Rooth (1992, 1996), where its primary function is the evoking of alternatives. The focusing of a constituent raises the assumption of the existence of an alternate set to the one expressed. This alternate set may be overt in the context or presupposed.

Figure 3 is a simple overview of a production perspective of information dynamics. Given the meaning the speaker wants to express, her assumptions of the information state, and her intentions with the utterance, the information is partitioned in discourse functions and assigned discourse features that may be formalised in the LFG i-structure, here represented by an *i*. The partition leads to language specific mapping choices, choices that determine which information packaging (Vallduví 1992, Vallduví and Engdahl 1996; cf. Chafe 1976) is optimal for the communication of the speaker's intention to be felicitous. In felicitous communication, the discourse functions and features interpreted by the hearer matches those intended by the speaker.

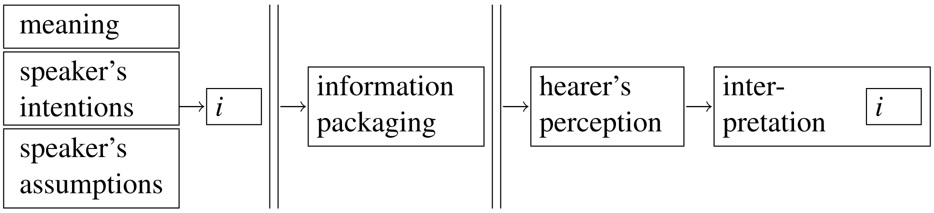


FIGURE 3: *Information dynamics: production perspective*

2.1 RHEME

In this article, the term RHEME (originally from the Prague school, cf. Firbas 1966) is defined as the information in a statement that is intended to increase the listener's knowledge. The definition coincides with the notion FOCUS in, for example, Vallduví (1992), Vallduví and Engdahl (1996) and Lambrecht (1996).

In a question-answer pair like the one in (2), the question of the listener reading something is brought up for discussion and the speaker requests information about the name of the item read. The elliptical answer *Kranes konditori* supplies the only information needed, the RHEME.

(2) a. Vad läser du?
what read-PRS *you*
'What are you reading?'
QUD: $\langle\ ?\lambda x\ (read\ (you, x\ ))\rangle$[3]
b. [RHEME Kranes konditori.]
... *Krane*-GEN *café*
'Krane's café'

In (3), the question of the listener's crying is brought up for discussion. Here the RHEME is not an elliptical answer, but consists of a full sentence: *Min undulat har dött*.

(3) a. Varför gråter du?
why cry-PRS *you*
'Why are you crying?'
QUD: $\langle\ ?\lambda Y\ (Y\ (cry\ (you)))\rangle$
b. [RHEME Min undulat har dött.]
... *my budgie have*-PRS *die*-SUPINE
'My budgie has died'

Which information is rhematic is not always a question about "old" vs. "new". Also information that is accessible in the context, and hence "old", may be part of the rhematic portion of a clause (cf. Vallduví and Engdahl 1996).

[3] QUD, see section 2.2, below.

In (4) the speaker requests information about who is going to accompany David to the Museum of World Culture. In the answer, the rhematic portion consists of the pronoun *jag* referring to the speaker, information that must be considered accessible as the referent is appearing in the situational context.[4]

(4) a. Vem ska följa med David till Världskulturmuseet?
who FUT *follow with David to Museum-of-World-Culture*-DEF
'Who's coming with David to the Museum of World culture?'
QUD:$\langle ?\lambda x\ (\textit{följa med David till V-museet}(x))\rangle$

b. [RHEME Jag].
... *I*
'I am.'

2.2 GROUND

As mentioned before, the answer to a question may consist only of a rheme, but it is also possible, and sometimes even necessary, to include some GROUND material, that is, constituents that relate the RHEME to questions that are under discussion, as in (5) and (6), below (for a more elaborate description of GROUND, see Vallduví 1992; Vallduví and Engdahl 1996; for the notion *under discussion* see Ginzburg 1996; forthc.).

(5) a. Vad läser du?
what read-PRS *you*
'What are you reading?'
QUD:$\langle\ ?\lambda x\ (\textit{read}\ (\textit{you}, x\ ))\rangle$

b. [GROUND Jag läser] Kranes konditori.
... *I read*-PRS *Krane*-GEN *café*
'I am reading Krane's café'

(6) a. Varför gråter du?
why cry-PRS *you*
'Why are you crying?'
QUD:$\langle\ ?\lambda Y\ (Y\ (\textit{cry}\ (\textit{you})))\rangle$

b. [GROUND Jag gråter för att] min undulat har dött.
... *I cry*-PRS *for that my budgie have*-PRS *die*-SUPINE
'I am crying because my budgie died'

When a speaker utters a sentence, this is done in relation to a context that she assumes is at least partly known to the listener. This context does not merely consist of the previous discourse, but comprises a wider range of circumstances as well as the actual words and sentences spoken previous to the utterance. World

[4] Erteschik-Shir (2007:17f.) states that the speaker and listener may be seen as "permanently available topics". This is does not imply that speaker and hearer can never be included in or constitute the rhematic portion of a clause, but merely that they must be regarded as accessible, even if they have not been overtly mentioned in the previous written or spoken context.

knowledge, memories from previous conversations, concrete items and/or events connected to the situational context, in short, all things that the speaker assumes are mentally accessible to her and the listener, may be considered "known". In this accessible context there is some information the speaker assumes she and the listener agree is under discussion.

Ginzburg (to appear) formalises these assumptions in his Dialogue Game Board as mental lists of *Questions Under Discussion*, QUD. Such QUDs exist in the mind and describe the information state of the speaker and the listener. They are formalised as ordered sets that are updated with information from the most recent utterance.

When the speaker enters the room in (6) and finds the listener in tears, this brings up the crying on the QUD. The question in (6 a) adds the question of the reason for the crying, and the answer in (6 b) adds the budgie and its death as the reason to the QUD. The QUDs shown in this article are a very simplified version of the speaker's QUD, included only to show a formalisation of what is assumed to be under discussion and what is not.

Dialogues are often used to show what is under discussion. But it is equally possible to analyse other text types. In example (7) below, the fact that a man was putting on clothes is brought to the reader's attention in the first sentence. Because of this, the first part of the second sentence, *Han tog på sig*, must be regarded as GROUND, while *grå kostym och en blå skjorta* is the rhematic portion.

(7) Han gick tillbaka till sovrummet och lyckades med viss möda
he go-PST *back to bedroom*-DEF *and succeed*-PST *with some effort*
klä sig. [GROUND Han tog på sig] [RHEME grå kostym och
dress REFL ... *he take*-PST *on* REFL ... *grey suit and*
en blå skjorta].
a blue shirt
'He went back to the bedroom and managed with some effort to get dressed. He put on a grey suit and a blue shirt.'

The GROUND portion of a clause consists of material that must be present in the clause for one or more reasons. They may be there to ensure that the RHEME is related to the right question under discussion. But sometimes there are also grammatical reasons for GROUND material not to be suppressed in a clause, like in Swedish, where clauses without a subject are mostly ungrammatical except in colloquial speech and in certain genres, such as diary and post-card writing (cf. Mörnsjö 2002, Magnusson 2007). In languages like Italian, on the other hand, it is a well known fact that GROUND subjects are generally left out, when not contrastive.

2.3 A brief note on accessibility

The page limit of this article does not allow more than a brief comment about accessibility and the activation of referents related to word order. The notion of accessibility or activation (cf. Gundel et al. 1993, Lambrecht 1996; on the discourse feature ACTVN, see O'Connor 2006) is closely related to the choice of linguistic expressions and to their positions in a clause, and elements with a high activation tend to appear early in a sentence.

Activation is nevertheless not inseparably connected to discourse functions, as we saw in example (4) above. Accessible information appearing early in a clause is hence not necessarily a consequence of accessible constituents being the GROUND of the sentence, even if GROUND by definition consists of accessible information.

In example (8) below, the referent of the pronoun is mentioned in the immediate context and thus accessible to the extent that it would be infelicitous to refer to her with a proper name. On the other hand, the pronoun is part of the rhematic portion of the clause. The information requested in the question is the reason for the listener not stopping, and the fact that Alma's waving is under discussion in the context does not make *her* part of the GROUND in the answer.

(8) a. Varför stannade du inte när Alma vinkade?
why *stop*-PST *you* NEG *when Alma wave*-PST
'Why didn't you stop when Alma waved?'
b. Jag [$_{\text{RHEME}}$ såg henne inte].
I ... *see*-PST *her* NEG
'I didn't see her.'

On the other hand, the accessibility of the object *henne* (which may be formalised as an +ACTVN feature in the i-structure) requests that it be placed as early as possible in the clause, and the pronoun is consequently shifted from its canonical position after the negation.

In a context where the referent Alma is not accessible, neither in the spoken text nor in person standing waving on the pavement, see (9) below, the proper name *Alma* has the feature –ACTVN and appears in the canonical object position in Swedish after the negation.[5]

(9) a. Varför stannade du inte?
why *stop*-PST *you* NEG
'Why didn't you stop?'
b. Jag [$_{\text{RHEME}}$ såg inte Alma].
I ... *see*-PST NEG *Alma.*
'I didn't see Alma.'

[5] For an object to appear before the negation in the F' domain in Swedish (i.e. *object shift*), an accessibility level that allows use of a pronoun is requested. A more elaborate analysis of the information dynamics and impact of the object's activation state in object shift will be performed within the post doc project *Pronominal Object Shift in Swedish and Danish* 2007–2008, at the University of Aarhus, Denmark, see <http://maia.andreasson.googlepages.com/objektsskifte>.

3 The discourse function SCENE

It is, as mentioned above, well known that GROUND material in general precedes the rhematic portion of a clause, and this is mostly the case also for constituents which have the possibility for word order variation in the F' domain in Swedish clauses. Interestingly, some constituents providing not previously mentioned but clearly not rhematic information show a somewhat different distribution. These are constituents that relate the proposition to a temporal, spatial, or circumstantial context, which is not under discussion. I call this discourse function SCENE (cf. Chafe 1976; Lambrecht 1996).

The corpus investigation in Andréasson (2007) shows that constituents denoting SCENE show a robust distributional pattern in relation to RHEME and GROUND in the F' domain. They align to the right of any GROUND constituents, but to the left of rhematic constituents, see (10) below, where < means 'appears before'.

(10) *F' domain*: [finite verb[6]] < GROUND < SCENE < RHEME

Example (11) below is from an article where the runner Marian Jones is under discussion. In this sentence the subject *Jones* is GROUND and appears immediately before an adverbial describing the temporal frame of the proposition *den senaste tiden* 'lately'.

(11) Enligt Guardian har [$_{\text{SUBJ}}$ Jones] den senaste tiden
according-to Guardian have-PRS ... *Jones* ART *latest time*
satts under hård press av sponsorn Nike,
put-SUPINE-PASSIVE *under hard pressure of sponsor*-DEF *Nike*
som betalar Jones runt 70 miljoner kronor för att hon
REL-PRON *pay*-PRS *Jones around 70 million*-PL *crown*-PL *for that she*
marknadsför företagets produkter.
market-PRS *company*-DEF-GEN *product*-PL
'According to the Guardian, Jones has lately been under hard pressure from the sponsor Nike, who pays Jones about 70 million Swedish crowns for marketing the company's products.'

In example (22 a), on the other hand, the same kind of information, denoted by the adverbial, *i höstas*, 'this autumn', instead appears immediately preceding a subject that is part of the rhematic portion of the clause.

(12) På Åbro bryggeri fattades i höstas [$_{\text{SUBJ}}$ beslutet att lägga
on Åbro Brewery take-PASSIVE *in autumn* ... *decision*-DEF *to lay*
ned produktionen med läsk i returglas].
down production-DEF *with soda in returnable bottles*

[6] The finite verb appears first in this domain of the clause for grammatical reasons, since Swedish is a V2 language.

‘This autumn, a decision was made at Åbro brewery to close down the production of soda in returnable bottles’

Lambrecht (1996) categorises adverbials appearing initially in a sentence, “scene-setting adverbials”, as part of his “topic” notion. In the discussion about example (13) (Lambrecht’s 4.2 d), Lambrecht states that the scene setting topic *After the children went to school* supplies information about the temporal conditions for the rest of the sentence, that it is presupposed, and cannot be regarded as part of what is asserted (Lambrecht 1996:121, 125f., 219).

(13) (John was very busy that morning.) After the children went to SCHOOL, he had to clean the house and go shopping for the party. (Lambrecht 1996:121)

If the event of the children’s departure to school is presupposed, as suggested in Lambrecht (1996), it may be seen as accessible. On the other hand, this does not necessarily mean that the event must be under discussion.

Lambrecht’s scene-setting adverbials are closely related to the notion of “stage topic” of Erteshik-Schir (2007:16f.). This notion builds on the spatio-temporal location always being a possible TOPIC, since it is indispensable for the evaluation of truth values. Both scene-setting and stage topics build on a TOPIC notion that differs from the concept of GROUND in this article. Even if SCENE material may be presupposed, it cannot be seen as a variety of GROUND since constituents of this category are by definition under discussion.

It is moreover not possible to define SCENE as a variety of RHEME either. Although SCENE material may be inaccessible, it does not really fill an informational gap. Yet another characteristic that separates SCENE from GROUND and RHEME is that it is not possible to focus constituents denoting SCENE.

Constituents that semantically denote the frame of a sentence may, but need not, be of the discourse function SCENE. In (14) below, the speaker puts the question of the listener’s activities during the upcoming weekend on the QUD. The expression *till helgen* in the question represents a set of several points in time, for example the days during the weekend. And when the listener answers, the frame setting expressions *på lördag* and *på söndag* are focused GROUND (cf. Vallduví and Engdahl 1996: *link*; Choi 1999: *topic*).

(14) a. Vad ska du göra till helgen?
what FUT *you do in weekend*
‘What will you do this weekend?’

b. [F-GROUND På lördag] ska jag skriva klart min artikel och
... *on Saturday* FUT *I write ready my article and*
[F-GROUND på söndag] ska jag måla om i sovrummet.
... *on Sunday* FUT *I paint* PRT *in bedroom*-DEF
‘On Saturday, I will finish writing my article, and on Sunday, I will repaint the bedroom’

In (15), on the other hand, the speaker requests information about the temporal frame for the event of the listener meeting with the mutual friend Alma. Here the rhematic portion of the clause is the constituent semantically denoting the frame: *På måndag klockan tre*. The answer may be elliptical or include reference to the event: *Det ska jag göra* [...].

(15) a. När ska du träffa Alma?
when FUT *you meet Alma*
'When are you going to meet Alma?

b. (Det ska jag göra) [$_{\text{RHEME}}$ På måndag klockan tre]
that FUT *I do ... on Monday clock*-DEF *three*
'(I will do that) On Monday at three.'

3.1 Setting the SCENE in a cleft construction

Expressions denoting SCENE are often placed early in a sentence. In the F' domain they appear before the RHEME and another common position is in the first position of the clause immediately before the finite verb (cf. Chafe 1976: 50f.; Lambrecht 1994:118; Teleman, Hellberg and Andersson 3:446, 3:492f., 4:432[7]). In news reports, constituents denoting a SCENE often appear in matrix clauses of cleft constructions; see example (16) below.

(16) Det var *sent på lördagskvällen* som ett gäng ungdomar
it be-PST *late in Saturday-night*-DEF *that a band young people*
enligt vittnesuppgifter helt oprovocerat attackerade
according to witness information totally unprovoked attack-PST
gående vid Stigbergstorget.
pedestrian-PL *by Stigbergstorget*
'It was *late Saturday evening* that, according to a witness, a band of young people made an unprovoked assault on pedestrians at Stigbergstorget.'

Cleft constructions are often otherwise used to mark a focused constituent. In this example, on the other hand, the frame setting adverbial *sent på lördagskvällen* is clefted, but not focused. The non-clefted portion of the clause in turn contains new information about an assault that is brought up for discussion in the preceding text and is not presupposed, as is the case when focused constituents are clefted (Rooth 1992, 1996).

4 Sentence adverbials and prominent information

Sentence adverbials (SADVL) are traditionally defined as 'clausal modifiers', that is modifiers of the proposition including the subject, as opposed to so called VP-adverbials, which modify only the verb and its complements. For an account of the

[7] Swedish SCENE may also be placed as the last and necessarily non-stressed adverbial in a clause.

differences between these two categories, see Dalrymple 2001:269–274.

It is, however, not unknown that adverbials that relate the proposition to a temporal, spatial, and circumstantial frame also modify the entire proposition, rather than only the verb phrase, even if these are not usually referred to as "sentence modifiers" (Nikula 1986). These adverbials are semantically comparable to some of the traditional sentence adverbials, namely those that affect the truth values of the sentence, since both these categories set the conditions under which the proposition is true.

What is it then that distinguishes sentence adverbials from other sentence modifiers? In the following, I will show that the defining characteristic for sentence adverbials seems to be their ability to function as information dynamic FOCUS operators.

In examples (17) and (18), sentence adverbials are used as FOCUS operators. The context of example (17) is a discussion about a violent handball game where the player Anders Franzén got beat up. In this sentence the SADVL *också*, 'also', serves as a focus operator, highlighting the rhematic constituent, *Mikael Franzén*.

(17) Inne på linjen fick *också Mikael Franzén* ta emot
in on line-DEF *get*-PST *also Mikael Franzén take*-INF *towards*
mycket stryk
much beating
'On the line, Mikael Franzén was also beaten up'

The context of the example in (18) is a dietician giving advice on infant diets. Here the pronoun *jag* referring to the writer is focused GROUND; by placing the pronoun after the SADVL, the writer aims to evoke the presupposition that there exists an alternate set of persons that are not of the same opinion.

(18) Om barnet går upp i vikt, ser i alla fall *inte jag*
if child-DEF *go*-PRS *up in weight see*-PRS *in all*-PL *case* NEG *I*
det som några problem om barnet äter vegetariskt.
that as any-PL *problem if child*-DEF *eat*-PRS *vegetarian*
'As long as the child is gaining weight, there is no apparent problem – in my opinion – if the child follows a vegetarian diet.'

It is not unknown that there are adverbs, like *only* and *even*, that function as focus operators (cf. Rooth 1996). These adverbs often appear in places where other SADVLs may not, for instance in NP:s, structurally adjoined to a focused element: *Even Alma sometime cooks.* But other SADVLs also relate to the focused part of a sentence and may function as FOCUS operators.

In example (19) the subject *Alma* is placed after the sentence adverbial *faktiskt*, 'actually', in the F′ domain. *Faktiskt* is syntactically restricted to appear only in propositional contexts. In this sentence, it modifies the sentence and is syntactically a sister to the subject in the F′ domain. The placement of *Alma* after *faktiskt* in (19) nevertheless evokes an interpretation where it is unexpected that Alma cooks and that an alternative set of one or several persons normally does the cooking.

(19) Ikväll lagade faktiskt Alma maten.
tonight cook-PST *actually Alma food*-DEF
'Tonight, it was actually Alma that cooked the meal.'

A clear indication of the FOCUS operator function of sentence adverbials is that in Swedish these, but not other adverbial categories, may be clefted with another constituent of a clause, as in (20) below (Andréasson 2007). This example shows that it is possible to cleft a constituent preceded by a sentence adverbial, like *faktiskt* 'actually', while this is not possible with a manner adverbial, *långsamt* 'slowly', or a frame adverbial, *igår* 'yesterday'.

(20) It is SADVL [focus domain] that [rest of sentence]
a. Det var *faktiskt* Alma som lagade maten.
it be-PST *actually Alma that cook*-PST *food*-DEF
'It was actually Alma that cooked the meal.'
b. *Det var *långsamt* Alma som lagade maten.
'It was *slowly* Alma that cooked the meal'
c. *Det var *igår* Alma som lagade maten.
'It was *yesterday* Alma that cooked the meal'

Interestingly, the English translations of (20 b) and (20 c) are also bad, even though a thorough investigation of the possibilities in English has not been carried out. An investigation of several languages is needed to decide whether the possibility to appear in a cleft construction with another constituent is a characteristic of sentence adverbials in other languages too.

A constituent that is clefted with a sentence adverbial, as in (20 a), is always interpreted as focused and can never be interpreted as the SCENE of the sentence, which it is in the cleft construction in (16) above. If an adverbial denoting a temporal frame, like *igår* in (20 c), is clefted with a sentence adverbial, the non-clefted portion of the clause is interpreted as presupposed and the frame adverbial as focused; see (21) below.

(21) Det var *faktiskt* igår som Alma lagade maten.
it be-PST *actually yesterday that Alma cook*-PST *food*-DEF
'It was actually yesterday that Alma cooked the meal'.

The construction in (20), *It is* SADVL *[focus domain] that [rest of sentence]*, serves as a test for sentence adverbials in Swedish and distinguishes this category from other propositional modifiers (Andréasson 2007).

5 The architecture of i-structure

To sum up, the attributes relevant for the i-structure in Swedish are on the one hand the discourse functions RHEME, GROUND, and SCENE and on the other hand the discourse features FOCUS and ACTIVATION. The DF:S have various possibilities of being focused; the DF SCENE is singled out from the other discourse functions by not being possible to focus. Furthermore, the discourse function GROUND is singled out from the others since it necessarily consists of information that is under discussion and hence active.

	FOCUS	ACTIVATION
RHEME	±	±
GROUND	±	+
SCENE	–	±

TABLE 1: *Discourse functions and discourse features*

In this section I will turn to the question of what consequences the conclusions in this article will have for the architecture of a separate i-structure in LFG.

5.1 Integrating SCENE

As discussed in section 3 above, there are reasons to believe that SCENE should be treated as a discourse function distinct from GROUND and RHEME. One consequence for the architecture of i-structure is then to integrate SCENE as an attribute with a possible value, as outlined in (22) below, where the sentence in example (22 a) is repeated. Here the SCENE of the sentence, the PP *i höstas*, is the value of the DF attribute SCENE in the i-structure.

(22) a. På Åbro bryggeri fattades i höstas [$_{\text{SUBJ}}$ beslutet att
on Åbro Brewery take-PASSIVE *in autumn ... decision*-DEF *to*
lägga ned produktionen med läsk i returglas].
lay down production-DEF *with soda in returnable bottles*
'This autumn, a decision was made at Åbro brewery to close down the production of soda in returnable bottles'

b.
$$\begin{bmatrix} \text{GROUND} & \left\{\begin{bmatrix}\textit{på Åbro bryggeri}\end{bmatrix}\right\} \\ \text{RHEME} & \left\{\begin{bmatrix}\textit{fattades beslutet att lägga ner}\\ \textit{produktionen med läsk i reurglas}\end{bmatrix}\right\} \\ \text{SCENE} & \left\{\begin{bmatrix}\textit{i höstas}\end{bmatrix}\right\}\end{bmatrix}$$

5.2 Integrating focused elements

FOCUS is a feature that may affect only part of the RHEME or GROUND of a sentence. A FOCUS attribute with a ± value within the attribute-value matrices representing the various discourse functions would hence not be a satisfactory solution to formalising FOCUS in the i-structure.

It is furthermore necessary to find a way to formalise the FOCUS operators in the i-structure. I propose that the FOCUS attribute of the i-structure take a FOCUS DOMAIN and a FOCUS OPERATOR as values. The value of the domain may be linked to one of the members in the GROUND or RHEME sets by structure sharing. The value of the operator in its turn may be linked to a sentence adverbial in some cases in, for example, Swedish. It may also be linked to the prosodic structure in speech, to information packaging constructions or c-structure positions, or to the morphological structure in languages that mark focus with morphemes.

In examples (23) and (24) the i-structures of examples (17) and (18) are outlined. In these i-structures the FOCUS domains and operators are linked to GROUND or FOCUS elements by structure sharing, marked with coindexation.

(23) Inne på linjen fick *också Mikael Franzén* ta emot
in on line-DEF *get*-PST *also Mikael Franzén take*-INF *towards*
mycket stryk
much beating
'On the line, Mikael Franzén was also beaten up'

$$\begin{bmatrix} \text{GROUND} & \left\{\textit{fick ta emot mycket stryk}\right\} \\ \text{RHEME} & \left\{\begin{array}{l}\textit{också}_i \\ \textit{Mikael Franzén}_j\end{array}\right\} \\ \text{SCENE} & \left\{\textit{inne på linjen}\right\} \\ \text{FOCUS} & \begin{bmatrix}\text{OPERATOR} & i \\ \text{DOMAIN} & j\end{bmatrix}\end{bmatrix}$$

(24) Om barnet går upp i vikt, ser i alla fall *inte jag*
if child-DEF *go*-PRS *up in weight see*-PRS *in all*-PL *case* NEG *I*
det som några problem om barnet äter vegetariskt.
that as any-PL *problem if child*-DEF *eat*-PRS *vegetarian*
'As long as the child is gaining weight, there is no apparent problem – in my opinion – if the child follows a vegetarian diet.'

$$\left[\begin{array}{ll} \textsc{ground} & \left\{ \textit{jag}_j \right\} \\ \textsc{rheme} & \left\{ \begin{array}{l} \textit{i alla fall} \\ \textit{inte}_i \\ \textit{ser det som några problem om...} \end{array} \right\} \\ \textsc{scene} & \left\{ \textit{om barnet går upp i vikt} \right\} \\ \textsc{focus} & \left[\begin{array}{ll} \textsc{operator} & i \\ \textsc{domain} & j \end{array} \right] \end{array}\right]$$

6 Conclusion

On the basis of Swedish data I have argued that the discourse function SCENE needs to be distinguished from RHEME and GROUND. I have furthermore proposed, following Andréasson (2007), that a characteristic that singles out sentence adverbials from other clausal modifiers is their ability to function as focus operators. Lastly, I have proposed a sketch for an LFG i-structure that makes use of these notions.

Most LFG-analyses of information dynamics so far have dealt with individual languages, making generalisations and proposing machinery based on these. This article is no exception. I have based my proposal on the information dynamics of Swedish, and – I might add – of a limited subset of Swedish, namely declarative main clauses and only concerning the constituent order in the F$'$ domain. The analysis of Swedish in this article is hence only one contribution to the jigsaw puzzle of the architecture of i-structure.

It is not clear to what extent the analysis in this article fits in with Cook and Paynes' (2006) recent analysis of information dynamics in German. Especially their notion of TOPIC infers an aboutness that is not directly related to information that is under discussion and hence not comparable to the QUD notion used in this article. O'Connor's (2006) analysis of spoken Serbo-Croatian makes use of the notion ACTIVATION that I have not yet included for Swedish, and he also proposes a mapping between i-structure (his *d-structure*) and the prosodic component of the grammar, the p-structure. Mycock (2007) discusses the notions of interrogative and non-interrogative FOCUS in her analysis of constituent questions, a distinction that has not been included here since I analyse declarative clauses.

Information dynamics is becoming more and more important today, having impact on analyses both in non-derivational and derivational frameworks. In my view, information dynamics is a field where it would be fruitful to see even more joint work in the future. The architecture of the LFG i-structure is still an open question and will probably remain so until several researchers with thorough and detailed insights in the information dynamics of various languages work together.

7 References

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

Börjars, Kersti, Elisabet Engdahl and Maia Andréasson 2003. Subject and Object Positions in Swedish. In: Butt, Miriam and Tracy Holloway King (eds.), *Proceedings of the LFG03 Conference*. Stanford: CSLI, p. 43–58. <http://csli-publications.stanford.edu/LFG/8/lfg03.html>.

Butt, Miriam and Tracy Holloway King 1996. Structural Topic and Focus without Movement. In: Butt, Miriam and Tracy Holloway King (eds.), *Proceedings of the First Annual LFG Conference*. Stanford: CSLI. <http://csli-publications.stanford.edu/LFG/1/lfg1.html>.

Butt, Miriam and Tracy Holloway King 2000. Null Elements in Discourse Structure. In: Subbarao, K.V. (ed.), *Papers from the NULLS Seminar*. Delhi: Motilal Banarasidas.

Chafe, Wallace L. 1976. Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In: Li, Charles N. (ed.), *Subject and topic.* New York: Academic Press, p. 27–55.

Choi, Hye-Won 1997. *Information Structure, Phrase Structure, and Their Interface.* In: Butt, Miriam and Tracy Holloway King (eds.), *Proceedings of the LFG97 Conference*. Stanford: CSLI. <http://csli-publications.stanford.edu/LFG/2/lfg 97.html>.

Choi, Hye-Won 1999. *Optimizing Structure in Context: Scrambling and Information Structure*. Stanford: CSLI.

Cook, Philippa 2001. *Coherence in German: An Information Structure Approach.* Doctoral dissertation. Manchester: University of Manchester.

Cook, Philippa and John Payne 2006. Information Structure and Scope in German. In: Butt, Miriam and Tracy Holloway King (eds.), *Proceedings of the LFG06 Conference*. Stanford: CSLI, p. 124–144. <http://csli-publications.stanford.edu /LFG/11/lfg06.html>

Dalrymple, Mary 2001. *Lexical Functional Grammar*. (Syntax and Semantics 34.) New York: Academic Press.

Erteschik-Shir, Nomi. 2007. *Information Structure: The Syntax-Discourse Interface*. Oxford Surveys in Syntax and Morphology. Oxford: Oxford University Press

Firbas, Jan 1966. On Defining the Theme in Functional Sentence Perspective. In: *L'École de Prague d'aujourd'hui*. (Travaux Linguistiques de Prague 1.) Paris: Editions Klincksieck, p. 267–280.

Ginzburg, Jonathan 1996. Interrogatives: Questions, Facts and Dialogue. In: Lappin, Shalom (ed.), p. 385–422.

Ginzburg, Jonathan (forthc.). *A Semantics for Interaction in Dialogue*. Stanford and Chicago: CSLI and the University of Chicago Press.

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

Halliday, Michael 1967. Notes on Transitivity and Theme in English, part II. Journal of Linguistics 3, p. 199–244.

King, Tracy Holloway 1995. *Configuring Topic and Focus in Russian*. Stanford: CSLI Publications.

King, Tracy Holloway 1997. Focus Domains and Information-Structure. In: Butt, Miriam and Tracy Holloway King (eds.), *Proceedings of the LFG97 Conference*. Stanford: CSLI. <http://csli-publications.stanford.edu/LFG/2/lfg97. html>.

Lambrecht, Knud 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.

Magnusson, Erik 2007. *Gränsöverskridande koordination. Syntaktisk förändring i äldre svenska*. (Nordistica Gothoburgensia 28.) Göteborg: Acta Universitatis Gothoburgensis.

Mörnsjö, Maria 2002. *V1 Declaratives in Spoken Swedish. Syntax, Information Structure, and Prosodic Pattern.* Lund: Lunds universitet.

Mycock, Louise (2007). *The Typology of Constituent Questions: A Lexical Functional Grammar Analysis of 'Wh'-Questions*. Doctoral dissertation. Manchester: University of Manchester.

Nikula, Henrik 1986. *Dependensgrammatik*. Malmö: Liber.

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

Rooth, Mats 1992. A Theory of Focus Interpretation. *Natural Language Semantics* 1, p. 75–116.

Rooth, Mats 1996. Focus. In: Lappin, Shalom (ed.) 1996. The Handbook of Semantic Theory. Oxford: Blackwell, p. 271–297.

Teleman, Ulf, Staffan Hellberg and Erik Andersson 1999. *Svenska Akademiens grammatik* (Swedish Academy Grammar). Stockholm: Norstedts.

Vallduví, Enric 1992. *The Information Component*. New York: Garland.

Vallduví, Enric and Elisabet Engdahl 1996. The Linguistic Realization of Information Packaging. *Linguistics* 34:3, p. 459–519.

PROJECTIONS AND GLUE FOR CLAUSE-UNION COMPLEX PREDICATES

Avery D Andrews
The Australian National University

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

The paper shows how glue-semantics can be integrated into the LFG architecture as an (almost) normal projection, so that it can do the work of 'argument-structure' in accounts of predicate-composition such as Alsina (1996, 1997), Butt et al. (1997) and Andrews and Manning (1999).

A significant innovation is that that the standard 'semantic projection' is abandoned in favor of a σ-projection that directly connects the f-structure and the meaning-structure, similar to the original proposal for a semantic projection in Kaplan (1987), but running many-to-one from the semantic structure to the f-structure.

In this paper, I will propose an analysis of clause-union complex predicates, such as Romance causatives, in which the glue proof plays the role of argument-structure in analyses such as those of Alsina (1996) and Andrews and Manning (1999), and functions like a normal level of structure in LFG, with a projection relating it to the f-structure. We call this projection σ, since it has the same position in the theory as the σ-projection of Kaplan (1987), but it is opposite in direction to this, and quite different in function to the σ-projection in standard presentations of glue.

1 Prefab Glue

I will formulate the analysis using a formulation of glue which I will call 'prefab glue', which can be regarded as a version of proof-nets (Fry (1999), Moot (2002), Andrews (2004)), reorganized along the lines of the structure of proof-terms, so that glue assembly produces proof-terms (which are essentially logical forms) directly rather than requiring some kind of 'semantic trip' (de Groote and Retoré (1996), Morrill (2005)) or similar conversion (such as the one proposed by Perrier (1999), used by Andrews (2004)) to do this. An extended and slow-paced account of prefab glue is provided in Andrews (2007); the presentation here will be quite concise, and will assume a good grasp of glue. The only substantive differences between prefab glue and previous formulations are:

(1) a. IOFU instantiation rather than linear universal quantification is used to account for quantifier scope variation (as also proposed by Lev (2007)); this simplifies the glue linear logic to propositional rather than higher order, or first-order quantificational (Kokkonidis to appear).

b. the standard 'semantic projection' is eliminated, and the σ-correspondence runs from atomic-type nodes of the glue-proof to the f-structure.

[†] I am indebted to Alex Alsina for various Catalan examples, the audience at LFG07 for some questions, an anonymous editor for some useful comments, and to Elisabeth Mayer for help with proof-reading. All errors remain my own.

Some discussion of both of these points is provided in Andrews (2007).

Technically, we derive the glue-side of a meaning-constructor in prefab glue 'structure tree' format from one in regular format as follows. We assume that the glue-sides are formulas of linear intuitionistic implication-conjuction ($\multimap$ $\otimes$) logic whose atomic formulas are pairs consisting of an f-structure designator (a label if the constructor is instantiated) and a semantic type. For semantic types, we will use e for 'entities', and p for 'propositions' (following Pollard (to appear)).

The first step is to label the whole meaning-side and its subformula instances with polarities $+/-$ as follows:[1]

(2) a. The polarity of an entire meaning-side is negative

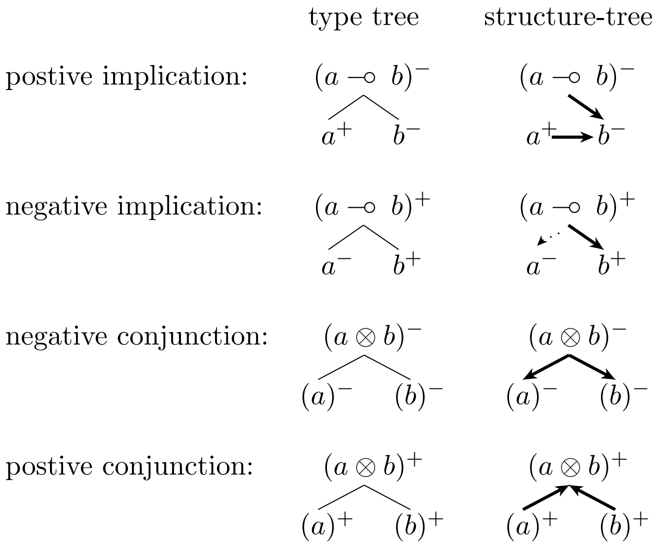
b. The polarity of the consequent of an implication is that of the entire implication.

c. The polarity of the antecedent of an implication is the opposite of that of the entire implication

d. The polarity of a component of a conjunction is that of the entire conjunction.

We next replace the original links with the 'dynamic graph'-links of de Groote (1999),[2] represented as bold arrows below, but retaining the original links between positive implications and their (negative) antecedents, represented as a dotted arrow below (the links are drawn upside-down to the usual orientation in the literature):

(3)



[1] The polarity rules go back at least to Jaśkowski (1963).

[2] The concept is originally due to Lamarche (1994), where it is called the 'essential net', but deGroote's paper is much more accessible (I must confess to understanding almost nothing of the Lamarche paper).

I will sometimes call the dotted links 'pseudo-daughter' links.

We now formulate assembly in the usual manner for proof-nets. The constructors to be assembled are taken as a collection of objects, and an additional structure is added consisting of a single positive polarity node $(f_p)^+$, where f is the label of the entire f-structure. This is essentially the same thing as a 'frame' in Type Logical Grammar (except that the logic is commutative).

Then we link negative to positive atomic formula occurrences with 'axiom-links', subject to the following rules:

(4) a. The linked pairs must be exhaustive and non-overlapping.

b. Members of a linked pair must have the same semantic type.

c. Members of a linked pair must have the same f-structure label.

The result of this is a 'proof-structure' in proof-net theory; to restrict proof-structures to ones constituting valid proofs, we need to impose a 'Correctness Criterion', which can be formulated like this (de Groote (1999), Moot (2002:94-95)), among many other ways:

(5) **Correctness Criterion:** The dynamic graph must be:

(a) rooted and acyclic.

(b) every dynamic graph path to the root that starts at the target of a dotted link must pass through the source of that link.

Note that the direction of the dynamic graph links, but not the pseudo-daughter links, is essential if the polarities are erased. A proof-structure that passes the Correctness Criterion is a proof-net, and represents a valid linear logic proof.

If the f-structure label information is ignored in the formation of the proof-structure, the constructors function somewhat like a numeration in Minimalism, and the possible proof-nets represent all possible ways of assembling the constructors consistent with their semantic types (Klein and Sag 1985).

For an example, here are instantiated constructors for the sentence *Bert likes everybody*:

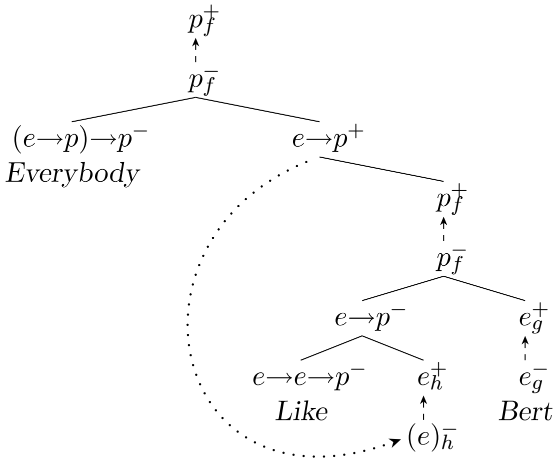
(6) *Everybody* : $(g_e \multimap f_p) \multimap f_p$

Bert : h_e

Like : $h_e \multimap g_e \multimap f_p$

Converted to structure-tree format, connected with axiom-links represented as dashed arrows, and arranged in a perspicuous manner, these constructors become:

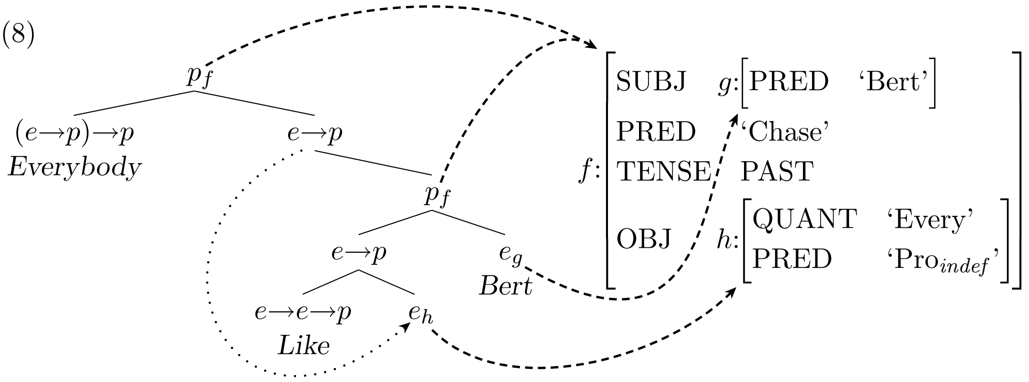
(7)



This looks very much like a structure-tree for a linear lambda-term, with the dotted pseudo-daughter link representing variable-binding.

The resemblance becomes essentially identity if we contract the axiom-links, and erase the polarities. Interpreting the f-structure label subscripting as a standard LFG correspondence relation (albeit opposite in direction to most of them), we get the following glue-structure f-structure pair for the sentence, where the heavy dashed lines represent the σ-correspondence:

(8)



This diagram is deliberately reminiscent of the ϕ-correspondence from c-structure to f-structure.

2 Glue as Argument-structure

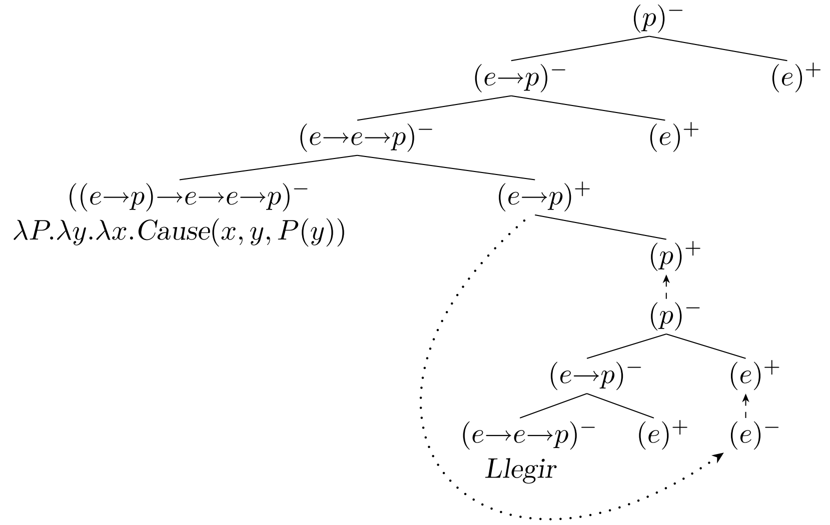
With meaning-constructors and proof-nets represented in this manner, it becomes apparent that glue-proofs have many of the properties of argument-structures as proposed by Alsina (1996) and many other works. Below is a meaning-constructor for the 'three place causative', without the syntactic information, whose meaning can be glossed as (b):

(9) a. $\lambda P.\lambda y.\lambda x.Cause(x, y, P(y))\ :\ (e \rightarrow p) \rightarrow e \rightarrow e \rightarrow p$

b. x does something to y. Because of this, y does P.

Combining the structure-tree format version of this with that for a transitive verb (here *Llegir* ‘read’ in Catalan), we get a structure like this:

(10)


$(p)^-$
$(e{\rightarrow}p)^-$
$(e)^+$
$(e{\rightarrow}e{\rightarrow}p)^-$
$(e)^+$
$((e{\rightarrow}p){\rightarrow}e{\rightarrow}e{\rightarrow}p)^-$
$\lambda P.\lambda y.\lambda x.Cause(x, y, P(y))$
$(e{\rightarrow}p)^+$
$(p)^+$
$(p)^-$
$(e{\rightarrow}p)^-$
$(e)^+$
$(e{\rightarrow}e{\rightarrow}p)^-$
$(e)^+$
$(e)^-$
Llegir

The Correctness Criterion will guarantee that the positive e in the property (innermost, $(e{\rightarrow}p)^+$) argument of the causative will link to an argument of the embedded verb, but not restrict it to the topmost one. Such a restriction seems plausible, and might be imposed by a semantic restriction that the controller of the property be its Agent, but we won’t look into this issue here.

Observe however that (9) has many similarities to the results of ‘predicate composition’ proposed by Alsina (1996:191):

(11) ‘cause<[P-A]$_3$ [P-P]$_2$ read <[P-A]$_2$ [P-P]$_1$>>’

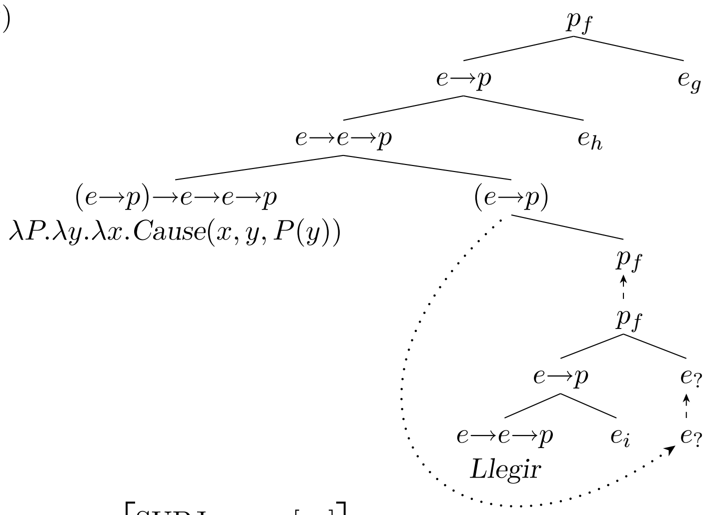
The subscripts represent (co-)linking to values of grammatical function in f-structure, roughly equivalent to our σ.

As discussed by Andrews and Manning (1999), the concept of predicate-composition and the associated structure (11) don’t fit very well into standard LFG architecture. But they go much better when glue is involved. The intent of (11) is that the Cause predicate has three arguments, one of which is a composite involving the caused predicate. This is directly expressed in (10). The arguments in (11) are also presented in a definite order, represented by the hierarchical nesting relationships in (10).

A difference is that the entity argument-positions in (11) are tagged with Dowty’s ‘Proto-Agent’ and ‘Proto-Patient’ labels. But this is a matter of the detailed formulation of linking theory, and there is no reason why meaning-constructor atomic formulas can’t have such information added to their lexical specification, if this proves to be empirically warranted.

Especially important is that in a glue-based approach, there is no reason why the meaning-constructors for the causative and caused predicates can't 'output' to the same level of f-structure, consistently with the many arguments for the monoclausality of Romance causatives. This is supported by the fact that the σ-correspondence, with the present directionality, is independently required to be many-to-one by constructions such as sentence-adverbials, and quantifiers. We illustrate this here for the causative by f-structural co-labelling:

(12)


p_f
$e \rightarrow p$
e_g
$e \rightarrow e \rightarrow p$
e_h
$(e \rightarrow p) \rightarrow e \rightarrow e \rightarrow p$
$\lambda P.\lambda y.\lambda x.Cause(x, y, P(y))$
$(e \rightarrow p)$
p_f
p_f
$e \rightarrow p$
$e_?$
$e \rightarrow e \rightarrow p$
e_i
$e_?$
Llegir

$$f{:}\begin{bmatrix} \text{SUBJ} & g{:}[\ \ ] \\ \\ \text{IOBJ} & h{:}[\ \ ] \\ \\ \text{OBJ} & i{:}[\ \ ] \end{bmatrix}$$

This many-to-one property is of course also a characteristic of the c-structure-to-f-structure correspondence ϕ. The '?' subscript to some of the e's represents an issue concerning what their f-structure correspondents ought to be.

The idea of predicate composition thus appears to fit into LFG+glue, but we do need to reconstrue our idea of how the PRED-features themselves work. This is because if the causative and causee verb both introduce a PRED-feature at the same level of f-structure, these will clash. Fortunately, as pointed out by Kuhn (2001), meaning-constructors are able to take on most of the functions of PRED-features, in particular, the management of the Completeness, Coherence and Predicate Uniqueness constraints. Andrews (to appear) however shows that PRED-features can still play a useful role in connecting irregular morphology to multiple meanings of verbs, such as the irregular forms *went* and *gone* with a wide range of different meanings such as *go off*, *go out*, *go crazy*, etc. But for this function, the features can be located on a 'morphological projection' such as proposed by Butt et al. (1996) and Butt et al. (1999). This projection shares less aggressively than ϕ, so that each verb can put its PRED-feature on a different level. We will return to this issue later, but now consider the specification of grammatical functions in the causative constructor.

An initial thought might be that the constructor would have to look something like this:

(13) $\lambda P.\lambda y.\lambda x.\mathit{Cause}(x, y, P(y))$:

$$((\uparrow \mathrm{?OBJ})_e \rightarrow \uparrow_p) \rightarrow (\uparrow \mathrm{?OBJ})_e \rightarrow (\uparrow \mathrm{SUBJ})_e \rightarrow \uparrow_p$$

'?OBJ' here represents whatever we need to do to accommodate the well known alternation between dative causees for transitive caused verbs, and accusative ones for intransitives. This can be accounted for in various ways, such as for example Falk's (2001:115) proposal that transitives take the causee as an OBJ$_\theta$, in effect the traditional 'indirect object' (IOBJ), while intransitives take it as an OBJ. The constructor synchronizes this GF between the object and controller-of-property positions, on the basis of the semantic relationship.

However, rather counterintuitively, this constructor will work as well:

(14) $\lambda P.\lambda y.\lambda x.\mathit{Cause}(x, y, P(y))$:

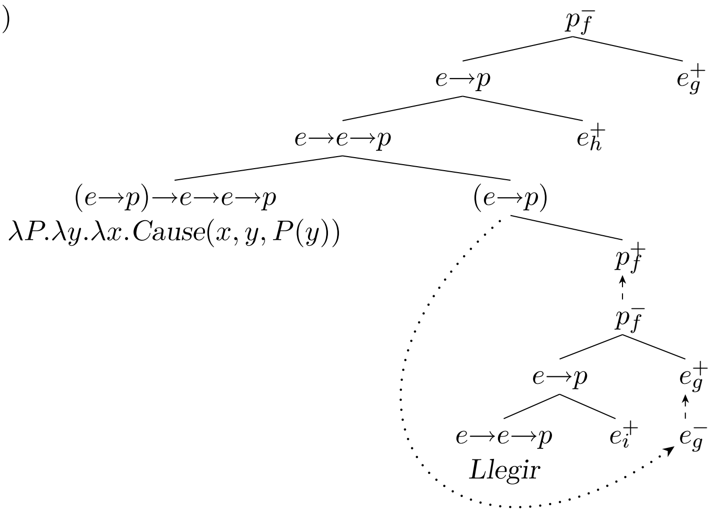
$$((\uparrow \mathrm{SUBJ})_e \rightarrow \uparrow_p) \rightarrow (\uparrow \mathrm{?OBJ})_e \rightarrow (\uparrow \mathrm{SUBJ})_e \rightarrow \uparrow_p$$

And it has the advantage that it will work with an unmodified constructor for the caused verb, requiring no linking theory:

(15) *Llegir* : $(\uparrow \mathrm{OBJ})_e \rightarrow (\uparrow \mathrm{SUBJ})_e \rightarrow \uparrow_p$

These two constructors will fit together to yield this assembly, with accompanying f-structure (σ represented with co-labelling):

(16)


p_f^-
$e \rightarrow p$
e_g^+
$e \rightarrow e \rightarrow p$
e_h^+
$(e \rightarrow p) \rightarrow e \rightarrow e \rightarrow p$
$\lambda P.\lambda y.\lambda x.\mathit{Cause}(x, y, P(y))$
$(e \rightarrow p)$
p_f^+
p_f^-
$e \rightarrow p$
e_g^+
$e \rightarrow e \rightarrow p$
e_i^+
e_g^-
Llegir

$$f{:}\begin{bmatrix} \text{SUBJ} & g{:}[\ \ ] \\ \\ \text{IOBJ} & h{:}[\ \ ] \\ \\ \text{OBJ} & i{:}[\ \ ] \end{bmatrix}$$

This works (note that since σ is many-to-one, there is no problem with the f-structure associated with (17) being monoclausal, as required for an analysis of complex predicates), even though the top e^+ argument of the caused verb and the e^- antecedent of the property argument of the controlled verb are associated with the causative subject f-structure g, which has nothing to do with the causee agent f-structure h. Note that this is not a specific property of the prefab glue formulation, but a consequence how glue premise-matching works.

So, although counter-intuitive, this is a somewhat tempting analysis of causatives, but that does not necessarily mean that it is the right thing to do. Next, I will argue that it isn't.

3 Problems with the Easy Analysis

I will present two kinds of problems, a general theoretical one, and a more concrete empirical difficulty.

The theoretical problem is that the technique employed in the analysis allows empirically wrong analyses of constructions which are standardly analysed in LFG with functional control. Consider the following meaning-constructor for *seem*:[3]

(17) $\lambda Px.Seem(P(x))$:
$$((\uparrow \text{XCOMP SUBJ})_e {\to} (\uparrow \text{XCOMP})_p) {\to} (\uparrow \text{SUBJ})_e {\to} {\uparrow_p}$$

Since we have already abandoned the usual Completeness and Coherence constraints in favor of glue assembly, the following f-structure, *without functional control*, can provide a satisfactory interpretation for a sentences such as *Bert seems to like Ernie*:

(18)
$$f{:}\begin{bmatrix} \text{SUBJ} & g{:}\begin{bmatrix} \text{PRED} & \text{`Bert'} \end{bmatrix} \\ \text{PRED} & \text{`Seem'} \\ \text{XCOMP} & h{:}\begin{bmatrix} \text{SUBJ} & i{:}\begin{bmatrix} \ \ \end{bmatrix} \\ \text{PRED} & \text{`Like'} \\ \text{OBJ} & j{:}\begin{bmatrix} \text{PRED} & \text{`Ernie'} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

[3] Partially inspired by some of the constructors in Asudeh (2002, 2005).

The constructor for the lower verb will construct the complement subject grammatical function, but the constructor (17) for *seem* will make it unnecessary for this to be functionally identified with anything else for the semantic interpretation to be found.

But although the analsyis works for this particular example, it leads to a variety of problems, such as the inability to account for 'long distance agreement' in languages such as Icelandic (Andrews 1982), or the narrow scope reading of examples like this, from Asudeh (2002, 2005):

(19) Every goblin seems to have pinched Merry

These phenomena provide evidence for functional control even though the basic semantic interpretation of simple examples doesn't require it. To explain these phenomena, LFG+glue ought therefore to contain some principle that would rule out the analysis without functional control, so that learners would adopt the standard analysis with functional control even without encountering the somewhat subtle evidence that motivates it.

The second, concrete, problem with the analysis (14–15) is that it fails to address Alsina's (1996) arguments that the causee agent is not a subject. For example, it is unable to host a floating quantifier, whereas the controlled subject of an Equi-construction can:

(20) Els metges ens deixen beure una cervesa cadascun
the doctors us let drink a beer each

a. Each of the doctors lets us drink a beer

b. *The doctors let each of us drink a beer

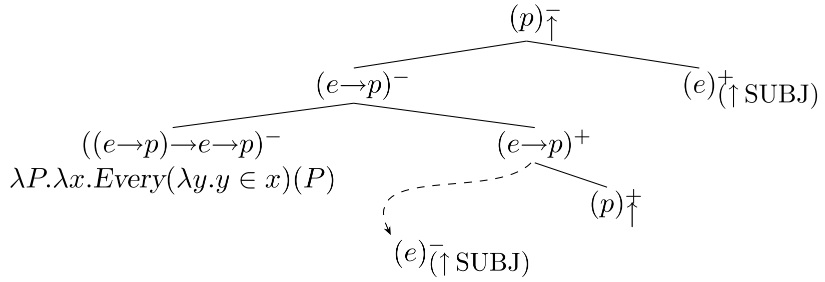
(Alsina 1996:217)

(21) Els metges$_i$ ens$_j$ han convençut de beure una cervesa cadascun$_{i/j}$
the doctors us have convinced of drink a beer each
The doctors each convinced us to drink a beer
The doctors convinced us to drink a beer each
(Alsina p.c.)

The following meaning-constructor seems appropriate for floating quantifiers which can only float off the subject, given in both standard (a) and structure-tree (b) format:

(22) a. $\lambda Px.\mathit{Every}(\lambda y.y \in x)(P) \ : \ ((\uparrow \mathrm{SUBJ})_e \rightarrow \uparrow_p) \rightarrow (\uparrow \mathrm{SUBJ})_e \rightarrow \uparrow_p$

b.


$(p)^{-}_{\uparrow}$
$(e{\rightarrow}p)^{-}$
$(e)^{+}_{(\uparrow \text{SUBJ})}$
$((e{\rightarrow}p){\rightarrow}e{\rightarrow}p)^{-}$
$\lambda P.\lambda x.\mathit{Every}(\lambda y.y \in x)(P)$
$(e{\rightarrow}p)^{+}$
$(p)^{+}_{\uparrow}$
$(e)^{-}_{(\uparrow \text{SUBJ})}$

What it does is abstract over the subject GF to create a property from the predicate, and applies a semantically distributed version of this property to a plural subject.

Combining it with the caused verb and abbreviated representation **b** of the object of (20), we get:

(23) $\lambda x.\mathit{Every}(\lambda y.y \in x)(\lambda z.\mathit{Beure}(\mathbf{b})(z))$

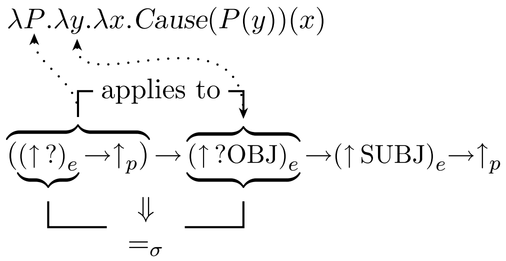
This can then combine with the causative verb and abbreviated representation of object to produce (24) with its β-reduction to the undesired reading of (20):

(24) $\lambda x.\mathit{Let}(x, \mathbf{ns}, (\lambda x.\mathit{Every}(\lambda y.y \in x)(\lambda z.\mathit{Beure}(\mathbf{b})(z)))(\mathbf{ns})) \Rightarrow_{\beta}$
$\lambda x.\mathit{Let}(x, \mathbf{ns}, \mathit{Every}(\lambda y.y \in \mathbf{ns})(\lambda z.\mathit{Beure}(\mathbf{b})(z))) \equiv$
$\lambda x.\mathit{Let}(x, \mathbf{ns}, \mathit{Every}(y, y \in \mathbf{ns}, \mathit{Beure}(\mathbf{b})(y)))$

To aid comprehension, we express the result of the reduction in two possible formats for the quantifier, first an 'Aristotelian' one where it relates two properties, then a '3-part' one where there is a variable and two formulas open on that variable.

To rule out these undesired analyses, I will suggest a constraint that rules out the intuitively odd property of the constructor (16), that it can in effect transmit a meaning via σ-linking to an f-structure that has nothing at all to do with that meaning. On its meaning-side, this constructor attributes the property expressed by the innermost argument to the entity expressed by the next-innermost one (the causee agent). This can be formulated in terms of the structural relationships within the meaning-side between the two lambda-variables corresponding to the arguments. The relationship that triggers the constraint is that the glue-subformula corresponding to the property has a (conditional, not anaphoric) antecedent of the same semantic type (e, in this case) as the one corresponding to the argument to which the property is applied, and the proposed constraint requires that, under these conditions, the σ-correspondents of these subformulas be the same as well. We can depict this constraint, which we will call Functional Consistency, diagrammatically as follows, where the material subtending the lower horizontal braces is what the constraint requires to hold, if the other material is present:

(25) Functional Consistency:


$\lambda P.\lambda y.\lambda x.Cause(P(y))(x)$
applies to
$((\uparrow ?)_e \rightarrow \uparrow_p) \rightarrow (\uparrow ?\mathrm{OBJ})_e \rightarrow (\uparrow \mathrm{SUBJ})_e \rightarrow \uparrow_p$
$\Downarrow$
$=_\sigma$

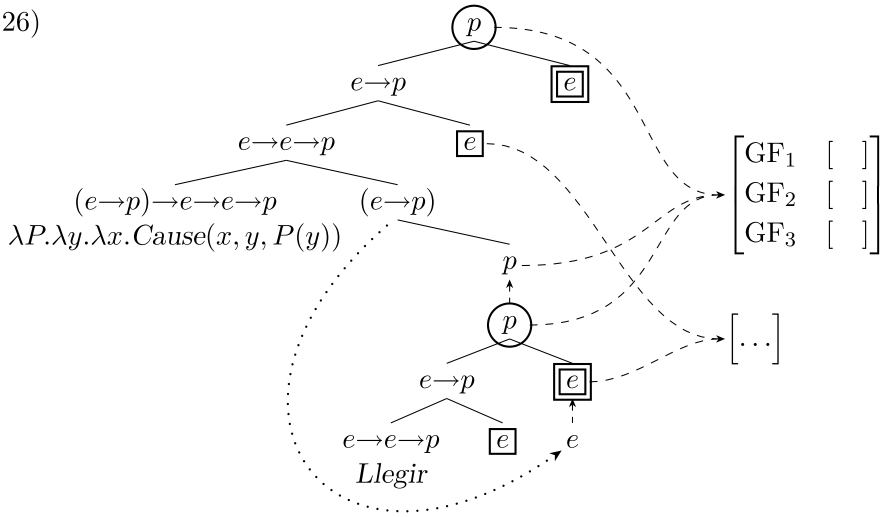
Functional Consistency will rule out the counterintuitive constructor (14), but allow one of the initially expected form (13). Likewise, the undesirably innovative (17) will be excluded, while conventional analyses using functional control will be allowed. In the absence of plausible alternatives for the allowed analyses that satisfy Functional Consistency, these can be regarded as required by the theory.

4 Linking Theory

Although it is in a sense good news that there are real reasons for ruling out the counterintuitive analyses, the accompanying bad news is that we will after all need a linking theory for the complex predicates. Fortunately, LFG+glue provides good support for producing such a theory. (26) below shows how notions such as (co-)argument, logical subject, and relative (semantic role-based) prominence can be formulated in terms of the structures.

One fundamental notion is the 'Final Output' of a meaning-constructor, which is the root node of the constructor in structure-tree format. These are circled (this concept might require adjustment if tensors are used in the formalism). Then 'basic arguments' are nodes of basic type that are daughters of nodes on the 'spine' from the meaning-bearing node to the final output. These are boxed. It is plausible that there is a typological division between languages that assign object grammatical functions to nodes of basic type other than e (such as Icelandic, where clausal complements are arguably NPs bearing ordinary object grammatical functions (Thráinsson 1979)), and those that don't, such as English and Dutch ((Koster 1978), (Bresnan 1994); see also Alsina et al. (2005)).

(26)



Basic arguments seem to behave differently from those of higher-order type, such as *e→p*. In particular, there seems to be a rather solid constraint that a predicate take only one higher-order argument.

Another important concept is role-based semantic prominence, expressed by the hierarchical relationships between the basic arguments. This is widely, although not universally, assumed to be necessary.[4] A problem with it is that it is largely predictable from semantic roles; if relative prominence is totally so predictable, then it should not be an independent notion of the theory. I suggest that the semantic-role assignment contrasts between verbs such as *predecease*, on the one hand, and *outlive* and *survive*, on the other, show that relative prominence is sometimes independent of semantic roles. An important concept based on relative prominence is 'logical subject', double-boxed in (26); the logical subject is the most prominent argument of a predicate.

A somewhat more complex notion is that of 'co-argument': co-arguments are arguments whose Final Outputs have the same f-structural correspondent. So all of the boxed and double-boxed positions in (26) are co-arguments, with the result that they are simultaneously subjected to the constraints of linking theory. But if the two type p Final Arguments had different f-structure correspondents, then the arguments would fall into two sets of co-arguments, each linking independently, as appropriate for multiclausal constructions, including those with functional control. We might also want to recognize 'immediate' co-arguments, which would be arguments sharing the same Final Output.

Next, we face the challenge of producing an actual linking theory. In (26), the argument positions aren't connected to any GF-values in the f-structure correspondent of the final outputs, so the intended effects of the

[4] See for example Zaenen (1993), Asudeh (2001) for proposals that dispense with it.

linking theory aren't represented. But the effects of Functional Consistency are represented, by linking the two relevant argument positions to the same piece of f-structure material, which is however not integrated into the f-structure of the Final Outputs. There have unfortunately accumulated a rather large number of options for linking theory in LFG, usefully surveyed by Butt (1999). I can't systematically investigate all of these here, so will merely propose something that works out for the case at hand, and doesn't seem immediately and unsalvageably hopeless from a typological point of view.

In the first place, we accept a basic distinction between 'core' and 'oblique' grammatical functions, with the latter pre-specified for a morphologically marked oblique grammatical function, typically marked by a preposition in Romance or Germanic languages (or semantic cases in many others). Oblique grammatical functions don't participate in causative grammatical function alternations, so we need consider them no further here (but would have to in a consideration of applicatives). The non-oblique argument positions will then be ranked in terms of relative prominence, for the linking principles to apply to.

Observe that the approach has already made an improvement on Andrews and Manning (1999) in that it has a specific proposal for oblique arguments. Now we propose that in the lexicon, core arguments are optionally and constructively assigned any of the core grammatical functions SUBJ, OBJ and OBJ_θ. To be a bit more precise about this, I propose a notation whereby $\upsquigarrow$ means 'the σ-correspondent of the Final Output of the meaning-constructor I am an annotation of', while $\downsquigarrow$ means 'the σ-correspondent of the argument-position I'm attached to' (the squiggle in the arrows is supposed to indicate that these arrows are *not* evaluated with respect to positions in a c-structure, but to positions somewhere else, namely within a glue-assembly). We can now write the constructive GF-assignment principle as follows:

(27) $(\upsquigarrow \text{SUBJ}|\text{OBJ}|\text{OBJ}_\theta) = \downsquigarrow$

This applies to all core argument positions, at least those of type e (leaving the treatment of other types aside, in this paper). (26) will now get the correct grammatical function assignment, as well as many incorrect ones.

We next have a constraint which requires the GFs of co-arguments to be assigned harmonically w.r.t. their relative prominence, with the GF's ranked:

(28) $\text{SUBJ} > \text{OBJ}_\theta > \text{OBJ}$

A biuniqueness constraint can prevent the argument positions that are identified by Functional Consistency from getting distinct governable functions; we formulate it as a condition preventing one f-structure from bearing two governable GFs to another (but, in order to allow functional control, we permit an f-structure to bear distinct GF's to different f-structures).

Only one GF-assignment will now be available for ditransitives and causatives of transitives, but so far there will be three for intransitives. We can rule this out by requiring that the maximally prominent co-argument be assigned SUBJ, if it gets any core GF at all (passives can be plausibly treated as not assigning a core GF to the maximal co-argument). For transitives, and causatives of intransitives, this leaves two possibilities for the other co-argument, OBJ, or OBJ_θ. The former appears to be the default, with the latter appearing with various non-Patientlike semantic roles, such as Addressee, Object of Obedience, etc. We can propose that OBJ is assigned to the least prominent argument-position, subject to a semantic-role-based restriction which blocks this for non-Patientlike roles, leaving OBJ_θ as the only option. A sharp characterization of what this restriction is would be highly desirable, but will not be attempted here.

5 The Morphological Projection and Respect for the Tree

Now we turn to the other significant problem for monoclausal structures, accounting for how each semantically higher verb determines the form of the following one, and the arrangement in the c-structure reflects the semantic organization (called 'respecting the tree structure' in Alsina (1997)). These problems are illustrated in these examples (Alsina p.c; adapted from Alsina (1997)):

(29) a. L' acabo de fer llegir al nen
It I.finish of make read to the boy
'I just made/I finish making the boy read it.'

b. La faig acabar de llegir al nen
It.F I.make finish of read to the boy
'I make the boy finish reading it (say, a map ([GND FEM])).'

The appearance of the direct object clitic semantically associated with the final verb in front of the first one shows that these are clause-union constructions, but we see that the order of verbs nevertheless reflects the meaning, and each verb determines the form of the one after it, suggesting some kind of complement-structure.

The form-determination problem is basically the same as arises with monoclausal analysis of auxiliaries, and for it we can use the same solution, the 'morphological projection' proposed by Butt et al. (1996) and Butt et al. (1999). However we will suggest a slightly different version of the architecture, in which the morphological projection ('m-structure') comes between the c-structure and the f-structure, similarly to the argument-structure of Butt et al. (1997), but not that of Andrews and Manning (1999). The motivation for this is to impose a principle that f-structure shares more aggressively than m-structure, rather than just differently.

The m-structure projection is governed by various principles, the most important of which is that it is shared between 'primary' (but not extended) X-bar projections and their heads. Therefore I and IP, and V and VP, will have the same m-structure correspondent, but the IP and VP levels will have different m-structure correspondents. But, as in the original m-structure proposals, the VP will be the m-structural DEP of the IP (an S-complement of IP will also share m- and therefore f- structure). VP-complements will furthermore have the option of being treated either as complements (biclausal), or as extended projections (monoclausal).

Although distinct at m-structure, extended projections will always be merged at f-structure, so that the f-structures of familiar constructions will for the most part look the same as in standard LFG, except for the location of certain attributes, which will be located at m-structure rather than f-structure (from which they can however be located by means of inverse projections, albeit in a functionally uncertain manner).

Amongst these attributes are of course the verbal form features distinguishing infinitives and participles, and the prepositional markers in (29), but also, innovatively the PRED-features, whose semantic functions have been taken over by glue, and whose most obvious and possibly only remaining function is to control the morphological spellout of lexical items, as discussed in Andrews (to appear). I will not now make any proposals concerning nominal features; the existence of two places in which they can be put seems promising in light of Wechsler and Zlatić (2003), but the details may well fail to work out.

The architecture so far can be diagrammed like this:

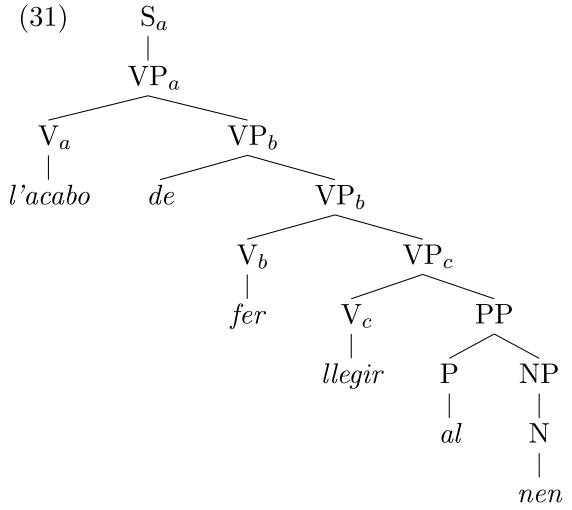
(30)

$$\begin{tikzcd}
 & & \boxed{\text{glue-structure}} \arrow[d, dashed, "\sigma"] \\
\boxed{\text{c-structure}} \arrow[r, dashed, "\mu"] & \boxed{\text{m-structure}} \arrow[r, dashed, "\psi"] & \boxed{\text{f-structure}}
\end{tikzcd}$$

$$\phi = \psi \circ \mu$$

and partial c-, m- and f-structures for (29a) presented as follows, where a, b, c are m-structure labels, prefixed to the m-structures they label, and postfixed to the f-structure that these m-structures correspond to under ψ:

(31)



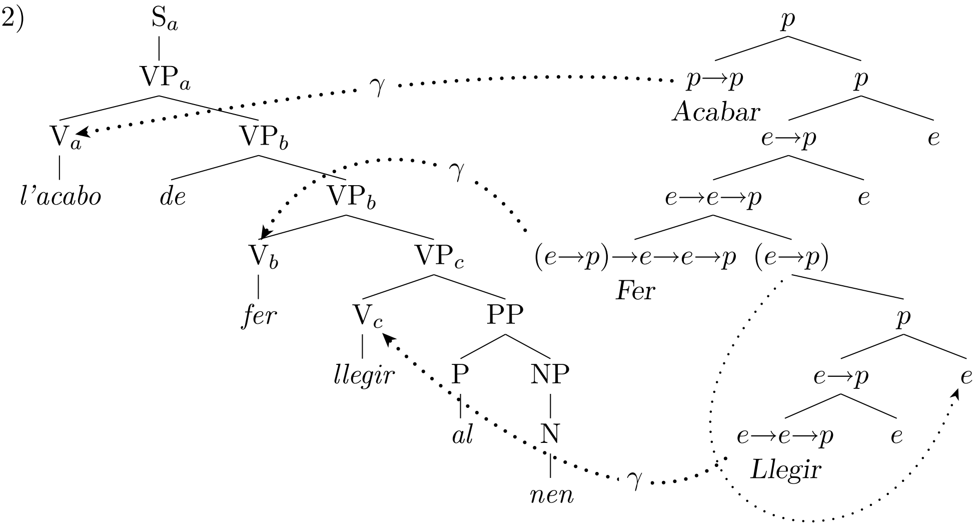
$$a{:}\left[\begin{array}{ll}\text{PRED} & \text{‘Acabar’}\\ \text{VFORM} & \text{FIN}\\ \text{DEP} & b{:}\left[\begin{array}{ll}\text{PRED} & \text{‘Fer’}\\ \text{VFORM} & \text{INF}\\ \text{VMARK} & \text{DE}\\ \text{DEP} & c{:}\left[\begin{array}{ll}\text{PRED} & \text{‘Lllegir’}\\ \text{VFORM} & \text{INF}\end{array}\right]\end{array}\right]\end{array}\right] \qquad \left[\begin{array}{ll}\text{SUBJ} & [\ ]\\ \text{IOBJ} & [\ ]\\ \text{OBJ} & [\ ]\end{array}\right]: a, b, c$$

We can now get the forms of the examples of (29), but each will have both meanings rather than the sole correct one.

Alsina (1997:237-238) addresses this issue with an informally stated constraint to the effect that predicate composition must mirror the c-structure. In effect, all of the predicates found under a VP in the c-structure must constitute a composite PRED-value which in some sense corresponds to that VP (Alsina's example (50)). This indicates the presence of another projection, which in the present context, would be most naturally construed as directly linking the glue-structure and the c-structure. I will call this projection γ, and construe it as running from the meaning-bearing nodes of the glue-structure (left terminal daughters) to the c-structure node that lexical item introducing the constructor appears under in the c-structure.[5] The result for (29a) will be:

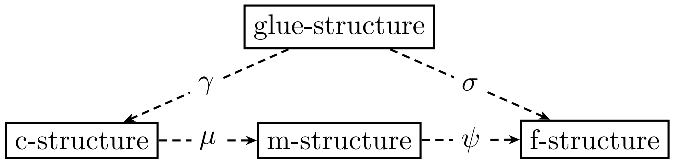
[5] This may need to be revised in light of idioms, and meaning-constructors introduced directly by PS-rules, if these latter exist.

(32)



The revised architecture will then be:

(33)



$\phi = \psi \circ \mu$ γ connects meaning nodes to their c-structure introducers

We then need some constraints which will assure that the relationships between meaning-constructors in the assembled glue-structure reflect the c-structure relationships between their introducers.

A problem which the constraint needs to be able to deal with is the ambiguous interpretation of adverbs in examples like:[6]

(34) He fet beure el vi a contracor a Maria
I-have made drink the wine against will to Mary
I made Mary drink the wine against her/my will
(Manning (1992), Andrews and Manning (1999:126), from Alex Alsina p.c.)

The constraint I suggests involves a glue-structure relationship that I will call 'Extended Argument of', and a c-structure relationship that I'll call 'β-command':

(35) **Extended Argument of**: Meaning-bearing glue node m is an extended argument of meaning-bearing glue-node n iff the dynamic path of m joins the dynamic path of n before the FinalOutput of n (= 'Feeds Into' from Andrews (to appear)).

[6] Andrews (2003) gets into trouble with this.

(36) **β-command:** c-structure node c β-commands node d iff every $\overline{\mathrm{X}}$ projection dominating c dominates d.

Note that in a complex predicate, the higher verb will β-command the lower, but not vice-versa, even if they are in an extended projection relation.

The constraint is then:

(37) **γ-harmony:** If $\gamma(m)$ β-commands $\gamma(n)$ but not vice-versa, and if $\phi(\gamma(m)) = \phi(\gamma(n))$, then n must be an extended argument of m (the condition on $\phi \circ \gamma$ is supposed to keep this from applying to adjuncts, so as to allow the ambiguity of (34)).

Perhaps more elegant formulations can be found, but (37) relates the levels of glue-structure and c-structure by means of a constraint that is plausibly universal, and intuitively iconic.

6 Conclusion

By construing meaning-constructors as being essentially the same thing as argument-structures, we have managed to capture many of the insights of Alsina's analysis of complex predicates in a more formalized framework, glue-semantics, that explicitly integrates argument-structure with a general account of semantic composition in LFG.

References

Alsina, A. 1996. *The Role of Argument Structure in Grammar.* Stanford, CA: CSLI Publications.

Alsina, A. 1997. A theory of complex predicates: Evidence from causatives in Bantu and Romance. In A. Alsina, J. Bresnan, and P. Sells (Eds.), *Complex Predicates*, 203–246. Stanford, CA: CSLI Publications.

Alsina, A., K. Mohanan, and T. Mohanan. 2005. How to get rid of the COMP. In M. Butt and T. H. King (Eds.), *Proceedings of the LFG05 Conference.* Stanford, CA: CSLI Publications. URL: `http://csli-publications.stanford.edu/LFG/10/lfg05.html`.

Andrews, A. D. 1982. The representation of case in Modern Icelandic. In Bresnan (Ed.).

Andrews, A. D. 2003. Glue logic, projections, and modifiers. ANU ms, URL: `http://arts.anu.edu.au/linguistics/People/AveryAndrews/Papers`.

Andrews, A. D. 2004. Glue logic vs. spreading architecture in LFG. In C. Mostovsky (Ed.), *Proceedings of the 2003 Conference of the Australian Linguistics Society*. URL: `http://www.als.asn.au/`.

Andrews, A. D. 2007. Prefab glue. ANU ms, URL: `http://arts.anu.edu.au/linguistics/People/AveryAndrews/Papers`.

Andrews, A. D. to appear. Generating the input in OT-LFG. In J. Grimshaw, J. Maling, C. Manning, and A. Zaenen (Eds.), *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*. Stanford CA: CSLI Publications. URL: `http://arts.anu.edu.au/linguistics/People/AveryAndrews/Papers`.

Andrews, A. D., and C. D. Manning. 1999. *Complex Predicates and Information Spreading in LFG*. Stanford, CA: CSLI Publications.

Asudeh, A. 2001. Linking, optionality and ambiguity in Marathi. In P. Sells (Ed.), *Formal and empirical issues in optimality-theoretic syntax*, 257–312. Stanford, CA: CSLI Publications.

Asudeh, A. 2002. A resource-sensitive semantics for Equi and Raising. In D. Beaver, S. Kaufmann, B. Clark, and L. Casillas (Eds.), *The Construction of Meaning*. Stanford, CA: CSLI Publications.

Asudeh, A. 2005. Control and resource sensitivity. *Journal of Linguistics* 41:465–511.

Bresnan, J. W. 1994. Locative inversion and the architecture of universal grammar. *Language* 70:72–131.

Bresnan, J. W. (Ed.). 1982. *The Mental Representation of Grammatical Relations*. Cambridge MA: MIT Press.

Butt, M. 1999. The development of linking theory in LFG. Handout for invited talk at ESSLLI99, URL: `http://citeseer.ist.psu.edu/butt99development.pdf`.

Butt, M., M. Dalrymple, and A. Frank. 1997. An architecture for linking theory in LFG. In T. H. King and M. Butt (Eds.), *Proceedings of the LFG97 Conference*, 1–16. Stanford, CA: CSLI Publications. URL: `http://csli-publications.stanford.edu`.

Butt, M., T. H. King, M.-E. Niño, and F. Segond. 1999. *A Grammar-Writer's Cookbook*. Stanford CA: CSLI Publications.

Butt, M., M. E. Niño, and F. Segond. 1996. Multilingual processing of auxiliaries within LFG. In D. Gibbon (Ed.), *Natural Language Processing and Speech Technology*. Berlin: Mouton de Gruyter.

Dalrymple, M. (Ed.). 1999. *Syntax and Semantics in Lexical Functional Grammar: The Resource-Logic Approach.* MIT Press.

Dalrymple, M., R. M. Kaplan, J. T. Maxwell, and A. Zaenen (Eds.). 1995. *Formal Issues in Lexical-Functional Grammar.* Stanford, CA: CSLI Publications.

de Groote, P. 1999. An algebraic correctness criterion for intuitionistic multiplicative proof-nets. *TCS* 115–134. URL: `http://www.loria.fr/~degroote/bibliography.html`.

de Groote, P., and C. Retoré. 1996. On the semantic reading of proof-nets. In G. G.-J. Kruijff and D. Oehrle (Eds.), *Formal Grammar*, 57–70, FOLLI Prague, August. URL: `citeseer.ist.psu.edu/degroote96semantic.html`.

Falk, Y. N. 2001. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax.* Stanford, CA: CSLI Publications.

Fry, J. 1999. Proof nets and negative polarity licensing. In M. Dalrymple (Ed.), *Syntax and Semantics in Lexical Functional Grammar: The Resource-Logic Approach*, 91–116.

Jaśkowski, S. 1963. Über Tautologien, in welchen keine Variable mehr als zweimal vorkommt. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 9:219–228.

Kaplan, R. M. 1987. Three seductions of computational psycholinguistics. In P. Whitelock, M.M.Wood, H. Somers, R. Johnson, and P. Bennet (Eds.), *Linguistics and Computer Applications*, 149–188. Academic Press. reprinted in Dalrymple et al. (1995), pp. 337-367.

Klein, E., and I. Sag. 1985. Type-driven translation. *Linguistics and Philosophy* 8:163–201.

Kokkonidis, M. to appear. First order glue. *Journal of Logic, Language and Information.* URL: `citeseer.ist.psu.edu/kokkonidis06firstorder.html`.

Koster, J. 1978. Why subject sentences don't exist. In S. J. Keyser (Ed.), *Recent Transformational Studies in European Languages*, 53–64. MIT Press.

Kuhn, J. 2001. Resource sensitivity in the syntax-semantics interface and the German split NP construction. In W. D. Meurers and T. Kiss (Eds.), *Constraint-Based Approaches to Germanic Syntax.* Stanford CA: CSLI Publications.

Lamarche, F. 1994. Proof nets for intuitionistic linear logic 1: Essential nets. Technical Report, Imperial College.

Lev, I. 2007. *Packed Computation of Exact Meaning Representations using Glue Semantics (with automatic handling of structural ambiguities and advanced natural language constructions).* PhD thesis, Stanford University. URL: `http://www.stanford.edu/~iddolev/pulc/current_work.html`.

Manning, C. D. 1992. Romance is so complex. Technical Report CSLI-92-168, Stanford University, Stanford CA. URL: `http://nlp.stanford.edu/~manning/papers/romance.ps`.

Moot, R. 2002. *Proof-Nets for Linguistic Analysis.* PhD thesis, University of Utecht. URL: `http://www.labri.fr/perso/moot/`.

Morrill, G. 2005. Geometry of language and linguistic circuitry. In C. Casadio, P. J. Scott, and R. Seely (Eds.), *Language and Grammar*, 237–264. Stanford, CA: CSLI Publications.

Perrier, G. 1999. Labelled proof-nets for the syntax and semantics of natural languages. *L.G. of the IGPL* 7:629–655. URL: `http://www.loria.fr/~perrier/papers.html`.

Pollard, C. to appear. Hyperintensions. To appear in *Journal of Logic and Computation.* URL: `http://www.ling.ohio-state.edu/~hana/hog/pollard2006-hyper.pdf`.

Thráinsson, H. 1979. *On Complementation in Icelandic.* Garland Press.

Wechsler, S., and L. Zlatić. 2003. *The Many Faces of Agreement.* Stanford, CA: CSLI Publications.

Zaenen, A. 1993. Unaccusativity in Dutch: Integrating syntax and lexical semantics. In J. Pustejovsky (Ed.), *Semantics and the Lexicon*, 128–161. Kluwer Academic Publications.

LINGUISTIC CONSTRAINTS IN LFG-DOP

Doug Arnold and Evita Linardaki
University of Essex

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

LFG-DOP (Bod and Kaplan, 1998, 2003) provides an appealing answer to the question of how probabilistic methods can be incorporated into linguistic theory. However, despite its attractions, the standard model of LFG-DOP suffers from serious problems of overgeneration because (a) it is unable to define fragments of the right level of generality, and (b) it has no way of capturing the effect of anything except simple positive constraints. We show how the model can be extended to overcome these problems.

1 Introduction

The question of how probabilistic methods should be incorporated into linguistic theory is important from both a practical, grammar engineering, perspective, and from the perspective of ‘pure’ linguistic theory. From a practical point of view such techniques are essential if a system is to achieve a useful breadth of coverage and avoid being swamped by structural ambiguity in realistic situations. From a theoretical point of view they are necessary as a response to the influence of probabilistic factors in human language behaviour (see e.g. Jurafsky, 2003, for a review).

Bod and Kaplan (1998, 2003) provide a very appealing and persuasive answer to this question in the form of LFG-DOP, where the linguistic representations of Lexical Functional Grammar (LFG) are combined with the probabilistic methods of Data Oriented Parsing (DOP). The result is a descriptively powerful, clear, and elegant fusion of linguistic theory and probability. However, it suffers from two serious problems, both related to generative capacity, which have the effect that the model overgenerates. This paper shows how these problems can be overcome.

The paper is structured as follows. Section 2 provides background, introducing the basic ideas of DOP. Section 3 describes the Bod and Kaplan (B&K) model, and introduces the first problem: the problem of defining DOP fragments with the right level of generality. Section 4 shows how this problem can be overcome. Section 5 describes the second problem (which arises because LFG-DOP fragments effectively encode only simple, positive, LFG constraints) and shows how it can be overcome. Section 6 discusses some issues and potential objections.

2 Tree-DOP

The central idea of DOP is that, rather than using a collection of rules, parsing and other processing tasks employ a database of *fragments* produced by decomposing a collection of normal linguistic representations (e.g. trees drawn from a

[†] We are grateful to the participants at LFG07 in Stanford, Ca, for insightful and stimulating discussion, in particular: Joan Bresnan, Aoife Cahill, Grzegorz Chrupała, Ron Kaplan, Jonas Kuhn, and Louisa Sadler.

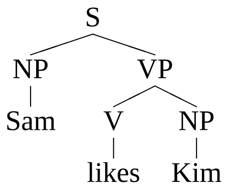


Figure 1: Treebank representation

treebank).[1] These fragments can be assigned probabilities (e.g. based on their relative frequency of appearance in the fragment database). Parsing a string involves, in effect, finding a collection of fragments which can be combined to derive it, i.e. provide a representation for it. These representations are assigned probabilities based on the probabilities of the fragments used. This general approach can, of course, be realized in many different ways, via different choices of basic representation, different decomposition operations, etc. So, standardly, specifying a DOP model involves instantiating four parameters: (i) representational basis; (ii) decomposition operations; (iii) composition operation(s); and (iv) probability model.

Specified in this way, Tree-DOP, the simplest DOP model, involves:

(i) a treebank of context free trees, such as Figure 1;
(ii) two decomposition operations: Root and $\mathit{Frontier}$;
(iii) a single composition operation: *Leftmost Substitution*;
(iv) a probability model based on relative frequency.

Fragments are produced from representations such as Figure 1 by two decomposition operations: Root and $\mathit{Frontier}$:

(i) Root selects any node n and makes it the root of a new tree, erasing all other nodes apart from those dominated by n.
(ii) $\mathit{Frontier}$ chooses a set of nodes (other than the root) and erases all subtrees dominated by these nodes.

Intuitively, Root extracts a complete constituent to produce a fragment with a new root. For example, the fragments in Figure 2 can be produced from the tree in Figure 1 by (possibly trivial) application of Root. $\mathit{Frontier}$ deletes part of a fragment to produce an 'incomplete' fragment — a fragment with a new frontier containing 'open slots' (i.e. terminal nodes labeled with a non-terminal category), as in Figure 3.

Leftmost Substitution involves substituting a fragment for the leftmost open slot. Figure 4 exemplifies one of the several ways in which a representation of *Kim likes Sam* can be derived.

[1] Standard references on DOP include, for example, Bod and Scha (1997), Bod (1998), and the papers in Bod et al. (2003). All of these contain presentations of Tree-DOP.

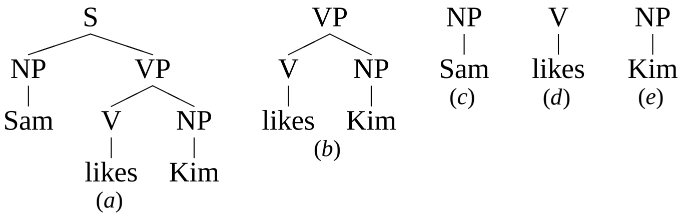


Figure 2: Fragments produced by the $Root$ operation

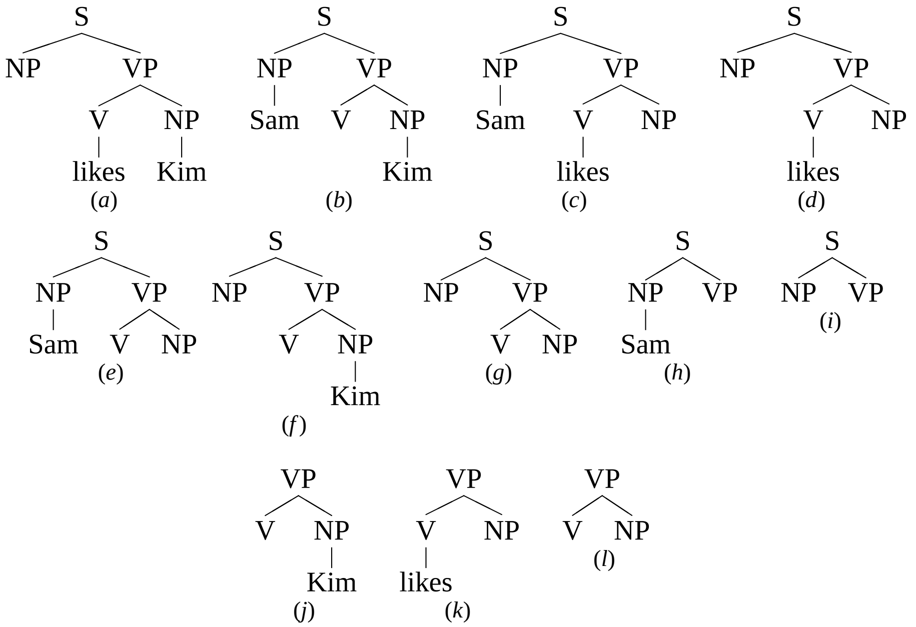


Figure 3: Fragments produced by the $Frontier$ operation

The following define a very simple probability model for this version of DOP.[2]

(1)
$$P(f_i) = \frac{|f_i|}{\sum\limits_{root(f)=root(f_i)} |f|}$$

(2)
$$P(d) = \prod_{i=1}^{n} P(f_i)$$

[2] Simple, and one should add, inadequate. This model is based on relative frequency estimation, which has been shown to be biased and inconsistent (Johnson, 2002). A number of alternatives have been proposed, e.g. assuming a uniform derivation distribution (Bonnema et al., 1999), backing-off (Sima'an and Buratto, 2003), and held-out estimation (Zollmann, 2004). Nothing in what follows depends on the choice of probability model, however.

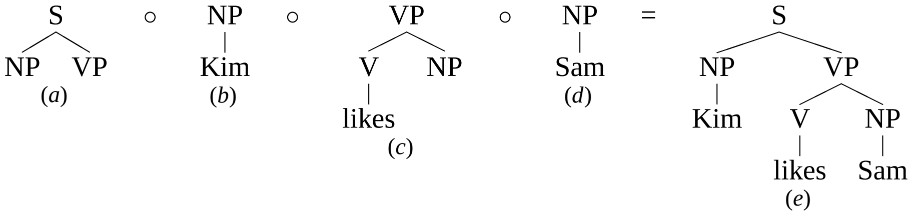


Figure 4: Fragment composition

(3) $$P(R) = \sum_{j=1}^{m} P(d_j)$$

Equation (1) says that the probability associated with a fragment f_i is the ratio of the number of times it occurs compared to the number of times fragments with the same root category occur. (2) says that the probability of a particular derivation d is the product of the probabilities of the fragments used in deriving it. (3) says that the probability associated with a representation (tree) is to be found by summing over the probabilities of its derivations.

Apart from its obvious simplicity, this version of DOP has numerous attractions. However, from a linguistic point of view it suffers from the limitations of the underlying linguistic theory (context-free phrase structure grammar), and for this reason does not provide a satisfactory answer to the question of how probabilistic and linguistic methods should be combined. A much better answer emerges if DOP techniques are combined with a richer linguistic theory, such as LFG.[3]

3 LFG-DOP

The idea of combining DOP techniques with the linguistic framework of LFG was first proposed in Bod and Kaplan (1998) (see also Bod and Kaplan, 2003; Way, 1999; Bod, 2000b,a; Hearne and Sima'an, 2004; Finn et al., 2006; Bod, 2006). As one would expect given the framework, representations are triples $\langle c, \phi, f \rangle$, consisting of a c-structure, an f-structure, and a 'correspondence' function ϕ that relates them (see Figure 5).[4]

Decomposition again involves the $Root$ and $Frontier$ operations. As regards c-structure, these operations are defined precisely as in Tree-DOP. However, the operations must also take account of f-structure and the ϕ-links: (i) when a node is erased, all ϕ-links leaving from it are removed, and (ii) all f-structure units that are not ϕ-accessible from the remaining nodes are erased.[5] (iii) In addition, $Root$

[3] Attempts to adapt DOP for other grammatical formalisms, notably HPSG, include Neumann (2003), Linardaki (2006), and Arnold and Linardaki (2007).

[4] Discussion of the key ideas of LFG can be found in e.g. Bresnan (1982), Dalrymple et al. (1995), Bresnan (2001), and Dalrymple (2001).

[5] A piece of f-structure is ϕ-accessible from a node n if and only if it is ϕ-linked to n or contained within a the piece of f-structure that is ϕ-linked to n.

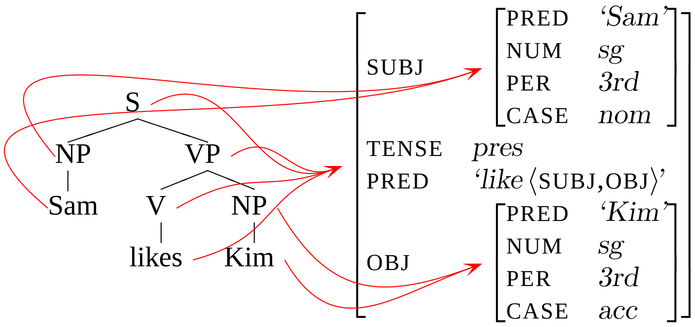


Figure 5: LFG-DOP Treebank representation.

deletes all semantic forms (PRED features) that are local to f-structures which are linked to erased nodes. (iv) *Frontier* also removes semantic forms from f-structures corresponding to erased nodes.

The intuition here is (a) to eliminate f-structure that is not associated with the c-structure that remains in a fragment, and (b) to keep everything else, except that a fragment should contain a PRED value if and only if the c-structure contains the corresponding word. Thus, from the representation in Figure 5, *Root* will produce (*inter alia*) fragments corresponding to the NPs *Sam* and *Kim* and the VP *likes Kim*, as in Figure 6. The cases of *Sam* and *Kim* are straightforward: all other nodes, and the associated ϕ-links have been removed; the only f-structures that are ϕ-accessible are the values of SUBJ and OBJ respectively, and these are what appear in the fragments. The case of the VP *likes Kim*, is slightly more complex: deleting the S and subject NP nodes does not affect ϕ-accessibility relations, because the S and VP nodes in Figure 5 are ϕ-linked to the same f-structure. However, deleting the subject NP removes the PRED feature the SUBJ value, as required by (iii). Notice that nothing else is removed: in particular, notice that person-number information about the subject NP remains.

Applying *Frontier* to Figure 6 (*c*) to delete *Kim* will produce a fragment corresponding to *likes NP*, as in Figure 7. Again, ϕ-accessibility is not affected, so the only effect on the f-structure is the removal of the PRED feature associated with *Kim*, as required by (iv).

The composition operation will not be very important in what follows. For the purpose at hand it can be just the same as that of Tree-DOP, with two provisos. First, we must ensure that substitution of a fragment at a node preserves ϕ-links and also unifies the corresponding f-structures. Second, we require the f-structure of any final representation we produce to satisfy a number of additional well-formedness conditions, specifically *uniqueness*, *completeness* and *coherence*, in the normal LFG sense (e.g. Dalrymple, 2001, pp35-39). Similarly, for the purpose of this discussion we can assume the probability model is the same as used in Tree-DOP. [6]

[6] In fact, a small extension of the probability model is needed. *Completeness* cannot be checked in

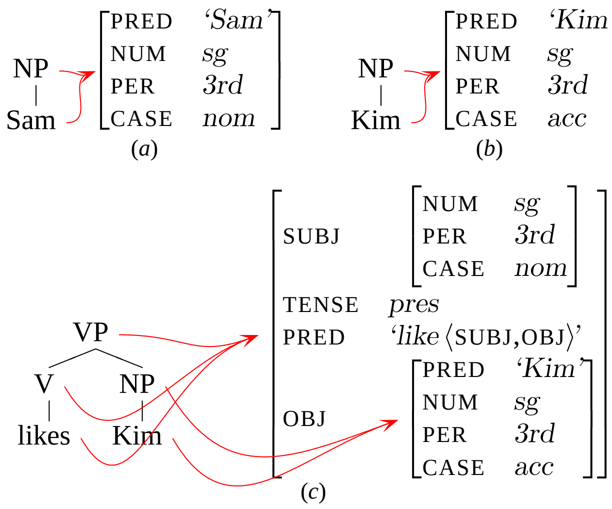


Figure 6: LFG-DOP *Root* fragments

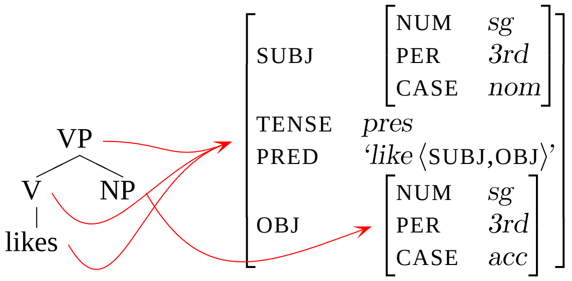


Figure 7: An LFG-DOP *Frontier* fragment

What is of central concern here is that the fragments produced by *Root* and *Frontier* are highly *undergeneral* (overspecific). In particular, the fragment for *Sam* is *nom*, the fragment for *Kim* is *acc*, and in the fragment for *likes NP* the direct object NP is third person and singular.

This will lead to under-generation (under-recognition). For example, it will not be possible to use the *Root* fragments for *Sam* and *Kim* in Figure 6 in analyzing a sentence like (4) where *Kim* appears as a subject, and *Sam* as an object, because they have the wrong case marking. Similarly, it will not be possible to use the *Frontier* fragment in Figure 7 to analyze (5), since it requires the OBJ to be 3rd person singular, which *us*, *them* etc. are not.[7]

the course of a derivation, but only on final representations, some of which will therefore be invalid. The problem is that the probability mass associated with such representations is lost. Bod and Kaplan (2003) address this issue by re-normalizing to take account of this wasted probability mass.

[7] Another way of thinking about this problem is as an exacerbation of the problem of *data sparsity*: an approach like this will require much more data to get an accurate picture of the contexts where words and phrases can occur. Data sparsity is one of the most pervasive and difficult problems for

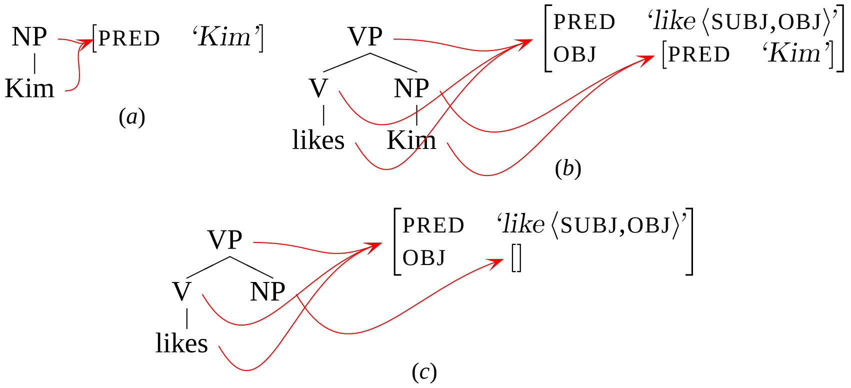


Figure 8: Overgeneral *Discard* fragments

(4) Kim likes Sam.
(5) Sam likes them/us/me/you/the children.

To deal with this problem, B&K introduce a further operation, *Discard*, which produces more general fragments by erasing features. *Discard* can erase any combination of features apart from PRED, and those features whose values ϕ-correspond to remaining c-structure nodes. As regards the fragments *Sam* and *Kim*, this means everything except the PRED can be removed, as in Figure 8 (*a*). In the case of *likes Kim* in Figure 6 (*c*), this means everything can be removed except for the value of PRED and the OBJ (and its PRED), see Figure 8 (*b*). In the case of *likes NP* in Figure 7, it means everything can be removed except the PRED and the OBJ (however, though the OBJ remains, the features it contains can be deleted), see Figure 8 (*c*).

Clearly, such fragments are *over*-general (under specific). For example, the fragment for *Kim* in Figure 8 (*a*) will be able to appear as subject of a non-third person singular verb, as in (6); the fragments for *likes NP* and *likes Kim* will allow non-third singular subjects (and subjects marked accusative), and the fragment for *likes NP* will also allow a nominative object, as in (7).

(6) *Kim were happy.
(7) *Them likes we.

To deal with this, B&K propose a redefinition of grammaticality: rather than regarding as grammatical anything which can be given an analysis, they regard an utterance as grammatical if it can be derived without using *Discard* fragments. For words with relatively high frequency (including common names such as *Kim* and *Sam* and verbs such as *likes*) this is likely to work. For example, every derivation of examples like (6) and (7) is likely to involve *Discard* fragments, so they will be correctly classified as ungrammatical. Equally, (4) will have a non-*Discard*

statistical approaches to natural language.

derivation, and be correctly classified as grammatical, so long as *Kim* appears at least once as a subject, and *Sam* appears at least once as an object, and (5) will have a non-$Discard$ derivation so long as *likes* appears with a sufficiently wide range of object NPs.

The reason this can be expected to work for high frequency words is that for such words the corpus distribution represents the true distribution (i.e. in the language as whole). Unfortunately, most words are *not* high frequency, and their appearance in corpora is not representative of their true distribution. In fact, it is quite common for more than 30% of the words in a corpus to appear only once — and of course this single occurrence is unlikely to reflect the true potential of the word.[8]

For example, in the British National Corpus (BNC) the noun *debauches* ('moral excesses') appears just once, as in (8), where it will be *acc.* Thus, the only way to produce (9) will be to use a $Discard$ fragment. But (8) and (9) are equally grammatical.

(8) [H]e . . . shook Paris by his wild debauches on convalescent leave.
(9) His wild debauches shook Paris.

Similarly, the verbs *to debauch* ('to corrupt morally') and *to hector* ('talk in a bullying manner') appear several times, but never with a first person singular subject: So analyzing (10) and (11) will require $Discard$ fragments, and they will be classified as ungrammatical. But both are impeccable.

(10) I never debauch anyone.
(11) I never hector anyone.

In short: there is a serious theoretical problem with the way LFG-DOP fragments are defined. Without $Discard$, the fragments are *under*general, and the model undergenerates, e.g. it cannot produce (4) and (5). There is a clear need for a method of producing more general fragments via some operation like $Discard$. However, as formulated by B&K, $Discard$ produces fragments that are *over*general, and the model overgenerates, producing examples like (6) and (7). Since B&K's attempt to avoid this problem via a redefinition of grammaticality does not help, we need to consider alternative approaches. The most obvious being to impose constraints on the way $Discard$ operates (cf Way, 1999).[9]

[8] Baroni (to appear) notes that about 46% of all words (types) in the written part of the British National Corpus (90 million tokens) occur only once (in the spoken part the figure is 35%, lower, but still above $1/3$). Of course, the BNC is not huge by human standards: listening to speech at normal rates (say, 200 words per minute) for twelve hours per day, one will encounter more than half this number of tokens each year ($200 \times 60 \times 12 \times 365 = 52,560,000$). But Baroni also observes that the proportion of words that appear only once seems to be largely independent of corpus size.

[9] A number of participants at LFG07 suggested alternative approaches based on 'smoothing', rather than $Discard$ (see also Hearne and Sima'an (2004)). Suppose, we have seen the proper name *Alina* just once, marked *nom* (*Alina*$_{nom}$). We 'smooth' the corpus data, by treating *Alina*$_{acc}$ as an 'unseen event' (e.g. we might assign it a count of 0.5). We can generalize this to eliminate

4 Constraining *Discard*

The problem with B&K's formulation of $Discard$— the reason it produces overgeneral fragments — is that it is indiscriminate. In particular, it does not distinguish between features which are 'inherent' to a fragment (that is, 'grammatically necessary' given its c-structure), and those which are 'contextual' or 'contingent' given its c-structure and are simply artifacts of structure that has been eliminated by the decomposition operations. The former must not be discarded if we are to avoid overgeneration; the latter can, and in the interest of generality should, be discarded. Consider, for example, the fragment for *likes NP* in Figure 7. Intuitively, the PER and NUM features on the object NP are just 'contextual' here — they simply reflect the presence of a third person singular NP in the original representation. On the other hand, the CASE feature on the object is grammatically necessary, as are the PER, NUM and CASE features on the subject NP (given that the verb is *likes*). Similarly, with fragments for NPs like *Sam* and *Kim*: PER and NUM features seem to be grammatically necessary, but CASE seems to be an artefact of the context in which the fragments occur (while with a fragment for *she* all three features would be grammatically necessary).

One approach would be to look for general constraints on $Discard$, e.g. to try to identify certain features as grammatically 'essential' in some way, and immune to $Discard$ (i.e. like PRED for B&K). While appealing, this seems to us unlikely to be sucessful, and certainly no plausible candidates have been proposed.[10]

We think this is not an accident. Rather, the difficulty of finding general constraints on $Discard$ is a reflection of a fundamental feature of f-structures, and LFG: the fact that f-structures do not record the 'structural source' of pieces of f-structure. This is in turn a reflection of an important fact about natural language — one for which constraint based formalisms provide a natural expression: that information at one place in a representation may have many different structural sources (in the case of agreement phenomena, many sources simultaneously). Consider, for

the need for $Discard$: we simply hypothesize similar unseen events for all possible attribute-value combinations. This is an interesting approach, but (a) it will overgenerate, and (b) we will still be unable to reconstruct any idea of grammaticality. To see this, consider that we will also treat *Alina* marked plural (*Alina*$_{pl}$) as an unseen event, and presumably assign it the same count as *Alina*$_{acc}$. We will now be able to derive **Aline run* (so we have overgeneration). Moreover, the same arguments that we used to show the inadequacy of $Discard$ as a basis for a notion of grammaticality apply here, equally (e.g. if we try to identify ungrammaticality with 'involving a smoothed fragment'). Notice it is not the case that grammatical sentences will receive higher probability on such an account: suppose that the probability of *NP run* is the same or higher than *We saw NP*: it is likely that the probability assigned to **Alina run* will be the same or higher than *We saw Alina*. (We are especially grateful to Ron Kaplan, Jonas Kuhn, and Grzegorz Chrupała for stimulating discussion on this point.)

[10] Way (1999), suggests it might be possible to classify features as 'lexical' or 'structural' in some general fashion (so the presence of 'lexical' features in fragments would be tied to the presence of lexical material in c-structures in the same way as PRED). He suggests PER and NUM might be lexical, and CASE might be structural, but notice that there are cases where CASE is associated with particular lexical items (e.g. pronouns *she*, *her*), and where PER and NUM values are associated with a particular structure (e.g. subject of a verb with a third person singular reflexive object, such as *NP criticized herself*).

example, the NUM:*pl* feature that will appear on the subject NPs in the following:

(12) These sheep used to be healthy.
(13) Sam's sheep are sick.
(14) Sam's sheep used to look after themselves.
(15) These sheep are able to look after themselves.
(16) Sheep can live in strange places.

In (12), this feature is a reflex of the plural determiner; in (13) it is a result of the form of the verb (*are*); in (14) it is a result of the reflexive pronoun; in (15) it comes from all these places at once; in (16) it is the *absence* of an article that signals that the noun is plural.

Thus, instead of trying to find general constraints, we propose that the production of generalized fragments should be constrained by the existence of what we will call 'abstract fragments'. Intuitively, abstract fragments will encode information about what is grammatically essential, and so provide an upper bound on the generality of fragments that can be produced by $Discard$. We will call this generalizing operation $cDiscard$ ('constrained $Discard$'). Furthermore, we propose that the knowledge underlying such abstract fragments be expressed using normal LFG grammar rules.

Formally, the key insight is that it is possible to think of a grammar and lexicon as generating a collection of (often very general) fragments, by constructing the minimal c-structure that each rule or lexical entry defines, and creating ϕ-links to pieces of f-structure which are minimal models of the constraints on the right-hand-side of the rule. We will call fragments produced in this way 'basic abstract fragments'.

For example, suppose that, in response to the problems discussed above, we postulate the rules and entries in (17). These rules can be interpreted so as to generate the basic abstract fragments in Figure 9.[11]

(17)	a.	S →	NP	VP
			(↑SUBJ CASE)=*nom*	↑=↓
	b.	VP →	V	NP
			↑=↓	(↑OBJ CASE)=*acc*
	c.	*Kim*	NP	(↑NUM)=*sg*
				(↑PER)=*3*
	d.	*she*	NP	(↑NUM)=*sg*
				(↑PER)=*3*
				(↑CASE)=*nom*
	e.	*her*	NP	(↑NUM)=*sg*
				(↑PER)=*3*
				(↑CASE)=*acc*

[11] Notice that we do not follow the normal LFG convention whereby the absence of f-structure annotation on category is interpreted as '↑=↓': absence of annotation means exactly an absence of f-structure constraints. Notice also that this means we are treating the ϕ-correspondence as a partial function in abstract fragments: in Figure 9 (a) the NP is not linked to any f-structure.

f. *likes* V (↑SUBJ NUM)=*sg*
(↑SUBJ PER)=*3*
(↑TENSE)=*pres*

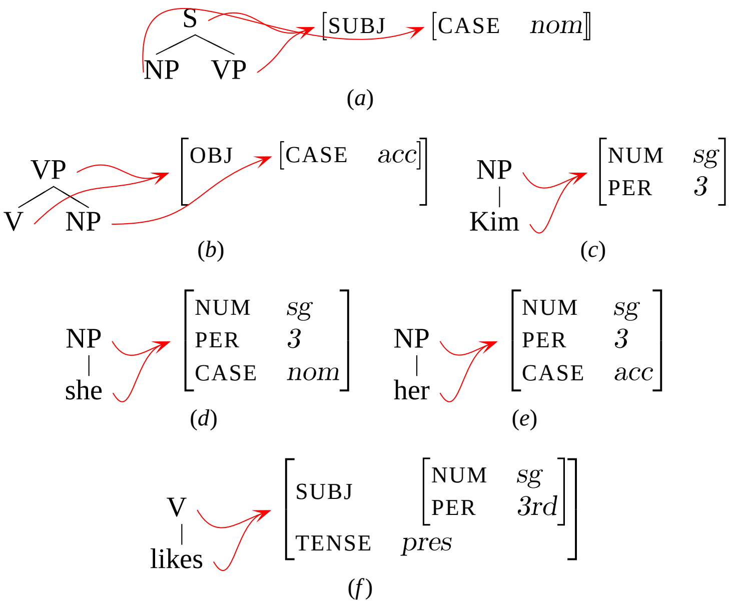


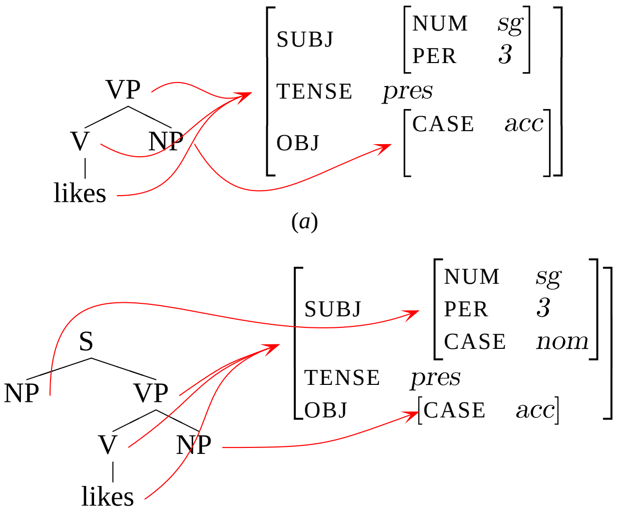
Figure 9: Basic abstract fragments generated by the grammar rules in (17)

Formally speaking, these are fragments in the normal sense, and they can be composed in the normal way. For example composing Figure 9 (*b*) and Figure 9 (*f*) will produce the 'derived' abstract fragment in Figure 10 (*a*). This in turn can be composed with Figure 9 (*a*) to produce Figure 10 (*b*). The idea is that such fragments can be used to put an upper bound on the generality of the fragments produced by $cDiscard$, by requiring the latter to be 'licensed' by an abstract fragment.

More precisely, we require that, for a fragment f, if $cDiscard(f)$ produces fragment f_d, then there must be some abstract fragment f_a which *licenses* f_d, which for the moment we take to mean f_a 'frag-subsumes' f_d. We will say that an abstract fragment f_a *frag-subsumes* a fragment f_d just in case:

1. the c-structures are isomorphic, with identical labels on corresponding nodes; and
2. the ϕ-correspondence of f_a is a subset of the ϕ-correspondence of f_d (recall that ϕ-correspondences are functions, i.e. sets of pairs); and
3. every f-structure in f_a subsumes (in the normal sense) the corresponding f-structure of f_d.[12]

[12] This desciption glosses over a small formal point: normal fragments contain an f-structure with a single root. For abstract fragments this will not always be the case. For example, a rule like



(*a*)

(*b*)

Figure 10: Derived abstract fragments

To see the effect of this, consider the *Root* and *Frontier* fragments in Figure 11 (*b*), (*d*) and (*f*), and the abstract fragments that would license possible applications of *Discard* to them, in Figure 11 (*a*), (*c*) and (*e*).

The abstract fragment in Figure 11 (*a*) will license the discarding of PER and NUM from the object slot of Figure 11 (*b*), but will not permit discarding of TENSE information, or information about the CASE of the subject or object, or PER and NUM information from the subject. Thus, we will have fragments of sufficient generality to analyze (18), but not (19):

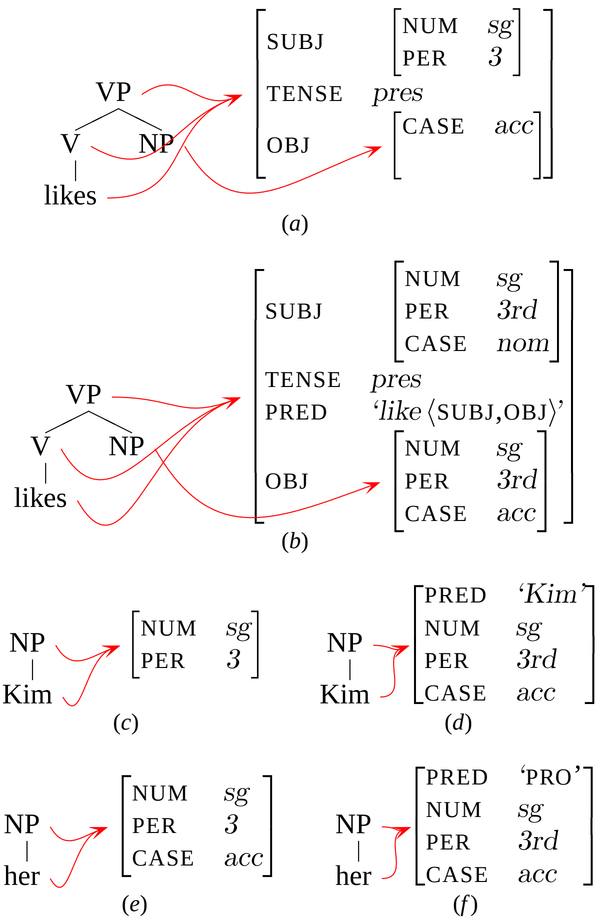
(18) Sam likes them/us/me/the children. [=(5)]
(19) *Them likes we. [= (7)]

Similarly, the abstract fragment in Figure 11 (*c*) will license generalized fragments for *Kim* from which CASE has been discarded, but will not allow fragments which from which PER or NUM information has been discarded. Thus, as we would like, we will be able to analyze examples where *Kim* is an object, but not where it is, say, the subject of a non-third person singular verb:

(20) Kim likes Sam. [= (4)]
(21) *Kim were happy. [= (6)]

On the other hand, the abstract fragment in Figure 11 (*e*) will not permit any features to be discarded from *her*, which will therefore be restricted to contexts which allow third person singular accusatives:

S →NP VP (without any constraints) should produce an abstract fragment with c-structure consisting of three nodes, each associated with a separate, empty, f-structure.



(*a*)

(*b*)

(*c*)

(*d*)

(*e*)

(*f*)

Figure 11: *Root*, *Frontier*, and abstract fragments

(22) Sam likes her.
(23) *Her likes Sam.

5 General Constraints

The previous section has shown how one source of overgeneration can be avoided. A second source of overgeneration arises from the fact that, while it provides a reasonable model of normal c- and f-structure constraints (i.e. defining equations), an LFG treebank is only a poor reflection of other kinds of constraint, e.g. negative constraints, functional uncertainty constraints, existential constraints, and constraining equations.[13] A treebank is a finite repository of positive information, and cannot properly reflect negative constraints, constraints with potentially infinite

[13] See Dalrymple (2001) for discussion and exemplification of such constraints.

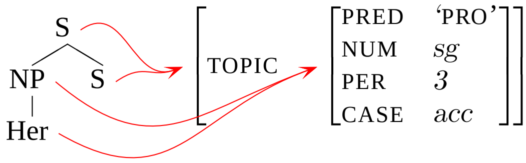


Figure 12: A *cDiscard Frontier* fragment

scope, or constraints whose essential purpose is information 'checking'. In this section we will show how the approach of the previous section can be extended to address this source of overgeneration. For reasons of space, we will focus on functional uncertainty constraints and negative constraints.

As an example of a functional uncertainty constraint, consider the need to 'link' topicalized constituents. Suppose the treebank contains representations of examples like (24) and (25).

(24) Her, Sam likes.
(25) Her, we think Sam likes.

As things stand, it will be possible to produce a fragment like Figure 12 from (24) by deleting the structure corresponding to *Sam likes* (and discarding a number of features like TENSE, which are not relevant here). Notice it will be possible to compose any complete sentence with this, and so derive ungrammatical examples like the following, in which the topicalized constituent *her* is not linked to any normal grammatical function.

(26) *Her, Sam likes Kim.

In a normal LFG grammar, examples like (26) are excluded by including a functional uncertainty constraint on the rule that produces topicalized structures:[14]

(27) $$\begin{array}{lcc} S \rightarrow & NP & S \\ & (\uparrow \text{TOPIC})=\downarrow & \uparrow=\downarrow \\ & (\uparrow \text{COMP* GF})=\downarrow & \end{array}$$

As things stand, the LFG-DOP model is unable to prevent examples like (26) being derived: there is no way of capturing the effect of anything like an uncertainty constraint.

As regards negative constraints, in Section 4 we expressed facts about subject verb agreement with *likes* by means of a positive constraint requiring its subject to

[14] In (27), GF is a variable over grammatical function names, such as OBJ and SUBJ, and COMP* is a regular expression meaning any number of COMPs (including zero). COMP is the grammatical function associated with complement clauses. Thus, the constraint requires the NP's f-structure to be the OBJ (or SUBJ, etc.) of its sister S, or of a complement clause inside that S, or a complement clause inside a complement clause (etc).

be 3rd person singular. This still leaves the problem of agreement for other forms. For example, we must exclude *like* appearing with a 3rd person singular form, as in (28).

(28) *Sam like Kim.

This can be expressed with a disjunction of normal constraints, but the most natural thing to say involves a negative constraint, along the lines of (29) (which simply says that the subject of *like* must not be third person singular). The existing apparatus provides no way of encoding anything like this.

(29) *like* V $\neg\Big($ ($\uparrow$SUBJ PER)=*3* ($\uparrow$SUBJ NUM)=*sg* $\Big)$

In fact, apparatus to avoid this sort of overgeneration is a straightforward extension of the approach described above.

- We add to fragments a fourth component, so they become 4-tuples: $\langle c, \phi, f, Constr\rangle$, where *Constr* is a collection of 'other' (i.e. non-defining) constraints.
- For basic abstract fragments the elements of $Constr$ are the 'other' constraints required by the corresponding rule or lexical entry.
- Combining abstract fragments involves unioning these sets of constraints.
- Licensing a fragment involves adding these constraints to the fragment (i.e. fragments inherit the Constraints of the abstract fragment that licenses them).
- The composition process is amended so as to include a check that these constraints are not violated (specifically, we require that, in addition to normal completeness and coherence requirements, the f-structure of any final representation we produce must satisfy all constraints in $Constr$).

The idea is that, given a grammar rule like (29), any basic abstract fragment for *like* will include a negative constraint on the appropriate f-structure, which will be inherited by any derived abstract fragment, and any fragment that is thereby licensed. So, for example, the most general $cDiscard$ fragment for *NP like Kim* will be as in Figure 13. While it will be possible to adjoin a 3rd person singular NP to the subject position of this fragment, this will not lead to a valid final representation, because the negative constraint will not be satisfied. Thus, as one would hope, we will be able to derive (30), but not (31).

(30) They like Kim.
(31) *Sam like Kim.

Similarly, the rule in (27) will produce abstract fragments which contain the uncertainty constraint given, and these will license normal fragments like that in Figure 14. Again, the only valid representations which can be constructed which satisfy this constraint will be ones which contain a 'gap' corresponding to the TOPIC. That is, as one would like, we will be able to produce (32), but not (33):

(32) Her, Sam (says she) likes.
(33) *Her, Sam (says she) likes Kim.

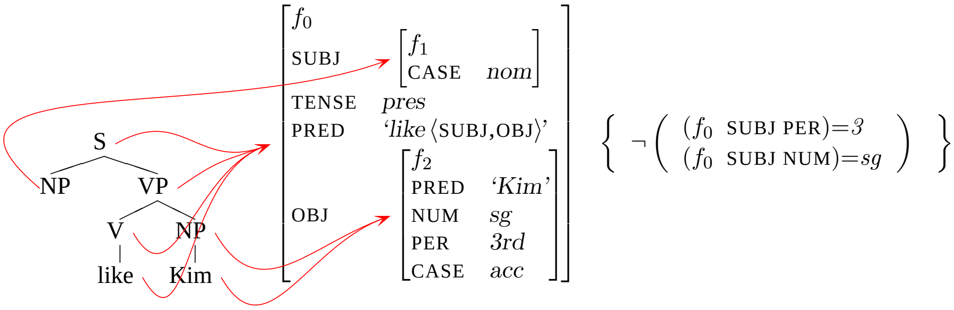

S
NP
VP
V
NP
like
Kim
f_0
SUBJ
f_1
CASE nom
TENSE pres
PRED 'like ⟨SUBJ,OBJ⟩'
OBJ
f_2
PRED 'Kim'
NUM sg
PER 3rd
CASE acc
$\{ \neg \left(\begin{array}{l} (f_0 \text{ SUBJ PER})=3 \\ (f_0 \text{ SUBJ NUM})=sg \end{array} \right) \}$

Figure 13: Fragment incorporating a negative constraint

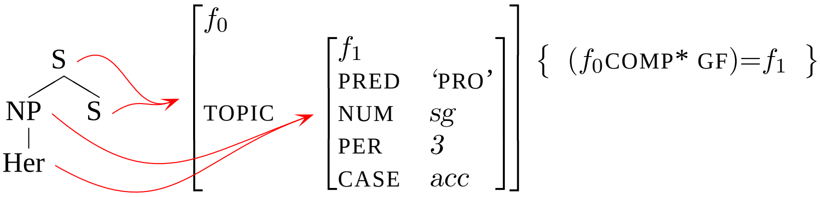


Figure 14: Fragment incorporating an uncertainty constraint

6 Discussion

The proposals presented in the previous sections constitute a relatively straightforward extension to the formal apparatus of LFG-DOP, but they are open to a number of objections, and they have theoretical implications of wider significance.

One kind of objection that might arise is a result of the relatively minor phenomena we have used for exemplification (case assignment and person-number agreement in English). This objection is entirely misplaced. First, because, in an LFG context, similar problems will arise in relation to any phenomenon whose analysis involves f-structure attributes and values. More generally, similar problems of fragment generality will arise whenever one tries to generalize DOP approaches beyond the context-free case, e.g. to deal with semantics.[15] More generally still, analogues of the problems we have identified with fragment generality and capturing the effect of 'general' constraints on the basis of a finite collection of example representations will arise with any 'exemplar' based approach.

A second source of objections might arise from the fact that we have focused

[15] At least, this is the case if one wants to preserve the idea that a treebank consists of representations in the normal sense. In the approach to semantic interpretation in DOP described in Bonnema et al. (1997) these problems are avoided at the cost of not using semantic representations in the normal sense. Rather than having semantic representations, the nodes of trees are annotated with an indication of how the semantic formula of the node is built up from the semantic formulae of its daughters, and hence how it should be decomposed. The 'fragment generality' problem is sidestepped by explicitly indicating on each and every node how its semantic representation should be decomposed as fragments are created.

on the problem of overgeneration: one might object (a) that in a practical, e.g. language engineering, setting this is not very important, and (b) that in a probabilistic setting, such as DOP, overgeneration can be hidden statistically (e.g. because ungrammatical examples get much smaller probability compared to grammatical ones).

As regards (a), the appropriate response is that a model which overgenerates is generally one which assigns excessive ambiguity (which is a pervasive problem in practical settings). Sag (1991) gives a large number of plausible examples. In relation to subject-verb agreement, he notes that the following are *un*ambiguous, but will be treated as ambiguous by any system that ignores subject-verb agreement: (34) presumes the existence of a unique English-speaking Frenchman among the programmers; (35) presumes there is a unique Frenchman among the English speaking programmers.

(34) List the only Frenchman among the programmers who understands English.
(35) List the only Frenchman among the programmers who understand English.

Similarly, a system which does not insist on correct linking of Topics will treat (36) and (37) as ambiguous, when both are actually unambiguous (in (36) *to them* must be associated with *contributed*, in (37) it must be associated with *appears*, because *contribute* requires, and *discover* forbids, a complement with *to*):

(36) To them, Sam appears to have contributed it.
(37) To them, Sam appears to have discovered it.

As regards (b), it is important to stress that the problem of overgeneration as we describe it has to do with the characterization of grammaticality (i.e. the characterization of a language), and grammaticality simply cannot be identified with relative probability (casual inspection of almost any corpus will reveal many simple mistakes, which are uncontroversially ungrammatical, but have much higher probability than perfectly grammatical examples containing, e.g., rare words).

A third objection would be that in avoiding overgeneration, we have also lost the ability to deal with ill-formed input (robustness). But there is no reason why the model should not incorporate, in addition to 'constrained $Discard$', an unconstrained operation like the original B&K $Discard$. Notice that this would now give a correct characterization of grammaticality (a sentence would be grammatical if and only if it can be derived without the use of unconstrained $Discard$ fragments).

A fourth, and from a DOP perspective very natural, objection would be that these proposals in some sense violate the 'spirit' of DOP — where an important idea is exactly to dispense with a grammar in favor of (just) a collection of fragments. A partial response to this is to note that to a considerable degree the sort of grammar we have described is implicit in the original treebank. For example, the set of c-structure rules can be recovered from the treebank by simply extracting all trees of depth one. This will produce a grammar without f-structure constraints, and abstract fragments with empty f-structures and constraint sets, which is exactly equivalent to the original B&K model. Taken as a practical proposal for grammar engineering, the idea would be that one can begin with such an unconstrained model, and simply add constraints to these c-structure rules to rule out overgeneration. This can clearly be done incrementally, and in principle, the full range of LFG rule notation should be available, so this should be a relatively straightforward and natural task for a linguist. It should be, in particular, much easier than writing a normal grammar.

However, it is also possible to take the proposal in a different way, 'theoretically', as describing an idea about linguistic knowledge, and human language processing and acquisition. Taken in this way, the suggestion is that a speaker has at her disposal two knowledge sources: a database of fragments (in the normal DOP sense), which one might think of as a model of grammatical usage, and a grammar (an abstract fragment grammar) which expresses generalizations over these fragments, which one might take to be a characterization of something like grammatical competence. Notice that on this view: (i) the grammar as such plays no role in sentence processing (but only in fragment creation, i.e. off-line); (ii) the task of the learner is only secondarily to construct a grammar (the primary task is the creation of the fragment database — learning generalizations over this is a secondary task); (iii) the grammar does not generate or otherwise precisely characterize the language (this is achieved by the fragment database with the composition operation), rather its job is to license or legitimize the fragments in the fragment database. Taken in this way, the model is an enrichment of the standard DOP approach.

References

Arnold, Doug and Linardaki, Evita. 2007. A Data-Oriented Parsing Model for HPSG. In Anders Søgaard and Petter Haugereid (eds.), *2nd International Workshop on Typed Feature Structure Grammars (TFSG'07)*, pages 1–9, Tartu, Estonia: Center for Sprogteknologi, Kobenhavens Universitet, Working Papers, Report No. 8.

Baroni, Marco. to appear. Distributions in Text. In Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, Berlin: Mouton de Gruyter.

Bod, Rens. 1998. *Beyond Grammar: An Experience-Based Theory of Language*. Stanford, California: CSLI Publications.

Bod, Rens. 2000a. An Empirical Evaluation of LFG-DOP. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING00)*, volume 1, pages 62–68, Saarbrüken.

Bod, Rens. 2000b. An Improved Parser for Data-Oriented Lexical-Functional

Analysis. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 61–68, Hong Kong.

Bod, Rens. 2006. Exemplar-Based Syntax: How to Get Productivity From Examples. *The Linguistic Review* 23(3), 291–320, (Special Issue on Exemplar-Based Models in Linguistics).

Bod, Rens and Kaplan, Ronald. 1998. A Probabilistic Corpus-Driven Model for Lexical Functional Analysis. In *Proceedings of COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 145–151, Montreal, Canada.

Bod, Rens and Kaplan, Ronald. 2003. A DOP Model for Lexical-Functional Grammar. In Rens Bod, Remko Scha and Khalil Sima'an (eds.), *Data-Oriented Parsing*, Chapter 12, pages 211–233, Stanford, California: CSLI Publications.

Bod, Rens and Scha, Remko. 1997. Data Oriented Language Processing. In Steve Young and Gerrit Bloothooft (eds.), *Corpus-Based Methods in Language and Speech Processing*, volume 2 of *Text, Speech and Language Technology*, pages 137–173, Dordrecht: Kluwer Academic Publishers.

Bod, Rens, Scha, Remko and Sima'an, Khalil (eds.). 2003. *Data-Oriented Parsing*. Stanford, California: CSLI Publications.

Bonnema, Remko, Bod, Rens and Scha, Remko. 1997. A DOP Model for Semantic Interpretation. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the EACL*, pages 159–167, Madrid, Spain.

Bonnema, Remko, Buying, Paul and Scha, Remko. 1999. A new probability model for Data Oriented Parsing. In Paul Dekker and Gwen Kerdiles (eds.), *Proceedings of the 12th Amsterdam Colloquium*, pages 85–90, Amsterdam, The Netherlands.

Bresnan, Joan (ed.). 1982. *The Mental Representation of Grammatical Relations*. Cambridge, Massachussets: MIT Press.

Bresnan, Joan. 2001. *Lexical-Functional-Syntax*. Blackwell Textbooks in Linguistics, Oxford: Blackwell.

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

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

Finn, Riona, Hearne, Mary, Way, Andy and van Genabith, Josef. 2006. GF-DOP: Grammatical Feature Data-Oriented Parsing. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG06 Conference*, Stanford, California: CSLI Publications.

Hearne, Mary and Sima'an, Khalil. 2004. Structured Parameter Estimation for LFG-DOP. In N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (eds.), *Recent Advances in Natural Language Processing III: Selected papers from the RANLP 2003*, volume 260 of *Current Issues in Linguistic Theory*, pages 183–192, Amsterdam: John Benjamins.

Johnson, Mark. 2002. The DOP Estimation Method is Biased and Inconsistent. *Computational Linguistics* 28, 71–76.

Jurafsky, Dan. 2003. Probabilistic Modeling in psycholinguistic comprehension and production. In Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probabilistic Linguistics*, Chapter 3, pages 39–96, Cambridge, Massachusetts: MIT Press.

Linardaki, Evita. 2006. *Linguistic and statistical extensions of Data Oriented Parsing*. PhD thesis, University of Essex.

Neumann, Günter. 2003. A Data-Driven Approach to Head-Driven Phrase Structure Grammar. In Rens Bod, Remko Scha and Khalil Sima'an (eds.), *Data-Oriented Parsing*, Chapter 13, pages 233–251, Stanford, California: CSLI Publications.

Sag, Ivan A. 1991. Linguistic Theory in Natural Language Processing Language Processing. In Ewan Klein and Frank Veltman (eds.), *Natural Language and Speech*, pages 69–84, Berlin: Springer Verlag.

Sima'an, Khalil and Buratto, Luciano. 2003. Backoff Parameter Estimation for the DOP Model. In Nada Lavrac, Dragan Gamberger, Hendrik Blockeel and Ljupco Todorovski (eds.), *Proceedings of the European Conference on Machine Learning*, Lecture Notes in Artificial Intelligence, pages 373–384, Berlin: Springer.

Way, Andy. 1999. A Hybrid Architecture for Robust MT using LFG-DOP. *Journal of Experimental and Theoretical Artificial Intelligence* 11, 441–471, (Special Issue on Memory-Based Learning).

Zollmann, Andreas. 2004. *A Consistent and Efficient Estimator for the Data-Oriented Parsing Model*. Masters Thesis, ILLC, Amsterdam, The Netherlands.

LEXICAL SHARING AND NON-PROJECTING WORDS: THE SYNTAX OF ZAPOTEC ADJECTIVES

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

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

A Zapotec attributive adjective forms a single phonological word with the noun that it modifies. This N+Adjective combination is an instance of an element that corresponds to one word in phonology, but two words in syntax. These mismatches can be successfully captured in the lexical sharing approach of Wescoat (2002).

1 Introduction[1]

Sadler and Arnold (1994), Sadler (2000) and Toivonen (2001, 2003) have introduced the idea of non-projecting words into LFG, focusing on data from Welsh and Swedish. In both Welsh and Swedish, the non-projecting elements are phonologically independent words. However, Toivonen (2001, 2003) argues that the criteria of syntactic projection and phonological dependence are separable, so it should be possible for non-projecting words to form a phonological unit with another word.

This paper argues for such an analysis in San Dionicio Ocotepec Zapotec (SDZ), an Otomanguean language of Oaxaca, Mexico. In this language, an attributive adjective forms a single phonological word with a preceding noun. I argue that Âdj is a non-projecting word which adjoins to N, and that the two are instantiated as a single word, using the lexical sharing hypothesis of Wescoat (2002).[2]

SDZ is a head-initial language, as shown in NP (1) and S (2)

1) X-quéét Juààny
p-tortilla Juan
'Juan's tortilla'

2) Ù-dàw bè'cw gèèt.
com-eat dog tortilla
'The dog ate the tortilla.'

Topic and focal phrases frequently appear preverbally. I have italicized the gloss corresponding to such phrases to mark their special discourse function.

[1] I thank Joan Bresnan, Michael Galant, Tracy Holloway King, and Michael Wescoat for useful comments on this paper. Special thanks are due to Luisa Martínez, who supplied all the data for this paper.

The orthography for San Dionicio Ocotepec Zapotec is adapted from the practical orthographies for other Zapotec languages spoken in the Valley of Oaxaca. In the SDZ orthography, <x> = /ʒ/ before a vowel and /ʃ/ before a consonant, <xh> = /ʃ/, <dx> = /ʤ/, <ch> = /ʧ/, <c> = /k/ before back vowels, <qu> = /k/ before front vowels, <e> = /ɛ/ and <ey> = /e/. Doubled vowels are long. SDZ is a language with four contrastive phonation types: breathy <Vj>, creaky <V'V>, checked <V'>, and plain <V>. High tone is marked with an acute accent, low with a grave. Nominal tones are affected by position within the intonational phrase, and so nouns may show slightly varying tones from example to example. Please note that the representation of /ɛ/ and /e/ in the practical orthography which is found in this paper differs from that found in my previous publications on SDZ.

Ordinary affixes are separated from the stem by the hyphen; clitics are separated by =, and the compound boundary is shown by +. Glosses use the following abbreviations: aff = affirmative, com = completive aspect, def = definite future aspect, hab = habitual aspect, neg = negative, p = possessed, pot = potential aspect, pred = predicative, 1s =1st person singular, 3 = 3rd person human (ordinary respect level), 3i = 3rd person inanimate.

[2] I follow Toivonen (2003) in indicating a non-projecting word with a circumflex over its part of speech label.

3) Bè’cw ù-dàw gèèt.
dog com-eat tortilla
‘*The dog* ate the tortilla.’

Adjectives follow nouns in NP, and the N+Âdj combination forms a single, compound-like structure:

4) Ù-dàw Juáàny gèt+ró’
com-eat Juan tortilla+big
‘Juan ate the big tortilla.’

While the attributive adjectives form a single word with the noun they modify, predicative adjectives are independent words:

5) Ró’ gèèt.
big tortilla
‘The tortilla is big.’

In this paper, I will give an account of the syntactic and morphological relationship between predicative and attributive adjectives that crucially relies on the notions of non-projecting word and lexical sharing.

2 Evidence for single-word status

2.1 Phonological evidence

In SDZ, attributive adjectives form a word with the preceding noun.[3] This phonological union has a number of consequences, all ultimately related to stress placement.

First, because stress regularly occurs on the final syllable of a word, the final adjective in such sequences is stressed and the noun is unstressed. Second, the unstressed vowel of the N is now short.[4] (In the following examples, the stressed syllable is underlined.)

6) Ù-dàw Juáàny gèèt
com-eat Juan tortilla
‘Juan ate the tortilla.’

7) Ù-dàw Juáàny gèt+ró’
com-eat Juan tortilla+big
‘Juan ate the big tortilla.’

SDZ has four contrastive phonation types in stressed syllables – plain (V), breathy (Vj), checked (V’), and creaky (V’V):

[3] The earliest explicit claim that the noun and adjective form a single phonological word in some varieties of Zapotec seems to be Pickett (1997), for Isthmus Zapotec.

[4] It is probably most accurate to say that vowels with plain phonation lengthen in stressed syllables; because the vowel in *gèt* is now in an unstressed syllable, it remains short. However, a number of words borrowed from Spanish seem to have underlying long vowels which do not vary in length according to stress, e.g. *sóóp* ‘soup’.

8)	bààl	‘bullet’	(plain)
	bêjld	‘fish’	(breathy)
	bè’él	‘meat’	(creaky)
	bê’ld	‘snake’	(checked)

Phonation type contrasts are reduced in unstressed contexts. Because adjectives cause stress-shift, the addition of an adjective often causes a change in phonation type. In the following examples, breathy vowels become plain when unstressed:[5]

9)	bèjl	‘flame’
	bèl+ró’	‘big flame’

10)	bàjd	‘dried maguey leaf’
	bàd+ró’	‘big dried maguey leaf’

In the same context, creaky vowels become checked:[6]

11)	dù’ú	‘rope’
	dù’+ró’	‘big rope’

12)	bè’él	‘meat’
	bè’l+ró’	‘big meat’

The stress-related shifts seen in these examples are like those seen in clear cases of compounding. Compare the vowel shortening in the following example:

13)	gèèt	‘tortilla’	
	xtììlly	‘Spanish’	(< Span. *Castellano*)
	gèt+xtííly	‘bread’	

There are also tonal effects which are related to stress. SDZ has a floating H tone which docks to the first stressed vowel of a initial focussed phrase.[7] As a result of the stress shift in N+Âdj sequences, we see a third phonological change — the stressed Âdj attracts the floating H tone and the unstressed N receives default L tone.

Compare the following, in which the object has been fronted to the focus position. In (14) the floating H tone docks to *géét* ‘tortilla’, but in (15), it docks to the adjective *ró’* ‘big’ instead:

[5] Similar phonation type shifts are documented in Mitla Zapotec (Briggs 1961:9-10).

[6] Checked vowels show more complex behavior; some remain checked, and some become plain. Clearly, more needs to be said about the phonology of such words, but I will not pursue that issue in this paper.

[7] More accurately, the H docks to the first stressed vowel of the first intonational phrase within the focussed material. In the examples under consideration here, the focussed phrase is relatively light and shows only one possible phrasing. When the focussed phrase is heavier and more syntactically complex, there is often more than one way to construct the intonational phrases. See Broadwell (2000) for discussion.

14) H

[$_{\text{FocP}}$ Géét] ù-dàw Juáàny.
tortilla com-eat Juan
'Juan ate *the tortilla.*'

15) H

[$_{\text{FocP}}$ Gèt+ró'] ù-dàw Juáàny.
tortilla+big com-eat Juan
'Juan ate *the big tortilla.*'

As a result of all these phonological changes, *géét* in (14) is long, stressed, and high-toned, but *gèt* in (15) is short, unstressed, and low-toned.

2.2 Clitic placement

The N+Âdj structure also acts like a single word for the purposes of clitic placement. SDZ has a set of 2nd position clitics, which may occur after the first word or the first constituent of the phrase within their domain. I will give examples using two of these clitics as tests. One such clitic is *=chà' ~ =dxà'* 'maybe'; another is the affirmative clitic *=cà ~ = àc.*[8] Such a clitic appears after the first word or constituent of CP.

If the initial constituent is a topicalized or focused [N NP], then two positions for the clitic are possible:

16) a.) [X-quèèt]=**chà'** Juáàny ù-dàw bè'cw.
poss-tortilla=maybe Juan com-eat dog

b.) [X-quèèt Juáàny]**=dxà'** ù-dàw bè'cw.
poss-tortilla Juan=maybe com-eat dog
'Maybe the dog ate *Juan's tortilla.*'

17) a.) Éè, [x-quèèt]=**cà** Juáàny ù-dàw bè'cw.
yes p-tortilla=aff Juan com-eat dog

b.) Éè, [x-quèèt Juáàny] =**cà** ù-dàw bè'cw.
yes p-tortilla Juan=aff com-eat dog
'Yes, the dog ate *Juan's tortilla.*'

This flexibility in clitic position is found with almost every type of noun phrase. However, an initial [N+Âdj] combination may never be split up by a clitic:

[8] These clitics show the following allomorphy: For the 'maybe'clitic, *=chà'*[ʧàʔ] is found after voiceless segments; *=dxà'* [ʤàʔ] after voiced segments. For the affirmative clitic, *=cà* is found after a consonant and *=àc* after vowels.

18) a. [Gèt+ró’]=**dxà’** ù-dàw bè’cw
tortilla+big=maybe com-eat dog
‘Maybe the dog ate *the big tortilla*.’

b.) *[Gèt=**chà’** ró’] ù-dàw bè’cw
tortilla=maybe big com-eat dog

Furthermore, phrases like the following show us that the N+Âdj combination may count as the first word in a more complex NP:

19) [X-qùeht+ró’**]=dxà’** Juáàny ù-dàw bè’cw
poss-tortilla+big=maybe Juan com-eat dog
‘Maybe the dog ate *Juan’s big tortilla*.’

Thus the evidence for the [N+Âdj] combination as a single phonological word is strong. This implies that there must be a productive lexical rule joining N+Âdj together.

3 Lexical sharing

We need a lexical rule which combines a N and a Âdj of the following type (using the conventions of Wescoat 2002):

8.) Φ← N, Ψ ← Âdj ⇨ [Φ - Ψ] ← N Âdj

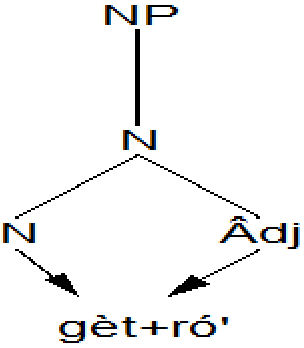


Figure 1 A lexical sharing configuration

This rule is interpreted as follows ‘If Φ instantiates a N and Ψ instantiates Âdj, then Φ - Ψ is a word which instantiates N Âdj.’ This points toward an analysis of Zapotec where attributive adjectives are non-projecting words, adjoined to N, as in figure 1. The lexically shared instantiation is shown with arrows from both N and Âdj pointing to the word *gèt+ró’,* indicating that it instantiates both these terminal nodes.

4 Why have two syntactic nodes?

4.1 Scope of adjectives

Although the adjective is part of the same phonological word as the preceding noun, it has scope properties that suggest syntactic independence. Consider the following examples:

20) R-yúlàz=à' càfé cùn téy+nààxh.
hab-like=1s coffee and tea+sweet
'I like sweet [tea and coffee].' (both are sweet)
'I like [sweet tea] and [coffee].' (only tea is sweet)

21) Ù-dàw=à' gàmòn cùn dzìtbéédy+nàxíí.
com-eat=1sg ham and egg+salty
'I ate salty [eggs and ham].' (both are salty)
'I ate [salty eggs] and [ham].' (only eggs are salty)

These sentences have two readings – one in which the adjective takes scope only over the immediately preceding noun, and one in which it takes scope over both nouns. The wide scope reading suggests a c-structure like that shown in Figure 2:

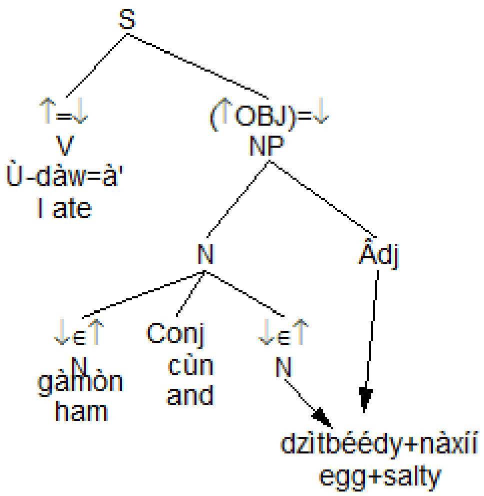


Figure 2 Lexical sharing and coordination

If N+Âdj compounds were purely lexical, we would not expect such scopal properties. Compare English sentences like the following, where *black* is unambiguous in scope when in a compound (a), but ambiguous as an attributive adjective (b):

22) a.) I saw blackbirds and squirrels.
b.) I saw black birds and squirrels.

Thus SDZ N+Âdj combinations show behavior like independent attributive adjectives in English, and not like the adjective portion of an English compound.

4.2 Adjectives with complements

A second argument for the c-structure representation of the adjectives is found in the behavior of adjectives with complements. Though the combination of N+Âdj into a single word is obligatory with a single-word adjective, the facts change if the Adj heads a phrase.

One case in which Adj heads AdjP is in the comparative:

23) Ngìw góórrd=rù quèy nàà’ b-èèny gáàn.
man fat=more than me com-do win
‘*The man fatter than me* won.’

24) R-yùlààz=à’ sóóp nàxìì=rù quèy bè’l.
hab-like-1s soup salty=more than meat
‘I like the soup that is saltier than the meat.’

In these cases, the N and Adj no longer form a single word, as shown by both the phonological evidence and the clitic placement tests.

Looking first at the phonological evidence, we see that in the following example, *bèjl* ‘flame’ has breathy phonation in isolation. This reduces to plain phonation when followed by a non-projecting Âdj:

25) bèjl ‘flame’
bèl+ró’ ‘big flame’

However, if the Adj is necessarily projecting, then the phonation change does not occur:

26) tòyby bèjl ró’=rù quèy stòyby=nì
a fire big=more than other=3i
‘a fire bigger than the other one’

This shows that the N and Adj do not form the ordinary compound in this case.

Similarly, clitic placement tests also show that the N and the following Adj are now different words, and that a clitic may be placed between them:

27) Éèy, ngìw=cà góórrd=rù quèy nàà’ b-èèyny gáàn.
yes man=aff fat=more than me com-do win
‘Yes, *the man fatter than me* won.’

28) Éèy, sóóp=cà nàxìì=rù quèy bè’l r-yùlààz=à’
yes soup=aff salty=more than meat hab-like-1sg
‘Yes, I like *the soup that is saltier than the meat.*’

We can contrast these sentences with those where the adjective has no complement. In such cases, the N+Âdj combination is still a single word, which cannot be penetrated by a clitic:

29) a. Éèy, ngìw góórrd=cà b-èèyny gáàn.
yes man fat com-do win
‘Yes, the fat man won.’

b. *Éèy, ngìw=cà góórrd b-èèyny gáàn.
yes man=aff fat com-do win

So the correct tree for the N followed by a non-projecting attributive adjective is as in Figure 3:

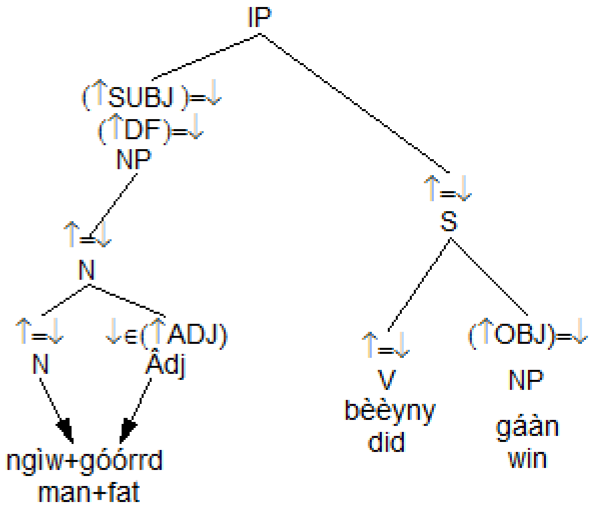


Figure 3 Lexical sharing with a non-projecting adjective

However, when the adjective has a complement, the tree is instead as in Figure 4:

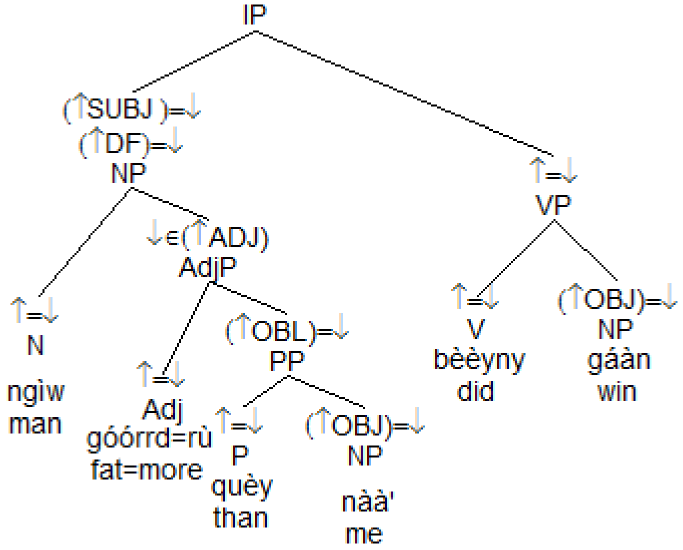


Figure 4 No lexical sharing with a projecting adjective.

These facts show us that the lexical rule combining noun and adjective only applies to non-projecting adjectives. Thus a coherent lexical-sharing analysis needs to make use of the non-projecting word hypothesis.

One additional consideration. Since lexical sharing is obligatory for a non-projecting adjective, we need to rule out a tree like the following, where the Adj projects a AdjP, rather than appearing as a non-projecting word, as in Figure 5:

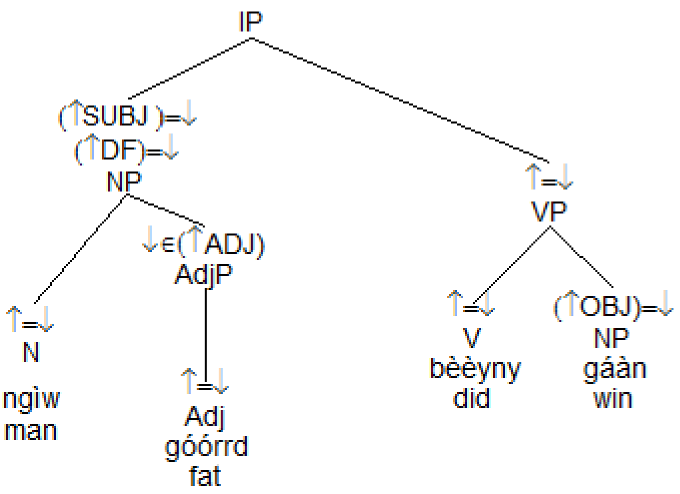


Figure 5 Violation of Economy of Expression

Following Toivonen (2003), I will assume that a tree of this sort is suboptimal relative to the tree with a non-projecting Âdj, due to Economy of Expression (Bresnan 2001), since it contains an additional phrasal node (AdjP).

5 Predicative and and attributive adjectives

5.1 Morphological background[9]

In the examples (4) and (5) above (repeated below), the adjective *ró'* serves as both a predicative and attributive adjective with no change.

30) Ù-dàw Juáàny gèt+ró'
com-eat Juan tortilla+big
'Juan ate the big tortilla.'

31) Ró' gèèt.
big tortilla
'The tortilla is big.'

Adjectives of this type, which are identical in their predictive and attributive forms, I will label Group A (Invariable) adjectives. Some other examples of native Zapotec adjectives from Group A:

[9] The account given here of morphologically defined subgroups of predicative and attributive adjectives is influenced by the treatments of similar phenomena in two related Zapotec languages – Mitla Zapotec (Briggs 1961:67-70; Stubblefield and Stubblefield 1991:208-210) and San Lucas Quiaviní Zapotec (Munro 2002; Munro and Lopez 1999; Lee 1999; Galant 1998).

32)	ldàà’	‘loose, slack’
	mèw	‘dirty’
	dè’	‘narrow’
	chííny	‘skinny’
	ldíí	‘straight; upright’
	cúújxh	‘squint-eyed’
	nàjxh	‘sweet’
	bí’ch	‘small’
	nnà’á	‘heavy’
	gòòp	‘mute’
	méèxh	‘blond’
	gííby	‘stingy’
	lèèt	‘empty’

It appears that all adjectives borrowed from Spanish also go into Group A:

32)	máàl	‘bad’	(<Span. *malo*)
	còchíìn	‘filthy, disgusting’	(<Span. *cochino*)
	lííèst	‘ready, intelligent’	(<Span. *listo*)
	tràbáàgw	‘difficult’	(<Span *trabajoso*)
	plòòj	‘lazy’	(<Span. *flojo*)
	súújl	‘blue’	(<Span. *azul*)

Group A (Invariable) appears to be the open, productive class of adjectives in SDZ.

However, many adjectives show different forms in their predicative and attributive uses. Adjectives which show a morphological change between their predicative and attributive uses, I will label Group B (Variable) adjectives. The most frequent change is the addition of *na-*:

33) a. **Ná-dxè’ch**=dù’úxh ngìw=gà
pred-irritable=very man=that
‘That man is very irritable.’

b. Ngìw+**dxè’ch**=dù’úxh Juáàny.
man+irritable=very Juan
‘Juan is a very irritable man.’

Here are some examples of adjectives from Group B:

34)	*Attributive*	*Predicative*	*Gloss*
	dxè’ch	nà-dxè’ch	‘quick-tempered; irritable’
	yààn	nà-yààn	‘spicy’
	bííèz	nà-bííèz	‘dry’

A few adjectives appear to contain a ‘frozen’ *n-* or *na-* prefix, which appears in both predicative and attributive forms in SDZ. They are thus synchronically Group A (invariable) adjectives in SDZ. Adjectives in this group include *ngààs* ‘black’ and *ngàjts* ‘yellow’.

35) a.) Ngààs bè'cw.
black dog
'The dog is black.'

b.) bè'cw+ngààs
dog+black
'black dog'

c.) *bè'cw+gààs

Comparison with nearby Zapotec languages (Mitla Zapotec, SLQZ) shows that many of these adjectives are Group B (Variable) in those languages. Thus the diachronic change is that some adjectives in SDZ have moved from the lexically restricted Group B (Variable) into the open class Group A (Invariable).

There are also a few adjectives that seem to still be in the process of changing from Group B to Group A. For these adjectives, the predicate must have the na- prefix, but this prefix is optional in the attributive:

36)

Predicative	*Attributive*	*Gloss*
nàldàj	ldàj ~ nàldàj	'bitter'

37) a. Nà-ldáj sèrbèjs.
pred-bitter beer
'The beer is bitter.'

b. *Ldàj sèrbèjs.
bitter beer

38) Ííty r-yùlááz=tì=à' sèrbèjs+(nà-)ldàj.
not hab-like=neg=1s beer+(pred-)bitter
'I don't like bitter beer.'

The reverse pattern is also found for a few adjectives:

39)

Predicative	*Attributive*	*Gloss*
xú'ny ~ nàxú'ny	xú'ny	'wrinkled'

40) a. (Nà-)xú'ny x-cùtòòny=á'.
(pred-)wrinkled p-shirt=1s
'My shirt is wrinkled.'

b. R-àp=á' x-cùtòòny+(*nà-)xú'ny
hab-have=1s p-shirt+(pred-)wrinkled
'I have a wrinkled shirt

5.2 The syntax of predicative adjectives

The examples given below show Group A (Invariable) and Group B (Variable) predicative adjectives acting as the sole predicate of a sentence:

41) Nà-ldáj sèrbẻjs.
pred-bitter beer
'The beer is bitter.'

42) Péncw yààg.
bent tree
'The tree is bent.'

Adjectival predicates show a different syntax than most verbal predicates. I argued in Broadwell (2002, 2005) that the clausal syntax of San Dionicio Ocotepec Zapotec has two X^0 positions for verbal predicates, and this will be important for understanding the syntax of predicate adjectives. Let me briefly review that argument before returning to adjectives.

5.2.1 The definite future

SDZ, like other Valley Zapotec languages, has two different aspects which are translated into the future in English/Spanish. The definite future is marked with *s-* or *z-*; the potential has a number of allomorphs, the most common of which is *g-*:

43) S-àw báád bèld yù'ù.
def-eat duck snake earth
'The duck is going to eat a worm.'

44) G-âw báád bèld yù'ù.
pot-eat duck snake earth
'The duck is going to eat a worm.'

The difference between these two is subtle and Lee (1999) has done the most careful investigation of the semantics. The names of the definite future reflects its use with future events that are more certain and also perhaps closer in time. The potential is appropriate with a wider range of future events and shows less of a speaker commitment to the certainty or proximity of the event.

Despite the close semantics, verbs in the potential and future aspects show strikingly different syntactic properties, and most of these properties follow from the assumption that a verb in the definite instantiates both the Infl and V positions, while a verb in the potential remains in the ordinary V position.[10] Evidence for this is discussed in the following sections.

5.2.2 Lack of internal topic/focus in the definite future

SDZ has a preverbal position for elements which bear a discourse function such as TOPIC or FOCUS.[11]

[10] My analysis here is slightly altered from that in Broadwell (2005), where I did not employ lexical sharing. My analysis is also clearly influenced by Lee (1999), in which SLQZ verbs in the definite future move into [Spec, FocP].

[11] In Broadwell (2002), I call this the internal prominence (i-prom) position, to distinguish it from a CP-adjoined position for external topics (e-topic). In that paper, I also give more detailed argumentation

This preverbal position is not possible when the verb is in the definite future aspect (*s-/z-*). In contrast, this position is possible when the verb is in the potential aspect.

45) S-àw báád bèld yù'ù.
def-eat duck snake earth
'The duck is going to eat a worm.'

*Báád s-àw bèld yù'ù. *TOP/FOC definite future
duck def-eat snake earth

46) G-âw báád bèld yù'ù.
pot-eat duck snake earth
'The duck is going to eat a worm.'

✓Báád g-âw bèld yù'ù. ✓TOP/FOC potential
duck pot-eat snake earth
'*The duck* is going to eat a worm.'

5.2.3 Manner adverbs and the definite future

Manner adverbs ($\text{Adv}_{\text{Manner}}$) must not precede a verb in the definite future, though these adverbs may precede a verb in other aspects.

47) a.) Dìáp g-ú'ld Màrìì. ✓$\text{Adv}_{\text{Manner}}$ Potential
strongly pot-sing Maria
'Maria will sing strongly/loudly.'

b.) *Dìáp s-ù'ld Màrìì. *$\text{Adv}_{\text{Manner}}$ Definite Future
strongly def-sing Maria

c.) S-ù'ld Màrìì dìàp.
def-sing Maria strongly

d.) G-ú'ld Màrìì dìàp.
def-sing Maria strongly

Pursuing this latter approach, the examples above will have the following (simplified) representations:[12]

for the multiple discourse roles of elements that occupy the i-prom position.

[12] For expository purposes, the trees shown in this figure show potential positions for focused and adverbial positions in parentheses. The excluded positions in the definite future are shown with strike-out to emphasize their unavailability.

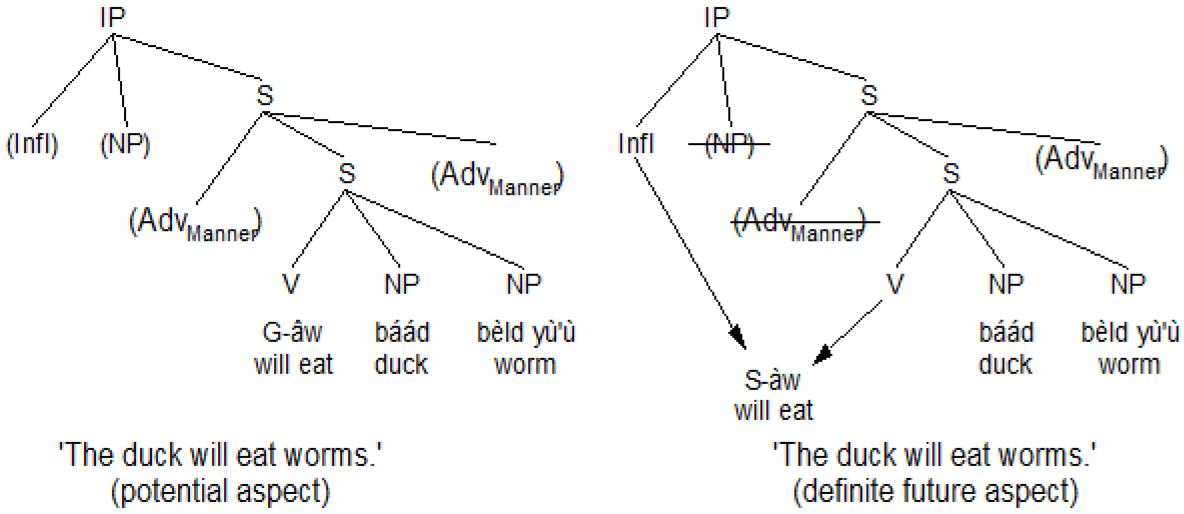


Figure 6 The syntax of potential and definite future aspects compared

These trees show that when the verb is in the definite future aspect, it instantiates both the V and Infl positions. It thus precludes words in the $\text{Adv}_{\text{Manner}}$ (manner adverb) position and the [Spec, IP] (internal TOP/FOC) position. This is an example of what Wescoat (2002:24-30) calls intermediate constituent suppression, whereby normally available phrase-structure positions become unavailable in cases of lexical sharing.

5.2.4 Predicate adjectives and phrase structure

Predicate adjectives show a syntax very similar to that of verbs in the definite future aspect. In particular, the internal TOP/FOC position is unavailable:[13]

48) a.) Ngáás gìich+ìicy=à'
black hair+head=1s
'My hair is black.'

b.) *Gìich+ìicy=à' ngáás.
hair+head=1s black.

We can capture the similarity between verbs in the definite future and predicate adjectives by writing a lexical rules of the following sort:

49) /Φ / ← [POS V] ⇨ /s-Φ/ ← [POS V+Infl]
[VCLASS 1] [ASP DEF-FUT]

[13] My language consultant rejects manner adverbs with adjectival predicates, regardless of their position. This is presumably because of semantic incompatibly. For this reason, it is not possible to test the availability of the initial Manner Adverb position.

50) /Φ / ← [POS V] ⇨ /z-Φ/ ← [POS V+Infl]
[VCLASS 2] [ASP DEF-FUT]

These rules say that for a verb instantiated as /Φ/, there is also a form /z-Φ/ or /s-Φ/ which realizes the definite future aspect and that such a form instantiates both the V and Infl nodes. (The difference between the two morphological classes is shown by a VCLASS feature.)

For predicate adjectives, we want similar rules, along the following lines:

51) / Φ / ← [POS Adj] ⇨ /na-Φ/ ← [POS V+Infl]
[ADJCLASS B]

52) / Φ / ← [POS Adj] ⇨ /Φ/ ← [POS V+Infl]
[ADJCLASS A]

These two rules take an Adj and change its part of speech category to the portmanteau V+Infl category. The first rule prefixes /na-/ to adjectives of Class B and the second is a phonologically null derivation for adjectives of Class A.

5.2.5 Lexical entries for irregular adjectives

The adjectives which fall outside the main patterns will be listed in the lexicon. Some, like ‘wrinkled’ (predicative *xú’ny ~ nàxú’ny*; attributive *xú’ny*) can be listed as variable as to ADJCLASS. Others like ‘bitter’ (predicative *nàldàj*; attributive *ldàj ~ nàldàj*) seem to have alternate underlying forms. Lexical entries for these adjectives would be along the following lines:

53) xú’ny ← [POS ADJ], [ADJCLASS A|B], [PRED ‘wrinkled <SUBJ>’]

nàldàj ← [POS ADJ], [ADJCLASS A], [PRED ‘bitter <SUBJ>’]
or ldàj ← [POS ADJ], [ADJCLASS B], [PRED ‘bitter <SUBJ>’]

5.2.6 Inflection of predicate adjectives and the use of copulas

I have called the part of speech category for the derived predicate adjectives V+Infl because that is the position that they seem to occupy in the syntax. Still it is not the case that predicate adjectives are identical to verbs in terms of their inflectional possibilities.

Ordinary verbs generally show inflection for six aspects. Five of these are shown below with their most frequent allomorphs:[14]

54)

completive	(g)u-/bi-
continuative	cá(y)-
potential	ì-/gú-
habitual	r-/rr-
definite future	s-/z-

[14] There is also a prefix known as negative aspect, which shows up after certain negative predicates and adverbs. In the interests of space, I omit discussion of it here.

The completive, continuative, habitual, and potential aspect markers are shown for the following fairly regular verb *-ù'ld* 'to sing':

55) bì-'ld=bí 'S/he sang.'
com-sing=3

cáy-ù'ld=bí 'S/he is singing.'
con-sing=3

r-ù'ld=bí 'S/he sings.'
hab-sing=3

gú-'ld=bí 'S/he will sing.'
pot-sing=3

s-ú'ld=bí 'S/he will sing.'
def-sing=3

Predicate adjectives do not show this range of inflection. In SDZ, group B adjectives show the *nà-* prefix in what is called neutral aspect. For adjectives, this is the most normal translation of present tense sentences in English or Spanish.[15]

If the clause is to be interpreted in some other aspect, such as completive or potential, then an overt copula is necessary, and the adjective is adjoined to it as a non-projecting word:

56) Gùùc+sàlààd x-cómììd=à'.
com:be+salty p-food=1s
'My food was salty.'

57) Gáác+sàlààd x-cómììd=à'
pot:be+salty p-food=1s
'My food will be salty.'

58) Cáyààc+sàlààd x-cómììd=à'
con:be+salty p-food=1s
'My food is becoming salty.'

The pattern of non-verbal predicates which require an overt copula in non-present contexts is fairly common crosslinguistically.

We can capture this restriction by including an aspect specification in the lexical rule that creates the predicative adjectives:

[15] An aspect labelled 'neutral' also appears with verbs, but is restricted to a few semantic categories – primarily verbs of position and speech. See Munro (2002) for a discussion of the relationship between the adjectival and verbal morphological categories.

59) / Φ / ← [POS Adj] ⇨ /na-Φ/ ← [POS V+Infl]
[ADJCLASS B] [ASP NEUTRAL]

60) / Φ / ← [POS Adj] ⇨ /Φ/ ← [POS V+Infl]
[ADJCLASS A] [ASP NEUTRAL]

Because the predicative adjectives that result from this rule already have an aspectual value, they are not eligible to undergo additional aspect morphology. So changing their part of speech to V+Infl does not imply that they are eligible for the full range of verbal morphology.

The combination of copula and non-projecting adjective counts as a single word by the clitic placement tests:

61) a. Gùùc+sàlááàd=cà x-cómììd=à'.
com:be+salty=aff p-food=1s
'Yes, my food was salty.'

b. *Gùùc=cà sàlááàd x-cómììd=à'.
com:be=aff salty p-food=1s
'Yes, my food was salty.'

However, these combinations of copula and adjective do not preclude a preceding topic:

62) a. Gùùc+ngáás gììch+ììcy=à'.
com:be+black hair+head=1s
'My hair was black.'

b. Gììch+ììcy=à' gùùc+ngáás.
hair+head=1s com:be+black

Contrast this last example with the same pair in neutral/unmarked aspect (repeated from above):

63) a.) Ngáás gììch+ììcy=à'
black hair+head=1s
'My hair is black.'

b.) *Gììch+ììcy=à' ngáás.
hair+head=1s black.

We thus need additional lexical rules which produce the combination of copula and adjective. However, these rules need to yield a V, rather than a V+Infl:[16]

[16] I have let these morphological rules directly spell out the phonological realizations of the different aspectual forms of the Copula+Âdj combination. A more elegant morphological rule could use a rule of referral to point to the forms of the copula already present in the lexicon.

64) / Φ / ← [POS Adj] ⇨ /gùùc-Φ/ ← [POS V]
[ASP COM]

/ Φ / ← [POS Adj] ⇨ /gáác-Φ/ ← [POS V]
[ASP POT]

Note the interesting contrast between these rules which yield a.) a Copula+Âdj with the part of speech V and b.) the rules that make adjectives predicative, which yields a word of the V+Infl type. The latter type will entail lexical sharing and intermediate constituent suppression, while the former will not.

Unlike the N+Adj combination, there is no good evidence that the Copula+Âdj combination needs to be represented at c-structure. Because only the copula combines via this rule, it is not possible to construct examples that show a scope ambiguity comparable to that seen with nouns and adjectives.

However, it is possible to have sentences where the adjectival portion of the Copula+Âdj compound has a complement:

65) Gùùc+ró'=rú gèèt quèy gètgù'.
com:cop+big=more tortilla than tamale
'The tortilla was bigger than the tamale.'

However, it is impossible to have an order in which the adjective forms a constituent with its complement:

66) *Gùùc+ró'=rú quèy gètgù' gèèt.
com:cop+big=more than tamale tortilla
'The tortilla was bigger than the tamale.'

67) *Gùùc gèèt ró'=rú quèy gètgù'.
com:cop tortilla big=more than tamale
'The tortilla was bigger than the tamale.'

Thus the Copula+Âdj combination is unlike the N+Âdj combination; the Copula+Âdj is always a single word, while N and Âdj are not.

Thus we see evidence of lexical sharing with the attributive adjectives and with predicative adjectives in neutral aspect as well. Predicate adjectives compounded with a copula, however, act like simple verbs in syntax, and show no evidence of lexical sharing.

This is a complex set of facts, but a carefully articulated inventory of lexical rules, lexical sharing, and non-projecting words allows a satisfying explanation of the syntax of Zapotec adjectives

6 Conclusion

Zapotec attributive adjectives are persuasive examples of non-projecting words which form a single phonological word with the words to which they adjoin. An LFG analysis of such constructions in terms of non-projecting words and lexical sharing successfully captures the fact that the Zapotec construction acts as two words syntactially, but a single word in phonology. This analysis relies on the distinction between projecting and non-projecting words introduced by Sadler and Arnold (1994), Sadler (2000) and Toivonen (2001, 2003). It also lends support to the lexical sharing hypothesis of Wescoat (2002) in which a single phonological word may instantiate more than one than one syntactic terminal.[17]

[17] See also Kim, Sells, and Wescoat (2004) for an HPSG analysis of Korean using lexical sharing.

References

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

Briggs, Elinor. 1961. *Mitla Zapotec grammar*. Mexico City: Instituto Lingüístico de Verano.

Broadwell, George A. 2000a. On the phonological conditioning of clitic placement in Zapotec. Workshop on Structure and Constituency of the Languages of the Americas 5. Toronto, Ontario, March 24, 2000.

Broadwell, George A. 2000b. Coordination, clitic placement, and prosody in Zapotec. Berkeley Linguistic Society 26.

Broadwell, George A. 2002. Preverbal positions and phrase boundaries in Zapotec. Paper presented at the annual meeting of the Linguistic Society of America.

Broadwell, George A. 2005. It ain't necessarily S(V)O: Two kinds of VSO languages. In Miriam Butt and Tracy Holloway King (eds), *Proceedings of the LFG 05 Conference*. http://csli-publications.stanford.edu/LFG/10/lfg05.html. Stanford, CSLI Publications.

Galant, Michael. 1998. *Comparative constructions in Spanish and San Lucas Quiaviní Zapotec*. Ph. D. thesis, UCLA.

Kim, Jong-bok, Peter Sells, and Michael Wescoat 2004. Korean Copular Constructions: A Lexical Sharing Approach. 2004. To appear in M. Hudson, S.-A. Jun and P. Sells (eds.) *Proceedings of the 13th Japanese/Korean Linguistics Conference*. Stanford, CSLI Publications.

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

Munro, Pamela and Felipe Lopez. 1999. *Dicyonaary X:tèe'n Dìi'zh Sah Sann Lu'uc: San Lucas Quiaviní Zapotec Dictionary: Diccionario Zapoteco de San Lucas Quiaviní*. Chicano Studies Research Center, UCLA.

Munro, Pamela. 2002. Aspects of stativity in Zapotec. Paper presented at the annual meeting of the Linguistic Association of the Southwest (LASSO).

Pickett, Velma. 1997. When is a Phrase a Word? Problems in Compound Analysis in Isthmus Zapotec. Presented at the UCLA American Indian Linguistics Seminar.

Sadler, Louisa. and Douglas Arnold. 1994. Prenominal adjectives and the phrasal/lexical distinction. *Journal of Linguistics* 30:187-226.

Sadler, Louisa. 2000. Noun phrase structure in Welsh. In Miriam Butt and Tracy Holloway King (editors), *Argument Realization*. Stanford, CA: CSLI Publications.

Stubblefield, Morris and Carol Miller de Stubblefield. 1991. *Diccionario zapotec de Mitla, Oaxaca*. Mexico City: Instituto Lingüístico de Verano

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

Toivonen, Ida. 2003. *Non-Projecting Words. A Case Study of Swedish Particles*. Dordrecht:Kluwer.

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

Wescoat, Michael T. 2007. Preposition-determiner Contractions: an Analysis in Optimality-theoretic Lexical-functional Grammar with Lexical Sharing. In Miriam Butt and Tracy Holloway King (eds), *Proceedings of the LFG 07 Conference*. http://csli-publications.stanford.edu/LFG. Stanford, CSLI Publications.

URDU CORRELATIVES: THEORETICAL AND IMPLEMENTATIONAL ISSUES

Miriam Butt, Tracy Holloway King and Sebastian Roth
Universität Konstanz, PARC, Universität Konstanz

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

The inclusion of South Asian languages in multilingual grammar development projects that were initially based on European languages has resulted in a number of interesting extensions to those projects. Butt and King (2002) report on the inclusion of Urdu in the Parallel Grammar Project (ParGram; Butt et al. (1999, 2002)) with respect to case and complex predicates. In this paper, we focus on a possible integration of *correlatives* into the computational analysis. Hindi/Urdu correlative clauses have received various analyses in the past that treat them as distinct from other strategies of relativization. We follow Bhatt (1997), who argues that the syntax and semantics of single-headed correlative clauses strongly resemble those of free relative clauses in European languages, but we analyze these as specifiers of a DP, rather than as adjuncts.

1 Introduction

This paper aims at introducing the discussion of so-called *correlative* constructions, a special strategy of relativization commonly found in a number of Indo-European, especially Indo-Aryan, languages, into LFG analyses. In particular, we look at correlatives from within the context of creating broad-coverage grammars as part of the Parallel Grammar project (ParGram; Butt et al. 1999, 2002). Among the aims of the ParGram project is to test the LFG formalism for its universality and coverage limitations and to see how far crosslinguistic parallelism at f-structure can be maintained. Where possible, the analyses produced by the grammars for similar sentences in each language are parallel. The standardization of the analyses has the computational advantage that the grammars can be used in similar applications and it can simplify cross-language applications such as machine translation. Parallelism, however, is not maintained at the cost of misrepresenting the language. Given this context, the phenomenon of correlatives is particularly interesting as it is a puzzling construction from the perspective of most European languages.

The pattern in correlatives is that a demonstrative pronoun, which also functions as determiner in Urdu/Hindi[1], in this case *vo*, always occurs in correlation with a relative pronoun, *jo*. In fact, the language employs a series of such pronouns: e.g., *jɪs/ʊs* ‘which/that’ (oblique), *jahã/vahã* ‘where/there’ (distal), *jɪdər/ɪdər* ‘where/there’ (proximal).

We base our analysis in large part on Srivastav (1991), who argues convincingly that correlative constructions in Hindi fall into two classes: one in which the relative *jo* clause appears to the right of the *vo* head noun ((1a,b)) and one in which the *jo* clause precedes the *vo* noun ((1c)). Srivastav, whose analysis is primarily semantic, identifies the former as straightforward relative clauses, the latter as true correlatives.

[1] Urdu and Hindi are structurally essentially identical. For the sake of brevity, we only refer to Urdu, but all observations in this paper apply to Hindi as well.

(1) a. [**vo** lɑṛki [**jo** kʰɑṛ-i hɛ]
that girl.F.Sg.Nom which stand-Perf.F.Sg be.Pres.3.Sg

hɑs rɑh-i hɛ]
laugh stay-Perf.F.Sg be.Pres.3.Sg
'Who is standing, that girl is laughing.' (Srivastav 1991:642)

b. [**vo** lɑṛki hɑs rɑh-i hɛ
that girl.F.Sg.Nom laugh stay-Perf.F.Sg be.Pres.3.Sg

[**jo** kʰɑṛ-i hɛ]]
which stand-Perf.F.Sg be.Pres.3.Sg
'That girl is laughing, who is standing.' (Srivastav 1991:642)

c. [**jo** kʰɑṛ-i hɛ]
which stand-Perf.F.Sg be.Pres.3.Sg

[**vo** lɑṛki hɑs rɑh-i hɛ]
that girl.F.Sg.Nom laugh stay-Perf.F.Sg be.Pres.3.Sg
'Who is standing, that girl is laughing.' (Srivastav 1991:642)

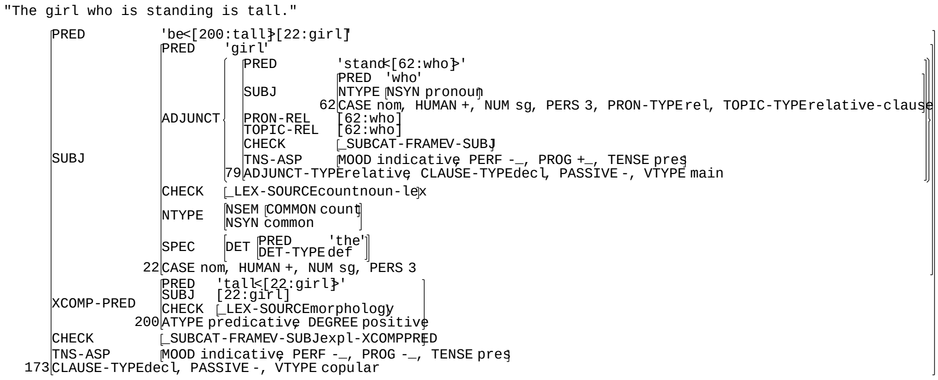
With respect to true correlatives as in (1c), no standard LFG analysis exists to date. Here, we depart from Srivastav's analysis and instead follow Bhatt (1997), who argues that Hindi correlatives must be understood as the equivalent of free relative clauses in European languages. Unlike Bhatt, however, we do not treat the correlative as an adjunct, but as a specifier of DP. The relevant evidence comes from the interaction with quantifiers and demonstratives, topicalization and the behavior of multi-head correlatives. Our analysis therefore builds on existing argumentation from a primarily semantic perspective (Srivastav 1991) and from within Minimalism (Bhatt 1997), but ultimately differs in the syntactic treatment of correlatives.

2 Standard Analyses of Relative Clauses

Linguistic typology (e.g., Lehmann 1984) generally distinguishes three classes of relative clauses: free and bound relative clauses, with the latter divided into restrictive and non-restrictive relative clauses. Bound relative clauses appear either adjacent to the phrase that they modify or extraposed at the end of the sentence. Within the ParGram project, bound relative clauses are analyzed as NP-modifying adjuncts of the c-structure category *CPrel* (Butt et al. 1999): The lexical requirements of the embedded finite verb must be fulfilled, meaning that arguments corresponding to the verb's subcategorization frame must be provided. A sample ParGram f-structure analysis of an English simple (non-extraposed) bound relative is shown in (2).

The relative pronoun, which must be an argument of the relative clause's predicate or the argument of a prepositional adjunct modifying the relative clause's predicate, is encoded as the TOPIC-REL of the relative clause if it appears in preposed (topicalized) position, such as in English, German or French relatives. The functional structure projected by the relative clause is encoded as an adjunct (with ‘ADJUNCT-TYPE relative’) of the relative head.[2] Extraposed bound relative clauses are adjoined at f-structure to a single NP via functional uncertainty.

(2)



```
"The girl who is standing is tall."
     PRED          'be<[200:tall]>[22:girl]'
                   PRED    'girl'
                              PRED        'stand<[62:who]>'
                                          PRED  'who'
                              SUBJ        NTYPE NSYN pronoun
                           62 CASE nom, HUMAN +, NUM sg, PERS 3, PRON-TYPE rel, TOPIC-TYPE relative-clause
                   ADJUNCT    PRON-REL    [62:who]
                              TOPIC-REL   [62:who]
                              CHECK       _SUBCAT-FRAME V-SUBJ
     SUBJ                     TNS-ASP     MOOD indicative, PERF -_, PROG +_, TENSE pres
                           79 ADJUNCT-TYPE relative, CLAUSE-TYPE decl, PASSIVE -, VTYPE main
                   CHECK   _LEX-SOURCE countnoun-lex
                   NTYPE   NSEM COMMON count
                           NSYN common
                   SPEC    DET PRED      'the'
                               DET-TYPE def
                22 CASE nom, HUMAN +, NUM sg, PERS 3
                   PRED   'tall<[22:girl]>'
                   SUBJ   [22:girl]
     XCOMP-PRED    CHECK  _LEX-SOURCE morphology
               200 ATYPE predicative, DEGREE positive
     CHECK         _SUBCAT-FRAME V-SUBJ expl-XCOMPPRED
     TNS-ASP       MOOD indicative, PERF -_, PROG -_, TENSE pres
  173 CLAUSE-TYPE decl, PASSIVE -, VTYPE copular
```

Non-restrictive relative clauses, such as (3) receive essentially the same structural analysis as restrictives, but are marked with an additional feature ‘RESTR −’ in order to flag them for a different semantic interpretation.

(3) Mary, who is standing there, is tall.

Since non-restrictive relative clauses, like appositives, perform a different illocutionary act than the proposition signified by the matrix clause, they do not have any truth-conditional value regarding the interpretation of their relative head, whereas restrictive relative clauses and their heads form a semantic unit via set intersection. Note that it is not always trivial to classify a relative clause as restrictive or not. In English, non-restrictive relative clauses are often marked by distinct punctuation, changes in intonation or special lexical items (e.g., *incidentally* or *by the way*).

Within ParGram, free relatives are analyzed quite differently from bound relative clauses. In English, free relatives have the distribution of an NP and are thus treated as such, whereas in German they cannot, like other finite clauses, appear clause internally and are thus treated as a special category *CPfreerel*. The relative pronoun in both languages takes the double function of relative clause head (i.e. the relative clause predicate is attached at f-structure as an ADJUNCT) as well as that of an argument of the matrix clause predicate. The existence of an empty argument of the matrix verb is deduced at f-structure from information provided either by the c-structure construction, as in German, or by the lexical entry of the

[2] The term ‘relative head’ refers to the noun phrase that is modified by the relative clause. We separate externally-headed clauses from internally-headed clauses. Both external and internal heads are denoted as ‘relative heads’ in our paper.

free relative pronoun, such as English *whoever* (Butt et al. 1999:96).[3] This allows the relative pronoun to take the grammatical function of the missing matrix argument and project the feature 'PRON-TYPE free' to its f-structure. (4) provides an example, again for English.

(4)

```
"Whoever is driving the tractor is laughing."

    PRED     'laugh<[24:whoever]>'
             PRED     'whoever'
                         PRED      'drive<[39-SUBJ:null_pro], [133:tractor]>'
                         SUBJ      PRED 'null_pro'
                                   CASE nom, NUM sg, PERS 3, PRON-TYPE rel
                                   PRED  'tractor'
                                   CHECK [_LEX-SOURCE countnoun-lex]
                                   NTYPE  NSEM [COMMON count]
                         OBJ              NSYN common
    SUBJ     ADJUNCT
                                   SPEC   DET  PRED      'the'
                                               DET-TYPE  def
                               133 CASE obl, NUM sg, PERS 3
                         PRON-REL  [39-SUBJ:null_pro]
                         TOPIC-REL [39-SUBJ:null_pro]
                         CHECK     [_SUBCAT-FRAME V-SUBJ-OBJ]
                         TNS-ASP   [MOOD indicative, PERF -_, PROG +_, TENSE pres]
                       39 CLAUSE-TYPE decl, PASSIVE -, VTYPE main
             NTYPE    [NSYN pronoun]
          24 CASE nom, HUMAN +, NUM sg, PERS 3, PRON-TYPE free
    CHECK    [_SUBCAT-FRAME V-SUBJ]
    TNS-ASP  [MOOD indicative, PERF -_, PROG +_, TENSE pres]
174 CLAUSE-TYPE decl, PASSIVE -, VTYPE main
```

Note that free relatives in English as well as German are semantically ambiguous between singular definite and generic readings. The f-structure encoding of the free relative does not handle this semantic ambiguity although the f-structure provides all necessary information for further semantic processing: for example, tense/aspect information that disambiguates the reference. A generic reading is unavailable if the verb does not license it, which would be the case with progressive aspect, as for example in (5), in which the *-ever* does not lead to a free choice generic reading, but simply implies the uncertainty or irrelevance of the subject's identity. It is assumed that such constraints are handled within a separate semantic projection. A semantic consideration, however, that does have syntactic import is that free relatives do not allow stacking of further restricting or appositive relative clauses, as shown in (6).

(5) Whoever is driving the tractor is laughing.

(6) *Whoever drives the tractor, who is happy, is laughing.

Furthermore, no non-restrictive interpretation of the free relative clause itself is possible. These restrictions are significant within the context of our paper as they apply to Urdu correlative clauses as well (section 4.1).

[3] Note that not only *-ever* pronouns can function as free relative pronouns in English: (i) is also an example for a typical free relative.

i. I eat what you eat.

3 Relativization in Urdu

Urdu, the national language of Pakistan and an official language of India, is an Indo-Aryan language spoken by around 60-80 million native speakers today. It is an SOV-language with relatively free word order, a split-ergative case system and correlative clauses. Since its grammar is largely identical to that of Hindi and large portions of vocabulary are shared, Hindi-Urdu is commonly regarded by linguists as a single language, in contrast to their constitutional status.

Urdu, like Sanskrit, preserves the old Indo-European distinction between relative (Urdu j-class), interrogative (k-class), proximal demonstrative (y-class) and distal demonstrative (v-class) pronouns. It furthermore retains a remnant of the correlative clausal structure that, in Sanskrit, was used to express all kinds of clausal relations, such as relatives, conditionals or sentential complementation in a paratactic manner. Although it has seen some modification and appears in a more constrained distribution than the Sanskrit correlative (for a comparison, see Davison 2006), the Urdu correlative nevertheless retains some of the properties that separate it from the English-type postposed relative clauses, which also exist in Urdu.

Modern Urdu left-adjoined relatives as in (1c), repeated in (7), are generally called correlative clauses after their Sanskrit ancestors and are found in a number of Indo-Aryan languages, such as Bengali, Sindhi, Punjabi, Marathi, Gujarati and Urdu, but also in Hittite, Latin, Ancient Greek, Medieval Russian, and Old English as well as modern Hungarian, Bulgarian and Serbo-Croatian (Bhatt 2003).

(7) [**jo** k^{h}ɑṛ-i hɛ]
which stand-Perf.F.Sg be.Pres.3.Sg

[**vo** lɑṛki hɑs rɑh-i hɛ]
that girl.F.Sg.Nom laugh stay-Perf.F.Sg be.Pres.3.Sg
'Who is standing, that girl is laughing.' (Srivastav 1991:642)

As already mentioned in the introduction, whereas the Urdu embedded and right-adjoined relatives manifest typical properties of restrictive relative clauses and are covered by the same analyses proposed for these structures in the other ParGram languages, previous analyses provide evidence that left-adjoined relatives form a distinct class, rather than being an instance of left-extraposition of an NP modifier. In the generative literature, there has been much discussion of whether so-called preposed, embedded and postposed relative clauses derive from the same underlying structure (e.g., Subbarao 1984) or are base-generated in their respective positions (McCawley 2004). We follow Srivastav (1991) in considering embedded and postposed relative clauses of the *vo*-*jo* pattern as being structurally and functionally identical, whereas the *jo*-*vo* correlative pattern in (7) is analyzed as a different construction, based on the following evidence: 1) felicity of internal heads; 2) requirement of an overt demonstrative/quantifier; 3) compatibility with the inclusive focus particle b^hi; 4) strictly non-restrictive interpretation; 5) impossibility of relative clause stacking; 6) multiple relativization.

We go through some of the relevant data in the next sections and then, in section 4 proceed to analyze correlatives as in (7) on a par with free relatives (following Bhatt 2003) that appear to be situated in SpecDP (contra Bhatt 2003).[4]

3.1 Structural Differences — Headedness

Embedded and extraposed restrictive relative clauses modify an external head, which means that the head NP is not allowed to appear in the relative clause itself. This is demonstrated in (8) and (9), where the relevant NP head(s) are underlined.[5] In contrast, correlative clauses may realize the full head NP in either clause, neither clause, or both clauses. This is shown in (10).

Normal Relative Clause

(8) a. **vo** lɑṛki [**jo** kʰɑṛ-i hɛ]
that girl.F.Sg.Nom which stand-Perf.F.Sg be.Pres.3.Sg
lɑmbi hɛ
tall.F.Sg be.Pres.3.Sg

b. ***vo** [**jo** lɑṛki kʰɑṛ-i hɛ]
that which girl.F.Sg.Nom stand-Perf.F.Sg be.Pres.3.Sg
lɑmbi hɛ
tall.F.Sg be.Pres.3.Sg

c. ***vo** lɑṛki [**jo** lɑṛki kʰɑṛ-i hɛ]
that girl.F.Sg.Nom which girl.F.Sg.Nom stand-Perf.F.Sg be.Pres.3.Sg
lɑmbi hɛ
tall.F.Sg be.Pres.3.Sg
'The girl who is standing is tall.'

Extraposed Relative Clause

(9) a. **vo** lɑṛki lɑmbi hɛ [**jo** kʰɑṛ-i hɛ]
that girl.F.Sg.Nom tall.F.Sg be.Pres.3.Sg which stand-Perf.F.Sg be.Pres.3.Sg

[4] Most of the data discussed in this paper is based on the previous discussions of Hindi correlative structures found in Srivastav (1991), Dayal (1996) and Bhatt (2003). Additionally, we have checked the data with three Pakistani doctoral students at Konstanz, all native speakers of Urdu.

[5] Mahajan (2000:9) does not agree with Srivastav's judgment with respect to the relative clause in (8c) and considers it grammatical, a view shared by our informants. However, this does not make Srivastav's generalization, nor our argumentation here, invalid. Srivastav's analysis deals with restrictive relative clauses, but the sentence-initial *vo* in (8c) has a clear deictic interpretation: Its reference is fixed and further intersective import (e.g., by restrictive relative clauses) is not admissible. Thus the relative clause in this case must have non-restrictive meaning if the sentence is to be grammatical. The fact that non-restrictive relative clauses can be internally headed is confirmed by McCawley (2004).

b. *__vo__ lɑmbi hɛ [__jo__ lɑṛki kʰɑṛ-i hɛ]
that tall.F.Sg be.Pres.3.Sg which girl.F.Sg.Nom stand-Perf.F.Sg be.Pres.3.Sg

c. *__vo__ lɑṛki lɑmbi hɛ
that girl.F.Sg.Nom tall.F.Sg be.Pres.3.Sg

[__jo__ lɑṛki kʰɑṛ-i hɛ]
which girl.F.Sg.Nom stand-Perf.F.Sg be.Pres.3.Sg
'The girl is tall, who is standing.'

Correlative Clause

(10) a. [__jo__ kʰɑṛ-i hɛ] __vo__ lɑṛki lɑmbi hɛ
which stand-Perf.F.Sg be.Pres.3.Sg that girl.F.Sg.Nom tall.F.Sg be.Pres.3.Sg

b. [__jo__ lɑṛki kʰɑṛ-i hɛ] __vo__ lɑmbi hɛ
which girl.F.Sg.Nom stand-Perf.F.Sg be.Pres.3.Sg that tall.F.Sg be.Pres.3.Sg

c. [__jo__ lɑṛki kʰɑṛ-i hɛ]
which girl.F.Sg.Nom stand-Perf.F.Sg be.Pres.3.Sg

__vo__ lɑṛki lɑmbi hɛ
that girl.F.Sg.Nom tall.F.Sg be.Pres.3.Sg
'Which girl is standing, that girl is tall.'

With respect to headedness, the *jo*-*vo* (correlative) and *vo*-*jo* (relative) patterns thus differ quite markedly.

3.2 Correlative as an Operator — The Demonstrative Requirement

Coming from a primarily semantic perspective, Dayal (1996:181) analyzes the correlative *jo*-clause as an operator that locally binds a variable in the main clause. The variable must be contained in the interpretation of the determiner of the external head NP. Her reasons for this analysis build on Subbarao's (1984:13) initial observation that the relative clause cannot be adjoined to the left if the main clause NP is indefinite, as shown in (11a). In fact, Dayal shows that correlatives have to observe a more stringent requirement. Given that in Urdu bare NPs can in principle always also function as definites (Dayal 2003), Dayal formulates a 'demonstrative requirement' (Srivastav 1991:649): the matrix clause must contain a demonstrative. This demonstrative can either be overt as in (11b), or can be analyzed as being there implicitly in the presence of quantifiers such as *sab* 'all' ((11c)), *dono* 'both' ((11d)) or *tino* 'all three'.

(11) a. *[__jo__ lɑṛkiyã kʰɑṛ-i hɛ̃] lɑṛkiyã
which girl.F.Pl.Nom stand-Perf.F be.Pres.3.Pl girl.F.Pl.Nom

lɑmbi hɛ̃
tall.F.Sg be.Pres.3.Pl
'Girls that are standing are tall.'

b.

[**jo**	laṛkiyã	kʰaṛ-i	hɛ̃]	**vo**	laṛkiyã	
which	girl.F.Pl.Nom	stand-Perf.F	be.Pres.3.Pl	those	girl.F.Pl.Nom	

lɑmbi	hɛ̃
tall.F.Sg	be.Pres.3.Pl

‘The girls that are standing are tall.’

c.

[**jo**	laṛkiyã	kʰaṛ-i	hɛ̃]	**sɑb**	laṛkiyã
which	girl.F.Pl.Nom	stand-Perf.F	be.Pres.3.Pl	all	girl.F.Pl.Nom

lɑmbi	hɛ̃
tall.F.Sg	be.Pres.3.Pl

‘All (the) girls that are standing are tall.’

d.

[**jo**	laṛkiyã	kʰaṛ-i	hɛ̃]	**dono**	laṛkiyã
which	girl.F.Pl.Nom	stand-Perf.F	be.Pres.3.Pl	both	girl.F.Pl.Nom

lɑmbi	hɛ̃
tall.F.Sg	be.Pres.3.Pl

‘Both (the) girls that are standing are tall.’

Srivastav (1991) and Dayal (1996) thus analyze correlative clauses as generalized quantifiers (Cooper 1983) that bind a position inside an IP. Syntactically, she posits the structure in (12) for correlatives as in (11).

(12) [[. . .REL-XP. . .]$_{\text{CP}}$ [. . .DEM-XP. . .]$_{\text{IP}}$]$_{\text{CP}}$

4 Correlatives as Free Relative Specifiers of DP

As shown above, Dayal’s (1996) seminal work on relatives and correlatives provides a clear basis for analysis. However, Bhatt (1997, 2003) looks at additional empirical evidence and argues that the *jo*-*vo* correlatives are better understood as being like free relatives. We discuss his reasons briefly in section 4.1, and adopt his view of correlatives as free relatives, but in a slightly different manner. We present an alternative analysis in section 4.2 by which we analyze the *jo*-*vo* correlatives as being situated in SPECDP, rather than being adjoined to IP (Srivastav 1991) or to the demonstrative phrase (Bhatt 2003).

4.1 DP Adjunction and Free Relatives

Dayal’s analysis of the *jo*-*vo* correlative as a quantifier within a CP that adjoins to and binds a position inside an IP is challenged by several facts. For example, consider (13), where the correlative clause can appear directly to the left of the external head inside the main clause. This construction is not uncommon, and indicates that an analysis of direct adjunction to DP should be considered (Bhatt 1997, 2003).

(13) hɑsan=ne [**jo** kɪtab tara=ne lɪk^h-i]
Hassan=Erg which book.F.Sg.Nom Tara=Erg write-Perf.F.Sg

vo pɑsɑnd k-i
that liking do-Perf.F.Sg
'Hassan liked the book which Tara wrote.'

Further evidence for DP-adjunction of the correlative clause comes from question-answer pairs as in (14), which show that the *jo*-*vo* clause in combination with the required demonstrative makes a perfect short answer to a question, just like a simple DP/NP would.

(14) kɔ̃ a-yi?
who come-Perf.F.Sg?
'Who came?'

[**jo** lɑṛki vɑhã rɑh-ti hɛ] **vo**
Which girl.F.Sg.Nom there stay-Impf.F.Sg be.Pres.3.Sg that
'Which girl lives there, she'/'The girl who lives there' (Dayal 1996)

Indeed, Wali (1982) already used this fact to argue for DP-adjunction. But Dayal rejects the DP-adjunction analysis in favor of a unified treatment of single correlatives and those with multiple heads (see section 5), by which IP-adjunction is treated as the basic phenomenon and DP-adjunction is analyzed as a case of crosscategorial quantification (Dayal 1996:206).

In contrast, Bhatt (1997) situates the correlative clause primarily within the DP (but see section 4.3 on topicalization facts) and, in particular, as having the functional properties of a free relative clause. As clearly demonstrated by Bhatt (1997), the properties of correlatives and free relatives are strikingly similar. For example, only free relatives (as opposed to restrictives) in English can feature the inclusive focus item *-ever*, a property which we also find in Urdu, where the focus particle *b^{h}i* 'also' cannot modify a restrictive relative clause ((15a)), but is admissible in correlatives in order to bring out the unspecified identity of the internal head, as shown in (15b).

(15) a. *vo lɑṛki [jo **b^{h}i** vɑhã k^{h}ɑṛ-i hɛ]
that girl.F.Sg.Nom which also there stand-Perf.F.Sg be.Pres.3.Pl

nadya=ki sɑheli hɛ
Nadya=Gen.F.Sg friend.F.Sg be.Pres.3.Sg
'That girl, whichever is standing there, is Nadya's friend.'

b. [jo **b^{h}i** lɑṛki vɑhã k^{h}ɑṛ-i hɛ]
which also girl.F.Sg.Nom there stand-Perf.F.Sg be.Pres.3.Pl

vo nadya=ki sɑheli hɛ
that Nadya=Gen.F.Sg friend.F.Sg be.Pres.3.Sg
'Whichever girl is standing there is Nadya's friend.'

c. jo **b^{h}i** lɑṛki mehnɑt kɑr-ti hɛ
which also girl.F.Sg.Nom effort do-Impf.F.Sg be.Pres.3.Sg

vo safal ho-ti hɛ
that successful be-Impf.F.Sg be.Pres.3.Sg
‘Whichever girl makes an effort is successful.’

Note that the presence of *b^{h}i*, just like *-ever*, forces a generic reading if one is possible, as shown in (15c). The acceptability of the generic reading, however, is dependent on the aspect of the relative clause predicate. If information from tense/aspect does not provide information about genericity, the standard interpretation of the correlative clause is definite, regardless of the presence of the focus item (as is generally the case with free relatives; Jacobson 1995).

As Bhatt (1997, 2003) further shows, the correlative can even take the form of a true free relative without a demonstrative ‘correlate’ if the case marking of the internal as well as the external head is unmarked (nominative in Urdu). The demonstrative cannot be omitted if either the demonstrative or the correlate is overtly marked by a case clitic (e.g., *ne, ko*), but can be left out if the surface form matches or, in the case of Urdu, has no surface form, as shown in (16).

(16) [**jo** lɑṛki kʰɑṛ-i hɛ] hɑs rɑh-i
which girl.F.Sg.Nom stand-Perf.F.Sg be.Pres.3.Sg laugh stay-Perf.F.Sg

hɛ
be.Pres.3.Sg
‘Which girl is standing, is laughing.’

This form of surface matching is known from German free relative constructions, which also require a resumptive demonstrative/determiner if the case marking differs overtly, as is the case in (17a), but not in (17b,c).

(17) a. **Wer** dich nicht mag, ***(den)** mag ich
who.Nom you.Acc not like.Pres.3.Sg. **that.Acc** like.Pres.1.Sg I.Nom

auch nicht.
also not
‘Who doesn’t like you, I don’t like either.’

b. **Wen** du magst, **(den)** will ich auch
who.Acc you.Nom like.Pres.2.Sg **that.Acc** want.Pres.1.Sg. I.Nom also

treffen.
meet
‘I also want to meet the one who you like.’

c. **Was** du magst **(das)** gefällt mir auch.
what.Acc you.Nom like.Pres.2.Sg **that.Nom** please.Pres.3.Sg I.Dat also
‘Whatever you like also pleases me.’

In sum, correlative clauses in Urdu have a number of semantic and morphosyntactic properties that are familiar from free relatives in German and English. Within the context of the ParGram project, this points towards the need to find a common underlying analysis for free relatives and correlatives in these languages.

4.2 Specifier of DP

In order to account for similarities between free relatives and correlatives, we treat the correlative clause plus demonstrative constituent as a DP with an f-structure analogous to free relatives in English, since these have comparable semantics and distribution. However, instead of analyzing the relative clause predicate as an adjunct, as Butt et al. (1999) did for free relative clauses, we consider correlative clauses as occupying a specifier position and thus contributing a SPEC attribute to the f-structure. This is done for the following reasons:

- Correlatives cannot be stacked: Whereas normal relative clauses project into an adjunct set, a DP can only be modified by a single correlative.

  (18) *[jo gari tez hɛ] [jo lal hɛ]
  which car.F.Sg.Nom fast be.Pres.3.Sg which red be.Pres.3.Sg
  vo gari sundɑr hɛ
  that car beautiful be.Pres.3.Sg
  'Which car is fast, which car is red, that car is beautiful.'

- Semantically, correlatives function as quantifiers. Thus, they cannot have a non-restrictive interpretation and cannot modify, for example, proper nouns, as this would result in vacuous quantification.

  (19) *[jo vɑhã k^{h}ɑṛ-a hɛ] ram lɑmba hɛ
  who there stand-Perf.M.Sg be.Pres.3.Sg Ram tall.M.Sg be.Pres.3.Sg
  'Who is standing there, Ram is tall.'

- Correlatives appear in complementary distribution with other SPEC material, such as possessors.

  (20) *[jo lal hɛ] [yonas=ki] gari
  which red be.Pres.3.Sg Jonas=Gen.F.Sg car.F.Sg.Nom
  sundɑr hɛ
  beautiful be.Pres.3.Sg
  'Which car is red, Jonas's car is beautiful.'

All of the evidence presented so far points to a DP-internal, non-adjunct analysis of correlatives. In particular, we situate the correlative clause in a specifier position directly left-adjacent to the DP that it quantifies over. However, there is a further set of data that remain to be accounted for.

4.3 Topicalization

In addition to finding correlatives that are directly left-adjacent to the modified constituent, instances of discontinuous correlatives also occur. In both of the examples in (21), the correlative clause is in the regular sentence-initial position, but the demonstrative is embedded further inside the main clause ((20a)) or even embedded within a sentential complement ((20b)).

(21) a. [jo lɑṛki$_i$ vɑhã hɛ]
which girl.F.Sg.Nom there be.Pres.3.Sg

ram=ne ʊs=ko$_i$ pɑsɑnd ki-ya
Ram=Erg that=Acc liking do-Perf.3.Sg
'Which girl is there, Ram likes her.'

b. [jo lɑṛki$_i$ ga rɑh-i hɛ]
which girl.F.Sg.Nom sing stay-Perf.F.Sg be.Pres.3.Sg

sita soc^h-ti hɛ [kɪ vo$_i$ sundɑr hɛ]
Sita think-Impf.F.Sg be.Pres.3.Sg that that beautiful be.Pres.3.Sg
'Which girl is singing, Sita thinks that she is beautiful.'

We propose to analyze this dislocation as an instance of standard topicalization of the correlative clause. This analysis is reasonable, given that Urdu is a discourse-configurational language with basic SOV order that makes heavy use of word order permutations to syntactically encode information structure (cf. Butt and King 1996, Kidwai 2000). The TOPIC function is generally associated with the initial item of the utterance, located in SPECIP. Since Urdu allows not only arguments to be topicalized, but also, unlike English, SPECNP content such as genitive possessors (Mohanan 1994), as shown in (22), it is predicted that correlative clauses should also be able to undergo this dislocation.

(22) [ram=ki] sundɑr hɛ gari
Ram=Gen.F.Sg beautiful be.Pres.3.Sg car.F.Sg.Nom
'Ram's car is beautiful.'

As Bhatt (2003) shows, topicalized correlatives may connect into sentential complements, but are sensitive to island effects, and thus cannot be topicalized from within adjuncts or complex NPs. Furthermore, only one correlative clause can be topicalized. If any other DP in the sentence is modified by a correlative clause, none of these may additionally appear in the front, but must be located in non-initial position within the relevant SPECDP. The only admissible structure that allows two correlative clauses at the beginning of the sentence is one where one correlative is topicalized and the other occupies the specifier position of a sentence-inital DP, as illustrated by (23).

(23) [jo$_i$ kɪtab mez=pɑr tʰ-i]
which book.F.Sg.Nom table.F.Sg=on be.Past.F.Sg

[[[[jo$_j$ talɪb‘ɪlm hɛ] ʊs$_j$] lɑrke=ne] vo$_i$ lɪkʰ-i]
which student.Nom be.Pres.3.Sg that boy.M.Obl=Erg that write-Perf.F.Sg
‘The boy who is a student wrote the book that was on the table.’

4.4 The LFG Analysis

DP internal correlative clauses are linked to their heads via the functional description expanding the DP node. As topicalized correlatives are discontinuous from their heads, this case is more interesting. As is standard with respect to long-distance dependencies within LFG, topicalized correlatives are linked to their heads via functional uncertainty paths (Kaplan and Zaenen 1989).[6] However, as correlative clauses are not quite standard topics, a little more work needs to be done. The external head of the topicalized correlative is found via the disjunction in (24) (defined in XLE’s regular expression notation; Crouch et al. 2006):

(24) CorFunc =
{Subj ∨ {Xcomp ∨ Comp}* {Obj ∨ Obj-Goal } (Adj-Gen)}

This means that the function CorFunc is assigned a grammatical function that is either Subj or an Obj or Obj-Goal which may be embedded in zero to infinitely many (signified by the Kleene Star) Comps or Xcomps, or a genitive possessor of any of these. In the c-structure rule that licenses the topicalized correlatives shown in (25), this function is given a local name (%Cor-Head) in order to formulate the constraints that must be simultaneously satisfied, such as the demonstrative/quantifier requirement (cf. section 3.2) as well as number and oblique/nominative agreement (coded under NMorph). The rule in (25) also includes the possibility of a topicalized KP (Urdu Kase Phrase, featuring a DP plus optional case clitic, Butt and King 2004).

(25) SpecIP ⟶
CPcorr ∨ KP
(↑CorFunc)=%Cor-Head (↑Topic)=↓
(%Cor-Head Spec Corr)=↓ @GF
(%Cor-Head Num)=(↓Topic-Rel Num)
(%Cor-Head NMorph)=(↓Topic-Rel NMorph)
{(%Cor-Head Spec Det Pron-Type)=c dem
(%Cor-Head Spec Det Deixis)=c distal
∨ (%Cor-Head Spec Quant Quant-Type)=c universal}

The CPcorr category is defined in analogy to the standard relative clause, CPrel, with the exception that in CPcorr the Topic-Rel may include a contentful NP. The function Spec Corr, which encodes the correlative clause itself,

[6] Topicalized possessors as in (22) are, of course, also linked via functional uncertainty equations.

is proposed as an interim solution since the facts presented in Jacobson (1995) and others call for a consequent reanalysis of free relative clauses in the ParGram languages that departs from the ADJUNCT solution proposed in Butt et al. (1999), along with a unified analysis that provides enough information in the f-structure to lead to the correct semantic representation for both correlatives and free relatives.

An unusual but positive aspect to the analysis in (25) is that the subject of the main clause and that of the correlative do not stand in a direct functional relation other than noun agreement. Since the correlative is allowed to contain a full noun phrase in its internal head, it is in principle possible for the internal head and the external head to contain diffent nouns. And precisely this possibility is required by data as in (26).

(26) [**jo** **catr** vahã k^{h}ɑṛ-a hɛ]
which student.M.Sg.Nom there stand-Perf.M.Sg be.Pres.3.Sg

vo **lɑṛka** mera dost hɛ
that boy.M.Sg.Nom I.Gen.M.Sg friend.M.Sg.Nom be.Pres.3.Sg
'Which student is standing there, that boy is my friend.'
(McCawley 2004:300)

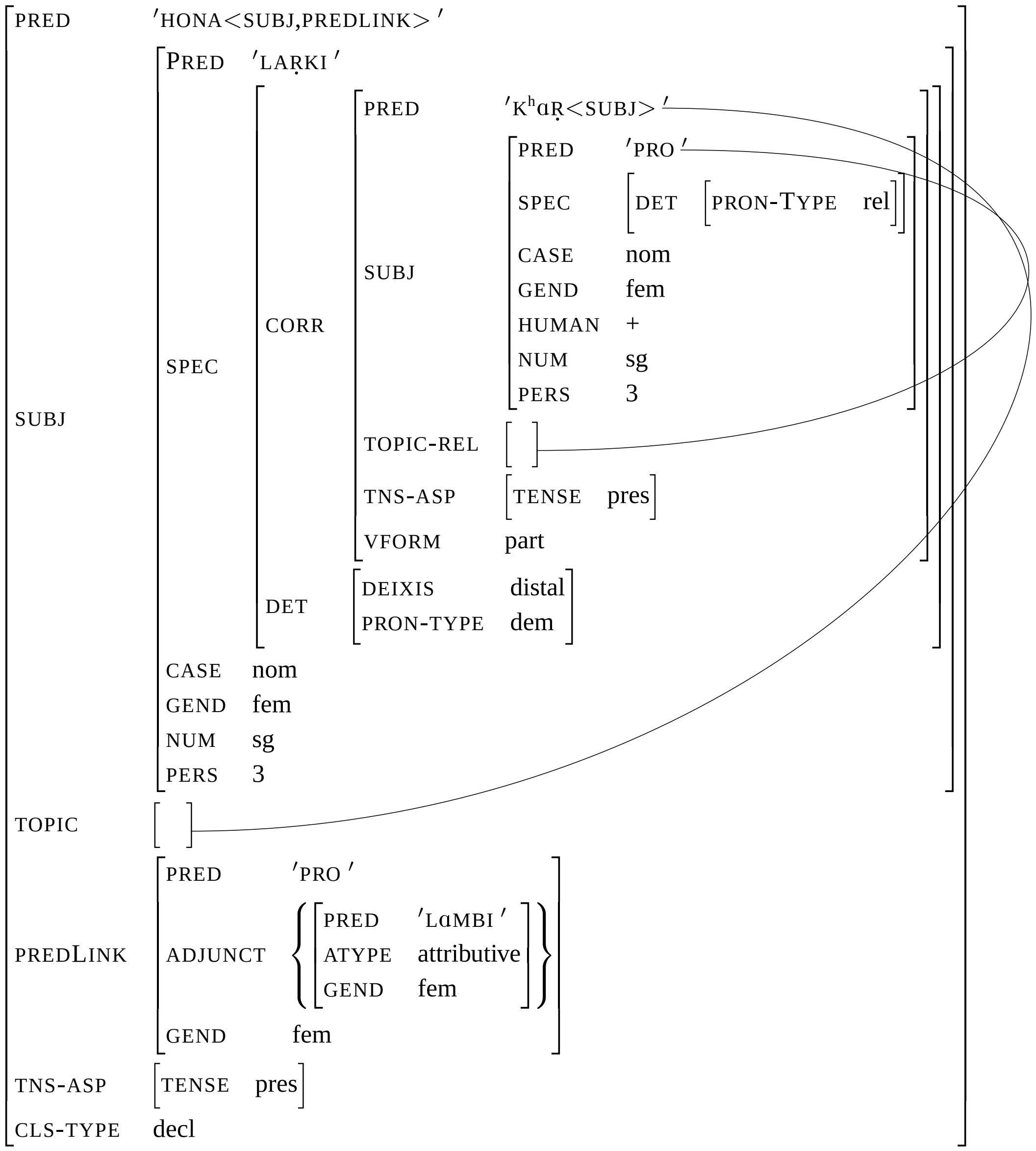
There seem to be semantic constraints on the felicitous choice of the two different nouns (involving synonymy or hyponymy), but previous analyses of these (cf. Dayal 1996:196 and McCawley 2004:300) as well as judgements of our informants leave an inconclusive picture of what relations are acceptable. With respect to our analysis, since the constraints are purely semantic, they are not handled by the syntactic c-structure and f-structure components.

In (28) and (29) we present a sample c-structure and f-structure analysis for the example in (27). The representation of noun phrase structure departs from previous analyses in the Urdu grammar by postulating a DP structure above NP that holds the determiner *vo* (which, in its use as a personal pronoun, accompanies an empty noun head) as well as a SPECDP position potentially containing the correlative.

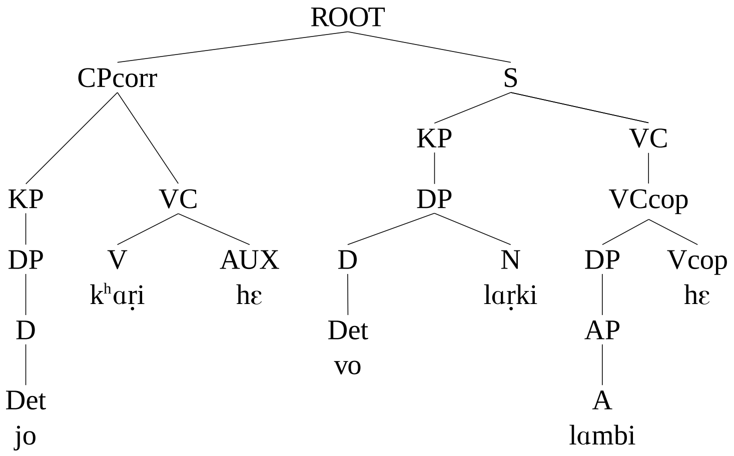
Note that the same sentence can also receive another analysis by which the correlative clause is not topicalized and is contained inside the sentence-inital DP instead of SPECIP. This option, which then lacks the TOPIC function at f-structure, is dispreferred through the use of OT marks (Frank et al. 2001).

(27) [**jo** k^{h}ɑṛ-i hɛ] [**vo** lɑṛki lɑmbi hɛ]
which stand-Perf.F.Sg be.Pres.3.Sg that girl.F.Sg.Nom tall.F.Sg be.Pres.3.Sg
'Who is standing, that girl is tall.'

(28) F-Structure:



(29) C-structure:



5 Further Issues: Multi-Head-Correlatives

So far we have presented an analysis for single-head correlatives (SHC). However, as shown in (30), one of the striking features of Urdu correlatives is that they can appear with more than one relativized element, containing multiple relative pronouns linked to multiple correlate demonstratives in the main clause.

(30)

[jɪs$_i$	lɑṛki=ne	jɪs$_j$	lɑṛke	ke satʰ	kʰel-a]
which.Obl	girl.F.Sg=Erg	which.Obl	boy.M.Obl	with	play-Perf.M.Sg

ʊs=ne$_i$	ʊs=ko$_j$	hɑra-ya
that.Obl=Erg	that.Obl=Acc	defeat-Perf.M.Sg

‘Which girl played with which boy, she defeated him.’ (Dayal 1996)

This correlative clause cannot be attached to any single correlate at f-structure, since both internal heads are equally governed by the relative clause predicate, and the functional projection of the predicate cannot be attached to both external heads with the same internal structure. The correlative clause cannot be said to determine either argument of the matrix, but rather determines both by specifying a relation between them. This can be expressed semantically by arguing that multi-head-correlatives (MHC) quantify over ordered tuples rather than individuals (as proposed by Lehmann 1984:344). It can be expressed syntactically by attaching the f-structure of the correlative clause directly to the main clause predicate rather than to one of its arguments. Analogously, Srivastav (1991) and Bhatt (1997) argue for base-generation of the MHC in a position adjoined to IP. Within Dayal’s account this means that she presents a unified analysis of SHC and MHC, within Bhatt’s account this means that MHC and SHC receive a differing syntactic analysis.

With respect to this issue, we again propose to follow Bhatt’s analysis and treat MHC as a separate class for which no analogous construction (such as free relatives for SHC) exists in languages that do not feature correlatives. Consequently, these sentences cannot be translated straightforwardly into English. Andrews (1975), for example, proposes to translate MHC as conditionals, which gives adequate results as long as the correlative can have a generic interpretation, but this is not always the case. Another suggestion, propably first proposed by Delbrück (1900) for Sanskrit MHC, is to use an indefinite in place of the second relativized phrase, which is anaphorically picked up in the matrix clause (*Whichever girl played with a boy defeated him.*). This translation would also be faithful to the semantics of the construction, but does not do justice to the differing syntactic constraints. As shown in (31), Urdu MHC cannot appear with a matrix predicate of less arity, whereas relatives-cum-indefinite can.

(31) a. [*jɪs lɑṛki jɪs lɑṛke ke sat^h k^{h}el-egi]
which.Obl girl.F.Sg.Nom which.Obl boy.M.Obl with play-Fut.F.Sg

vo jit-egi
that win-Fut.F.Sg
'Which girl will play with which boy, she will win.'

b. Whichever girl will play with a boy will win.

MHC are also less constrained in contrast to SHC when it comes to the resumptive pronoun requirement. Even in cases where the demonstrative accompanying a correlative clause could not be dropped, i.e. if there is overt case-marking, they may be dropped with MHC (Bhatt 1997), as shown in (32).

(32) [*jɪs lɑṛke=ne jɪs lɑṛki=ko dek^h-a]
which.Obl boy.M.Sg.Obl=Erg which.Obl girl.F.Sg=Acc see-Perf.M.Sg

(ʊs=ne ʊs=ko) pɑsɑnd ki-ya
that.Obl=Erg that.Obl=Acc liking do-Perf.M.Sg
'Which boy saw which girl, he liked her.'

Since the exact nature of the interaction between the constraints of correlative formation and the rampant pro-drop that is generally possible in Urdu (Neeleman and Szendroi 2007) is not yet well understood, our analysis is rather minimal. At c-structure, we assume MHC to be located adjoined above IP. At f-structure, the correlative clause projects an ADJUNCT to the main clause predicate. The anaphoric relation between the relativized elements and possible correlates in the main clause is left to the semantic processing component, which may be tackled once a better understanding of the structure is reached.

6 Conclusion

Building on previous insights by Srivastav/Dayal and Bhatt, we distinguish between relative clauses (*vo-jo*) and correlatives (*jo-vo*), and account for their different internal structure and semantic interpretation. Correlatives are treated as quantifiers that appear either in the specifier position of the DP they modify or in a topicalized position at the left periphery. At f-structure, they differ from normal relative clauses by projecting to a SPEC structure rather than an adjunct set, which goes along with their quantifier interpretation and their inability to stack. The parallels to free relative clauses suggest that a similar analysis might be argued for in the case of German and English free relatives, which currently receive the same ADJUNCT treatment as standard relatives. The advantages and disadvantages of such a parallel analysis, as well the issue with multi-head-correlatives, can hopefully be understood once a standardized semantic representation has been agreed on within ParGram, and once the existing analysis has been incorporated into the main Urdu grammar in order to investigate interactions with other phenomena, such as pro-drop.

References

Andrews, Avery D. 1975. *Studies in the Syntax of Relative and Comparative Clauses*. Ph. D.thesis, Massachusetts Institute of Technology.

Bhatt, Rajesh. 1997. Matching Effects and the Syntax-Morphology Interface: Evidence from Hindi Correlatives. In *MIT Working Papers in Linguistics*, volume 31, MIT Press.

Bhatt, Rajesh. 2003. Locality in Correlatives. *Natural Language and Linguistic Theory* 21, 485–541.

Butt, Miriam, Dyvik, Helge, King, Tracy Holloway, Masuichi, Hiroshi and Rohrer, Christian. 2002. The Parallel Grammar Project. In *COLING Workshop on Grammar Engineering and Evaluation*.

Butt, Miriam and King, Tracy Holloway. 1996. Structural Topic and Focus without Movement. In *Proceedings of the First LFG Conference*.

Butt, Miriam and King, Tracy Holloway. 2002. Urdu and the Parallel Grammar Project. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*.

Butt, Miriam and King, Tracy Holloway. 2004. The Status of Case. In Veneeta Dayal and Anoop Mahajan (eds.), *Clause Structure in South Asian Languages*, pages 153–198, Berlin: Springer Verlag.

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

Cooper, Robin. 1983. *Quantification and Syntactic Theory*. Dordrecht.

Crouch, Richard, Dalrymple, Mary, Kaplan, Ronald M., King, Tracy Holloway, Maxwell III, John T. and Newman, Paula. 2006. XLE Documentation, Palo Alto Research Center, on-line.

Davison, Alice. 2006. Correlative Clause Features in Sanskrit and Hindi/Urdu, Paper presented at the 9th Diachronic Syntax Conference, Trieste.

Dayal, Veneeta. 1996. *Locality in WH Quantification. Questions and relative clauses in Hindi*. Kluwer Academic Publishers.

Dayal, Veneeta. 2003. Bare Nominals: Non-specific and Contrastive Readings under Scrambling. In Simin Karimi (ed.), *Word Order and Scrambling*, Oxford: Blackwell.

Delbrück, Berthold. 1900. Vergleichende Syntax der indogermanischen Sprachen. Dritter Theil. In Karl Brugmann and Berthold Delbrück (eds.), *Grundriß der vergleichenden Grammatik der Indogermanischen Sprachen*, Trübner.

Frank, Anette, King, Tracy Holloway, Kuhn, Jonas and Maxwell III, John T. 2001. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality Theoretic Syntax*, pages 367–397, CSLI Publications.

Jacobson, Pauline. 1995. On the Quantificational Force of English Free Relatives. In Emond Bach, Eloise Jelinek, Angelika Kratzer and Barbara Partee (eds.), *Quantification in Natural Language*, Kluwer Academic Publishers.

Kaplan, Ron and Zaenen, Annie. 1989. Long-distance dependencies, constituent structure, and functional uncertainty. In Mark Baltin and Andy Kroch (eds.), *Alternative Conceptions of Phrase Structure*, Chicago University Press.

Kidwai, Ayesha. 2000. *XP-Adjunction in Universal Grammar: Scrambling and Binding in Hindi-Urdu*. Oxford: Oxford University Press.

Lehmann, Christian. 1984. *Der Relativsatz*. Gunter Narr Verlag.

Mahajan, Anoop. 2000. Relative Asymmetries and Hindi Correlatives. In Artemis Alexiadou, Paul Law and Chris Wilder (eds.), *The Syntax of Relative Clauses*, John Benjamins.

McCawley, John D. 2004. Remarks on Adsentential, Adnominal and Extraposed Relative Clauses in Hindi. In Veneeta Dayal and Anoop Mahajan (eds.), *Clause Structure in South Asian Languages*, Berlin: Springer Verlag.

Mohanan, Tara. 1994. *Argument Structure in Hindi*. CSLI Publications.

Neeleman, Ad and Szendroi, Kriszta. 2007. Radical pro drop and the morphology of pronouns. *Linguistic Inquiry* 38(4), 671–714.

Srivastav, Veneeta. 1991. The Syntax and Semantics of Correlatives. *Natural Language and Linguistic Theory* 9(4).

Subbarao, K. V. 1984. *Complementation in Hindi Syntax*. Academic Publications.

Wali, Kashi. 1982. Marathi Correlatives: A Conspectus. *South Asian Review* 6.

DESIGNING FEATURES FOR PARSE DISAMBIGUATION AND REALISATION RANKING

Aoife Cahill Martin Forst and Christian Rohrer
Universität Stuttgart

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

We present log-linear models for use in the tasks of parse disambiguation and realisation ranking in German. Forst (2007a) shows that by extending the set of features used in parse disambiguation to include more linguistically motivated information, disambiguation results can be significantly improved for German data. The question we address in this paper is to what extent this improved set of features can also be used in realisation ranking. We carry out a number of experiments on German newspaper text. In parse disambiguation, we achieve an error reduction of 51%, compared to an error reduction of 34.5% with the original model that does not include the additional features of Forst (2007a). In realisation ranking, BLEU score increases from 0.7306 to 0.7939, and we achieve a 10 point improvement in exact match over a baseline language model. This being said, our results also show that further features need to be taken into account for realisation ranking in order to improve the quality of the corresponding model.

1 Introduction

Statistical disambiguation of syntactic structures has been extensively studied in recent years. Riezler et al. (2002) have successfully applied a log-linear model based on features referring to simple, mostly locally restricted c- and f-structure configurations to the task of LFG parse disambiguation for English. However, recent studies suggest that these types of features are not sufficient for the disambiguation of languages with relatively free word order, such as Japanese (Yoshimura et al., 2003) or German.

Forst (2007a) shows that by extending the set of features used in parse disambiguation to include more linguistically motivated information, disambiguation results can be significantly improved for German data. The question we address in this paper is to what extent this improved set of features can also be used in realisation ranking.[1] It is clear that some features designed for parse disambiguation will not be useful for realisation ranking and vice versa. For example, features that capture lexical dependencies will not be useful in generation ranking, since lexical dependencies are given in this task. Conversely, the log-linear model for realisation ranking, where the task is to determine the most natural sounding sequence of words, will need features that refer (only) to the surface string, and those features are, of course, not interesting for parse disambiguation. Nevertheless, it is reasonable to assume that c-structure features or features that refer to c-structure and f-structure simultaneously are useful for both tasks, and that taking the angle of both tasks may help to identify relevant features.

We present a model for realisation ranking similar to that of Velldal and Oepen (2005). The main differences between our work and theirs is that we are working

[1] The work described in this paper has been carried out as part of the COINS project of the linguistic Collaborative Research Centre (SFB 732) at the University of Stuttgart, which is funded by the German Research Foundation (DFG). Furthermore, we would like to thank John Maxwell of the Palo Alto Research Center for being so responsive to our requests for extensions of the XLE generator functionalities, some of which were crucial for our work.

within the LFG framework and concentrating on a less configurational language: German.

2 System Setup

2.1 A Broad-Coverage LFG for German

For the construction of our data, we use the German broad-coverage LFG documented in Dipper (2003) and Rohrer and Forst (2006). It is a hand-crafted grammar developed in and for the LFG grammar development and processing platform XLE (Crouch et al., 2006). It achieves parsing coverage of about 80% in terms of full parses on newspaper text, and for sentences out of coverage, the robustness techniques described in Riezler et al. (2002) (i.e. fragment grammar, ‘skimming’) are employed for the construction of partial analyses. The grammar is reversible, which means that the XLE generator can produce surface realisations for well-formed input f-structures.

2.2 Parse Disambiguation

We use a standard log-linear model for carrying out parse disambiguation (Toutanova et al., 2002; Riezler et al., 2002; Miyao and Tsujii, 2002; Malouf and van Noord, 2004; van Noord, 2006; Clark and Curran, 2004). A key factor in the success of these models is feature design. As a baseline, we design features based on the property set used for the disambiguation of English ParGram LFG parses (Riezler et al., 2002; Riezler and Vasserman, 2004). These properties are based on thirteen property templates, which can be parameterised for any combination of c-structure categories or f-structure attributes and their values. Forst (2007a) shows that by extending this set of features used in parse disambiguation to include more linguistically motivated information, disambiguation results can be significantly improved for German data.

2.3 Surface Realisation

As XLE comes with a full-fledged generator, the grammar can be used both for parsing and for surface realisation.[2] Figure 2 shows the set of 18 strings that are generated from the f-structure in Figure 1. In this case, the German parser only produces one parse, and so there is no parse disambiguation necessary. However there is some work to be done in ranking the alternative string realisations for the input f-structure. Note that all of the surface realisations are grammatical; however, some of them are clearly more likely or unmarked than others.

[2] At the moment it is not possible to generate from packed structures where ambiguity is preserved. However, in the future we hope to be able to do so. This would be particularly useful in an application such as machine translation, where some ambiguities transfer across languages.

(1) Die Nato werde nicht von der EU geführt.
The NATO is not from the EU led.
‘NATO is not led by the EU.’

```
"Die Nato werde nicht von der EU geführt."

    PRED        'führen<[249:von], [21:Nato]>'
                PRED   'Nato'
                CHECK  [_SPEC-TYPE [_COUNT +, _DEF +, _DET attr]]
                       [_INFL      strong-det                   ]
    SUBJ        NTYPE  [NSYN proper]
                SPEC   [DET [PRED     'die']]
                            [DET-TYPE def  ]
             21 CASE nom, GEND fem, NUM sg, PERS 3
                PRED   'von<[283:EU]>'
                       PRED   'EU'
                       CHECK  [_SPEC-TYPE [_COUNT +, _DEF +, _DET attr]]
                              [_INFL      strong-det                   ]
    OBL-AG      OBJ    NTYPE  [NSYN proper]
                       SPEC   [DET [PRED     'die']]
                                   [DET-TYPE def  ]
                   283 CASE dat, GEND fem, NUM sg, PERS 3
            249 PSEM dir, PTYPE sem
    ADJUNCT     {    [PRED          'nicht']}
                {215 [ADJUNCT-TYPE  neg    ]}
                [_AUX-FORM  <werden-pass_>       ]
    CHECK       [_VLEX      [_AUX-SELECT sein]   ]
                [_VMORPH    [_PARTICIPLE perfect]]
    TNS-ASP     [MOOD subjunctive, PASS-SEM dynamic_, TENSE pres]
    TOPIC       [21:Nato]
128 CLAUSE-TYPE decl, PASSIVE +, STMT-TYPE decl, VTYPE main
```

Figure 1: F-structure for (1)

Just as hand-crafted grammars, when used for parsing, are only useful for most applications when they have been complemented with a disambiguation module, their usefulness as a means of surface realisation depends on a reliable module for realisation ranking. A long list of arbitrarily ordered output strings is useless for practical applications such as summarisation, question answering, machine translation, etc.

```
Die Nato werde von der EU nicht geführt.
Die Nato werde nicht von der EU geführt.
Nicht von der EU geführt werde die Nato.
Nicht werde von der EU die Nato geführt.
Nicht werde die Nato von der EU geführt.
Nicht geführt werde von der EU die Nato.
Nicht geführt werde die Nato von der EU.
Von der EU nicht geführt werde die Nato.
Von der EU werde die Nato nicht geführt.
Von der EU werde nicht die Nato geführt.
Von der EU geführt werde nicht die Nato.
Von der EU geführt werde die Nato nicht.
Geführt werde die Nato nicht von der EU.
Geführt werde die Nato von der EU nicht.
Geführt werde nicht von der EU die Nato.
Geführt werde nicht die Nato von der EU.
Geführt werde von der EU nicht die Nato.
Geführt werde von der EU die Nato nicht.
```

Figure 2: The set of strings generated from the f-structure in Figure 1

Very regular preferences for certain realisation alternatives over others can be implemented by means of so-called optimality marks (Frank et al., 2001), which are implemented in XLE both for the parsing and the generation direction. For ranking string realisations on the basis of ‘soft’ and potentially contradictory constraints, however, the stochastic approach based on a log-linear model, as it has previously been implemented for English HPSGs (Nakanishi et al., 2005; Velldal and Oepen, 2005), seems more adequate.

3 Feature Design

3.1 Feature Design for Parse Disambiguation

Feature design for parse disambiguation is often carried out in a semi-automatic manner, i.e. by designing feature templates that are then instantiated automatically. Although the number of features built this way is often in the hundreds of thousands, nothing guarantees that the information relevant for disambiguation is actually captured by some feature(s). This is particularly true when the feature templates have been designed with little attention to typical ambiguities in the language under consideration. Forst (2007a) shows that linguistically motivated features that capture, e.g., the linear order of grammatical functions, the (surface and functional uncertainty path) distance of an extraposed constituent to its f-structure head, the nature of a DP in relation to its grammatical function (pronominal vs. full DP, animate vs. inanimate) etc. allow for a significantly improved disambiguation

Name of feature template and parameters	Explanation
Features used for the disambiguation of English ParGram LFG parses (Riezler et al., 2002; Riezler and Vasserman, 2004)	
fs_attrs <attrs>	counts number of occurrences of attribute(s) <*attrs*> in the f-structure
cs_label <cat>	counts number of occurrences of category <*cat*> in the c-structure
fs_attr_val <attr> <val>	counts number of times f-structure attribute <*attr*> has value <*val*>
cs_num_children <cat>	counts number of children of all nodes of category <*cat*>
fs_adj_attrs <attr1> <attr2>	counts the number of times feature <*attr2*> is immediately embedded in feature <*attr1*>
fs_sub_attrs <attr1> <attr2>	counts the number of times feature <*attr2*> is embedded somewhere in <*attr1*>
cs_adjacent_label <cat1> <cat2>	counts the number of <*cat1*> nodes that immediately dominate <*cat2*> nodes
cs_sub_label <cat1> <cat2>	counts the number of <*cat1*> nodes that (not necessarily immediately) dominate <*cat2*> nodes
cs_embedded <cat> <Depth>	counts the number of <*cat*> nodes that dominate (at least) <*Depth*> other <*cat*> nodes
cs_conj_nonpar <Depth>	counts the number of coordinated c-structures that are not parallel at <*Depth*> levels under the coordinated constituent
lex_subcat <Lemma> <SCFs>	counts the number of times <*Lemma*> occurs with one of the subcategorisation frames in <*SCFs*>
Additional Linguistically Motivated Features	
ADD-PROP MOD1 <Lemma>	counts the number of times a given lemma occurs as a member of a MOD set
ADD-PROP F2 <Lemma> <PoS>	counts the number of times a given lemma occurs as a particular <*PoS*>
ADD-PROP ACTIVE/PASSIVE <Lemma>	counts the number of times a (verb) lemma occurs in active/passive voice
ADD-PROP isCommon/Def/ Pronoun/... <GF>	determines whether a DP with function <*GF*> is common, definite, pronominal, etc.
ADD-PROP DEP11 <PoS1> <Dep> <PoS2>	counts the number of times a sub-f-structure of type <*PoS2*> is embedded into a (sub-)f-structure of type <*PoS1*> as its <*Dep*>
ADD-PROP PATH	counts given instantiations of functional uncertainty paths
ADD-PROP PRECEDES <GF1> <GF2>	counts the number of times a <*GF1*> precedes a <*GF2*> of the same (sub-)f-structure
DISTANCE-TO-ANTECEDENT %X	distance between a relative clause and its antecedent
ADD-PROP DEP12 <PoS1> <Dep> <PoS2> <Lemma2>	counts the number of times a sub-f-structure of type <*PoS2*> and with <*Lemma2*> as its PRED is embedded into a (sub-)f-structure of type <*PoS1*>as its <*Dep*>
ADD-PROP DEP21 <PoS1> <Lemma1> <Dep> <PoS2>	counts the number of times a sub-f-structure of type <*PoS2*> is embedded as its <*Dep*> into a (sub-)f-structure of type <*PoS1*> and with <*Lemma1*> as its PRED
ADD-PROP PRECEDES <Lemma> <GF1> <GF2>	counts the number of times a <*GF1*> subcategorised for by a PRED <*Lemma*> precedes a <*GF2*> subcategorised for by the same PRED
ADD-PROP MOD2 <Lemma1> <Lemma2>	counts the number of times <*Lemma2*> occurs in the MOD set of a (sub-)f-structure with <*Lemma1*> as its PRED
ADD-PROP VADJUNCT_PRECEDES <Prep1> <Prep2>	counts the numbers of times an ADJUNCT PP headed by <Prep1> precedes an ADJUNCT PP headed by <Prep2>, both being in an f-structure with a VTYPE
ADD-PROP DEP22 <PoS1> <Lemma1> <Dep> <PoS2> <Lemma2>	counts the number of times a sub-f-structure of type <*PoS2*> and with <*Lemma2*> as its PRED is embedded as its <*Dep*> into a <*Dep*> into a <*Dep*> into a (sub-)f-structure of type <*PoS1*> and with <*Lemma1*> as its PRED

Table 1: Feature templates used for semi-automatic feature construction for parse disambiguation

of German LFG parses. Many of these features are inspired by studies on “soft” syntactic constraints, which are most often formulated within an OT framework (Aissen, 2003; Bresnan et al., 2001), but can also be captured as features of more general probabilistic models (Snider and Zaenen, 2006). Table 1 gives a description of the main types of features used in parse disambiguation.

The evaluation of the log-linear model for parse disambiguation is described in more detail in Forst (2007a) and Forst (2007b), so here we will be brief. The model is trained on 8,881 partially labelled structures and tested on a test set of 1,497 sentences (with 371 sentences held out to fine-tune the log-linear model parameters). Table 2 gives a summary of the results broken down by dependency. The overall F-score is significantly better with the disambiguation model that includes the linguistically motivated additional features than the disambiguation model that relies on the XLE template-based properties only. Overall error reduction increases from 34.5% to 51.0%.

3.2 Feature Design for Realisation Ranking

Most traditional approaches to stochastic realisation ranking involve applying language model n-gram statistics to rank alternatives. However, n-grams alone are often not a good enough measure for ranking candidate strings. For example, for the f-structure associated with the string *Verheugen habe die Worte des Generalinspekteurs falsch interpretiert.* (‘Verheugen had wrongly interpreted the words of the inspector general’.), 144 strings can be generated. The original string is ranked 7th among all candidate strings by our language model. There are several features in the input f-structure that we can use to improve the ranking of the desired string. The following features could be useful: (1) Linear order of functions (SUBJ generally precedes OBJ), (2) Adjunct position (sentence beginning, distance from the verb, etc.), (3) Partial VP fronting (generally marked and thus dispreferred).

(2)	Verheugen	habe	die	Worte	des	Generalinspekteurs	falsch
	Verheugen	had	the	words	the-GEN	inspector-general	wrongly
	interpretiert.						
	interpreted.						

‘Verheugen had mis-interpreted the words of the inspector-general.’

	upper	stoch. select.		stoch. select.		lower
grammatical relation/	bound	all properties		templ.-based pr.		bound
morphosyntactic feature	F-sc.	F-sc.	err. red.	F-sc.	err. red.	F-sc.
all	85.50	83.01	51.0	82.17	34.5	80.42
PREDs only	79.36	75.74	46.5	74.69	31.0	72.59
`app` (close apposition)	63	60	63	61	75	55
`app_cl` (appositive clause)	53	53	100	52	86	46
`cc` (comparative complement)	28	19	-29	19	-29	21
`cj` (conjunct of coordination)	70	68	50	67	25	66
`da` (dative object)	67	63	67	62	58	55
`det` (determiner)	92	91	50	91	50	90
`gl` (genitive in specifier position)	89	88	75	88	75	85
`gr` (genitive attribute)	88	84	56	84	56	79
`mo` (modifier)	70	63	36	62	27	59
`mod` (non-head in compound)	94	89	29	89	29	87
`name_mod` (non-head in compl. name)	82	80	33	81	67	79
`number` (number as determiner)	83	81	33	81	33	80
`oa` (accusative object)	78	75	77	69	31	65
`obj` (argument of prep. or conj.)	90	88	50	87	25	86
`oc_fin` (finite clausal object)	67	64	0	64	0	64
`oc_inf` (infinite clausal object)	83	82	0	82	0	82
`op` (prepositional object)	57	54	40	54	40	52
`op_dir` (directional argument)	30	23	13	23	13	22
`op_loc` (local argument)	59	49	29	49	29	45
`pd` (predicative argument)	62	60	50	59	25	58
`pred_restr` (lemma of nom. adj.)	92	87	62	84	38	79
`quant` (quantifying determiner)	70	68	33	68	33	67
`rc` (relative clause)	74	62	20	59	0	59
`sb` (subject)	76	73	63	71	38	68
`sbp` (logical subj. in pass. constr.)	68	63	62	61	46	55
`case`	87	85	75	83	50	79
`comp_form` (complementizer form)	74	72	0	74	100	72
`coord_form` (coordinating conj.)	86	86	100	86	100	85
`degree`	89	88	50	87	0	87
`det_type` (determiner type)	95	95	–	95	–	95
`fut` (future)	86	86	–	86	–	86
`gend` (gender)	92	90	60	89	40	87
`mood`	90	90	–	90	–	90
`num` (number)	91	89	50	89	50	87
`pass_asp` (passive aspect)	80	80	100	79	0	79
`perf` (perfect)	86	85	0	86	100	85
`pers` (person)	85	84	83	82	50	79
`pron_form` (pronoun form)	73	73	–	73	–	73
`pron_type` (pronoun type)	71	70	0	71	100	70
`tense`	92	91	0	91	0	91

Table 2: F-scores (in %) in the 1,497 TiGer DB examples of our test set

```
"Verheugen habe die Worte des Generalinspekteurs falsch interpretiert

 PRED     'interpretieren<[1:Verheugen], [106:Wort]>'
              PRED 'falsch<[279-SUBJ:pro]>'
                   PRED      'pro'
 ADJUNCT {    SUBJ NTYPE     [NSYN pronoun]
                   PRON-TYPE null
          279 ATYPE adverbial, DEG-DIM pos, DEGREE positive }
          PRED       'Wort'
                     PRED  'Inspekteur'
                     MOD   {-12[PRED 'General']}
                     NTYPE NSEM [COMMON count]
          ADJ-GEN          NSYN common
                     SPEC  DET PRED     'die'
                               DET-TYPE def
 OBJ             229 CASE gen, GEND masc, NUM sg, PERS 3
          NTYPE      NSEM [COMMON count]
                     NSYN common
          SPEC       DET PRED     'die'
                         DET-TYPE def
      106 CASE acc, GEND neut, NUM pl, PERS 3
          PRED  'Verheugen'
 SUBJ     NTYPE NSEM [PROPER [PROPER-TYPE name]]
                NSYN proper
        1 CASE nom, NUM sg, PERS 3
 TNS-ASP [MOOD subjunctive, PERF +_, TENSE pres]
21 CLAUSE-TYPE decl, PASSIVE -, STMT-TYPE decl, VTYPE main
```

Figure 3: F-structure for (2)

```
1. Falsch interpretiert habe die Worte des
      Generalinspekteurs Verheugen.
2. Falsch interpretiert habe die Worte des
      Generalinspekteures Verheugen.
3. Die Worte des Generalinspekteurs falsch
      interpretiert habe Verheugen.
5. Die Worte des Generalinspekteurs habe Verheugen
      falsch interpretiert.
7. Verheugen habe die Worte des Generalinspekteurs
      falsch interpretiert.
11. Falsch interpretiert habe Verheugen die Worte des
      Generalinspekteurs.
15. Die Worte des Generalinspekteurs interpretiert habe
      Verheugen falsch.
17. Interpretiert habe die Worte des Generalinspekteurs
      Verheugen falsch.
```

Using the feature templates presented in Riezler et al. (2002), Riezler and Vasserman (2004) and Forst (2007a), we construct a list of 186,731 features that can be used for training our log-linear model.[3] Out of these, only 1,471 actually occur in our training data. In the feature selection process of our training regime (see Subsection 4.2), 360 features are chosen as the most discriminating; these are used to rank alternative solutions when the model is applied. Table 3 gives a list of the types of features used for realisation ranking.

We divide the features into three distinct categories: language model features (LM), c-structure features (CF) and additional features (AF). For realisation ranking, we do not use f-structure features, since the f-structure is given in the input. Examples of c-structure features are the number of times a particular category label occurs in a given c-structure, the number of children the nodes of a particular category have, or the number of times one particular category label dominates another. Examples of features that take both c- and f-structure information into account are the relative order of grammatical functions (e.g. 'SUBJ precedes OBJ'). As in Velldal and Oepen (2005), we incorporate the language model score associated with the string realisation for a particular structure as a feature in our model.

[3] For technical reasons, we were not able to include all the additional features we would have liked to include. For example, we could not use features that capture the relative order of ADJUNCT PPs headed by given prepositions.

Name of feature template and parameters	Explanation
C-structure Features	
cs_label <cat>	counts number of occurrences of category <*cat*> in the c-structure
cs_right_branch	counts number of right children
cs_num_children <cat>	counts number of children of all nodes of category <*cat*>
cs_adjacent_label <cat1> <cat2>	counts the number of <*cat1*> nodes that immediately dominate <*cat2*> nodes
cs_sub_label <cat1> <cat2>	counts the number of <*cat1*> nodes that (not necessarily immediately) dominate <*cat2*> nodes
cs_embedded <cat> <Depth>	counts the number of <*cat*> nodes that dominate (at least) <*Depth*> other <*cat*> nodes
cs_conj_nonpar <Depth>	counts the number of coordinated c-structures that are not parallel at <*Depth*> levels under the coordinated constituent
Additional Linguistically Motivated Features	
ADD-PROP PATH	counts given instantiations of functional uncertainty paths
ADD-PROP PRECEDES <GF1> <GF2>	counts the number of times a <*GF1*> precedes a <*GF2*> of the same (sub-)f-structure
ADD-PROP PRECEDES <Lemma> <GF1> <GF2>	counts the number of times a <*GF1*> subcategorised for by a PRED <*Lemma*> precedes a <*GF2*> subcategorised for by the same PRED
DISTANCE-TO-ANTECEDENT %X	distance between a relative clause and its antecedent
Language Model Features	
GEN_NGRAM_SCORE %X	3-gram language model score assigned to the generated sentence
GEN_WORD_COUNT %X	number of words in the generate sentence

Table 3: Feature templates used for semi-automatic feature construction in realisation ranking

4 Realisation Ranking Experimental Setup

4.1 Data

We use the TIGER Treebank (Brants et al., 2002) to train and test our model. It consists of just over 50,000 annotated sentences of German newspaper text. The sentences have been annotated with morphological and syntactic information in the form of functionally labelled graphs that may contain crossing and secondary edges.

We split the data into training and test data using the same data split as in Forst (2007a), i.e. sentences 8,001–10,000 of the TIGER Treebank are reserved for evaluation. Within this section, we have 422 TIGER-annotation-compatible f-structures, which are further divided into 86 development and 336 test structures. We use the development set to tune the parameters of the log-linear model. Of the 86 heldout sentences and the 336 test sentences, 78 and 323 respectively are of length >3 and hence are actually used for our final evaluation.

For training, we build a symmetric treebank of 8,609 packed c/f-structure representations in a similar manner to Velldal et al. (2004). We do not include structures for which only one string is generated, since the log-linear model for realisation ranking cannot learn anything from them. The symmetric treebank was established using the following strategy:

1. Parse the input sentence from the TIGER Treebank.
2. Select all of the analyses that are compatible with the TIGER Treebank annotation.
3. Of all the TIGER-compatible analyses, choose the most likely c-/f-structure pair according to the log-linear model for parse disambiguation.
4. Generate from the f-structure part of this analysis.
5. If the input string is contained in the set of output strings, add this sentence and all of its corresponding c-/f-structure pairs to the training set. The pair(s) that correspond(s) to the original corpus sentence is/are marked as the intended structure(s), while all others are marked as unintended.

Theoretically all strings that can be parsed should be generated by the system, but for reasons of efficiency, punctuation is often not generated in all possible positions, therefore resulting in an input string not being contained in the set of output strings. Whenever this is the case for a given sentence, the c-/f-structure pairs associated with it cannot be used for training. Evaluation can be carried out regardless of this problem, but it has to be kept in mind that the original corpus string cannot be generated for all input f-structures. In our test set, it is generated for only 62% of them.

Tables 4 and 5 give information about the ambiguity of the training and test data. For example, in the training data there are 1,206 structures with more than 100 string realisations. Most of the training and test structures have between 2 and 50 possible (and grammatical) string realisations. The average sentence length of the training data is 11.3 and it is 12.8 for the test data.[4] The tables also show that the structures with more potential string realisations correspond to longer sentences than the structures that are less ambiguous when generating.

4.2 Training

We train a log-linear model that maximises the conditional probability of the observed corpus sentence given the corresponding f-structure. The model is trained in a (semi-)supervised fashion on the 8,609 (partially) labelled structures of our training set using the `cometc` software provided with the XLE platform. `cometc` performs maximum likelihood estimation on standardised feature values and offers

[4] This is lower than the overall average sentence length of roughly 16 in TIGER because of the restriction that the structure produced by the reversible grammar for any TIGER sentence be compatible with the original TIGER graph. As the grammar develops further, we hope that longer sentences can be included in both training and test data.

String Realisations	# of Strings	Average # of Words
> 100	1206	18.3
$\geq 50, < 100$	709	14.3
$\geq 10, < 50$	3029	11.8
$> 1, < 10$	3665	7.6
Total	8609	11.3

Table 4: Number of structures and average sentence length according to ambiguity classes in the training set

String Realisations	# of Strings	Average # of Words
> 100	61	23.7
$\geq 50, < 100$	26	13.5
$\geq 10, < 50$	120	11.6
$> 1, < 10$	129	7.8
Total	336	12.8

Table 5: Number of structures and average sentence length according to ambiguity classes in the test set

several regularisation and/or feature selection techniques. We apply the combined method of incremental feature selection and l_1 regularisation presented in Riezler and Vasserman (2004), the corresponding parameters being adjusted on our heldout set.

For technical reasons, the training was carried out on unpacked structures. However, we hope to be able to train and test on packed structures in the future, which will greatly increase efficiency.

5 Analysis of Results by Feature Type

Given the three distinct types of features in Table 3, we carry out a number of smaller experiments on our heldout set, only training on a subset of features each time. This is done in order to see what effect each group of features has on the overall performance of the log-linear model, and to see what combination of feature types performs best. We evaluate the most likely string produced by our system in terms of two metrics: **exact match** and **BLEU score** (Papineni et al., 2002). Exact match measures what percentage of the most probable strings are exactly identical to the string from which the input structure was produced. BLEU score is a more relaxed metric which measures the similarity between the selected string realisation and the observed corpus string.

The results are given in Table 6. The results show that training on c-structure features alone achieves the worst exact match and BLEU score. This is possibly due to the nature of the c-structure features used, which were initially designed for parse disambiguation. Therefore, future work is required to investigate whether

	Exact Matches (%)	BLEU Score
Baseline	24	0.7291
LM	23	0.7034
CF	22	0.6824
AF	23	0.7060
LM + CF	27	0.7529
LM + AF	33	0.7705
CF + AF	33	0.7303
LM + CF + AF	35	0.7808

Table 6: Results on the heldout set of training only on subsets of feature types

c-structure features more appropriate for realisation ranking can be devised. Training on language model features alone, or additional features alone, also does not achieve very high results. Surprisingly, the log-linear model trained on language model features alone performs worse than the baseline language model applied directly. We cannot be sure what causes this, but one possible reason is that the number of words is taken into account as a feature in the log-linear model, while the language model does not use this feature. Another reason might be that because we are working with unpacked structures, we loose a lot of precision with the log-linear model, so that often more than one solution is ranked highest. When this happens, we choose a solution at random, which may not always reflect the original language model scores. This problem generally does not arise with the language model which assigns more precise scores. However, the combination of language model features and additional features is the one that leads to the greatest improvement in exact match and BLEU scores. It achieves a BLEU score of 0.7705, which is only a little less than the best result achieved by combining all three feature types. The results thus suggest that the language model features and the additional features contribute most to the model, while the c-structure features contribute less. Nevertheless, the c-structure features are beneficial, since the best results are achieved by combining the three feature types.

6 Final Evaluation

We first rank the generator output with a language model trained on the Huge German Corpus (a collection of 200 million words of newspaper and other text) using the SRILM toolkit. The results are given in Table 7, achieving exact match of 27% and BLEU score of 0.7306 on the test set. In comparison to the results reported by Velldal and Oepen (2005) for a similar experiment on English, these results are markedly lower, presumably because of the relatively free word order of German.

We then rank the output of the generator with our log-linear model as described

Exact Match Upper Bound	62%
Exact Matches	27%
BLEU score	0.7306

Table 7: Results on the test set with the language model

above and give the results in Table 8. There is a noticeable improvement in quality. Exact match increases from 27% to 37%, which corresponds to an error reduction of 29%,[5] and BLEU score increases from 0.7306 to 0.7939.

Exact Match Upper Bound	62%
Exact Matches	37%
BLEU score	0.7939

Table 8: Results on the test set with the log-linear model

There is very little comparable work on realization ranking for German. Gamon et al. (2002) present work on learning the contexts for a particular subset of linguistic operations; however, no evaluation of the overall system is given. The work that comes closest to ours is that of Filippova and Strube (2007) who present a two-step algorithm for determining constituent order in German. They predict the surface order of the major non-verbal constituents in a German sentence, given its dependency representation. They do not predict the position of the verb or the order within constituents, nor do they generate word forms from lemmas followed by morphological tags. Training and evaluation is carried out on Wikipedia data and their algorithm outperforms four baseline models. They achieve an exact match metric of 61%, i.e. for 61% of their corpus sentences, the order of the major constituents generated matches the original order. At first sight, this result looks very superior to the exact match metric of 37% we achieve, but when we take into account that our upper bound for exact match is 62% as opposed to theirs of 100%, the results become comparable. Furthermore, it has to be taken into account that many of the mismatches that we are penalized for result from generated word forms that diverge from the forms in the corpus, a problem Filippova and Strube (2007) do not deal with at all. This being said, this recent publication provides us with many useful ideas of how to design further features relevant for the task of realization ranking.

[5] Remember that the original corpus string is generated from only 62% of the f-structures of our test set, which fixes the upper bound for exact match at 62% rather than 100%.

7 Error Analysis

We had initially expected the increase in BLEU score to be greater than 0.0633, since German is far less configurational than English and therefore we thought the syntactic features used in the log-linear model would play an even greater role in realisation ranking. However, in our experiments, the improvement was only slightly greater than the improvement achieved by Velldal and Oepen (2005). In this section, we present some of the more common errors that our system still produces.

Word Choice Often there is more than one surface realisation for a particular sequence of morphological tags. Sometimes the system chooses an incorrect form for the sentence context, and sometimes it chooses a valid, though marked or dispreferred, form. For example, from the structure in Figure 3, the system chooses the following string as the most probable.

Verheugen	habe	die	**Wörter**	des	**Generalinspekteures**	falsch	interpretiert.
Verheugen	had	the	**words**	of the	**inspector-general**	wrongly	interpreted.

There are two mismatches in this output string with respect to the original corpus string. In the first case the system has chosen *Wörter* as the surface realisation for the morpheme sequence *Wort+NN.Neut.NGA.Pl* rather than the, in this case, correct form *Worte*. The difference between the two realisations is semantic; they both translate as *words* in English, but *Worte* is a more abstract concept referring to a meaningful stretch of text or speech, whereas *Wörter* is more concrete and can refer, e.g., to the words in a dictionary.

In the second (less critical) case, the system has chosen to mark the genitive case of *Generalinspekteur* with *es* rather than the *s* that is in the original corpus sentence. This is a relatively frequent alternation that is difficult to predict, and there are other similar alternations in the dative case, for example.

The second case is merely a phonological variation and does not alter the projected meaning. The first case, however, is completely incorrect and should not be generated. To correctly generate only *Worte* in this instance, the morphological component of the system needs to be improved. The most obvious solution is to have different lemmas for the different senses of (the plural of) *Wort*. In order to improve the selection of the most natural variant of the genitive and dative markings, one solution might be to try and learn the most frequent variant for a given lemma based on corpus statistics.

Placement of adjuncts Currently, there is no feature that captures the (relative) location of particular types of adjuncts. In German, there is a strong tendency for temporal adjuncts to appear early in the sentence, for example. Since the system was not provided with data from which it could learn this generalisation, it generated output like the following:

Frauenärzte haben die Einschränkung umstrittener Antibabypillen
Gynaecologists have the restriction controversial birth control pills
wegen erhöhter Thrombosegefahr **am Dienstag** kritisiert.
because of increased risk of thrombosis **on Tuesday** critisised.
‘Gynaecologists criticised the restriction on controversial birth control pills due to increased risk of thrombosis on Tuesday.’

where the temporal adjunct *on Tuesday* was generated very late in the sentence, resulting in a highly marked utterance.

Discourse Information In many cases, the particular subtleties of an utterance can only be generated using knowledge of the context in which it occurs. For example, the following sentence appears in our development corpus:

Israel stellt den Friedensprozess nach Rabins Tod nicht in Frage
Israel puts the peace process after Rabin’s death not in question
‘Israel does not challenge the peace process after Rabin’s death’

Our system generates the string:

Nach Rabins Tod stellt Israel den Friedensprozess nicht in Frage.
After Rabin’s death puts Israel the peace process not in question.

which, taken on its own, gets a BLEU score of 0. The sentence produced by our system is a perfectly valid sentence and captures essentially the same information as the original corpus sentence. However, without knowing anything about the information structure within which this sentence is uttered, we have no way of telling where the emphasis of the sentence is.

7.1 Additional Features

It is clear from the errors outlined above that further features are required in order to achieve improved realisation ranking. For example, a feature is required that captures the placement of adjunct types so that the tendency of temporal adjuncts to appear before locatives is captured correctly. Including information structure features is also necessary for the improvement of the overall system. The work described in this paper is part of a much larger project, and future research is already planned to integrate information structure into the surface realisation process. It is yet to be seen whether these features could also be useful in parse disambiguation.

8 Conclusion

In this paper, we have presented the features used in log-linear models for parse disambiguation and realisation ranking for a large-scale German LFG. We train both parse disambiguation and realisation ranking systems on over 8,000 partially labelled structures and test on a heldout section of almost 2,000 sentences. In the parse disambiguation experiments, we achieve an increase in error reduction of 16.5 points with the additional features over the simple template-based features used in the parse disambiguation of English (Forst, 2007b). In the task of realisation ranking, we achieve an increase in exact match score from 27% to 37% and an increase in BLEU score from 0.7306 to 0.7939 over a baseline language model trained on a large corpus of German. We thus show that linguistically motivated features that were initially developed for the task of parse disambiguation carry over rather well to the task of realisation ranking. Despite these encouraging results, an error analysis of the realisation ranking shows that further features are required by the log-linear model in order to improve the quality of the output strings. It is also unclear how suitable the BLEU score as an evaluation metric is, and further research into other metrics and a comparison with human evaluation is necessary.

References

Aissen, Judith. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* .

Brants, Sabine, Dipper, Stefanie, Hansen, Silvia, Lezius, Wolfgang and Smith, George. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

Bresnan, Joan, Dingare, S and Manning, Christopher. 2001. Soft Constraints Mirror Hard Constraints. In *LFG 2001*.

Clark, Stephen and Curran, James R. 2004. Parsing the WSJ using CCJ and Log-Linear Models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, Barcelona, Spain.

Crouch, Dick, Dalrymple, Mary, Kaplan, Ron, King, Tracy, Maxwell, John and Newman, Paula. 2006. XLE Documentation. Technical Report, Palo Alto Research Center, CA.

Dipper, Stefanie. 2003. *Implementing and Documenting Large-scale Grammars – German LFG*. Ph. D.thesis, IMS, University of Stuttgart.

Filippova, Katja and Strube, Michael. 2007. Generating Constituent Order in German Clauses. In *Proceedings of the 45th Annual Meeting of the Association of*

Computational Linguistics, pages 320–327, Prague, Czech Republic: Association for Computational Linguistics.

Forst, Martin. 2007a. *Disambiguation for a Linguistically Precise German Parser*. Ph. D.thesis, University of Stuttgart.

Forst, Martin. 2007b. Filling Statistics with Linguistics – Property Design for the Disambiguation of German LFG Parses. In *Proceedings of the ACL Workshop on Deep Linguistic Processing,* Prague, Czech Republic.

Frank, Anette, King, Tracy Holloway, Kuhn, Jonas and Maxwell, John T. 2001. Optimality Theory Style Constraint Ranking in Large-Scale LFG Grammars. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality Theoretic Syntax*, pages 367–397, Stanford, CA: CSLI Publications.

Gamon, Michael, Ringger, Eric, Corston-Oliver, Simon and Moore, Robert. 2002. Machine-learned contexts for linguistic operations in German sentence realization. In *Proceedings of ACL 2002,* pages 25–32, Philadelphia, PA.

Malouf, Robert and van Noord, Gertjan. 2004. Wide Coverage Parsing with Stochastic Attribute Value Grammars. In *Proceedings of the IJCNLP-04 Workshop "Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses"*.

Miyao, Yusuke and Tsujii, Jun'ichi. 2002. Maximum entropy estimation for feature forests. In *Proceedings of the Human Language Technology Conference*, San Diego, CA.

Nakanishi, Hiroko, Miyao, Yusuke and Tsujii, Jun'ichi. 2005. Probabilistic models for disambiguation of an HPSG-based chart generator. In *Proceedings of IWPT 2005*.

Papineni, Kishore, Roukos, Salim, Ward, Todd and Zhu, WeiJing. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002,* pages 311–318, Philadelphia, PA.

Riezler, Stefan, King, Tracy Holloway, Kaplan, Ronald M., Crouch, Richard, Maxwell, John T. and Johnson, Mark. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of ACL 2002*, Philadelphia, PA.

Riezler, Stefan and Vasserman, Alexander. 2004. Gradient feature testing and l_1 regularization for maximum entropy parsing. In *Proceedings of EMNLP'04*, Barcelona, Spain.

Rohrer, Christian and Forst, Martin. 2006. Improving coverage and parsing quality of a large-scale LFG for German. In *Proceedings of LREC-2006*, Genoa, Italy.

Snider, Neil and Zaenen, Annie. 2006. Animacy and Syntactic Structure: Fronted NPs in English. In *Intelligent Linguistic Architectures – Variations on Themes by Ronald M. Kaplan*, CSLI Publications.

Toutanova, Kristina, Manning, Christopher D., Shieber, Stuart M., Flickinger, Dan and Oepen, Stephan. 2002. Parse Disambiguation for a Rich HPSG Grammar. In *First Workshop on Treebanks and Linguistic Theories (TLT2002)*, pages 253–263.

van Noord, Gertjan. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister and Patrick Watrin (eds.), *TALN06. Verbum Ex Machina. Actes de la 13e conférence sur le traitement automatique des langues naturelles*, pages 20–42, Leuven, Belgium.

Velldal, Erik and Oepen, Stephan. 2005. Maximum Entropy Models for Realization Ranking. In *Proceedings of the 10th MT Summit*, pages 109–116, Thailand.

Velldal, Erik, Oepen, Stephan and Flickinger, Dan. 2004. Paraphrasing Treebanks for Stochastic Realization Ranking. In *Proceedings of TLT Workshop*, pages 149–160, Tübingen, Germany.

Yoshimura, H, Masuichi, Hiroshi, Ohkuma, Tomoko and Sugihara, Daigo. 2003. Disambiguation of F-Structures based on Support Vector Machines. *Information Processing Society of Japan SIG Notes* pages 75–80.

USING VERY LARGE CORPORA TO DETECT RAISING AND CONTROL VERBS

Grzegorz Chrupała and Josef van Genabith
National Center for Language Technology
Dublin City University

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

The distinction between raising and subject-control verbs, although crucial for the construction of semantics, is not easy to make given access to only the local syntactic configuration of the sentence. In most contexts raising verbs and control verbs display identical superficial syntactic structure. Linguists apply grammaticality tests to distinguish these verb classes. Our idea is to learn to predict the raising-control distinction by simulating such grammaticality judgments by means of pattern searches. Experiments with regression tree models show that using pattern counts from large unannotated corpora can be used to assess how likely a verb form is to appear in raising vs. control constructions. For this task it is beneficial to use the much larger but also noisier Web corpus rather than the smaller and cleaner Gigaword corpus. A similar methodology can be useful for detecting other lexical semantic distinctions: it could be used whenever a test employed to make linguistically interesting distinctions can be reduced to a pattern search in an unannotated corpus.

1 Introduction

In this paper we investigate to what degree very large unannotated corpora can be useful in acquiring detailed specifications of verbal subcategorization: specifically we attempt the task of detecting *raising* and *subject control* verbs.

The task of data-driven lexical acquisition is interesting from at least two points of view. First it can shed light on the process of lexical learning from linguistic input in humans. Second, it is relevant for Natural Language Engineering, where detailed information on subcategorization requirements of lexical items is useful for parsing.

Distinguishing between raising and control verbs is a small but interesting and seldom investigated aspect of automatically acquiring verbal lexical resources. In this paper we propose to make a somewhat non-standard use of large unannotated corpora to aid lexical acquisition. We extract features associated with raising and control verbs in a large unannotated corpus, learn a model which distinguishes the two classes using a small annotated (gold) corpus, and then verify how well our model predicts the two classes in a held-out portion of the gold corpus.

The errors our model makes may be partly be due to the limitations of the method we use, i.e. the features we extract or the learning mechanism we employ. More interestingly, they may also reveal mistakes or omissions in the small gold manually constructed resource when contrasted with usages in large amounts of naturally occurring data. In Section 6 we discuss those issues in more detail.

The structure of the paper is as follows: In Section 2 we briefly describe the raising-control distinction and its treatment in LFG. In Section 3 we briefly discuss previous work. In Section 4 we describe the methodology and resources used, while in Section 5 we present the experimental evaluation. Finally in Section 6 we discuss the implications of our results and present our conclusions.

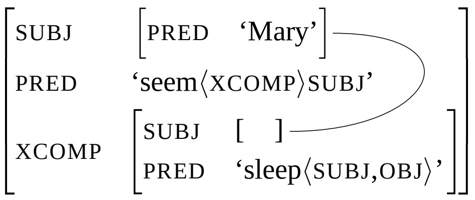


Figure 1: F-structure for *Mary seems to sleep* (raising - functional control)

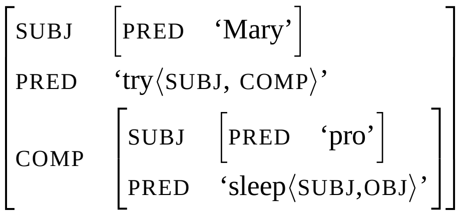


Figure 2: F-structure for *Mary tries to sleep* (anaphoric control)

2 Raising and control verbs

In English *raising* verbs are verbs such a *seem*. They require a syntactic subject which does not correspond to a semantic argument.

Subject control verbs are matrix verbs such as *try* one of whose arguments is shared with the the subordinate verb’s SUBJ. In Dalrymple (2001) they receive a treatment in terms of obligatory anaphoric control, where the COMP’s SUBJ’s PRED value is bound to the matrix verb’s SUBJ (see Fig. 2).

In Bresnan (2001) subject control verbs are treated in terms of functional control similar to raising verbs (see Fig. 3). In this type of analysis the only thing distinguishing raising constructions from control constructions is the subcat frame (semantic form): the fact that the subject argument is not a semantic argument of the raising verb is indicated notationally by putting it outside the angle brackets: ‘seem⟨XCOMP⟩SUBJ’.

Whichever analysis one adopts, the distinction between raising and control verbs is important as it affects meaning: the predicate encoded by *seems* is unary whereas the one encoded by *try* is binary. Thus it is crucial when constructing the semantic argument structure for a verb with a non-finite complement.

There are a number of constructions which distinguish between those two verb classes:

(1) a. It seemed to rain.

b. There seems to be a problem.

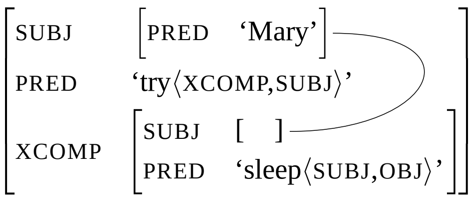


Figure 3: F-structure for *Mary tries to sleep* (functional control)

c. Did she leave? *She seemed.

(2) a. * It tried to rain.
b. * There tried to be a problem.
c. Did she leave? She tried.

English raising verbs appear with dummy subjects as in examples (1a) and (1b). They do not admit VP drop (1c). Control verbs exhibit the opposite behavior as shown in (2).

3 Previous work

In most contexts, raising verbs and control verbs display identical superficial syntactic structure. Many resources meant to provide training and evaluation material for data-driven computational methods do not encode the raising-control distinction in any way; examples include the Penn Treebank (Marcus et al., 1994), or the PARC 700 Dependency Bank (King et al., 2003). O'Donovan et al. (2005) implement a large scale system for acquiring LFG semantic forms using the Penn Treebank but do not differentiate between frames for raising and control verbs.

Briscoe and Carroll (1997) mention in passing that the fact that argument slots of different subcategorization frames for the same verb share the same semantic restrictions could be used to learn about alternations the verb participates in and thus make inferences about raising and control facts. However to our knowledge neither they nor other researchers have followed on these ideas and there have been no studies specifically focusing on acquiring the raising/control distinction.

In the following sections we investigate whether frequency counts from very large corpora can be used to reliably distinguish those two verb classes.

4 Methods

The raising-control distinction is not easy to make given access to only the local syntactic configuration of the sentence. However, speakers have little difficulty

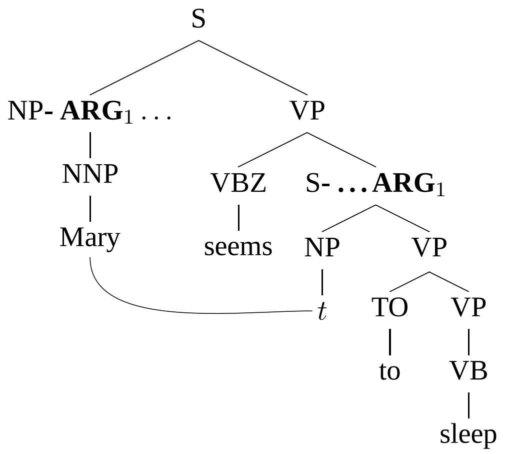


Figure 4: Propbank-style annotation for the raising construction with *seem*

in applying grammaticality tests such as those in example (1) to distinguish these verb classes. Our idea is to simulate making those grammaticality judgements. We hypothesize that the absence of evidence approximates evidence of absence: a simple construction, if it is grammatical, is bound to show up in a sufficiently large amount of naturally occurring language data. So a grammaticality test reduces to a pattern search in a corpus.

There are two complicating factors:

- the need for a very large corpus to minimize the chance that the absence of matches is accidental rather than systematic

- the inevitable presence of noise in the form of false positive matches, for example caused by misspellings, interlinguistic interference or automatically generated pseudo-language.

These two factors have to be traded off against each other: a corpus with carefully selected text samples is likely to be mostly free of noise but will probably be too small to avoid false negatives. Conversely, a terabyte-scale corpus will almost inevitably contain some proportion of false positives due to noise.

We use two types of corpora in our study. First we use a relatively small corpus annotated with syntactic structure and semantic roles, namely the English Propbank (Palmer et al., 2005). This contains the same text as the English Penn Treebank. Each verb form is annotated with the labeled semantic arguments it governs. The semantic roles are to a large extent verb-specific and are numbered as ARG$_0$ through ARG$_5$. In general ARG$_0$ can be said to correspond to a prototypical Agent (Dowty, 1991) and ARG$_1$ is the prototypical Patient. The higher-numbered roles are completely verb specific and no generalizations can be made about them.

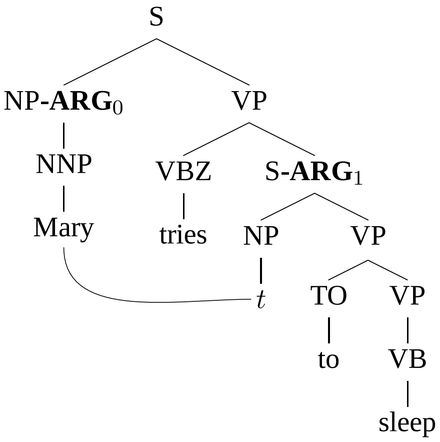


Figure 5: Propbank-style annotation for the control construction with *try*

Thanks to the information about semantic roles which Propbank annotations add to Penn treebank trees, it is possible to distinguish raising and control constructions. In Figures 4 and 5 we present the analyses that example raising and control verbs receive in Propbank. In the case of the raising construction with *seem* there is a single (discontinuous) semantic argument ARG_1. In contrast, in a control construction the verb *try* has two arguments ARG_0 and ARG_1.

We use the English Propbank to extract verb forms which appear at least 3 times in contructions with non-finite complements.[1]. For each verb form we also extract the form of the complement (to-infinitive or gerund). To each verb form v we assign the maximum-likelihood estimate of its *raising probability* $P_R(v)$, i.e. the proportion of times it appears in raising constructions. We take the presence of the ARG_0 semantic argument to indicate a subject control construction and its lack to indicate a raising construction. The resulting list of 120 verbs forms is randomly divided into a training set and test set of equal sizes.

The second type of resource we use is a large-scale unannotated corpus of English text. We experiment with two such corpora Gigaword (Graff, 2003) (1.7 billion words of newswire) and the English web pages indexed by Yahoo!.

Those large corpora are used to extract frequencies of occurrence of the verb forms in context that are indicative of the degree to which they can appear in raising contructions (i.e. $P_R(v)$). From those frequency counts we derive features used to train regression models that will predict $P_R(v)$ for each verb form.

There are a number of choices as to how to extract the most informative occurrence frequency counts. In this study we decided to try to mimic grammaticality

[1] The extraction is not 100% reliable, due to annotation errors in the Penn Treebank. For example in several cases the participle use of *said* as in *X is said to Y* is mistagged as past tense, which is why *said* appears among our 120 verb forms.

tests used by linguists in distinguishing between raising and control constructions. The assumption which enables us to approximate grammaticality judgements by corpus searches is that any simple grammatical construction is very likely to occur in a sufficiently large corpus. There are some important qualifications that need to be made about its validity. The construction in question should be as simple as possible and ideally contain high frequency lexical items. The semantics associated with it should be plausible. The search pattern itself should be possible to run on un-annotated data and still be resistant to noise.

Those are quite strict prerequisites and it can be hard to build search patterns that satisfy all of them. For example it is challenging to come up with a template based on the grammaticality test in (1a) and (2a) which will not suffer from some shortcomings: *it X to rain* depends on the lexical item *rain* which is not high frequency enough for most corpus sizes. Even in combination with the most common raising verb, *it seemed to rain* only occurs in two unique sentences in Gigaword. For the test in (1c) and (2c), with access just to un-annotated data it would very hard to detect those sentence-final strings such as "seemed" which are VP-drop. An additional complication is that Web search indexes such as Yahoo! do not typically include punctuation which makes it impossible to detect sentence boundaries. Thus in the experiments described below we use the search patterns based on the test b vs b, which we deemed the most robust.

For each verb form V tested, we build patterns using the following templates:

(3) a. there V to be
    b. there V being

(4) a. V to be
    b. V being

Version (a) or (b) is chosen depending on the complement type the verb takes. String (3) is our test pattern which is meant to check whether verb form X is grammatical in raising constructions. String (4) is the background frequency of verb form V with a non-finite complement. The ratio of (3) to (4) gives us the maximum likelihood estimate of the probability of dummy-*there* in nonfinite complement contexts.

Gigaword contains articles or portions of articles that are repeated: to correct for inflated counts caused by this we remove duplicate lines from the corpus in a preprocessing step. We match patterns by ignoring upper/lower case.

In the case of the Web we use the Yahoo! search API – we restrict the search to English-language pages, thus relying on Yahoo!'s language-detection method, and use the *total result available* number as our frequency count, thus trusting the estimate Yahoo! provides. All the web frequency counts were collected on a single day (July 1 2007) and stored to ensure consistency between experiments.

5 Experiments

We performed experiments with two corpora: Gigaword and the Web. We search for occurrences of the pattern strings (3) and (4) and for each verb form we gather the following scores:

- $C_1(v) =$ frequency of pattern (3)
- $C_2(v) =$ frequency of pattern (4)
- $C_1(v)/C_2(v)$

5.1 Models

We experiment with two baselines and a regression tree model to learn to predict $P_R(v)$ from training examples. As a metric for evaluating the quality of the models, both during cross-validating and for final evaluation, we use the Mean Squared Error (MSE). For the list of gold scores $\mathbf{v}$ and the list of predicted scores $\hat{\mathbf{v}}$ for n verb forms, this metric is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(\mathbf{v}_i - \hat{\mathbf{v}}_i)^2 \tag{5}$$

Mean This is a very simple baseline: for each verb we form predict $P_R(v)$ to be the mean $\overline{P}_R$ in the training set.

Linear regression This baseline is the linear regression model fitted to training data using $C_1(v)/C_2(v)$ as the sole explanatory variable. The model for Gigaword data is $P_R = 13.2936 \times C_1(v)/C_2(v) + 0.2741$, while the Web model has the form $P_R = 11.5011 \times C_1(v)/C_2(v) + 0.2547$.

Regression tree This is the model obtained by inducing a regression tree. A regression tree is simply a type of decision tree where the response at each leaf is a real number. The tree is built using the recursive partitioning method of Breiman et al. (1984), as implemented in the *rpart* R package (Therneau et al., 2007; Therneau and Atkinson, 2000).

We chose this model because of its relative simplicity and transparency. At this stage our main goal was to gain insight from our data rather than simply maximize performance.

The algorithm starts by grouping all training examples in a single node. At each step a split (i.e. a value of one of the features) is chosen to partition the training examples at the current node T in such a way as to maximize the splitting criterion:

$$SS_T - (SS_L + SS_R) \tag{6}$$

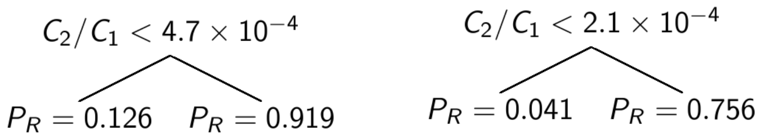

$C_2/C_1 < 4.7 \times 10^{-4}$
$P_R = 0.126$
$P_R = 0.919$
$C_2/C_1 < 2.1 \times 10^{-4}$
$P_R = 0.041$
$P_R = 0.756$

Figure 6: The regression tree model: left for Gigaword data, right for Web data

Model	Gigaword MSE	Yahoo Web MSE
Mean	0.194	0.194
Linear regression	0.165	0.164
Regression tree	**0.134**	**0.110**

Table 1: Evaluation results on the test set

SS_T is the within node sum of squares for the current node T, where y_i is the output value for the i^{th} training example at node T and $\overline{y}$ is the mean of the outputs of examples at node T:

$$SS_T = \sum_i (y_i - \overline{y})^2 \tag{7}$$

SS_L and SS_R are sums of squares for the left and right child given by the split under consideration.

The same step is applied recursively to both children nodes until the maximum number of splits is reached or no further splits are possible. For each node the predicted response is the mean of the instances in this node. The tree constructed in this fashion is then pruned using leave-one-out cross-validation in order to find the tree which minimizes Mean Squared Error.

In our experiments we start with all three features but the resulting pruned trees only use the ratio feature $C_1(v)/C_2(v)$: trees with more depth increase cross-validated error. Figure 6 shows the regression trees for both experiments. For the Gigaword tree the top node is split at $C_1(v)/C_2(v) < 4.7 \times 10^{-4}$ and for the Web tree at $C_1(v)/C_2(v) \geq 2.1 \times 10^{-4}$.

5.2 Results

In Table 1 we report the Mean Squared Error score on the test set for counts extracted from the Gigaword and the Yahoo Web achieved by the models.

Our results show that for **regression tree** the Web counts give models with lower error on test data in comparison to the Gigaword-based model.

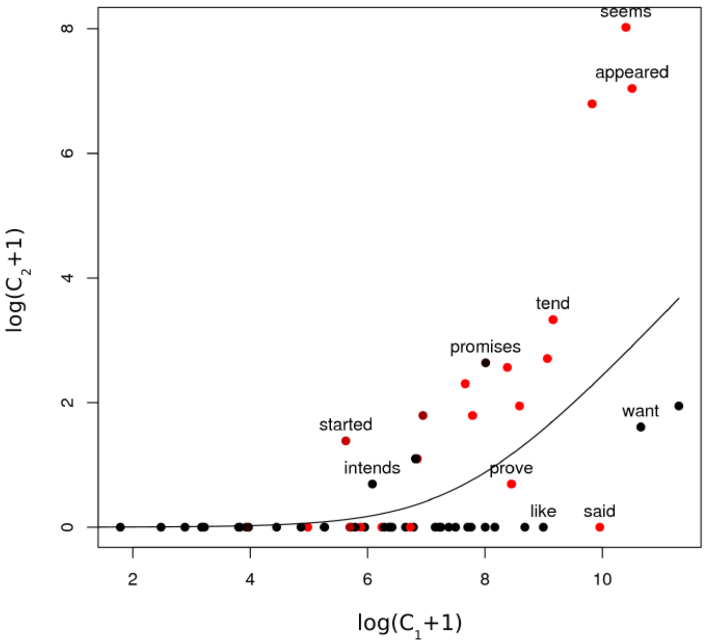


Figure 7: Results for Gigaword regression tree

Since both regression trees are of depth 2, in effect both trees partition verb forms into two classes: predominantly raising verbs and predominantly control verbs. Figures 7 and 8 illustrate how well that partition separates verb forms in the test data. Both figures plot C_2 against C_1 on a logarithmic scale. Each dot represents a verb form; the varying color indicates the following: black stands for gold $P_R(v)$ = 0 and red for $P_R(v)$ = 1, with intermediate colors encoding values between 0 and 1. The black curve on each plot separates points in the same fashion as the top node in the regression tree model, i.e. $C_2(v) = 4.7 \times 10^4 \times C_1(v)$ for the Gigaword tree and $C_2(v) = 2.1 \times 10^4 \times C_1(v)$ for the Web tree.

The complete results obtained by the regression tree models trained with the Gigaword and Web counts for the verb forms in the Propbank-derived test set are included in Tables 2 and 3. Column three shows the values of $P_R(v)$ estimated from Propbank; the following two columns show the predictions of the Gigaword model, the squared errors for that prediction, and analogous numbers for the Web model in the last two columns.

Among the 60 verb forms in the test set, the Gigaword regression tree has squared errors larger than 0.25 for 10 verb forms. The corresponding Web model has squared errors above 0.25 for 8 verb forms.

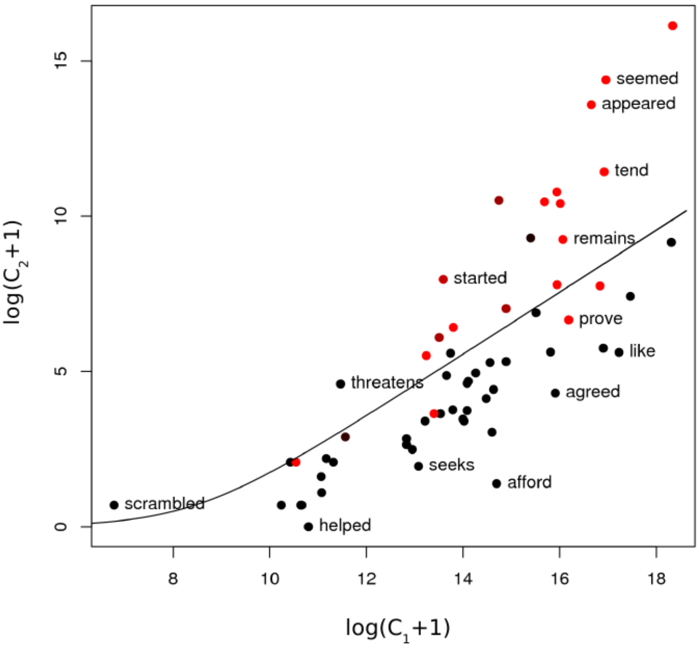


Figure 8: Results for Web regression tree

In some cases where the models disagree with the Propbank-derived gold standard they are not necessarily wrong. For example both the regression tree models give a high $P_R(promised)$ based on occurrences of strings such as *At $300 apiece there promised to be a tremendous profit in the thing* which seem genuine raising usages. However, all the uses of *promised to* in Propbank are classified as control, which results in a gold $P_R(promised) = 0$.

In our experiments we did not group all the inflected forms of each verb together – rather we treat each verb-form as a separate example. This means that we have more training and test examples; but also that there are fewer frequency counts for each individual example. Grouping the verb forms together might change our numbers somewhat but we do not expect this effect to be large.

6 Discussion

The experiments show that using pattern counts from large corpora can be used to assess how likely a verb form is to appear in raising vs. control constructions. We evaluated two simple models and showed that they perform much better than the baseline.

Table 2: Regression tree results on test set - part 1

Form	Complement	Gold P_R	Giga	Giga SE	Web	Web SE
afford	TO	0	0.126	0.0159	0.041	0.0017
agreed	TO	0	0.126	0.0159	0.041	0.0017
aims	TO	0	0.126	0.0159	0.041	0.0017
appeared	TO	1	0.919	0.0065	0.756	0.0594
attempt	TO	0	0.126	0.0159	0.041	0.0017
began	TO	0.609	0.919	0.0966	0.756	0.0218
begin	TO	1	0.126	0.7634	0.756	0.0594
came	TO	1	0.919	0.0065	0.756	0.0594
chose	TO	0	0.126	0.0159	0.041	0.0017
decide	TO	0	0.126	0.0159	0.041	0.0017
decline	TO	0	0.126	0.0159	0.041	0.0017
declined	TO	0	0.126	0.0159	0.041	0.0017
declines	TO	0	0.126	0.0159	0.041	0.0017
expected	TO	0	0.126	0.0159	0.041	0.0017
failed	TO	1	0.126	0.7634	0.756	0.0594
get	TO	0.667	0.919	0.0639	0.756	0.0080
happen	TO	1	0.919	0.0065	0.756	0.0594
helped	TO	0	0.126	0.0159	0.041	0.0017
hesitate	TO	0	0.126	0.0159	0.041	0.0017
hope	TO	0	0.126	0.0159	0.041	0.0017
hoped	TO	0	0.126	0.0159	0.041	0.0017
include	VBG	1	0.126	0.7634	0.041	0.9193
intend	TO	0	0.126	0.0159	0.041	0.0017
intended	TO	0	0.126	0.0159	0.041	0.0017
intends	TO	0	0.919	0.8454	0.041	0.0017
like	TO	0	0.126	0.0159	0.041	0.0017
likes	TO	0	0.126	0.0159	0.041	0.0017
moved	TO	0.2	0.126	0.0054	0.041	0.0252
offer	TO	0	0.126	0.0159	0.041	0.0017

Table 3: Regression tree results on test set - part 2

Form	Complement	Gold P_R	Giga	Giga SE	Web	Web SE
plan	TO	0	0.126	0.0159	0.041	0.0017
planned	TO	0	0.126	0.0159	0.041	0.0017
prefer	TO	0	0.126	0.0159	0.041	0.0017
prepared	TO	0	0.126	0.0159	0.041	0.0017
promised	TO	0	0.919	0.8454	0.756	0.5718
promises	TO	0.111	0.919	0.6534	0.756	0.4161
proposed	TO	0	0.126	0.0159	0.041	0.0017
prove	TO	1	0.126	0.7634	0.041	0.9193
refuse	TO	0	0.126	0.0159	0.041	0.0017
remains	TO	1	0.919	0.0065	0.756	0.0594
said	TO	1	0.126	0.7634	0.041	0.9193
scrambled	TO	0	0.126	0.0159	0.756	0.5718
seeks	TO	0	0.126	0.0159	0.041	0.0017
seemed	TO	1	0.919	0.0065	0.756	0.0594
seems	TO	1	0.919	0.0065	0.756	0.0594
serve	TO	0	0.126	0.0159	0.041	0.0017
served	TO	0	0.126	0.0159	0.041	0.0017
start	TO	0.667	0.126	0.2920	0.756	0.0080
started	TO	0.778	0.919	0.0201	0.756	0.0005
stood	TO	1	0.126	0.7634	0.041	0.9193
struggles	TO	0	0.126	0.0159	0.041	0.0017
tend	TO	1	0.919	0.0065	0.756	0.0594
threatens	TO	0	0.126	0.0159	0.756	0.5718
tries	TO	0	0.126	0.0159	0.041	0.0017
turn out	TO	1	0.919	0.0065	0.756	0.0594
turns out	TO	1	0.919	0.0065	0.756	0.0594
vote	TO	0	0.126	0.0159	0.041	0.0017
voted	TO	0	0.126	0.0159	0.041	0.0017
want	TO	0	0.126	0.0159	0.041	0.0017
wish	TO	0	0.126	0.0159	0.041	0.0017
worked	TO	0	0.126	0.0159	0.041	0.0017

It also seems that for this task it is beneficial to use the much larger but also noisier Web corpus rather than the relatively small and clean Gigaword. The method we used is to a certain extent robust to noise and benefits from the sheer quantity of data available on the web.

Similar methodology might be useful for detecting other lexical semantic distinctions: it could be used whenever a test employed to make linguistically interesting distinctions can be reduced to a pattern search in an unannotated corpus.

Acknowledgements

We gratefully acknowledge support from Science Foundation Ireland grant 04/IN/I527 for the research presented in this paper.

References

Breiman, Leo, Friedman, Jerome H., Olshen, R. A. and Charles J., Stone. 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth.

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

Briscoe, Ted and Carroll, John. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the fifth conference on Applied natural language processing*, pages 356–363.

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

Dowty, David R. 1991. Thematic Proto-Roles and Argument Selection. *Language* 67(3), 547–619.

Graff, David. 2003. LDC English Gigaword Corpus.

King, Tracy Holloway, Crouch, Richard, Riezler, Stefan, Dalrymple, Mary and Kaplan, Ron. 2003. The PARC 700 Dependency Bank. *Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)* pages 1–8.

Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.

O'Donovan, R., Cahill, A., Way, A., Burke, M. and Van Genabith, J. 2005. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. *Computational Linguistics* 31(3), 329–365.

Palmer, Martha, Gildea, Daniel and Kingsbury, Paul. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1), 71–106.

Therneau, Terry M., Atkinson, Beth and Ripley, Brian. 2007. *The rpart Package: Recursive Partitioning*.

Therneau, Terry M. and Atkinson, Elizabeth J. 2000. An introduction to recursive partitioning using the RPART routines. Technical Report.

ANALYTIC NOUN INCORPORATION
IN
CHUJ AND K'ICHEE' MAYAN

Lachlan Duncan
University at Albany, SUNY

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

In this paper, the noun incorporation (NI) construction in Chuj and K'ichee' Mayan is examined. Formal explanations are proposed using the non-projecting semantic argument (NPSA) within the Lexical Functional Grammar (LFG) architecture. Derivational morphology indicates that NI is an analytical construct, and by inference, is post-lexically formed. Traditional NI semantic analyses, although productive, fall short of a full accounting. Consequently I show that the incorporated noun (INCORPORATE) represents a hybrid category of grammatical function (GF) that displays a mix of properties acquired from prototypical subcategorized GFs and non-subcategorized ADJUNCTs.

1 Introduction

In this paper, I examine the noun incorporation (NI) construction in Chuj and K'ichee' Mayan.[1] Of particular interest in Chuj is the NI construction's striking dialectical variation, and its variety of stranded modifiers unknown in other Mayan languages. In K'ichee's NI construction, an apparent anomaly exists in that the incorporated noun can control verb agreement in an otherwise standard intransitive predicate.

This paper addresses two issues. The first concerns the NI construction's morphosyntax. Traditionally two fundamentally opposing approaches have been pursued based on the following assumptions. Is NI a morpholexical construct (Di Sciullo and Williams 1987; Mithun 1984), a syntactic construct (Baker 1988; Sadock 1986), or is it both (Ball 2005; Van Geenhoven 1998a)? It is apparent that the two opposing approaches are overly reliant on theory-internal assumptions, and ultimately, remain artifacts of a syntactic-semantic isomorphism. In addition, Van Geenhoven's (1998) semantic incorporation, although productive, is a strictly semantic account,

[†] I wish to thank Ash Asudeh for his assistance particularly with the semantics and George Aaron Broadwell for his assistance as well. I also wish to thank Doug Ball, Emily Bender, Joan Bresnan, Miriam Butt, Paul Kroeger, and Michael Wescoat for their questions and helpful comments. Finally I particularly wish to thank my K'ichee' Maya consultant, Rafael of Sololá, Guatemala. All the usual disclaimers apply.

[1] Chuj is spoken in the towns of San Sebastián Coatán, Nentón, and San Mateo Ixtatán all located in Guatemala's Cuchumatán mountains (Hopkins (1967:Intro.); Maxwell (1976:Fn.1); Williams and Williams (1966:219)). The Unified Mayan Alphabet (UMA), as adopted by the Academy of Mayan Languages of Guatemala, is used in this paper and not IPA symbols. Unless otherwise indicated, the K'ichee' data are from the author's field work. Note the following abbreviations: first, second, third person = 1, 2, 3, absolutive agreement marker = ABS, actor focus = AF, (absolutive) antipassive = AP, clitic = CL, completive = COM, derived transitive verb = DT, determiner = DET, ergative agreement marker = ERG, genitive = GEN, incompletive aspect = INC, independent pronoun = INDPRO, intransitive = INTR, interrogative = INT, irrealis = IRR, negative = NEG, nominalizing suffix = NOM, noun-incorporation = NI, passive = PAS, transitive/intransitive phrase final marker = T/IPF, plural = PL, preposition = P, relational noun (phrase) = RN(P), singular = S, lexical stem forming vowel = SFV, transitive = TRA.

and fails to adequately explain NI's unique morphosyntax. My discussion moves away from the prototypical approaches, due, in some part, to the availability of more fine-grained semantic analyses, LFG's monostratal architecture (Kaplan and Bresnan 1982), and the atomicity of lexical integrity (Bresnan and Mchombo 1995).

The second issue concerns the incorporated noun's representation and the complex cluster of morphosyntactic and semantic properties associated with it. The incorporated noun's representation includes its syntactic structure, agreement behavior, scopal properties, and semantic expression and composition. Van Geenhoven's (1998) semantic incorporation is a promising point of departure for a truth-conditional analysis of NI. I argue, however, that to adequately account for all types of NI requires the recognition of a syntactic element in the form of a grammatical function, the INCORPORATE, as first discussed in Asudeh (2007) and Asudeh and Ball (2005). As a non-set valued, non-subcategorized ADJUNCT (Asudeh 2007), the INCORPORATE links, I propose, to a thematic role in the argument structure, making the INCORPORATE indispensable to a principled explanation of analytic NI.[2]

The remainder of this paper is ordered in the following way. I review Chuj and K'ichee' NI data paying attention to the unusual dialectical variation of NI in Chuj. A discussion follows about the semantics of bare indefinites from various authors, and the non-projecting semantic argument (NPSA) (Asudeh 2007; Asudeh and Ball 2005). Following that is a presentation of the Chuj and K'ichee' NI data within the LFG framework. The paper ends with an elaboration of the INCORPORATE as a part-argument, part-adjunct GF.

2 Noun incorporation in Chuj

NI in Chuj occurs when the direct object, in the form of an unmarked noun stem, is 'incorporated' into the verb. When N incorporates, the NI verb detransitivizes. As an intransitive, the noun incorporating verb is uncontroversial because of multiple indicators of intransitivity in the verb morphology.

2.1 About Chuj

An active transitive clause with VOS word order is shown in example (1). England (1991:463-464) claims that in the San Mateo Ixtatán dialect, both VSO and VOS are permitted. But in San Sebastián Coatán, basic word order is VSO only:[3]

[2] Elsewhere I suggest an alternate, lexicon-based analysis for synthetic NI found in, for example, the lowland Mayan languages of Ch'orti', Itzaj, and Yukatek Mayan.

[3] (SS) refers to the San Sebastián Coatán dialect of Chuj while (SM) refers to the San Mateo Ixtatán dialect of Chuj as spoken in Guatemala.

(1) Ix-s-mak' waj Xun ix Malin CHUJ
COM-3SERG-hit NC John NC Mary
'Mary hit John (Dayley 1981:35).'

Example (2) shows a root intransitive (Maxwell 1976:131). Intransitives in Chuj are characterized by having a single agreement marker, called the Set B absolutive (ABS), mark the SUBJ (Maxwell 1976:128):

(2) Tz-onh-b'ey-i (SS)
INC-1PLABS-walk-IPF
'We walk (Maxwell 1976:130).'

Let us now look at the NI construction in Chuj, as shown in (3):[4]

(3) a. Ix-ach-mak'-w-i anima (SM)
COM-2SABS-hit-NI-IPF people
'You hit people (Dayley 1981:35).'

b. Ix-in-al-w-i ab'ix (SM)
COM-1SABS-tell-NI-IPF stories
'I told stories (Maxwell 1976:131).'

With regards to verb agreement, the absolutive also marks subject agreement in the intransitive NI verb (Maxwell 1976:135). Thus it can be safely assumed that Chuj's incorporated noun never controls verb agreement.

2.2 Restrictions on Chuj's incorporated nouns

This section reviews all the restrictions on the incorporated nouns of the San Mateo and San Sebastián dialects of Chuj.

2.2.1 Generic restrictions on Chuj's incorporated nouns

Maxwell (1976:133) distinguishes two divergent forms of the incorporated noun in Chuj. Let us begin our review with elements of the incorporated noun common to the dialects of Guatemalan Chuj.

In both Chuj's dialects, generic limitations on incorporated nouns include, but are not limited to, the following restrictions. The incorporated noun may not be modified by a determiner (4a), by a number, or by a noun classifier. In addition, it cannot be possessed (4b) (Maxwell 1976:132):

(4) a. *Ix-in-kuy-w nik anma' (SM/SS)
COM-1SABS-teach-NI DET people
(*'I taught the people (Maxwell 1976:132).')

[4] Interlinear glosses of the Chuj data are drawn mainly from Dayley (1981), Hopkins (1967), Maxwell (1976), Robertson (1980, 1992), and Williams and Williams (1966).

b. *Ix-in-ten-w he-lu'um (sm/ss)
com-1sAbs-mash-ni 2sPoss-dirt
(*'I mashed your dirt (Maxwell 1976:133).')

2.2.2 Further restrictions on San Mateo's incorporated noun

San Mateo's incorporated noun has two further restrictions, neither of which apply to San Sebastián's (Maxwell 1976:133). Post-nominal modifying adjectives (5a), and relativization (5b), are disallowed in San Mateo:

(5) a. *Ix-in-al-w-i ab'ix kuseltak (sm)
com-1sAbs-tell-ni-ipf story sad
(*'I told a sad story (Maxwell 1976:133).')

b. *Ix-in-kuy-w-i anima s-mun-l-aj t'atik (sm)
com-1sA-teach-ni-ipf people 3sErg-work-tra-int here
(*'I taught the people who work here (Maxwell 1976:133).')

Nonetheless a limited number of adjectives precede incorporated nouns in San Mateo, although these adjectives form adjective-noun compounds (6a) (Maxwell 1976:133–4). San Mateo's incorporated noun can also be a noun-noun (n-n) compound (Maxwell 1976:fn.4):

(6) Ix-in-pak-w-i takinh-awal (sm)
com-1sAbs-bend-ni-ipf adjec:dry-cornstalks
'I bent dry cornstalks (Maxwell 1976:134).'

With regards to verb agreement, the subject (agent) of the transitive controls the verb's ergative agreement marker while the object (patient) controls the verb's absolutive agreement marker (Maxwell 1976:135). However with regards to the ni verb, the subject controls the verb's sole agreement marker, the absolutive. In sum, only bare indefinites and adjective-noun and noun-noun compounds can function as incorporated nouns in San Mateo.

2.2.3 Fewer restrictions on San Sebastián's incorporated noun

In contrast to San Mateo's, San Sebastián's incorporated noun is far less constrained, differing in two fundamental ways. San Sebastián's incorporated noun allows prenominal (non-compounding) adjectives and postnominal adjectives, and limited types of relative clauses (Maxwell 1976:135, 137).

San Sebastián's incorporated noun allows 'some preceding adjectives,' but crucially these prenominal adjectives, like *al* 'heavy' in (7), appear not to form adjective-noun compounds (Maxwell 1976:135). And adjectives, also like *al* 'heavy' in (7), can appear postnominally (Maxwell 1976:136):

(7) Hin-man-w {al líwru, líwru al} (SS)
1SABS-buy-NI heavy book, book heavy
'I bought a heavy book (Maxwell 1976:135).'

Secondly the two restrictions on San Mateo's incorporated noun, as shown in (5a, b) do not apply to San Sebastián's incorporated noun (Maxwell 1976:136–7). But not all relative clauses are allowed, as shown in (8c):

(8) a. Ix-in-awt-w hunh ix-il-c[ha]j-i (SS)
COM-1SABS-read-NI paper COM-see-PAS-IPF
'I read the paper (that) was seen (Maxwell 1976:137).'

b. Hin-man-w lum ajtil x-in-el-a (SS)
1SABS-buy-NI land where COM-1SABS-see-TPF
'I bought the land where you saw me (Maxwell 1976:137).'

c. *Ix-in-awt-w hunh ix-w-il-a (SS)
COM-1SABS-read-NI paper COM-1SERG-see-TPF
(*'I read the paper I saw (Maxwell 1976:137).')

In sum, only San Sebastián allows adjectives as non-compounding prenominal modifiers, and adjectives and relative clauses as postnominal modifiers.

2.2.4 Noun incorporation in Chuj's agentives and instrumentals

Finally let us examine the 'incorporation of objects into NPs,' the agentives (9a), and the instrumentals (9b) (Maxwell 1976:138):

(9) a. Tz'ib'-m hu'unh (SS)
write-AGT paper
'Writer of papers (Maxwell 1976:138).'

b. Tz'ib'-l-ab' hu'unh (SS)
write-NOM-INSTR paper
'Writing tool for paper (Maxwell 1976:138).'

The *–(u)m* suffix (SM) represents the nominalizing actor morpheme (Hopkins 1967:92–3, 257), while the *–ap'* suffix (SM) represents the instrument morpheme (Hopkins 1967:85, 253). Note that the bare indefinites of the nominalized forms are constrained in exactly the same manner as are the incorporated nouns of the NI verb construction.

The nominalization data afford us an important insight into the formation of NI constructions. The initial word in the two-word construction is marked for the appropriate agentive or instrumental nominalization, and not the second word, the incorporated noun. On the assumption that derivational processes occur only in the lexicon, we are able to conclude from the data that NI is an analytic construction. Accordingly, we can reasonably infer that NI is post-lexically formed.

2.3 The semantics of noun incorporation

As noted above, the NI construction and its structural aspects have been the subject of a long and contentious debate in American linguistics. However in recent years, NI has received increased attention from semanticists. The semantic focus has been primarily on argument structure, on NI's unique scopal properties, and on the incorporated noun's role as a discourse antecedent. In this section, I briefly examine the semantics of NI, beginning with the seminal research of Van Geenhoven (1995; 1996; 1997; 1998a,b) followed by the more recent analyses of Chung and Ladusaw (2004).

2.3.1 The semantics of the bare indefinite

In this section, I review Van Geenhoven's structural and semantic approaches to NI in West Greenlandic. Van Geenhoven considers the historical debate about NI as either a lexically or syntactically formed construction to be the result of an uncritical acceptance of the theta criterion (Chomsky 1981). From a truth-conditional perspective, Van Geenhoven reasons that the theta criterion cannot adequately account for the syntactic expression of the argument structure of an incorporating verb. Instead she suggests that lexical and syntactic explanations can co-exist. Accordingly she recommends a structural representation of morphological NI word formation that is a 'syntactically base generated' sub-phrasal construction.

The fallacy in Van Geenhoven's structural analysis rests on the assumption that lexical categories can only participate in lexical operations. Nonetheless the focus of Van Geenhoven's analysis of NI is predominantly semantic. She analyzes West Greenlandic incorporated nouns, English and West Germanic bare plurals, German split topics, and existentials as instances of narrow scope indefinites. Essentially Van Geenhoven identifies incorporated nouns as predicative indefinites, interpreting them and most other narrow scope indefinites as property-denoting descriptions. She claims that incorporated nouns provide a predicate that is absorbed by the incorporating verb as a restriction on the internal argument of the incorporating verb. Van Geenhoven refers to the semantic process of the absorption of predicative indefinites as *semantic incorporation*, which, during the process, generates the narrow scope of the incorporated noun. For type theory, predicative indefinites are type $\langle e,t \rangle$, while free variables are type $\langle e \rangle$ (cf. Partee 1987).

Example (10a) shows a West Greenlandic standard transitive, and (10b) shows its predicate logic analysis by Van Geenhoven (1998b:243):

(10) a. Nuka-p iipili neri-v-a-a. W. GREENLANDIC
Nuka-ERG apple-ABS eat-IND-[+TR]-3SG.3SG
'Nuka ate a particular apple (Van Geenhoven 1998b:243).'

b. $\lambda y_e \lambda x_e$ [eat(x, y)]

c. Nuka iipili-tur-p-u-q. WEST GREENLANDIC
Nuka-ABS apple-eat-IND-[−TR]-3SG
'Nuka ate an apple/apples (Van Geenhoven 1998b:240).'

d. $\lambda P_{<e,t>} \lambda x_e \exists y \; [\text{eat}(x, y) \wedge P(y)]$

Example (10c) shows a West Greenlandic NI verb, and in (10d), the predicate logic analysis of semantic incorporation by Van Geenhoven (1998b:240). The crucial change from (10b) to (10d) is that the incorporated noun includes the symbolic representation of $P(y)$. This just means that the variable y, which represents the restriction of the meaning of the original syntactic object, has as its new function a property, P. In other words, the semantics of the restricted free variable, the object, has changed to that of a predicative indefinite, which here is an incorporated noun. This analysis, more or less, forms the basis of most current approaches to the semantics of NI.

The overall response to Van Geenhoven's theory of semantic incorporation is somewhat mixed. On the one hand, Farkas and de Swart (2003:10–11) accept semantic incorporation's core assumption that incorporated nouns are property-denoting arguments, or predicate modifiers. Chung and Ladusaw (2004:14–18) also adhere to the Property theory of indefinites, which holds that some or all indefinite DPs can be interpreted semantically as properties of the type $\langle e, t \rangle$. On the other hand, Farkas and de Swart (2003:2–4) reject Van Geenhoven's purely semantic view of (noun) incorporation. The reasons include that the incorporated noun has a special morphosyntax, and that the incorporated noun is syntactically invisible in intransitive NI constructions (Farkas and de Swart 2003:3, 11). In general, Farkas and de Swart (2003:156–7) do not accept that semantic incorporation can account for both incorporated nouns and all other narrow scope indefinites and existentials.

West Greenlandic's incorporated noun also has adnominal or stranded modifiers, which Van Geenhoven (1998a:17–22, 146–159) refers to as (discontinuous) external modifiers. They include adjectives, numerals, *wh*-words, other nouns, and even relative clauses. Van Geenhoven offers two important insights into the semantics of external modifiers. Incorporated nouns and their external modifiers are predicates of the same variable, and that it is unnecessary to semantically interpret the incorporated noun and the external modifier as a single syntactic unit. However Van Geenhoven's semantic incorporation of external modifiers requires a more complicated composition than the incorporated nouns they modify. Because of this, Van Geenhoven's analysis of stranded modifiers has not, on the whole, been well received (Chung and Ladusaw 2004:115–6; Farkas and de Swart 2003:156).

The approach of Chung and Ladusaw (2004) to incorporated nouns as property-denoting indefinites mirrors Van Geenhoven's semantic incorporation. The core difference between the two approaches is the mode of composition. That is, Chung and Ladusaw (2004:22) hypothesize that different modes of semantic composition of property-denoting indefinites of type

$\langle e,t \rangle$ will manifest different syntactic structures, assuming truth-conditional equivalency. The first mode of semantic composition of indefinites, called *Specify*, results in indefinites that are scopally unrestricted and that fully saturate the internal argument by function application (Chung and Ladusaw 2004:16). The second mode of composition of indefinites, called *Restrict*, restricts but does not saturate the internal argument. *Restrict* is very similar in spirit to semantic incorporation but the implementation and results differ somewhat. Thus incorporated nouns, stranded modifiers, and doubled DPs all compose with the variable of the verb's internal argument but do so using a variety of compositional modes.

2.3.2 The Non-Projecting Semantic Argument (NPSA)

In explaining NI in Niuean, Asudeh (2007) proposes the non-projecting semantic argument (NPSA), framed within LFG and Glue compositional semantics. Two assumptions underlie the NPSA: the existence of non-projecting words (cf. Toivonen 2003), and an explicit 'level of semantic structure.' The first assumption involves the proposition that, although the verb-incorporated noun (V-N̂) unit remains inseparable in the syntax, it does not form a single lexical item (Asudeh 2007:1). The second assumption involves the notion that an NP can possess an argument at semantic structure that remains invisible to syntactic processes. The incorporated noun is not a syntactic argument but is instead semantically related to the verb.

The incorporated noun can be modified by nominal elements that adjoin to the NP complement of the incorporating verb (Asudeh 2007:6; Asudeh and Ball 2005:2, 8). The INCORPORATE's phrasal part is called the remnant (Asudeh 2007), another term for a stranded modifier. At first glance, it might seem incongruous that the INCORPORATE can extend over several levels of X-bar structure. Yet this is an entirely acceptable practice in LFG and can be seen, for example, in the way that discontinuous constituents in Warlpiri unify in f-structure (cf. Bresnan 2001:326–7, 393–4).

2.4 Explaining noun incorporation in Chuj

This section provides explanations within the LFG framework for the NI construction in the San Mateo and San Sebastián dialects of Chuj.

2.4.1 Noun incorporation in San Mateo Chuj

I assume that Chuj's predicate initial clause is canonical and possesses the same phrase structure as that of Kaqchikel, a sister language to K'ichee':

(11) $[_{\mathrm{S}}\ \mathrm{V}^0\ \mathrm{XP}^*\ ]$ (Broadwell 2000)

To implement the NPSA, I begin with the San Mateo data in (3b), repeated here as (12). The clause consists of the NI verb complex *ixinalwi ab'ix*

'I told stories (lit. 'I story-told.').' Note that in Chuj, as in many Mayan languages, overt lexical subject and object NPs are optional ('pro-drop') because they are usually cross-referenced on the verb. Note also that the third person singular absolutive agreement marker is a zero anaphora and is thus never represented in LFG's c-structures. Because all the Chuj data cited in this paper are verb initial clauses, the phrase structure in (11) will suffice:

(12) Ix-in-al-w-i ab'ix (SM)
COM-1sABS-tell-NI-IPF stories
'I told stories (Maxwell 1976:13).

The derivational NI morphology *–wi* marks the verb, not the incorporated noun. This is an important point for NI theory development. It means that it is impossible for the incorporated noun to morpholexically incorporate into the verb complex to form a lexical N-V compound. The incorporated noun is morphologically and of course categorially distinct from the incorporating verb, and therefore, syntactically individuated. Therefore NI in Chuj is analytic, not synthetic. From this empirical observation, we infer that NI in Chuj is a post-lexical construct.

The annotated phrase structure for (12) is (13).The second line of phrase structure in (13) represents the c-structure rule for analytic NI. The incorporated noun's (N̂) first functional description indicates N̂ is an ARGUMENT in semantic-structure (σ–str) (Asudeh 2007). The second line indicates that N̂ is the grammatical function INCORPORATE in f-structure. Crucially N̂ is assimilated into the semantics in spite of it not being a subcategorized GF:

(13)
$$\begin{array}{lll} \text{S} \rightarrow & \text{V}^0 & \\ & \uparrow=\downarrow & \end{array}$$

$$\begin{array}{lll} \text{V}^0 \rightarrow & \text{V}^0 & \hat{\text{N}} \\ & \uparrow=\downarrow & (\uparrow_\sigma \text{ARGUMENT})=\downarrow_\sigma \\ & & (\uparrow \text{INCORP})=\downarrow \end{array}$$

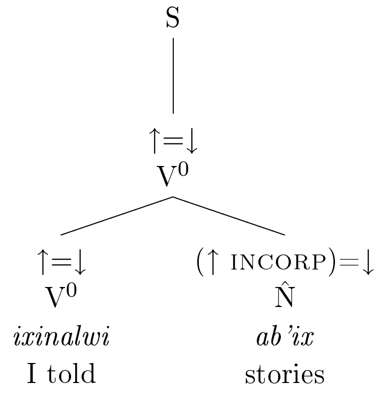
It is assumed that some lexical rule converts the transitive verb to the intransitive NI verb. (14) is the lexical entry for the incorporating verb *ixinalwi*:

(14)
$$\begin{array}{lll} \textit{ixinalwi}: & \text{V}^0 & (\uparrow \text{PRED}) = \text{`tell}\langle\text{SUBJ}\rangle\text{'} \\ & & (\uparrow \text{ASP}) = \text{COM} \\ & & (\uparrow \text{SUBJ PRED}) = \text{`Pro'} \\ & & (\uparrow \text{SUBJ CASE}) = \text{ABS} \\ & & (\uparrow \text{SUBJ NUM}) = \text{SG} \\ & & (\uparrow \text{SUBJ PER}) = 1 \end{array}$$

It is also assumed that a lexical rule converts a projecting noun (N^0) into a non-projecting noun (N̂). Context free rules (Chomsky 1986), such as phrase structure rules, determine the syntactic grouping of words according to the words' syntactic category. Only unary lexical rules can convert or derive syntactic categories, like an $\hat{\mathrm{N}}$ from an N^0, context free rules cannot.

The c-structure for example (12) is (15a), and its f-structure is (15b):

(15) a.



b.

$$\left[\begin{array}{ll} \textsc{Pred} & \text{`tell}\langle\textsc{Subj}\rangle\text{'} \\ \textsc{Subj} & \left[\begin{array}{ll} \textsc{Pred} & \text{`Pro'} \\ \textsc{Per} & 1 \\ \textsc{Num} & \textsc{sg} \end{array}\right] \\ \textsc{Incorp} & \left[\begin{array}{ll} \textsc{Pred} & \text{`stories'} \end{array}\right] \end{array}\right]$$

2.4.2 Noun incorporation in San Sebastián Chuj

I begin this section with the canonical phrase structure rule for the predicate-initial clause S(entence) of San Sebastián Chuj. Crucially the San Sebastián Chuj dialect licenses an incorporate remnant, which is (vacuously) adjoined to the NP complement of V^0:

$$(16)\quad \mathrm{S} \rightarrow \begin{array}{ccc} \mathrm{V}^0 & \mathrm{NP} & \mathrm{DP} \\ \uparrow=\downarrow & (\uparrow \textsc{incorp})=\downarrow & (\uparrow \textsc{subj})=\downarrow \end{array}$$

The relative clause as stranded modifier As stranded modifiers of incorporated nouns, Chuj's relative clauses are unusual because of their structural range and complexity. In (8a), repeated below as (17), the relative *ajtil xinela* 'where you saw me,' a bivalent clause with pro-drop headed by the relativizing adverb *ajtil* 'where,' modifies the incorporated nominal.

(17) Hin-man-w lum ajtil x-in-el-a (ss)
1sAbs-buy-ni land where com-1sAbs-see-tpf
'I bought the land where you saw me (Maxwell 1976:137).'

The phrase structure for example (17) is (18):

(18)

$$\begin{array}{llll} \mathrm{S} \rightarrow & \mathrm{V}^0 & \mathrm{NP} \\ & \uparrow=\downarrow & (\uparrow \textsc{incorp})=\downarrow \end{array}$$

$$\begin{array}{llll} \mathrm{V}^0 \rightarrow & \mathrm{V}^0 & \hat{\mathrm{N}} \\ & \uparrow=\downarrow & (\uparrow \textsc{incorp})=\downarrow \end{array}$$

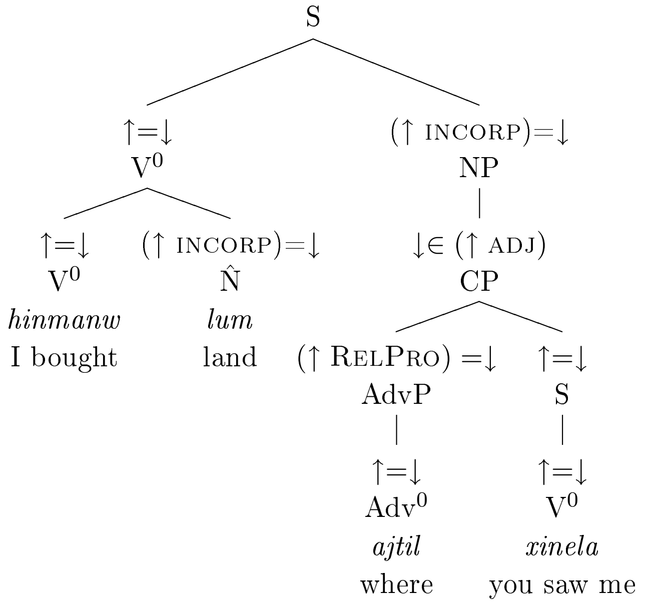
$$\begin{array}{llll} \mathrm{NP} \rightarrow & \mathrm{NP} & \mathrm{CP} \\ & \uparrow=\downarrow & (\uparrow \textsc{incorp})=\downarrow \end{array}$$

$$\begin{array}{llll} \mathrm{CP} \rightarrow & \mathrm{AdvP} & \mathrm{S} \\ & (\uparrow \mathrm{RelPro})=\downarrow & \uparrow=\downarrow \end{array}$$

$$\begin{array}{ll} \mathrm{AdvP} \rightarrow & \mathrm{Adv}^0 \\ & \uparrow=\downarrow \end{array}$$

The c-structure in (19a) represents (17), while its f-structure is (19b). It is essential to keep in mind that the INCORPORATE is an non-governable, non-subcategorized modifying ADJ of the incorporating verb in f-structure, but is a full argument of the incorporating verb in sem-structure:

(19) a.



b. $$\begin{bmatrix} \textsc{Pred} & \text{`buy}\langle\textsc{Subj}\rangle\text{'} \\ \textsc{Subj} & \begin{bmatrix} \textsc{Pred} & \text{`Pro'} \\ \textsc{Pers} & 1 \\ \textsc{Num} & \textsc{sg} \end{bmatrix} \\ \textsc{Incorp} & \begin{bmatrix} \textsc{Pred} & \text{`land'} \\ \textsc{Adj} & \left\{ \begin{bmatrix} \textsc{Pred} & \text{`see}\langle\textsc{Subj},\ \textsc{Obj}\rangle\text{'} \\ \textsc{Subj} & \begin{bmatrix} \textsc{Pred} & \text{`Pro'} \\ \textsc{Per} & 2 \\ \textsc{Num} & \textsc{sg} \end{bmatrix} \\ \textsc{Obj} & \begin{bmatrix} \textsc{Pred} & \text{`Pro'} \\ \textsc{Per} & 1 \\ \textsc{Num} & \textsc{sg} \end{bmatrix} \\ \textsc{RelPro} & \begin{bmatrix} \textsc{Pred} & \text{`where'} \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix} \end{bmatrix}$$

The prenominal adjective as modifier The following NI construction includes a prenominal adjective. Example (7), repeated here as (20) revised, shows the prenominal, non-compounding adjective *al* 'heavy':

(20) Hin-man-w al líwru (SS)
1sAbs-buy-ni heavy book
'I bought a heavy book (Maxwell 1976:135).'

The prenominal adjective *al* 'heavy,' which modifies the incorporated noun, is, I believe, a non-projecting adjective (Â) that head-adjoins to the incorporated noun. The prehead modifying adjective has the lexical entry in (21a). Example (20) is represented by the phrase structure in (21b):

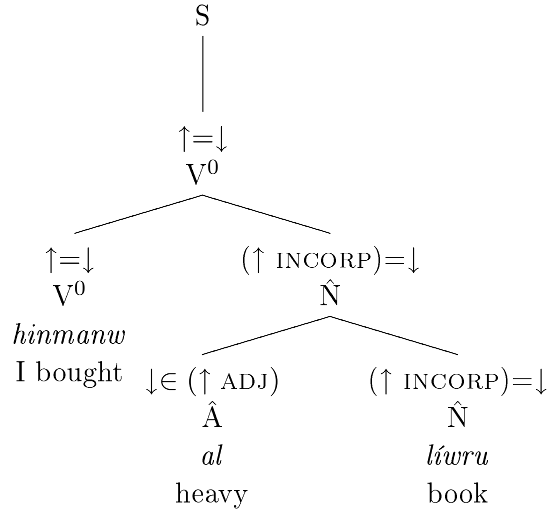
(21) a. *al* : Â (↑ Pred) = 'heavy'

b. $$\begin{array}{llll} \text{S} \rightarrow & \text{V}^0 & & \\ & \uparrow=\downarrow & & \\ \\ \text{V}^0 \rightarrow & \text{V}^0 & \hat{\text{N}} & \\ & \uparrow=\downarrow & (\uparrow \textsc{incorp})=\downarrow & \\ \\ \hat{\text{N}} \rightarrow & \hat{\text{A}} & \hat{\text{N}} & \\ & \downarrow\in(\uparrow \textsc{adj}) & (\uparrow \textsc{incorp})=\downarrow & \end{array}$$

Note that the two adjunction structures in (21b) are licensed by the *Adjunction Identity* condition (Toivonen 2003), which simply states that, 'Same adjoins to same.' This suggests that both X^0 and $\hat{\text{X}}$ can dominate lexical material. In other words, the non-projecting adjective may adjoin to the non-projecting noun, according to Adjunction Identity.

I suggest that example (20) can be represented by the c-structure in (22a). Its f-structure is shown in (22b). Crucially the adjective type in (22b) is attributive, not predicative:

(22) a.



b.

$$\begin{bmatrix} \textsc{Pred} & \text{‘buy}\langle\textsc{Subj}\rangle\text{’} \\ \textsc{Aspect} & \textsc{com} \\ \textsc{Subj} & \begin{bmatrix} \textsc{Pred} & \text{‘Pro’} \\ \textsc{Per} & 1 \\ \textsc{Num} & \textsc{sg} \end{bmatrix} \\ \textsc{Incorp} & \begin{bmatrix} \textsc{Pred} & \text{‘book’} \\ \textsc{Adj} & \left\{ \begin{bmatrix} \textsc{Atype} & \textsc{attrib} \\ \textsc{Pred} & \text{‘heavy’} \end{bmatrix} \right\} \end{bmatrix} \end{bmatrix}$$

3 Noun incorporation in K’ichee’ Mayan

K’ichee’ also has the NI construction, and it is identical to Chuj’s, except for one important difference. Whereas in Chuj the subject of the NI verb controls agreement, in K’ichee’ either the subject or the INCORPORATE can control agreement. In this section, I offer empirical support for the INCORPORATE as a type of GF, and in the process, account for K’ichee’s NI construction.

3.1 The incorporated noun in K’ichee’

The morphosyntax of the NI construction is subject to significant restrictions, and, in particular, the form and distribution of the incorporated noun. Before addressing agreement, let us first review adjacency and extraction data for the incorporated noun in K'ichee'.

Subject DPs, as in (23), and adjuncts, such as manner adverbs and prepositional phrases, cannot occur between the verb and the incorporated noun:

(23) *Utz k-at-paj-ow at atz'aam
well INC-2sAbs-weigh-NI 2sIndPro salt
(*'You measure salt well.')

The incorporated noun cannot be extracted to a preverbal position immediately before the verb. Normally it is quite acceptable for a bare nominal to occupy this immediate preverbal location, the generic focus position:

(24) *Utz atz'aam k-at-paj-ow at
well salt INC-2sAbs-weigh-NI 2sIndPro
(*'You measure salt well.')

The incorporated noun cannot extract to the sentence-initial topic position in Spec,CP. One should keep in mind that this is not an entirely unexpected result because Mayan topics are subject to a specificity restriction:

(25) *Carro na utz ta k-a-b'iin-i-sa-n lee achii
car NEG good IRR INC-EPE-drive-SFV-CAUS-NI DET man
(*'The man car-drives badly.')

We have noted above that the NI construction is subject to obligatory narrow scope. Because the incorporated noun is definite, (26) is ungrammatical:

(26) *At utz k-at-b'iin-i-sa-n lee carro
2sIndPro well INC-2sAbs-drive-SFV-CAUS-NI DET car
(*'You drive the car very well.')

3.2 Verb agreement in K'ichee's NI construction

Verb agreement in K'ichee's NI construction is quite unexpected in light of verb agreement in K'ichee's standard transitive. Agreement in the NI construction is based not on grammatical functions, as in the active transitive, but on the person hierarchy of the arguments themselves. The person hierarchy is defined as local person outranks non-local and plural outranks singular, so that the argument higher on the person hierarchy controls verb agreement. I will refer to the hierarchy as person-salience. We first look at examples of subject agreement, and then, incorporated noun agreement.

If the subject is either 1st person or informal 2nd person, then it is always cross-referenced by the verb agreement marker:

(27) Utz k-at-paj-ow atz'aam at
well INC-2SABS-weigh-NI salt 2SINDPRO
'You measure salt well.'

However if the subject is 3rd person singular or formal 2nd person singular or plural and the incorporated noun is 3rd person plural, then the incorporated noun—not the subject—controls verb agreement (Mondloch 1981). Continuing on, (28a) shows that the incorporated noun *ak'* 'chicken' is plural while the subject DP *lee ixoq* 'the woman' is singular. Thus the incorporated noun *ak'* controls verb agreement. Again in (28b), the plural incorporated noun *ak'alaab'* 'children' controls agreement because the formal 2nd person singular subject *la* 'you' can never control agreement (Mondloch 1981):

(28) a. Naj k-ee-pil-ow ak' lee ixoq
long.time INC-3PLABS-butcher-NI chicken DET woman
'It takes a long time for the woman to chicken-gut (M 1981:250).'

b. Utz k-ee-yuq'u-n la ak'al-aab'
well INC-3PLABS-take.care.of-NI 2SABSHON child-PL
'You child-care well (Mondloch 1981:250).'

The examples of K'ichee's NI construction in (28) highlight a serious disjunction for verb agreement in K'ichee'. If the incorporated noun does control agreement, an internal contradiction will result because non-subcategorized constituents like adjuncts never control agreement in K'ichee'. Assuming that agreement is systematized in f-structure, my configuration of it cannot account for control of agreement by the incorporated noun.

3.3 Explaining noun incorporation in K'ichee'

Two basic choices are available, regarding the agreement anomaly. The first interprets the NI construction as a bivalent transitive verb and the INCORPORATE as an OBJ. The second interprets NI as a monovalent intransitive and the INCORPORATE as an ADJ. Neither choice is without problems.

3.3.1 Noun incorporation in K'ichee' as transitive

The first approach interprets the NI verb as a bivalent transitive. This approach also retains head-adjunction of the incorporated noun. Agreement control by a subcategorized constituent is accounted for in f-structure along the usual lines. In sum, the ID and LP relations of the INCORPORATE and its mother (V^0) undergo substantial realignment from the canonical non-NI bivalent, transitive verb. Yet the INCORPORATE's grammatical relation with V^0 remains the same as the original direct object's. This is because the INCORPORATE functionally identifies with the grammatical object in the f-structure, and links directly to it.

There are advantages and disadvantages in the first approach. The advantage is the theory of agreement is entirely standard and does not introduce any new agreement mechanism into the theory. But a major disadvantage is that it identifies the verb as a transitive even though the verb morphology and morphosyntax is indisputably intransitive. Accepting this first approach incurs the rather disagreeable outcome of overturning long-held accounts of the transitive-intransitive dichotomy of Mayan verbs.

3.3.2 Noun incorporation in K'ichee' as intransitive

The second approach interprets the NI verb as a monovalent intransitive. The INCORPORATE will be a non-governable, non-subcategorized grammatical function, a non-set valued ADJUNCT in f-structure. The advantage with this approach is that the morphology is in complete compliance with long-held notions of (in)transitivity in Mayan linguistics. The most obvious disadvantage is the rather messy account of agreement that it engenders.

However there is another way to express agreement in LFG, other than in the f-structure, and that is in the lexical entry. Basically for INCORPORATE control of agreement, there are two sets of constraints required, one on the INCORPORATE and two on the subject. The constraint on the INCORPORATE is simply that it must be plural. The constraint on the subject is two part, but either one must hold for the INCORPORATE to control agreement. The first part requires that the subject be 3^{rd} person singular. Failing that, the subject must be the 2^{nd} person formal pronominal clitic, either singular or plural, because the formal pronominal clitic never controls agreement.

The lexical entries of the NI verb *keepilow* and its INCORPORATE *ak'* from (28) could be the following:

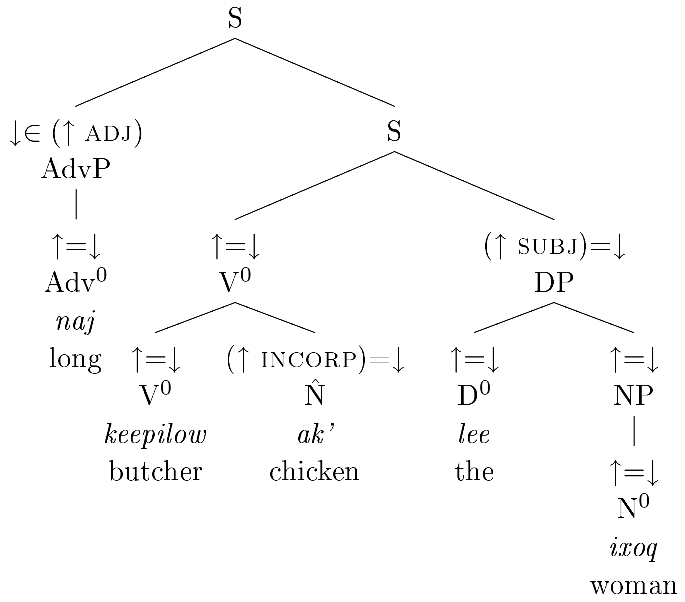
(29) a. *keepilow* : V^0 (↑ PRED) = 'butcher⟨SUBJ⟩'
(↑ SUBJ PRED) = 'PRO'
(↑ SUBJ NUM) = SG
(↑ SUBJ PER) = 3
(↑ INCORP NUM) = PL
(↑ INCORP CASE) = ABS

b. *ak'* : $\hat{N}$ (↑ PRED) = 'chicken'

The NI verb rather than the INCORPORATE should have the functional descriptions in its lexical entry to account for the constraints on the INCORPORATE and its agreement interaction with the SUBJ.

The c-structure in (30a) is identical except the INCORPORATE is annotated with (↑ INCORP)=↓, not (↑ OBJ)=↓, while its f-structure is (30b):

(30) a.



b.

$$\left[\begin{array}{ll} \textsc{Pred} & \text{‘butcher}\langle\textsc{Subj}\rangle\text{’} \\ \textsc{Asp} & \textsc{inc} \\ \textsc{Subj} & \left[\begin{array}{ll} \textsc{Pred} & \text{‘woman’} \\ \textsc{Num} & \textsc{sg} \end{array}\right] \\ \textsc{Incorp} & \left[\begin{array}{ll} \textsc{Pred} & \text{‘chicken’} \\ \textsc{Num} & \textsc{pl} \end{array}\right] \\ \textsc{Adj} & \left\{\left[\begin{array}{ll} \textsc{Pred} & \text{‘long’} \end{array}\right]\right\} \end{array}\right]$$

4 The INCORPORATE revisited: a new GF

The central issue at this point is how to account for the INCORPORATE as a grammatical function. In the NPSA, the INCORPORATE is invisible to syntactic processes but retains full argument status at semantic-structure. As we have seen from Van Geenhoven's predicate logic analysis in (10d), at the notional heart of the INCORPORATE is a property-denoting predicate that restricts the verb's internal argument. Although categorically an ADJ, the INCORPORATE is clearly not an ordinary, garden-variety ADJ. In fact, unlike the canonical ADJ in Table 1, the INCORPORATE maps to the argument structure as a set member. I assume that a–structure is a syntactic representation of the mapping of thematic roles to grammatical functions.

A binary-feature matrix can predict or reveal unknown or unrecognized grammatical relations, categories, or constructions. I propose that one of the defining properties of the matrix should be constituent selection by the syn-

		Syntactic Selection	
		+	−
Semantic Select	−	Raising GF *'Juan seems happy.'* [PRED 'seem⟨XCOMP⟩SUBJ] $\lambda P.\text{seem}(P)$	Adjunct *'Maria laughed loudly.'* [PRED 'laugh⟨SUBJ⟩'] [ADJ {[PRED 'loudly']}] $\lambda x.\text{laugh}(x)$
	+	Subcategorized GF *'Fido chased Fluffy.'* [PRED 'chase⟨SUBJ,OBJ⟩'] $\lambda y.\lambda x.\text{chase}(x, y)$	Incorporate *'I story-tell.'* [PRED 'tell⟨SUBJ⟩'] [INCORP ['story']] $\lambda P \lambda x.\exists y.[\text{tell}(x, y) \wedge P(y)]$

Table 1: Syntactic vs. semantic selection

tax, or more precisely, syntactic subcategorization. This attribute recognizes only argument functions. The second defining attribute should be argument structure encoded as semantic selection. This attribute represents thematic arguments that map to grammatical functions but excludes expletives and canonical ADJUNCTs.

Table 1 shows the division of the two properties or attributes of SEMANTIC and SYNTACTIC selection. Let us begin with the most obvious functions, the SUBCATEGORIZED GFs and the non-subcategorized ADJUNCTS. The former are represented in the a–structure as semantic roles that map to the syntactically selected core and non-core arguments. The latter, or the ADJUNCTs, are selected neither syntactically nor semantically. Next the subject and object RAISING functions are never semantically selected for because they are semantically vacuous, but are selected for syntactically. This we know because subjects of raising verbs control agreement.

Finally in Table 1 the fourth quadrant contains the category unselected for syntactically but selected for semantically. So the grammatical function predicted is the INCORPORATE, which possesses one selectional property but lacks the other. Thus the INCORPORATE fills an unexpected gap in the syntactic-semantic interface. And in a more technical sense, LFG does not seem to possess a dedicated mechanism with which to encode the INCORPORATE in the manner that the three other categories in Table 1 have.

5 Conclusion

In this paper, I reviewed the NI construction in Chuj and K'ichee' Mayan. I focused on the incorporated noun in K'ichee and in Chuj's two dialects spoken in Guatemala. I presented the data using the NPSA within LFG architecture. Based on derivational NI morphology, the data support the proposal that NI in Chuj and K'ichee' is analytically formed, and by inference, represents a post-lexical, syntactic construct. I have reviewed the semantics of NI and have concluded that semantics alone falls short of fully accounting for analytic NI. I suggested that the INCORPORATE is a GF unselected for syntactically but selected for semantically. It represents a hybrid category that exhibits a heterogeneous set of properties acquired from subcategorized GFs and non-subcategorized ADJs. The INCORPORATE presents as the following: a non-optional ADJ that structurally manifests lexical head-adjunction as a non-projecting word, obligatory narrow scope, non-extraction, non-iterability, optional control of verb agreement on the basis of the person-salience of arguments, no derivational options (eg. as possessum) except for very limited N–N or A–N compounding, no functional modification, number inflection, and restricted pre-head adjectival and post-head adjectival and relative clause modification functioning in a detransitivized clause.

In the end, I have identified an intermediate linguistic space, as illustrated in Table 1 of this paper. It is occupied by the INCORPORATE, but potentially available to other, similar in kind hybrids. Grimshaw (1990:109–132), for example, long ago introduced the notion of argument suppression manifested as an argument adjunct (*a–adjunct*). Passives, for example, suppress an external argument (EA) a–structure position with the result that it is not θ–marked. But contrarily, the a–adjunct is still related to or licensed by the a–structure. More recently, Rákosi (2006) has proposed a refinement of the generic athematic category of ADJs. As way of explaining circumstantial PPs, Rákosi (2006) introduces the thematic adjunct (ADJ$_\theta$), suggesting that certain types of adjuncts link thematically to a–structure. But it differs from the INCORPORATE in that the use of ADJ$_\theta$ remains optional.

It is also conceivable to consider the INCORPORATE as just a representational expedience. Nonetheless acceptance of an argument-adjunct category, and an inclusive one at that, should make accessible a greater number of previously unexplained linguistic inconsistencies.

References

Asudeh, Ash. 2007. Some notes on pseudo–noun incorporation on Niuean. Unpublished manuscript.

Asudeh, Ash, and Douglas Ball. 2005. Niuean incorporated nominals as non-projecting nouns. In *2005 LSA Annual Meeting*. Handout. Oakland, Ca.

Ball, Douglas. 2005. Phrasal Noun Incorporation in Tongan. In *Proceedings of AFLA XII.* UCLA Working Papers in Linguistics No. 12.

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

Broadwell, George Aaron. 2000. Word Order and Markedness in Kaqchikel. In *The Proceedings of the LFG '00 Conference.* CSLI Publications.

Chung, Sandra, and William A. Ladusaw. 2004. *Restriction and Saturation.* Linguistic Inquiry Monograph 42. Cambridge, Ma.: The MIT Press.

Dayley, Jon P. 1981. Voice and ergativity in Mayan languages. *Journal of Mayan Linguistics* 2(2). 3–82.

Di Sciullo, Anna-Maria, and Edwin Williams. 1987. *On the definition of the word.* Linguistic Inquiry Monographs. Cambridge, Ma.: The MIT Press.

England, Nora C. 1991. Changes in Basic Word Order in Mayan Languages. *IJAL* 57(4). 446–486.

Farkas, Donka F., and Henriëtte E. de Swart. 2003. *The Semantics of Incorporation: From Argument Structure to Discourse Transparency.* Stanford Monographs in Linguistics. Stanford, Ca.: CSLI Publications.

Grimshaw, Jane B. 1990. *Argument structure.* Linguistic Inquiry Monographs. Cambridge, Ma.: The MIT Press.

Hopkins, Nicholas Arthur. 1967. The Chuj language. Ph.D dissertation. Austin, Tx.: University of Texas Press.

Maxwell, Judie. 1976. Chuj intransitives: or when can an intransitive verb take an object?. In *Mayan Linguistics I.* (Ed.) Marlys Stefflre McClaran. 128–140. Los Angeles, Ca.: AISC, University of California.

Mithun, Marianne. 1984. The Evolution of Noun Incorporation. *Language* 60(4). 847–894.

Mondloch, James Lorin. 1981. Voice in Quiché–Maya. Ph.D dissertation. SUNY at Albany, Albany, NY.

Rákosi, György. 2006. On the Need for a More Refined Approach to the Argument-Adjunct Distinction: The Case of Dative Experiencers in Hungarian. In *The Proceedings of the LFG '06 Conference.*

Sadock, Jerrold M. 1986. Some Notes on Noun Incorporation. *Language* 62(1). 19–31.

Toivonen, Ida. 2003. *Non-projecting Words: A Case Study of Swedish Verbal Particles.* Dordrecht, The Netherlands: Kluwer Academic.

Van Geenhoven, Veerle. 1998a. On the Argument Structure of Some Noun Incorporating Verbs in West Greenlandic. In *The Projection of Arguments: Lexical and Compositional Factors.* (Ed.) Miriam Butt and Wilhelm Geuder. 225–263. CSLI Lecture Notes, No. 83. CSLI Publications.

— 1998b. *Semantic Incorporation and Indefinite Descriptions: Semantic and Syntactic Aspects of Noun Incorporation in West Greenlandic.* Dissertations in Linguistics. Stanford, Ca.: CSLI Publications.

DO WE WANNA (OR HAFTA) HAVE EMPTY CATEGORIES?

Yehuda N. Falk

The Hebrew University of Jerusalem

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

This paper revisits the relevance of *wanna* contraction for the existence of empty categories. An analysis of *wanna* contraction is proposed under which empty categories have a role in blocking contraction. It is then shown that alternative analyses (subject sharing, local c-command, and morpholexical analyses) are inadequate. It is suggested that empty categories are a last resort, which can only appear in non-subject LDD constructions because these have to be licensed by inside-out functional uncertainty equations, and the empty categories are needed as the c-structure positions on which these equations are annotated.

1. Prologue

The existence of empty categories is very controversial in LFG. So it is important to search for empirical evidence that bears on the existence of empty categories. While in the LFG literature weak crossover has been the focus of the argument over empty categories (as in Bresnan 1995 and Dalrymple, Kaplan, and King 2001), the most enduring such construction in the broader syntactic literature is contraction, particularly the contraction of *want to* to *wanna*. In this paper, we will evaluate the contraction argument for empty categories.[1]

Looking ahead, we will show that other attempts to account for the contraction facts are untenable, and therefore that contraction does provide evidence for empty categories. However, the same evidence that shows that empty categories do exist in a limited set of constructions shows that they do not exist in many other places in which they have been hypothesized in the transformational literature: in particular, they do not exist in long-distance dependency constructions involving subjects, nor do they exist in non-long-distance-dependency contexts. Finally, we will explain why it is not so terrible to have empty categories, and how they are constrained.

2. The Claim

The facts about contraction which are alleged to be relevant for the existence of empty categories were brought to light by Lakoff (1970: 632), who credited Larry Horn with the observation. Simplifying the examples somewhat,[2] Lakoff observed that contraction of *want to* to *wanna* is possible in (1) but (for most speakers) not in (2).

(1) a. Who do you want to see?
b. Who do you wanna see?

(2) a. Who do you want to see Pnina?
b. *Who do you wanna see Pnina?

The basic observation, abstracting away from the specific theoretical assumptions made by Lakoff, is that in (2) the preposed *wh* element bears the function of object of *want*. The canonical structural position of the object of *want* intervenes between *want* and *to* and it is this, Lakoff claimed, that causes contraction to be blocked. In (1), on the other hand, *who* is the object of *see*, and thus has a canonical structural position which does not block the contraction. We refer to this as the Lakoff/Horn Generalization. Using the term "locally licensed function" to refer to a non-discourse grammatical function, we can state the generalization as (3).

[1] This paper was presented at the LFG07 conference. Thanks to Irit Meir, Yael Ziv, and participants in LFG07 for comments. This paper is dedicated to the memory of my father Paul Falk ז״ל, who passed away 2 Tevet 5767 (22 December 2006). This research was supported in part by the Israel Science Foundation (grant No. 207/07).

[2] Lakoff presented the examples in terms of ambiguity of sentences, using an optionally transitive verb. Simple grammaticality is more straightforward.

(3) **Lakoff/Horn Generalization**
Want to cannot contract to *wanna* if the canonical position of the locally licensed function of a preposed element intervenes between *want* and *to*.

If the Lakoff/Horn Generalization is correct, linguistic theory needs a way to express it. Ideally, such an expression should be one in which the Lakoff/Horn Generalization results naturally from the system, rather than having to be stipulated. Empty categories provide a way to do this. The central idea is that the canonical structural position of the locally licensed function of a preposed element is occupied by an empty category, conventionally represented *e*.

(4) a. Who do you want to meet *e*? (=(1))
b. Who do you want *e* to meet Pnina? (=(2))

In (4b), *want* and *to* are not adjacent; the empty category intervenes. On the plausible assumption that contraction requires adjacency (but see footnote 4), the lack of contraction in (2) follows from the postulation of the empty category.

An empty-category-based approach provides an elegant expression of the Lakoff/Horn Generalization. For this reason, much of the literature which is ostensibly about empty categories focuses on the correctness of the generalization. Following our fleshing-out of an empty-category-based analysis in the next section, we will review the alternative descriptions that have been proposed, and discover that they are all flawed.

3. An Analysis

3.1. *To* Attachment

We begin with the infinitival *to*. As discussed by Jacobson (1982), and in more detail by Zwicky (1982), *to* is phonologically subordinate to adjacent material, with this subordinate status manifesting itself either as becoming part of the phonological phrase (what Zwicky refers to as leaning) or part of the phonological word (cliticization). Zwicky, in particular, shows that various puzzles about the distribution of *to* can be accounted for under such an analysis. We will use the term *attachment* as a way of referring to this phonological subordination which is neutral between leaning and cliticization. Zwicky takes the position that while *to* ordinarily functions as a leaner, in *wanna* contraction it cliticizes. The behavior of *to* in *wanna* contraction is, under this analysis, merely an extreme case of the normal behavior of *to*.

Under normal conditions, *to* attaches to the right, but it is also possible for it to attach to the left, primarily when it is stranded. (Parentheses here indicate phonological phrasing.)

(5) a. (We're nót) (to léave).
b. (We're nót to).

We propose the following constraint on infinitival *to*:

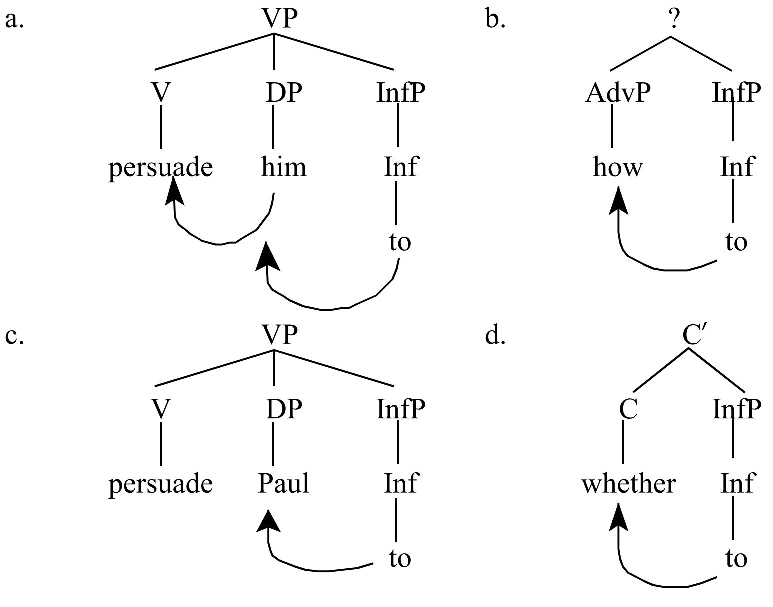
(6) ***To* Attachment** (first approximation)
Infinitival *to* must attach to an adjacent element. It may attach
(a) to the right (the usual situation)
or
(b) to the left

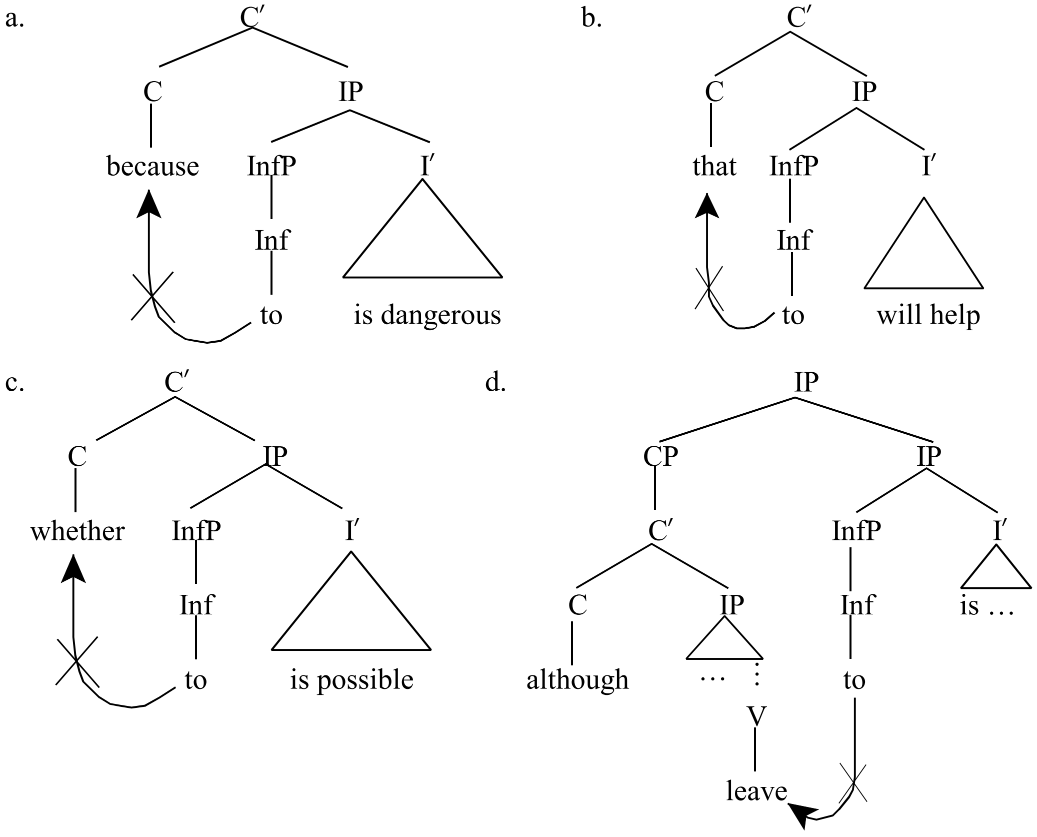
This statement of *to* attachment needs to be refined somewhat. Zwicky observes that there are various constraints on the ability of *to* to attach to the left.

(7) a. I don't know if Paul wants to buy the present, but I think we can (persuáde him to).
b. I might whittle a polar bear out of Ivory soap, but I don't know (hów to).
c. I don't know if he wants to buy the present, but I think we can persuade (Pául to).
d. I might whittle a polar bear out of Ivory soap, but I don't know (whéther to).

(8) a. *You shouldn’t play with rifles, (becáuse to) is dangerous.
b. *You can try to plead with him, but I doubt (thát to) will help.
c. *She’d like to surprise him, but I don't know (whéther to) is possible.
d. Although it would distress us for you to leave, to leave/*∅ is what I’d advise you to do.

Note the structures involved here, with arrows indicating cliticization. (For the sake of neutrality, we label the clause headed by *to* as an InfP, and label some other nodes “?”.)

(9)


(10)


In the grammatical cases, *to* attaches to the left to an element in a very specific structural configuration: the host c-commands *to* and there is no maximal-level phrasal category intervening between the host and the InfP which *to* heads. We will refer to this as local c-command, and revise the statement of *To* Attachment accordingly.

(11) ***To* Attachment**
The complementizer *to* must attach to an adjacent element. It may attach
(a) to the right (the usual situation)
or
(b) to the left onto a locally c-commanding host

(12) **Local C-Command**
X locally c-commands Y iff X c-commands Y and no phrasal node other than the maximal projection of Y intervenes.

As noted by Aoun and Lightfoot (1984), many of the non-long-distance-dependency structures in which *wanna* contraction is barred can be blocked if a local c-command condition is placed on *to* attachment. For example, Carden (1983) observes that the inability of *to* to contract with a *want* which is part of a coordinate structure is not mirrored by other contraction rules, such as the contraction of ...*t you* to [čə].

(13) a. I don't need or want to hear about it. *wanna
b. I don't expect or want you to get involved. ✓čə

The presence of a local c-command condition on the leftward attachment of *to* will rule out *wanna* contraction in this case. Unlike adjacency, which is a consequence of the concept of attachment, we take local c-command to be a rule-specific stipulation.

3.2. On *Want*

We turn now to the verb *want*. From the perspective of coming to an understanding of the nature of *wanna* contraction, there are two crucial facts: *to* cliticizes onto *want*, instead of merely leaning on it; and the /t/ of *want* deletes.

It is observed by Jacobson (1982) that deletion of /t/ is not a peculiarity of *wanna* contraction; rather, it is a general property of the verb *want*. She observes that the /t/ frequently deletes in forms like *wanted* and *wanting*. Strikingly, the phonetic sequence we are representing as *wanna* is not only a realization of *want to*; it is also a realization of *want a*:

(14) a. I wanna play. *wanna*=*want to*
b. I wanna toy. *wanna*=*want a*

The only situation in which the /t/ of *want* is obligatory is in the present subjunctive:

(15) a. I demand that you want a part in the play.
b. *I demand that you wanna part in the play.

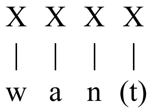
As noted by Brame (1981: 286 fn 13), *want to* cannot contract to *wanna* in the subjunctive either.

(16) a. The director requires that all of the actors want to give their most.
b. *The director requires that all of the actors wanna give their most.

This confirms our view that the deletion of /t/ in *wanna* is no different than deletion of /t/ in other uses of *want*.

While we do not presume to propose a full phonological analysis of the deletability of /t/, a possible analysis would give *want* a phonological representation in which the last skeletal position is only optionally filled by /t/.

(17)

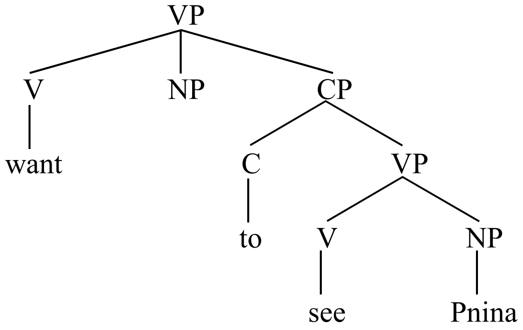


It is possible that it is the availability of an empty skeletal position if the /t/ is not included that drives cliticization of *to* to *want*.

3.3. Empty Categories and Contraction

We hypothesize that the canonical position of the locally licensed function of a fronted element is (sometimes, at least) marked by an empty category, an unfilled phrasal node. For example, the VP headed by *want* in (2) has the following structure:

(18)



The critical question is what happens in (18) if *to* attempts to attach to the left. The local c-command condition does not block the leftward attachment, since *want* does locally c-command *to*. The condition that is relevant is adjacency, which, as already noted, we take to be an integral property of attachment phenomena. In this case, the applicability of the adjacency condition is not entirely clear. Phonologically, *want* and *to* are adjacent; no phonological material intervenes between them. They are also adjacent at the level of terminal elements in the constituent structure; there is no terminal element that intervenes between them. However, at higher levels of structure they are not adjacent: the unfilled NP intervenes. That is to say, the interpretation of the adjacency condition in a structure with an empty category depends on which part of the structure is relevant. Given the ambiguous status of adjacency in this case, one might expect variability between speakers, with some treating *want* and *to* in (18) as adjacent, and others as not adjacent. Such an expectation would be well founded.

It is well known in the literature on *wanna* contraction that not all speakers share the judgment in (2). For some, often referred to as speakers of the liberal dialect, *wanna* contraction is possible in sentences of this kind. Zwicky (1982) takes this one step further. He notes that, even in the absence of contraction, speakers differ on the acceptability of B's utterance in the following discourse (Zwicky 1982: 26):[3]

(19) A: Who do you want *e* to vanish?
B: %I don't know; who do you want *e* to?

Since *to* must attach to something, and it is stranded on the right, it must attach to the left. If it cannot at least lean, the sentence is ungrammatical. Apparently, some speakers allow *to* to lean onto *want* in this case and others do not; Zwicky reports a 50-50 split among speakers he consulted. The number of speakers who accept *wanna* in these environments is apparently less: Zwicky suggests that not all

[3] For what it's worth, the author of this paper finds B's response crashingly bad.

speakers who allow leaning allow cliticization.[4]

Empty categories, by their very nature, have an ambiguous status in terms of adjacency. This ambiguity leads to a situation where the Lakoff/Horn Generalization is valid for most speakers, but not all. Strikingly, while the existence of the liberal dialect has been cited by Pullum and Postal (1979) as an embarrassment for an empty-category account of the Lakoff/Horn Generalization, it actually provides exactly the right tools for explaining this inter-speaker variation.

4. Alternative Analyses

4.1. Subject Sharing

The first challenge to the Lakoff/Horn Generalization came from Postal and Pullum (1978). Much of their argument concerns not the Lakoff/Horn Generalization, but rather the implementation of an empty category analysis in the Extended Standard Theory literature of the time. Many of their arguments against this kind of analysis are valid, but irrelevant to other implementations, including ours. However, in the course of arguing against "trace theory," they propose an alternative to the Lakoff/Horn Generalization.

As Postal and Pullum observe, *want* is not the only verb which contracts with *to*. The following is presented as a complete list by Pullum (1997: 81).

(20)	*want*	*wanna*
	prospective *go*	*gonna*
	habitual *used*	*usta*
	have (necessity/obligation)	*hafta*
	got (necessity/obligation)	*gotta*
	ought	*oughta*
	supposed	*supposta* or *sposta*

Postal and Pullum observe that, other than *want*, these are all are Raising-to-Subject verbs. *Want*, when contraction is possible, is a Subject Equi verb. What unites all of these cases is that the two clauses share a subject. They therefore propose that contraction is possible when the two clauses share a subject, or, in LFG terms, functional control.

The subject-sharing alternative to the Lakoff/Horn Generalization appears to have never been subjected to critical scrutiny in the literature. A close look reveals several problems. For example, it is only with *want* that there is a contrast between environments that allow contraction and those that do not. It is therefore difficult to draw conclusions about the source of the *wanna* contraction facts from these other verbs. Therefore, contra Postal and Pullum, it is not clear that the other verbs are relevant for determining the conditions under which contraction does and does not occur. Another problem is that the analysis is essentially arbitrary; unlike the empty category analysis, which is based on the idea that an intervening element breaks the contiguity necessary for contraction, there is no inherent relation between subject sharing and contraction.

However, the biggest problem with the subject sharing analysis is that it is empirically incorrect. The context in which *want to* can contract to *wanna* includes cases in which the reference of the subject of *want* is a subset of the reference of the subject of the *to* clause; i.e. cases which cannot be analyzed as functional control.

[4] Postal and Pullum (1978) mention a different idiolectal treatment, which appears to dispense with the adjacency condition. In this idiolect, the following is grammatical:

(i) I wanna very much go to the game tomorrow.

Without further information, it is difficult to know what lies behind judgments such as this. A very preliminary speculation would be that *to* undergoes prosodically motivated movement, much as second-position clitics do in some languages. On the movement of clitics, see Halpern (1995) and, in LFG, Kroeger (1993).

(21) a. *I met on Sunday at 10:00
b. I wanna meet on Sunday at 10:00.
c. *I hafta meet on Sunday at 10:00.
d. *I tried to meet on Sunday at 10:00.

As (21a) shows, *I* is not a possible subject for the intransitive verb *meet*, since it requires a plural subject in this subcategorization frame. Thus in (21b), the subject of *meet* cannot be *I*, but rather some group including *I*. This contrasts with a Raising verb like *hafta* (21c), which has to be functional control. This overlapping reference is also not a necessary property of Equi constructions: in the case of *try* there is functional control, resulting in the ungrammaticality of (21d). The overlapping reference that is possible in the case of *want* requires an anaphoric control analysis.

We conclude, therefore, that functional control (subject sharing) cannot be the property that licenses *wanna* contraction. Subject sharing is therefore not a possible alternative to the Lakoff/Horn Generalization.

4.2. Local C-Command

Another alternative that has been suggested in the literature (e.g. Bouchard 1984 and Barss 1995) is that the condition that *want* locally c-command *to* obviates the need for the Lakoff/Horn Generalization. While we have adopted such a condition on the leftward cliticization of *to*, the process that underlies *wanna* contraction, the claim by Bouchard and Barss is problematic.

This proposal has been made within the context of Government/Binding (or Principles and Parameters) theory, in which local c-command, under the name government, is taken to be one of the fundamental structural relations in syntax. The local domain of c-command is said to be delimited by certain nodes. One of these "barriers to government" is the CP node, and it is this which is taken to block contraction when the fronted element functions as the subordinate subject.

(22) a. *Who do you wanna see Pnina? (=2b)
b. … want [$_{\text{CP}}$ [$_{\text{IP}}$ *e* to see Pnina]]

Under this analysis, contraction would involve the matrix verb *want* and the head of the IP embedded within the CP. The intervening CP node renders the c-command non-local, and contraction is therefore impossible. Under standard GB assumptions, however, the same configuration would obtain in the case of grammatical contraction.

(23) a. Who do you wanna see? (=1b)
b. … want [$_{\text{CP}}$ [$_{\text{IP}}$ PRO to see *e*]]

The local c-command analysis of contraction has to therefore make an additional assumption: namely that control complements are bare IPs:

(24) … want [$_{\text{IP}}$ PRO to see *e*]

In this structure, there is no CP barrier between *want* and the IP headed by *to*, and contraction is therefore possible.

The viability of the Bouchard/Barss local c-command analysis depends on the plausibility of the proposed structures, in particular the status of the controlled infinitive as a bare IP and the categorization of *to* as an infl. The latter, while standard in the transformational literature, is not obviously correct; Falk (2001: 154) argues that *to* is a complementizer. If *to* is a complementizer, there cannot be any structural difference between (22b) and (23b); ignoring the positions of possible empty categories, the structures in question would be:

(25) a. … want [$_{\text{CP}}$ to see Pnina]
b. … want [$_{\text{CP}}$ to see]

However, even assuming an infl analysis for *to*, the analysis of the controlled complement of *want* as a bare IP seems dubious: its distribution is that of a CP, not an IP.

(26) a. [To see Pnina] is what I want.
b. [$_{CP}$ That I might see Pnina] is what I said.
c. *[$_{IP}$ I might see Pnina] is what I said.

(27) a. I want very much [to see Pnina].
b. I said very loudly [$_{CP}$ that I would see Pnina].
c. *I said very loudly [$_{IP}$ I would see Pnina].

We therefore consider the Bouchard/Barss analysis to be untenable.

4.3. *Wanna* as a lexical item

It has also been proposed that *wanna* is a lexeme distinct from *want*, and that there is no actual contraction in *wanna* sentences. If this is correct, syntactic structures are irrelevant to *wanna*, and what matters is the nature of the morphological relation between *want* and *wanna*. Such analyses have been proposed by several researchers (for example, Brame 1981: 286 fn 13), but the most thorough argument for it is that of Pullum (1997). As we will show here, we find Pullum's argument unconvincing.

The heart of Pullum's argument is that *to* contraction is morphologically and phonologically idiosyncratic. In this respect, the argument mirrors the argument presented by Zwicky and Pullum (1983) that *n't* is an inflectional suffix in contemporary English, and not a contracted form of *not*. However, while the argument is quite compelling in the case of *n't*, it is more problematic with *wanna*. We will discuss the claim of morphological idiosyncrasy first, and then phonological idiosyncracy.

Pullum's argument for morphological idiosyncracy is that only a limited set of verbs can contract with *to*, the ones listed in (20) above. For example, while *ought to* contracts to *oughta*, *thought to* does not contract to **thoughta*. Idiosyncracy of this kind is typical of morphology, not of syntacto-phonological contraction. However, as we have seen, the situation is more complicated: *to* obligatorily attaches, and when it attaches to the left, some verbs allow cliticization. So while it is true that *to* does not cliticize to *thought* the way it does to *ought*, it does lean on it. Pullum's proposal that forms like *wanna* and *oughta* are lexically derived is incompatible with the analysis of *to* presented earlier. Treating *wanna* as derived by phonological contraction (cliticization) forms a more harmonic part of an overall analysis of the phonological properties of infinitival *to*.

The argument for phonological idiosyncracy is based on the presence of irregular phonological changes in the form of the host of *to*. For example, Pullum discusses the devoicing of /v/ in *hafta* and observes that in other cases (e.g. *Aztec*) such devoicing does not occur. However, Pullum also notes that /hæf/ appears to have become a new underlying form for the verb which forms the core of *hafta*, at least for some speakers.[5] Similarly, Andrews (1978: 267) suggests that *usta* and *supposta* have underlying voiceless /s/. He reports the following judgments:

(28) a. Did they [yuws(t)] not to eat pickles?
b. *Did they [yuwz] not to eat pickles?
c. *usen't* = [yuwsnt], *[yuwznt]
c. You're [səpowst] not to light the wick until it's wet.

In other words, at least some of the cases Pullum cites may involve lexical reanalysis of the host, rather than lexical attachment of *to*.

The only case which is relevant for testing the Lakoff/Horn Generalization is *wanna*. The phonological change in question is the deletion of /t/ after /n/. As Pullum notes, this is not an automatic phonological rule in English: for example, in *wont to* (Postal and Pullum 1978: 2) and *taunt* (Pullum

[5] He reports that some speakers pronounce *having to* as [hæfiŋtʊ]. Even for those who do not, however, an underlying form /hæf/ in allomorphic alternation with /hæv/ is not inconceivable; alternatively, there are distinct verbs *haf* and *have*.

1997: 90) there is no deletion of /t/. However, as Pullum himself observes, /nt/ does sometimes at least optionally reduce to [n] word-internally (as in *Santa*, *twenty*, etc.). More importantly, as we observed above, following Jacobson (1982), this deletion of /t/ is a general property of the verb *want*. The deletion of /t/ in *wanna* is no different than deletion of /t/ in other uses of *want*, and, contra Pullum, is not an idiosyncratic phonological result of adding *to* to *want*.

A further argument against an analysis in terms of derivational morphology is provided by Hudson (2006), who observes that the infinitive following *wanna* can be coordinated with a *to* infinitive, a property not shared with bare infinitives. (It is more felicitous without the *to*, coordinating just the VPs, which is why Hudson marks the example as "?".)

(29) a. I wanna go to sleep and (?to) not wake up until I feel better.
b. He let me go to sleep and (*to) not wake up until I feel better.[6]

The acceptability of the coordination with a *to* infinitive is unexpected if *wanna* is a verb that takes a bare infinitive complement. However, Hudson's alternative does not fare much better. Under Hudson's account, *wanna* is the morphophonological realization of the sequence *want to*, i.e. a morphological unit which realizes a sequence of two words. For most speakers, this realizational rule is limited to the variety of *want* that takes an infinitive and no object; this accounts for the Lakoff/Horn Generalization, but in a totally ad hoc manner. Hudson justifies the use of a lexical realizational rule on the grounds that phonological rules should express generalizations, not a phenomenon that is as restricted as *wanna* contraction; assuming Pullum's entire list (as he does) it is still a small number of verbs. The place for such phenomena is in the lexicon. However, as we have seen, the facts of *wanna* contraction are the consequence of the interaction of the attachment properties of *to*, the lexical phonological properties of *want* (which are independent of *wanna* contraction), and the operation of cliticization. Placing the entire phenomenon in the lexicon leads to loss of generalization, rendering other syntactic consequences of the attachment properties of *to* a distinct phenomenon.

The evidence therefore points to a phonosyntactic contraction (cliticization) analysis. There is no support for a lexical analysis.

5. Consequences

We conclude that, when all the facts are considered, *wanna* contraction can best be described in a framework in which the canonical position of the locally licensed function of a fronted element is occupied by an empty category—a phrasal node which dominates no terminal nodes. This empty element blocks the leftward cliticization of *to* onto *want* for most speakers by breaking the adjacency between them.

5.1. Other Empty Categories

It was in the early Principles and Parameters rhetoric that the idea that *wanna* contraction provides evidence for empty categories was first raised. However, already in some of the earliest literature on the matter, it was observed that not all postulated empty categories block *wanna* contraction. The structure assumed in Principles and Parameters for (1), for example, is (30).

(30) [who do you want [*e* to see *e*]]

The second empty category in (30) is the one we have been discussing. However, there is another empty element in this hypothesized structure, occupying the canonical position of the controlled subject of *see*.[7] This empty element, no less than the one in (2), intervenes between *want* and *to*. If the inability of most speakers to contract in (2) is evidence for the empty category, their ability to contract in (1)/(30) is

[6] Hudson has *not* and *to* in the other order, but this appears to be what he intends.

[7] This additional empty category is usually referred to as PRO.

evidence against the presence of an empty category in that position. Similarly, as originally noted by Postal and Pullum (1978), the other verbs listed in (20) are Raising verbs: in Raising constructions yet another empty category is hypothesized in Principles and Parameters, and this one also does not block contraction.

We must also consider auxiliary contraction, in particular the contraction of *is*. As shown in the following examples from Carden (1983: 45), auxiliaries can contract over the canonical position of the subject in long-distance dependency constructions.

(31) a. Who do you think's gonna win?
b. Jack is the man that I bet's gonna win.

On the assumption that long-distance dependency constructions always have an empty category in the canonical position of the locally licensed function, these sentences should have the following representations.

(32) a. Who do you think [*e* is gonna win]?
b. Jack is the man that I bet [*e* is gonna win].

However, the presence of empty categories such as these ought to block contraction for the same speakers for whom *wanna* contraction is blocked by empty categories. The possibility of contraction thus indicates the absence of an empty category preceding *is* in these sentences:

(33) a. Who do you think [is gonna win]?
b. Jack is the man that I bet [is gonna win].

The difference between these cases and the earlier ones involving *wanna* is that in the previous cases the locally licensed function of the fronted element is arguably object (assuming a Raising-to-Object analysis for *want*), while in the auxiliary contraction cases the fronted element bears no locally licensed function other than subject. The conclusion that we draw from these facts is that when the locally licensed function of a fronted element is subject, there is no empty category in its canonical position. To summarize, the motivated structures are the following:

(34) a. Who do you want [$_{\text{CP}}$ to see *e*]?
b. Who do you want *e* [$_{\text{CP}}$ to see Pnina]?
c. Who do you think [$_{\text{IP}}$ is going to win]?

The conclusion, then, is that long-distance dependency constructions involve the use of empty constituent structure in the canonical structural position of the locally licensed function, unless this function is "subject". This conclusion converges with similar proposals made in other studies, such as Gazdar (1981) and Falk (2006). However, it clashes sharply with the view in the P&P tradition.

To summarize: Pronominal empty categories and empty categories for "NP movement" constructions do not exist, and neither do empty categories in canonical subject position even for long-distance dependency constructions.

5.2. Constraint-Based Syntax

Constraint-based theories of syntax are not inconsistent with the existence of empty categories, as can be seen by examining such studies as Gazdar (1981), Kaplan and Bresnan (1982), Zaenen (1983), Gazdar, Klein, Pullum, and Sag (1984), Pollard and Sag (1994), Bresnan (1995, 2001), Falk (2001), and Culicover and Jackendoff (2005). However, there is a natural suspicion of empty categories among people working in such frameworks. As Dalrymple (2001: 415) puts it,

> Further work will reveal … whether incontrovertible evidence exists for traces, gaps, or empty phrase structure categories. In the absence of such evidence, a simpler and more parsimonious theory of long-distance dependencies results if traces [i.e. empty

categories] are not allowed.

This suspicion stems in part from a perception that the P&P tradition shows that once empty categories are recognized, there is a tendency for the inventory of empty categories to grow in an unconstrained fashion.[8] As with all elements that are not overt in the actual utterance, these empty categories are prone to such unconstrained proliferation, as well as posing issues of parsing and the like.[9] For this reason, many studies in constraint-based approaches have championed non-empty-category approaches to long-distance dependency constructions (see, for example, Kaplan and Zaenen 1989, Sag and Fodor 1994, Ginzburg and Sag 2000, Bouma, Malouf, and Sag 2001, Dalrymple 2001, and Dalrymple, Kaplan, and King 2001).

The empirical evidence confirms the suspicion of empty categories up to a point. It is striking that the evidence points to the nonexistence of empty categories in so many of the contexts in which they have been hypothesized. Our proposal is that while empty categories are allowed, they are a dispreferred last resort. As noted by Bresnan (2001), the last-resort status of empty categories is a consequence of the principle of Economy of Expression, under which syntactic nodes are present only when needed to license grammatical f-structures or for their semantic content. Empty categories have no semantic content, so their only possible role is in licensing grammatical f-structures. If an alternative method is available for licensing the same f-structure, the use of an empty category will be blocked by Economy of Expression. The last-resort nature of empty categories explains the empirical evidence of their rather limited distribution.

Most of the constructions for which empty categories have been proposed are lexical constructions, for which empty categories are unnecessary. The only confirmed cases of empty categories have been in long-distance dependency constructions. Long distance dependency constructions, unlike the other constructions for which empty categories have been proposed, are not argument-realization constructions and thus not lexical. Instead, they are multifunctionality constructions in which an argument is paired with a **non**-argument function. In sentences such as the following, for example, the sole argument function of the NP *Pnina* is as subject or oblique object of the verb *spoke*, just as it would be in an ordinary non-LDD sentence. The reason that the NP appears at the beginning of the sentence is because it has an additional discourse-related function. This additional function is not a result of the argument status of *Pnina*; instead, it is related to the discourse in which the sentence is embedded.

(35) a. Pnina, I said that you think spoke to us
b. Pnina, I said that you think we spoke to.

The fact that this is not an argumenthood-related phenomenon is reinforced by the fact that adjuncts can also be assigned this additional discourse function.

(36) a. When do you think the plane will arrive?
b. How did Yoni say he would fix the sink?

Assimilating LDD constructions to lexically-based constructions would force us to represent adjuncts as part of the lexical selectional properties of verbs, as is done by Bouma, Malouf, and Sag (2001) in an HPSG analysis.

Given that LDD constructions are not lexical and that Universal Grammar allows them, there must be some non-lexical mechanism for licensing them. One possibility would be for the clause in which the "fronted" element is located to include an outside-in equation stating that this discourse-function-bearing element bears some grammatical function in a clause arbitrarily far down: outside-in

[8] As a student of mine once blurted out in class: "If you assume traces, you may as well be doing GB!"

[9] We will not address the question of how speakers recognize empty categories if they cannot be heard. In our view it is tautological that if it can be shown that empty categories exist, the parsing mechanism for hypothesizing them must also exist. Similarly, we do not consider it a problem that in languages with freer constituent order than English the empty category could occupy more than one linear position in the c-structure. There may be principles that restrict its linear position (such as a principle that places lighter elements before heavier ones); if not, what results is a case of innocuous structural ambiguity.

functional uncertainty:

(37) a. (↑ FOCUS) = (↑ COMP* SUBJ)
b. (↑ FOCUS) = (↑ COMP* OBJ)

However, while (37a) is unproblematic, (37b) runs afoul of the theory of subjecthood proposed by Falk (2006). According to Falk's theory, functional equations are not free to reference any element in a lower clause: they are limited to referencing the subject.[10] This limitation, which is justified on conceptual grounds, is responsible for the restriction of functional controllees to subjects. Just as the lower element in a control equation is limited to SUBJ, the lower element in an LDD-licensing functional uncertainty equation is limited to SUBJ.

If this line of argumentation is correct, Universal Grammar faces a problem: how to license LDD constructions in which the locally-licensed grammatical function is not SUBJ. The only possibility left is an equation associated with an element of the lower clause, an inside-out equation, which would have to be associated with the node representing the locally-licensed function. Since there is no lexical content to fill such a node, the result is an empty category: the only means available to Universal Grammar to license the construction.

Despite our endorsement above of the view expressed by Bresnan (2001) that the last-resort status of empty categories is a consequence of LFG's Economy of Expression principle, our approach differs crucially from Bresnan's. For Bresnan, inside-out functional uncertainty is necessitated in languages like English to identify the locally licensed function of the fronted element because of the lack of morphological devices such as Case; languages in which such morphological marking exists do not use empty categories. Under the present approach, all languages need empty categories to license non-subject LDD constructions because of the restriction of the lower end of an outside-in designator to SUBJ, regardless of the morphological devices available in the language.

The last-resort view of empty categories, combined with a distinction between lexical and constructional phenomena and Falk's theory of subjecthood, results in a situation in which empty categories are present only in non-subject LDD constructions. This agrees with the results of our empirical investigation into the distribution of empty categories.

References

Andrews, Avery (1978) "Remarks on *to* Adjunction." *Linguistic Inquiry* 9: 261–8.

Aoun, Joseph, and David Lightfoot (1984) "Government and Contraction." *Linguistic Inquiry* 15: 465–473.

Barss, Andrew (1995) "Extraction and Contraction." *Linguistic Inquiry* 26: 681–694.

Bouchard, Denis (1984) *On the Content of Empty Categories*. Dordrecht: Foris.

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

Brame, Michael (1981) "Trace Theory with Filters vs. Lexically Based Syntax Without." *Linguistic Inquiry* 12: 275–293.

Bresnan, Joan (1982a) "The Passive in Lexical Theory." in Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 3–86.

Bresnan, Joan (1982b) "Control and Complementation." in Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 282–390.

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, ed., *Formal Issues in Lexical-Functional Grammar*. Stanford, Calif.: CSLI Publications. 241–274.

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

Carden, Guy (1983) "The Debate About *wanna*: Evidence from Other Contraction Rules." in John F. Richardson, Mitchell Marks, and Amy Chukerman, ed., *Papers from the Parasession on the Interplay of Phonology, Morphology, and Syntax*. Chicago: Chicago Linguistic Society. 38–49.

[10] More precisely, they must reference the pivot, roughly in the sense of studies of ergativity such as Dixon (1994). For the purposes of the present study, this distinction is unnecessary, so the text will refer to the more familiar concept of subject.

Chomsky, Noam (1976) “Conditions on Rules of Grammar.” *Linguistic Analysis* 2: 303–351.

Culicover, Peter W., and Ray Jackendoff (2005) *Simpler Syntax*. Oxford: Oxford University Press.

Dalrymple, Mary (2001) *Lexical-Functional Grammar* (Syntax and Semantics, Vol. 34). 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, ed., *Proceedings of the LFG01 Conference, University of Hong Kong*. On-line: CSLI Publications. 66–82. http://cslipublications.stanford.edu/LFG/6/lfg01.html

Dixon, R.M.W. (1994) *Ergativity*. Cambridge: Cambridge University Press.

Falk, Yehuda N. (2001a) *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Stanford, Calif.: CSLI Publications.

Falk, Yehuda N. (2006) *Subjects and Universal Grammar: An Explanatory Theory*. Cambridge: Cambridge University Press.

Gazdar, Gerald (1981) “Unbounded Dependencies and Coordinate Structure.” *Linguistic Inquiry* 12: 155–184.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, and Ivan Sag (1984) *Generalized Phrase Structure Grammar*. Oxford: Basil Blackwell.

Ginzburg, Jonathan, and Ivan A. Sag (2000) *Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives*. Stanford, Calif.: CSLI Publications.

Halpern, Aaron (1995) *On the Placement and Morphology of Clitics*. Stanford, Calif.: CSLI Publications.

Hudson, Richard (2006) "*Wanna* Revisited." *Language* 82: 604--627.

Jacobson, Pauline (1982) “Evidence for Gap.” in Pauline Jacobson and Geoffrey K. Pullum, ed., *The Nature of Syntactic Representation*. Dordrecht: Reidel. 187–228.

Jaeggli, Osvaldo A. (1980) “Remarks on *to* Contraction.” *Linguistic Inquiry* 11: 239–245.

Kaplan, Ronald M., and Joan Bresnan (1982) “Lexical-Functional Grammar: A Formal System for Grammatical Representation.” in Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 173–281.

Kaplan, Ronald M., and Annie Zaenen (1989) “Long-Distance Dependencies, Constituent Structure, and Functional Uncertainty.” in Mark R. Baltin and Anthony S. Kroch, ed., *Alternative Conceptions of Phrase Structure*. Chicago: University of Chicago Press. 17--42.

Kroeger, Paul (1993) *Phrase Structure and Grammatical Relations in Tagalog*. Stanford, Calif: CSLI Publications.

Lakoff, George (1970) “Global Rules.” *Language* 46: 627–639.

Lightfoot, David (1976) “Trace Theory and Twice-Moved NPs.” *Linguistic Inquiry* 7: 559–582.

Pollard, Carl, and Ivan A. Sag (1994) *Head-Driven Phrase Structure Grammar*. Stanford, Calif.: CSLI Publications.

Postal, Paul M. (1974) *On Raising: One Rule of English Grammar and Its Theoretical Implications*. Cambridge, Mass.: MIT Press.

Postal, Paul M., and Geoffrey K. Pullum (1978) “Traces and the Description of English Complementizer Contraction.” *Linguistic Inquiry* 9: 1–29.

Pullum, Geoffrey K. (1997) “The Morpholexical Nature of English *to* Contraction.” *Language* 73: 79–102.

Pullum, Geoffrey K., and Paul M. Postal (1979) “On an Inadequate Defense of “Trace Theory”.” *Linguistic Inquiry* 10: 689–706.

Radford, Andrew (1997) *Syntactic Theory and the Structure of English: A Minimalist Approach*. Cambridge: Cambridge University Press.

Sag, Ivan, and Janet D. Fodor (1994) “Extraction Without Traces.” *WCCFL* 13: 365–384.

Uriagereka, Juan (1998) *Rhyme and Reason: An Introduction to Minimalist Syntax*. Cambridge, Mass.: MIT Press.

Zaenen, Annie (1983) “On Syntactic Binding.” *Linguistic Inquiry* 14: 469–504.

Zwicky, Arnold (1982) “Stranded *to* and Phonological Phrasing in English.” *Linguistics* 20: 3–57.

Zwicky, Arnold M., and Geoffrey K. Pullum (1983) “Cliticization vs. Inflection: English *n’t*.” *Language* 59: 502–513.

A FORMAL ANALYSIS OF THE VERB COPY CONSTRUCTION IN CHINESE

Ji Fang Peter Sells

Palo Alto Research Center SOAS, University of London

Proceedings of the LFG07 Conference

Tracy Holloway King and Miriam Butt (Editors)

2007

CSLI Publications

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

Abstract

This paper presents a formal analysis of the verb copy construction in Modern Chinese. Unlike the previous analyses, in which this construction is analyzed as a single-headed structure with the second VP as the head and the first VP as an adjunct, our analysis treats the verb copy construction as a coordinated VP, with each VP as a co-head. We further propose that the first VP subsumes the following VPs in this construction. We also show that this alternative approach can successfully capture and explain all of the three key properties that characterize this construction.

1 Introduction

This paper presents a formal analysis of the verb copy construction (VCC) in the framework of LFG (Kaplan and Bresnan 1982, Bresnan 2001, Dalrymple 2001), and this section provides an overview of the VCC and its key properties in Modern Chinese.

The verb copy construction in Modern Chinese refers to a construction in which the verb must be duplicated before its post-verbal adjunct (such as an adverbial phrase), in the presence of at least one other post-verbal constituent (such as an object). For example:

(1) a.	张三	学	中文	学	得	很	好。
	ZhangSan	xue	zhongwen	xue	de	hen	hao
	ZhangSan	study	Chinese	study	DE[1]	very	well

‘ZhangSan studies/studied Chinese very well’

b.	*张三	学	中文	得	很	好。
	ZhangSan	xue	zhongwen	de	hen	hao
	ZhangSan	study	Chinese	DE	very	well

The contrast between (1a) and (1b) shows that the verb *xue* ‘study’ must be duplicated before its post-verbal adjunct *de hen hao* ‘very well’, in the presence of the object *zhongwen* ‘Chinese’.

(1a) and (2) below represent a typical type of VCC in Modern Chinese, with the word order Verb-Object-Verb-Post-Verbal Adjunct. The post-verbal adjunct can be an adverbial phrase (as illustrated in (1a)), but it can be another category as well. For example,

(2)	张三	学	中文	学	了	三	年。
	ZhangSan	xue	zhongwen	xue	le	hen	hao
	ZhangSan	study	Chinese	study	ASP	three	year

‘ZhangSan have studied Chinese for three years’

In (2), the post-verbal adjunct consists of a noun phrase *san nian* ‘three years’.

Previous studies (C. Li 1975, Huang 1982, Gouguet 2004, 2006, etc.) addressing this construction have focused exclusively on this particular type, in which the first verb is followed by an object, and the verb is duplicated only once. However, this is by no means the only type of the verb copy construction, as shown by (3).

[1] In this paper, DE is a marker for introducing post-verbal adjuncts in Modern Chinese; ASP stands for ‘aspect marker’; CL stands for ‘classifier’.

(3) a.	张三	玩	了	一	天	玩	得	很	累。
	ZhangSan	wan	le	yi	tian	wan	de	hen	lei
	ZhangSan	play	ASP	one	day	play	DE	very	tired

‘ZhangSan played for a day and was/is tired.’

b.	我	送	他	这	件	礼物	送	得	很	好。
	wo	song	ta	zhe	jian	liwu	song	de	hen	hao
	I	give	him	this	CL	gift	give	DE	very	well

‘I gave him this gift and it turned out to be a very good idea.’

c.	他	玩	游戏	玩	了	一	天	玩	得	很	累。
	ta	wan	youxi	wan	le	yi	tian	wan	de	hen	lei
	he	play	game	play	ASP	one	day	play	DE	very	tired

‘He played games for a day and was/is tired.’

d.*	我	送	他	送	这	件	礼物	送	得	很	好。
	wo	song	ta	song	zhe	jian	liwu	song	de	hen	hao
	I	give	him	give	this	CL	gift	give	DE	very	well

(3a-c) are all examples of the VCC. However, note that in (3a), both verbs are followed by a post-verbal adjunct instead of an object in the first VP. In (3b), the first verb is followed by two objects instead of just one, and in (3c), the verb is duplicated more than once.

Previous studies have also focused on the idea that the VCC in Chinese is motivated by a condition on VP that the verb can only have one complement (C. Li 1975, Y. Li 1985, Dai 1992). Examples such as (3b) apparently do not support this claim, and in fact, the verb cannot be duplicated before the second object, as shown in (3d).

Instead, the correct generalization is that the first VP in the VCC must contain ALL the overt internal arguments of the verb (if there is any internal argument at all), and the next VP(s) contain only single post-verbal adjuncts (Fang 2005). Therefore, the ditransitive verb *song* ‘give’ in (3b) cannot be duplicated before the second object, and both objects have to be contained in the first VP (see *(3d)). The internal argument *youxi* ‘game’ in (3c) has to be in the first VP as well (and not simply in any of the (other) VPs).

Similarly, because both the object *yi ben shu* ‘a book’ and the oblique *zai zhuo shang* ‘on the desk’ are internal arguments of the verb *fang* ‘put’ in (4), the verb cannot be duplicated before the oblique, as shown by the contrast between (4a) and (4b). However, it must be duplicated before the post-verbal adjunct if present, as shown in (4c).

(4)a.	张三	放	了	一	本	书	在	桌	上。
	ZhangSan	fang	le	yi	ben	shu	zai	zhuo	shang
	ZhangSan	put	ASP	one	CL	book	at	desk	top

‘ZhangSan put a book on the desk.’

b.*	张三	放	了	一	本	书	放	在	桌	上。
	ZhangSan	fang	le	yi	ben	shu	fang	zai	zhuo	shang
	ZhangSan	put	ASP	one	CL	book	put	at	desk	top

c. 张三 放 了 一 本 书 在 桌 上 放 了 很 久。
ZhangSan fang le yi ben shu zai zhuo shang fang le hen jiu
ZhangSan put ASP one CL book at desk top put ASP very long
'ZhangSan put a book on the desk, and he left it there for a long time.'

In contrast, other VPs in the VCC, which do not contain internal arguments, must contain only one post-verbal adjunct, and the verb must be duplicated before each post-verbal adjunct if there is more than one post-verbal adjunct, as shown in (3a) and (3c).

To summarize, the VCC in Modern Chinese can be schematized as follows.

(5) The Verb Copy Construction Schema

First VP: Verb Object(s)/Post-verbal Adjunct
then
Second VP: Verb Post-verbal Adjunct
iterating to
Nth VP: Verb Post-verbal Adjunct

It is also important to point out that this construction has the following three key properties (i-iii) below. The third property has not been mentioned yet, but is crucial in diagnosing the right analysis for the VCC.

(i) The VP involving object(s), if it occurs, must occur before any other VP(s) involving a post-verbal adjunct in the VCC. In other words, the order in the VCC exhibits a type of asymmetry, namely, object(s) must precede post-verbal adjunct(s), as shown by the contrast between (6a) and (6b).

(6) a. 张三 弹 钢琴 弹 得 很 好。
ZhangSan tan gangqin tan de hen hao
ZhangSan play piano play DE very well
'ZhangSan plays piano very well.'

b. *张三 弹 得 很 好 弹 钢琴。
ZhangSan tan de hen hao tan gangqin
ZhangSan play DE very well play piano

(ii) VCC can be extended to multiple verb copying cases, as shown in (3c), repeated below as (7).

(7) 他 玩 游戏 玩 了 一 天 玩 得 很 累。
ta wan youxi wan le yi tian wan de hen lei
he play game play ASP one day play DE very tired
'He played games for a day and was/is tired.'

(iii) Object extraction from the first VP is allowed in the VCC, but only if the first VP contains another object, as illustrated by the contrast between (8a) and (8b).

(8)a. 这 件 礼物 我 送 他 <gap> 送 得 很 好 。
zhe jian liwu wo song ta <gap> song de hen hao
this CL gift I give him <gap> give DE very good
'I gave him this gift and (it turned out to be) very good.'

b. * 钢琴 张三 弹 <gap> 弹 得 很 好。
gangqin ZhangSan tan <gap> tan de hen hao
piano ZhangSan play <gap> play DE very well

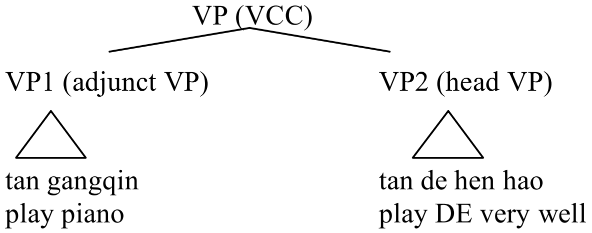
These three properties are key aspects of the VCC in Modern Chinese, and it is the goal of this paper to provide a formal analysis of this construction that can capture and explain all of these properties. Previous analyses of this construction are reviewed in section 2, and we then present our alternative analysis in section 3.

2 Previous Analyses

In previous approaches (Huang 1982, Gouguet 2004, 2006), the VCC is analyzed as a single-headed structure; the second VP is the head and the first VP is an adjunct, either base-generated or created as a movement copy. However, these approaches cannot capture and explain all of the three key properties of the VCC presented above.

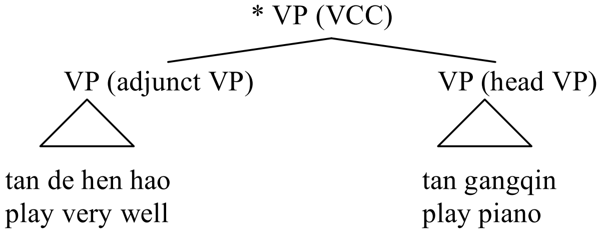
Huang (1982) proposes that the VCC in Modern Chinese is a single-headed construction, and the second VP is the main predicate. He also proposes that the first VP is the adjunct of the second VP in a VCC, as shown in (9).

(9)



Huang's analysis does not capture the three key properties of the VCC. First of all, it does not explain the word order asymmetry, namely, the object(s), if it occurs, must occur before the post-verbal adjunct(s). It is unclear how Huang's analysis would prevent the VP *tan gangqin* 'play piano' from being generated as the head VP and the VP *tan de hen hao* 'play very well' from becoming the adjunct, as shown in (10).

(10)



In fact, a syntactic representation like (10) would more closely match the semantics of the VCC, given that semantically, the VP with a post-verbal adjunct serves as the modifier of the VP with an object, and normally, the modifier of a VP would map into the adjunct position.

It is also unclear how Huang's analysis can accommodate cases such as (7), in which the verb is duplicated more than once. In his study, only cases in which the verb is duplicated once are considered.

Huang's analysis is further challenged by the object extraction facts illustrated in (8) (repeated below as (11)).

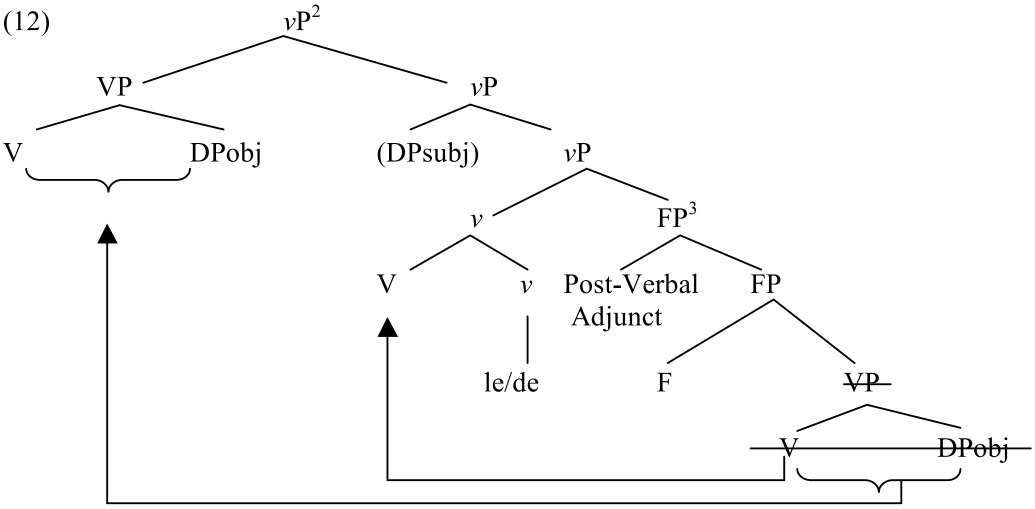
(11)a. 这 件 礼物 我 送 他 <gap> 送 得 很 好。
zhe jian liwu wo song ta <gap> song de hen hao
this CL gift I give him <gap> give DE very good
'I gave him this gift and (it turned out to be) very good.'

b. * 钢琴 张三 弹 <gap> 弹 得 很 好。
gangqin ZhangSan tan <gap> tan de hen hao
piano ZhangSan play <gap> play DE very well

(11a) contradicts the Adjunct Island Constraint (part of the CED, Huang 1982) which prohibits the extraction of an element from an adjunct. Furthermore, any movement approach that allows (11a) would have to allow (11b), which would obviously be an undesirable result. In fact, the contrast between (11a) and (11b) shows that any phrase structure constraint on VP in Modern Chinese makes reference only to overt structure (c-structure in LFG). As far as we can see, every movement-based analysis must incorrectly generate (11b): we know that movement out of the first VP is possible, somehow, due to (11a), and therefore we would also expect the base structure of (11b) to be generated by the grammar (it is grammatical if *gangqin* 'piano' remains in situ), with movement then giving the surface form in (11b).

Gouguet (2004, 2006) also treats the VCC as a single-headed construction. He proposes that the VCC in Modern Chinese is derived from VP movement and head movement, as illustrated below.

(12)



tan	gangqin	tan	de	hen hao
play	piano	play	DE	very well

According to Gouguet, (12) represents the structure and the derivational process of a VCC such as *tan gangqin tan de hen hao* 'play piano play very well'. The V moves first as a head, adjoining to *v*, and then the whole VP, including the V and the object, moves to the sister position (the adjunct position) of the *v*P. After these two movements, the original VP is deleted (unpronounced) and the new VCC is derived.

If the VCC is derived through the movement process described above, it is obvious that the object must precede the second verb and the post-verbal adjunct. Therefore, Gouguet's analysis does predict the asymmetry between the object and the post-verbal adjunct in the VCC. However, Gouguet's analysis ignores an important fact of the VCC, which is that a VCC in Modern Chinese does not necessarily involve a VP with an object. For example:

(13)								
他	玩	了	一	天	玩	得	很	累。
ta	wan	le	yi	tian	wan	de	hen	lei
he	play	ASP	one	day	play	DE	very	tired

'He played for a day and was/is tired.'

In this type of VCC, both the first and the second verb take a post-verbal adjunct. Because in Gouguet's analysis all post-verbal adjuncts appear in the SPEC (specifier) position of FP as shown in (12), it is unclear how this type of VCC can be derived through the movement mechanism described. Furthermore, the motivation for the VP movement remains unclear, as acknowledged by Gouguet himself.

[2] According to the Little *v* Hypothesis (Kratzer 1996), external arguments such as agent subjects are not assigned directly by a verb but rather by a silent "light verb" acting as a secondary predicate. This silent "light verb" is notated as a little *v* in syntactic theories adopting this hypothesis.

[3] In Gouguet (2004, 2006), post-verbal adjuncts merge into the SPEC (specifier) position of FP, which is a higher projection containing VP.

In addition, Gouguet's approach does not capture and explain the other two key properties of the VCC: multiple verb copying cases such as (7) and object extraction facts illustrated in (11)[4] pose the same challenges to Gouguet as they do to Huang (1982).

Our paper provides an alternative analysis in the framework of LFG for the VCC in Modern Chinese, which can capture and explain all of the relevant properties. Our analysis is discussed in the next section.

3 Our Approach

Based on historical evidence, facts of aspect attachment, adjunct distribution and negation scope in the VCC, we propose that the VCC should be analyzed as a double/multiple-headed coordinated VP, with each VP as a co-head. These four pieces of evidence are presented in 3.1.

3.1 Evidence Supporting VCC as a Coordinated VP

3.1.1 Historical Evidence

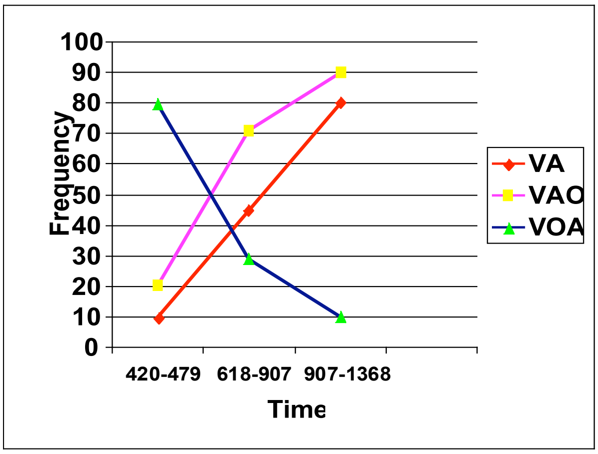
In the history of Chinese, the VCC (the '**V**(erb) **O**(bject) **V**(erb) (post-verbal) **A**(djunct)' pattern) did not emerge until the Early-Modern Chinese period (1001-1900). Instead, it was the '**V**(erb) **O**(bject) (post-verbal) **A**(djunct)' pattern that was commonly used until the 5th century. For example,

(14) 讀 書 百 遍（而義自見）。(VOA)
du shu bai bian
read book one hundred time
'Read a book a hundred times.'
(《三國志》*San Guo Zhi* (265-316))

However, this VOA pattern started to decline after the 5th century, and by the time the VOVA pattern emerged, the VOA had almost completely disappeared.

Fang (2006) proposes that the decline of the VOA pattern and the rise of the VOVA pattern (VCC) are triggered by the development of VA compound verbs (such as 打死 *da-si* 'beat to death'). According to Fang, this hypothesis is supported by the fact that the development of VA compound verbs coincides with the decline of the VOA and the rise of the VAO in Chinese history, as demonstrated by the graph below.

[4] According to Gouguet's analysis, the first VP is the adjunct and the second VP is the head VP in the VCC. This is because the second verb is derived through head movement whereas the first verb is derived through VP movement to an adjunct position.



Increased use of the VA compound verbs introduces a pattern pressure on A in the sense that the favored position for A is the position directly associated with the V, and this leads to the rise of VAO pattern. However, the VAO pattern is not the perfect replacement for the declining VOA pattern for the following two reasons: first, the VA in the VAO must be a compound verb, however, not all of the A in the VOA pattern can form a compound verb with the V; second, the A in the VAO is not the information focus whereas the A in the VOA is. The development of the VA compound verbs seems to introduce a syntactic and pragmatic conflict on the VOA pattern: syntactically the A is supposed to directly associated with the V, however, pragmatically the A is supposed to remain in final position, because it is the information focus, and the default position for information focus is final position.

Fang proposes that the reason of the VCC (VOVA pattern)'s emergence is precisely because the VOVA pattern can reconcile the syntactic and pragmatic conflicts developed on the VOA pattern due to the increased use of VA compound verbs. In the VOVA pattern, the A is directly associated with the V, and meanwhile, it remains in final position, the default position for information focus.

Fang also proposes that that the VCC in Modern Chinese emerged from two independently well-formed VPs in a context such as (15).

(15) (NP*i*) VO, t *i* V得*de* A

In (15), the two VPs serve as the main predicate of two different clauses, in which the subject of the second clause is pro-dropped and co-indexed with the subject of the first clause, as exemplified in (16).

(16) (當日武松歡喜)	飲	酒，	吃[5]	得	大	醉
	yin	jiu,	chi	de	da	zui
	drink	wine	drink	DE	very	drunk

‘(Wu, Song) drunk wine until he was very drunk.’ (《水滸傳》 *Shui Hu Zhuan* 13th -14th century)

Over time, the comma (the pause) separating these two VPs disappeared for reasons such as fast speaking speed, and the two clauses were reanalyzed as one. This reanalysis process produced a new pattern as shown in (17) below, in which the original two predicative VPs are

[5] 吃 *chi* ‘drink’ and 飲 *yin* ‘drink’ are synonyms in (16).

reanalyzed as elements of one predicate: a VCC. (18) is such an example in the same book in which (16) is found.

(17) VOV得*de* A

(18) (這胖和尚不時來我店中，)

吃	酒	吃	得	大	醉，
chi	jiu	chi	de	da	zui
drink	wine	drink	DE	very	drunk

'(This fat monk) drank wine until he was very drunk.' (《水滸傳》 *Shui Hu Zhuan* 13th -14th century)

Post-verbal adjuncts introduced by *de* 'DE' are the earliest type of post-verbal adjunct appearing in the VCC. Over time, other types of post-verbal adjuncts also started to occur in the VCC, as shown in (19). By the time of the late Qing Dynasty (1644-1911), the VCC had already been fully developed.

(19)

请	人	请	到	四	五	次
qing	ren	qing	dao	si	wu	ci
invite	people	invite	ASP	four	five	time

'(…) invited people four or five times.' (《卢太学诗酒傲公侯》*Lu taixue shi jiu ao gonghou* 14th -17th century)

This historical process clearly shows that the two VPs forming the VCC are equal elements of the VCC and thus supports our analysis of the VCC as a coordinated VP in Modern Chinese.

In addition to the historical evidence, facts regarding the aspect attachment, adjunct distribution and negation scope in the VCC also support analyzing it as a coordinated VP. These facts are discussed in the following sub-sections.

3.1.2 Aspect Attachment

The fact that the perfective aspect marker *le* can appear in either the first VP or the second VP or both in a VCC also suggests that each VP is a head, because the perfective *le* is only attached to heads in Modern Chinese. For example:

(20) a.

张三	弹	钢琴	弹	了	很	久。
ZhangSan	tan	gangqin	tan	le	hen	jiu
ZhangSan	play	piano	play	ASP	very	long

'ZhangSan played piano for a very long time.'

b.

张三	弹	了	一	天	弹	得	很	累。
ZhangSan	tan	le	yi	tian	tan	de	hen	lei
ZhangSan	play	ASP	one	day	play	DE	very	tired

'ZhangSan played for a day and was/is tired.'

c.

张三	弹	了	一	天	弹	了	一	百	遍。
ZhangSan	tan	le	yi	tian	tan	le	yi	bai	bian
ZhangSan	play	ASP	one	day	play	ASP	one	hundred	time

'ZhangSan played for a day and played one hundred times.'

It is true that the aspect marker *le* tends not to occur in the VP involving object(s) (as shown in (21)), which would be the first VP in the VCC if it occurs, and Huang (1982) view this fact as evidence that the first VP in the VCC is an adjunct rather than a head VP. However, as shown by (20b) and (20c), it is not true that the aspect marker cannot occur in the first VP in the VCC. We believe that the real reason why the aspect marker tends not to appear in the VP involving object(s) is because the VP involving object(s) must serve as a topic in the VCC (Cui 2003), and aspect markers do not normally appear in the topic.

(21) *张三 弹 了 钢琴 弹 很 久。
ZhangSan tan le gangqin tan hen jiu
ZhangSan play ASP piano play very long

3.1.3 Adjunct Distribution

Normally, an adjunct of a VP can only be distributed to the head of that VP, but not to another adjunct of that VP. For example:

(21) ZhangSan studied Chinese very well in Beijing.

In (21), the adjunct 'in Beijing' is distributed to the head VP 'studied Chinese', but not to the adjunct 'very well'.

However, the adjunct of a VCC in Modern Chinese must be distributed to all the VPs. For example:

(22) 张三 在 北京 学 汉语 学 得 很 好。
ZhangSan zai Beijing xue hanyu xue de hen hao
ZhangSan in Beijing study Chinese study DE very well
'ZhangSan studied Chinese very well in Beijing.'

(22) entails both 'ZhangSan studied Chinese in Beijing' and 'ZhangSan studied very well in Beijing', which suggests that the adjunct *zai Beijing* 'in Beijing' is distributed to both VPs. Thus both VPs are heads rather than one being an adjunct of the whole VCC: *xue hanyu xue de hen hao* 'study Chinese study very well'.

3.1.4 Negation

A negator such as *mei* 'not' cannot appear before the first VP in the VCC; however, it can appear before the second VP, as shown by the contrast between (23a) and (23b).

(23) a. *他 没 学 汉语 学 好。(*neg+VOVA)
ta mei xue hanyu xue hao
he not study Chinese study well

b. 他 学 汉语 没 学 好。(VO+neg+VA)
ta xue hanyu mei xue hao
he study Chinese not study well
'He studied Chinese, but did not study well.'

Analyzing the VCC as a coordinated VP provides an explanation for this contrast, for the following reason. When it appears before the coordination, m*ei* 'not' scopes over the entire construction, for example:

(24) 张三 没 批评 责备 李四。
ZhangSan mei piping zebei lisi
ZhangSan not criticize blame LiSi
'ZhangSan did not criticize and blame LiSi.'

As shown in (24), the negator *mei* 'not' is distributed to both *piping* 'criticize' and *zebei* 'blame', and (24) entails both 'ZhangSan did not criticize LiSi' and 'ZhangSan did not blame LiSi'.

Returning now to (23a), if it involves a coordinated VP, the negator *mei* 'not' will distribute to both *xue hanyu* 'study Chinese' and *xue hao* 'study well' in (23a); the ill-formedness is exactly due to this negator distribution. The first part, *ta mei xue hanyu* 'he not study Chinese' entails that 'he did not study at all'. However, *ta mei xue hao* 'he not study well' entails that 'he studied (but did not study well)', and these two entailments conflict. Therefore, (23a) is ill-formed due to an entailment conflict introduced by the negator distribution, which is in turn because the VCC is a coordinated VP construction.

By contrast, the negator is placed after the first VP and so only scopes over the second VP in (23b). Then both VPs in (23b) entail that 'he studied Chinese', and there is no entailment conflict.

3.2 VCC as a Special Type of Coordinated VP

Based on the evidence presented in section 3.1, we propose that the VCC is analyzed as a double/multiple-headed coordinated VP, with each VP as a co-head, as shown in (25).

(25)
$$\begin{array}{lll} \text{VP(VCC)} \rightarrow & \text{VP} & \text{VP +.} \\ & \downarrow\in\uparrow & \downarrow\in\uparrow \end{array}$$

We further propose that the first VP stands in a subsumption relation (Zaenen and Kaplan 2002, 2003) to every other VP. Making the first VP subsume other VPs makes the first VP more general than every other VP in the VCC, which captures an observation in previous studies (Cui 2003, etc.) that the first VP serves pragmatically as the secondary topic, and the other VPs involving post-verbal adjuncts serve as the comment to the first VP. For example,

(26) 张三 学 汉语 学 得 很 好。
ZhangSan xue hanyu xue de hen hao
ZhangSan study Chinese study DE very well
'ZhangSan studied Chinese very well.'

In (26), *xue hanyu* 'study Chinese' serves as the secondary topic and *xue de hen hao* 'studied very well' serves as the comment to the first VP and provides more specific information about the topic: the result of 'study Chinese'. In this sense, the first VP is more general and subsumes every other VP in the VCC.

Technically, this subsumption relation can be achieved by making the first VP the head of the entire VCC[6], as shown in (27).

(27)
$$\begin{array}{lll} \text{VP(VCC)} \rightarrow & \text{VP} & \text{VP +.} \\ & \downarrow\in\uparrow & \downarrow\in\uparrow \\ & \downarrow = \uparrow & \end{array}$$

(27) captures and explains all of the three key properties of the VCC discussed in section 1.

[6] Thanks to Ron Kaplan for this solution.

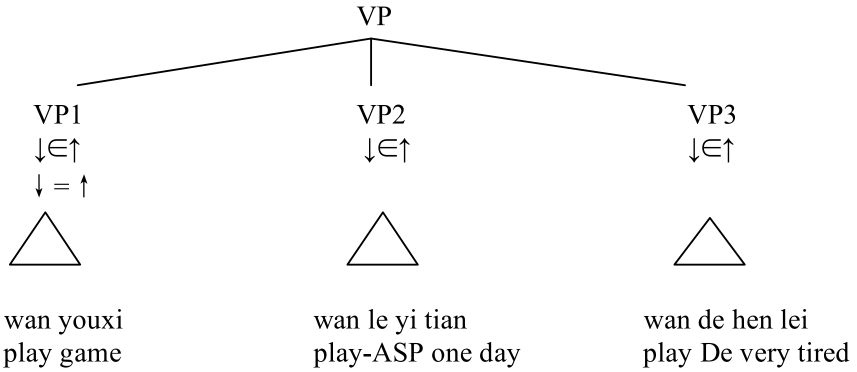
First of all, the VCC cases in which the verb is copied more than once (such as (28)) follow in a straightforward way.

(28)	他	玩	游戏	玩	了	一	天	玩	得	很	累。
	ta	wan	youxi	wan	le	yi	tian	wan	de	hen	lei
	he	play	game	play	ASP	one	day	play	DE	very	tired

‘He played games for a day and was/is tired.’

Following (27), (28) is simply a coordinated VP with three conjuncts, as illustrated in (29).

(29)



Second, our approach predicts the constituent order asymmetry in the VCC. Specifically, the VP involving object(s) must occur before any VP involving a post-verbal adjunct, as shown by (6), repeated below as (30).

(30) a.	张三	弹	钢琴	弹	得	很	好。
	ZhangSan	tan	gangqin	tan	de	hen	hao
	ZhangSan	play	piano	play	DE	very	well

‘ZhangSan plays piano very well.’

b.	*张三	弹	得	很	好	弹	钢琴。
	ZhangSan	tan	de	hen	hao	tan	gangqin
	ZhangSan	play	DE	very	well	play	piano

(27) requires that all of the verbs in the VCC must share the subcategorization frame of the first VP. Therefore, (30b) is ruled out either by the Completeness Condition of LFG (Kaplan and Bresnan 1982), as the first VP is locally incomplete, see (31); or by the Coherence Condition of LFG (Kaplan and Bresnan 1982), as the object in the second VP is ungoverned, see (32).

(31): blocked by (27)

$$\left\{\begin{array}{l}
\left[\begin{array}{ll}
\text{PRED} & \text{‘tan } <\text{SUBJ OBJ>’} \\
\text{SUBJ} & \text{[PRED ‘ZhangSan’]} \\
\text{ADJUNCT} & \text{[PRED ‘de hen hao’]} \\
\text{OBJ} &
\end{array}\right] \\
\qquad\qquad\qquad\qquad\text{Incomplete VP1} \\
\left[\begin{array}{ll}
\text{PRED} & \text{‘tan } <\text{SUBJ OBJ>’} \\
\text{SUBJ} & \text{[PRED ‘ZhangSan’]} \\
\text{OBJ} & \text{[PRED ‘gangqin’]} \\
\text{ADJUNCT} & \text{[PRED ‘de hen hao’]}
\end{array}\right]
\end{array}\right\}$$

(32): blocked by (27)

$$\left\{\begin{array}{l}
\left[\begin{array}{ll}
\text{PRED} & \text{‘tan } <\text{SUBJ >’} \\
\text{SUBJ} & \text{[PRED ‘ZhangSan’]} \\
\text{ADJUNCT} & \text{[PRED ‘de hen hao’]}
\end{array}\right] \\
\left[\begin{array}{ll}
\text{PRED} & \text{‘tan } <\text{SUBJ >’} \\
\text{SUBJ} & \text{[PRED ‘ZhangSan’]} \\
\text{OBJ} & \text{[PRED ‘gangqin’]} \\
\text{ADJUNCT} & \text{[PRED ‘de hen hao’]}
\end{array}\right] \\
\qquad\qquad\qquad\qquad\text{Incoherent VP2}
\end{array}\right\}$$

In contrast, even though the second verb *tan* ‘play’ in (30a) does not have a local object, its VP is complete, as the VP’s information is subsumed by that of the first VP, which is complete, as shown in (33).

(33): f-structure of (30a)

$$\left\{\begin{array}{l}
\left[\begin{array}{ll}
\text{PRED} & \text{‘xue } <\text{SUBJ OBJ>’} \\
\text{SUBJ} & \text{[PRED ‘ZhangSan’]} \\
\text{OBJ} & \text{[PRED ‘hanyu’]}
\end{array}\right] \\
\left[\begin{array}{ll}
\text{PRED} & \text{‘xue } <\text{SUBJ OBJ>’} \\
\text{SUBJ} & \text{[PRED ‘ZhangSan’]} \\
\text{OBJ} & \text{[PRED ‘hanyu’]} \\
\text{ADJUNCT} & \text{[PRED ‘de hen hao’]}
\end{array}\right]
\end{array}\right\}$$

Finally, the object extraction facts in the VCC (as shown in (8), repeated below as (34)) are explained by the C-structure constraints on Chinese VPs.

(34)a.	这	件	礼物	我	送	他	<gap>	送	得	很	好。
	zhe	jian	liwu	wo	song	ta	<gap>	song	de	hen	hao
	this	CL	gift	I	give	him	<gap>	give	DE	very	good

‘I gave him this gift and (it turned out to be) very good.’

b. *	钢琴	张三	弹	<gap>	弹	得	很	好。
	gangqin	ZhangSan	tan	<gap>	tan	de	hen	hao
	piano	ZhangSan	play	<gap>	play	DE	very	well

Each Chinese VP is internally as simple as possible (an Economy constraint; see Peck and Sells (2006)). Yet each VP must consist of V and a sister X where X can be internal argument(s) or one post-verbal adjunct (Fang 2005, 2006); a Chinese VP has a minimal condition that it contains some sister to V. Therefore, when the object *gangqin* 'piano' in (34c) appears in fronted position, and with no duplication of the verb, the VP *tan de hen hao* 'plays very well' satisfies the c-structure constraints on VP, and no VCC needs to be triggered. The f-structure is well formed – it is a single f-structure with an object and an adjunct. Note that this analysis crucially relies on the assumption that the "gap" in these examples has no status in c-structure: the long-distance dependency is only represented at f-structure (TOPIC=OBJ). In contrast, when the verb *tan* 'play' is duplicated as shown in (34b), the first VP consists of a bare verb, which violates the c-structure constraint on Chinese VPs mentioned above.

Sometimes, even with a topicalized object, the VCC is necessary: when the object *zhe jian liwu* 'this gift' in (34a) is fronted, the VCC must be triggered because otherwise **song ta de hen hao* violates the C-structure VP rules: it is too complex.

4 Conclusions

To summarize, we have presented an analysis for the VCC in Modern Chinese. Unlike the previous analyses, in which the VCC is analyzed as a single-headed structure with one VP as the head and the other VP as an adjunct, our analysis treats the VCC as a coordinated VP, with each VP as a co-head. We further propose that the first VP subsumes the following VPs in the VCC and technically, this can be achieved by making the first VP the head of the entire VCC.

We have shown that this proposed approach can successfully capture and explain all of the three key properties that characterize the VCC: it predicts that VO must precede VA in the VCC; it explains multiple verb copying cases in a straightforward way; and it correctly predicts the object extraction facts involving the VCC. It makes these predictions only because of the factorization of information into f-structure and c-structure: the constraints on VP are partly functional (the first VP must be locally complete and coherent) and partly purely structural: a VP consists of a V and at least one sister, but the VP must be as simple as possible (i.e., constrained by f-structure well-formedness).

5 Acknowledgements

We would like to thank Ron Kaplan (Powerset) and Tracy Holloway King (Palo Alto Research Center) for their generous help and suggestions. We also want to thank Ash Asudeh (Carleton University) and other reviewers of the 2007 LFG conference for their helpful comments and discussions of our proposals.

References

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

Cui, Huashan. 2003. *Xiandai Hanyu Chongdongju Dingliang Yanjiu* (*A Quantified Study of the Verb Copy Construction in Modern Chinese*). Master Thesis. Peking University, China.

Dai, John Xiang-ling. 1992. The Head in *WO PAO DE KUAI*. Journal of Chinese Linguistics 20, 84-119.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. Syntax and Semantics Volume 34. Academic Press.

Fang, Ji. 2005. The Verb Copy Construction and the Function of 得*de* Phrases in Modern Mandarin (Handout). The 79th Linguistic Society of America (LSA) Conference. Oakland, California, USA.

_____. 2006. *The Verb Copy Construction and the Post-Verbal Constraint in Chinese*. Ph.D. Dissertation. Stanford University, CA.

Gouguet, Jules. 2004. Verb Copying and the Linearization of Event Structure in Mandarin. Handout for GLOW, Thessaloniki, April 21, 2004.

_____. 2006. Adverbials and Mandarin Argument Structure. In O. Bonami and P. Cabredo Hofherr, eds., *Empirical Issues in Syntax and Semantics 6*, CNRS, 155-173.

Huang, James C-T. 1982. *Logical Relations in Chinese and the Theory of Grammar*. Ph.D. Dissertation. MIT, Cambridge, Mass.

Peck, Jeeyoung and Peter Sells. 2006. Preposition Incorporation in Mandarin: Economy within VP. In M. Butt and T.H. King, eds., *Proceedings of the LFG06 Conference*. CSLI On-line Publications.

Kaplan, Ronald M. and Bresnan, Joan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In J. Bresnan (ed.), *The Mental Representation of Grammatical Relations* 173-281. The MIT Press, Cambridge, MA.

Kaplan, Ronald M. and Zaenen, Annie. 2003. Things Are Not Aways Equal. In Alexander Gelbukh (ed.) *CICLing-2003 Computational Linguistics and Intelligent Text Processing*. Lecture Notes in Computer Science N 2588, pp. 12-22, Springer-Verlag.

Kratzer, Angelika. 1996. Severing the External Argument from its Verb. In J. Rooryck and L. Zaring (eds.) *Phrase Structure and the Lexicon*. Dordrecht, Kluwer Academic Publishers.

Li, Charles N. 1975. Synchrony VS. Diachrony in Language Structure. Language 51, 873-886.

Li, Yen-hui Audrey. 1985. *Abstract Case in Chinese*. Ph.D. Dissertation. University of Southern California, Los Angeles.

Sells, Peter. 1985. *Lectures on Contemporary Syntactic Theories*. CSLI, StanfordUniversity.

Zaenen, Annie and Kaplan, Ronald M. 2002. Subsumption and Equality: GermanPartial Fronting in LFG. In *Proceedings of the LFG02 Conference,* M. Butt and T. H. King (eds.), CSLI Publications, Stanford, CA.

_____. 2003. Subject Inversion in French: Equality and Inequality in LFG. In *Empirical Issues in Syntax and Semantics*, Vol. 4, C. Beyssade, O. Bonami, P. Cabredo Hofherr and F. Corblin (eds.), pp. 190-205, Presses Universitaires de Paris Sorbonne, Paris, France.

TREEBANK-BASED ACQUISITION OF LFG RESOURCES FOR CHINESE

Yuqing Guo[1], Josef van Genabith[1,2] and Haifeng Wang[3]
[1]NCLT, School of Computing, Dublin City University
[2]IBM Center for Advanced Studies, Dublin, Ireland
[3]Toshiba (China) R&D Center, Beijing, China

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

This paper presents a method to automatically acquire wide-coverage, robust, probabilistic Lexical-Functional Grammar (LFG) resources for Chinese from the Penn Chinese Treebank (CTB). Our starting point is the earlier, proof-of-concept work of Burke et al. (2004) on automatic f(unctional)-structure annotation, LFG grammar acquisition and parsing for Chinese using the CTB version 2 (CTB2). We substantially extend and improve on this earlier research as regards coverage, robustness, quality and fine-grainedness of the resulting LFG resources. We achieve this through (i) improved LFG analyses for a number of core Chinese phenomena; (ii) a new automatic f-structure annotation architecture which involves an intermediate dependency representation; (iii) scaling the approach from 4.1K trees in CTB2 to 18.8K trees in CTB version 5.1 (CTB5.1) and (iv) developing a novel treebank-based approach to recovering Non-Local Dependencies (NLDs) for Chinese parser output. Against a new 200-sentence good standard of manually constructed f-structures, the method achieves 96.00% f-score for f-structures automatically generated for the original CTB trees and 80.01% for NLD-recovered f-structures generated for the trees output by Bikel's parser.

1 Introduction

Automatically inducing deep, wide-coverage, constraint-based grammars from existing treebanks avoids much of the time and cost involved in manually creating such resources. A number of papers (van Genabith et al., 1999; Sadler et al., 2000; Frank, 2000; Cahill et al., 2002) have developed methods for automatically annotating treebank (phrase structure or c(onstituent)-structure) trees with LFG f(unctional)-structure information to build f-structure corpora to acquire LFG grammar resources.

In LFG, c-structure and f-structure are independent levels of representation which are related in terms of a correspondence function projection ϕ (Kaplan, 1995). In the conventional interpretation, the ϕ-correspondence between c- and f-structure is defined implicitly in terms of functional annotations on c-structure nodes, from which an f-structure can be computed by a constraint solver.

In one type of treebank-based LFG grammar acquisition approach, referred to as "annotation-based grammar acquisition", functional schemata are annotated either manually on the entire Context Free Grammar (CFG) rules automatically extracted from the treebank (van Genabith et al., 1999); or on a smaller number of hand-crafted regular expression-based templates representing partial and underspecified CFG rules (Sadler et al., 2000) which are applied to automatically annotate the CFG rules extracted from treebank trees; or, using an annotation algorithm traversing treebank trees, applying annotations to each node of a local c-structure subtree in a left/right context partitioned by the head node (Cahill et al., 2002).

An alternative grammar acquisition architecture for LFG, referred to as "conversion-based grammar acquisition", directly induces an f-structure from a c-structure tree, without intermediate functional schemata annotations on c-structure trees. An algorithm building on this architecture was developed in Frank (2000) by directly rewriting partial c-structure fragments into corresponding partial f-structures, using a rewriting system originally developed for transfer-based Machine Translation. As opposed to the CFG rule- and annotation-based architecture in which annotation principles are by and large restricted to local trees of depth one, this approach naturally generalises to non-local trees.

One of the challenges in both the annotation- and more direct conversion-based architectures is to keep the number of f-structure annotation/conversion rules which encode linguistic principles to a minimum, as their creation involves manual effort. Another challenge is to find automatic f-structure annotation/conversion architectures that generalise to different languages and treebank encodings.

A common characteristic of the work cited above is that all the methods are applied to English treebanks (Penn-II, Susanne and AP treebank) from which LFG resources are acquired for English. An initial attempt to extend the treebank- and annotation-based LFG acquisition methodology to Chinese data was carried out by Burke et al. (2004), which applied a version of Cahill et al. (2004)'s algorithm adapted to Chinese via the Penn Chinese Treebank version 2 (LDC2001T11) and was evaluated against a small set of 50 manually constructed gold-standard f-structures. The experiments were proof-of-concept and somewhat limited with respect to (i) the coverage of Chinese linguistic phenomena; (ii) the quality of the f-structures produced; (iii) parser output producing only 'proto' f-structures with non-local dependencies unresolved; (vi) the size of the treebank and gold standard.

In the present paper, we address these concerns and present a new f-structure annotation architecture and a new annotation algorithm for Chinese, which:

- combines aspects of both the annotation-based and conversion-based architectures described above;
- generates proper f-structures rather than proto-f-structures by resolving NLDs for parser output;
- scales up to the full Penn Chinese Treebank version 5.1 (LDC2005T01U01), whose size is more than 4 times of that of CTB2;
- is evaluated on a new extended set of Chinese gold-standard f-structures for 200 sentences.[1]

2 Automatic F-Structure Annotation of CTB5.1

2.1 Chinese LFG

Research on LFG has provided analyses for a considerable number of linguistic phenomena in Indo-European, Asian, African and Native American and Australian languages. However, Chinese is a language drastically different from such languages as English, German, French etc. which are often the focus of attention.

[1] Developed jointly with PARC.

The most distinctive linguistic properties of Chinese are: (i) very little inflectional morphology encoding tense, number, gender etc., resulting in the almost complete absence of agreement phenomena familiar from European languages; (ii) lack of case markers, complementisers etc., which often causes syntactic and semantic ambiguity; (iii) the tendency towards omission of constituents if they can be inferred from the context, which includes not only subject and object arguments, but also predicates and other heads of phrases, in some cases.

Though the main purpose of this paper is to address the technical issue of automatically inducing f-structures from the Penn Chinese treebank, an LFG account for various phenomena and constructions in Chinese is a prerequisite. Work addressing Chinese issues within LFG formalism has been carried out for a limited number of phenomena. For example, Fang (2006) provides a formal analysis for the verb copy construction in Chinese; Huang and Mangione (1985) offers an LFG account of post-verbal "得/DE" construction; Her (1991) presents a classification of Mandarin verbs by the subcategorised grammatical functions within LFG. In our research , we adopt some existing theoretical LFG analysis, but also provide our own solution to other Chinese core phenomena and disputable constructions due to the lack of standard LFG account for them.[2] To give a flavour of what the Chinese LFG likes look, below we illustrate the c-structure trees represented in the CTB and our analyses with the corresponding f-structures for a number of core linguistic phenomena characteristic of Chinese.

Classifiers are common in Chinese (and some other Asian languages) in that they cooccur with numerals or demonstrative pronouns to count things or persons (nouns) or indicate the frequency of actions (verbs). To provide a unified interpretation of classifiers, we treat a classifier as a grammatical function modifying the head noun (or verb) rather than for example as a feature attached to the determiner or head noun/verb, for the following reasons:

- classifiers have content meaning: standard classifiers such as "米/meter", "公斤/kilogram", "瓶/bottle" relate to distance, weight, volume, etc. and individual classifiers indicate prominent features of the noun they modify, for example "把/BA" which is derived from "handle" is used as a classifier for objects with a handle, as in (1).

  (1) 一 把 椅子
  one CLS chair
  'one chair'

- classifiers can function as the head within a phrase, as in (2).

  (2) 打 三 下
  hit three CLS
  'hit three times'

[2] Rather than providing a fully adequate LFG account in theory, our analysis is compromise and conservative in some respects for the practical reason and considering tree representations in the CTB.

- classifiers can be modified by adjectives, as in (3).

(3) 一 大 碗 饭
one big bowl/CLS rice
‘a big bowl of rice’

Figure 1 illustrates the CTB representation of a classifier and the corresponding schematic f-structure. A noticeable difference is that the determiner (DT) takes a quantifier phrase (QP) as its complement in the CTB constituent-tree, whereas in our f-structure the determiner and quantifier are parallel functions both specifying the head noun predicate.

(4) 这 五 个 学生
these five CLS student
‘these five students’

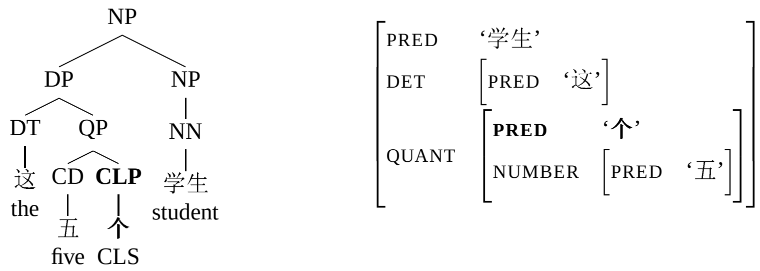
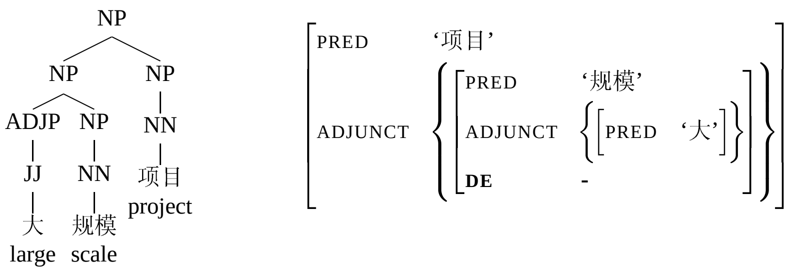


Figure 1: The CTB tree and our f-structure analysis of the classifier

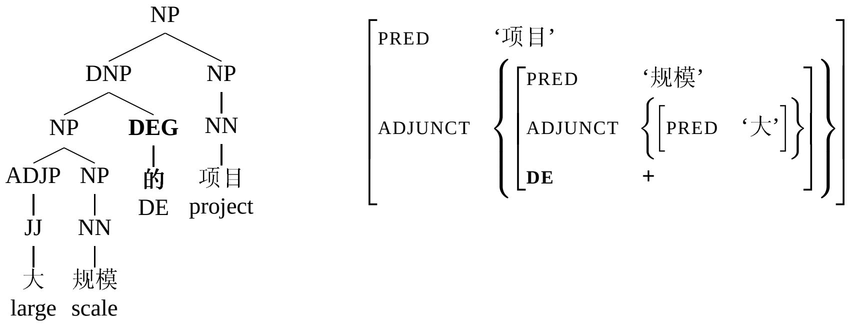
DE Phrases are formed by the function word “的/DE” attached to various categories, such as possessive phrases, noun phrases, adjective phrases or relative clauses. DE has no content other than marking the preceding phrase as a modifier of NP. Different from the original f-structure annotation algorithm and the 50-sentence gold-standard f-structures developed in Burke et al. (2004), we choose the content word rather than DE as head of the modifier, because all the other words in the modifier phrase will depend on the head, and moreover DE has no content and thus may be omitted in examples such as (5a). Therefore, in our analysis we treat DE as an optional feature attached to the modifier as exemplified in Figure 2. What is noticeable here is that the grammatical function of the DE-phrase in (5b) is an attributive modifier (ADJUNCT) while in (6) it is a possessor (POSS), even though the constituent structures are the same for both, due to the absence of any case marking. The difference is in fact lexical and due to the head word of the adjunct which is a common noun (NN) in (5), and the head word of the possessor which is a proper noun (NR) in (6).

BEI-Constructions are commonly considered approximately equivalent to passive voice in English. However we do not treat “被/BEI” as just a passive voice

(5) a. 大 规模 项目
large scale project



b. 大 规模 的 项目
large scale DE project
‘a large-scale project’



(6) 张三 的 书
ZhangSan DE book
‘ZhangSan’s book’

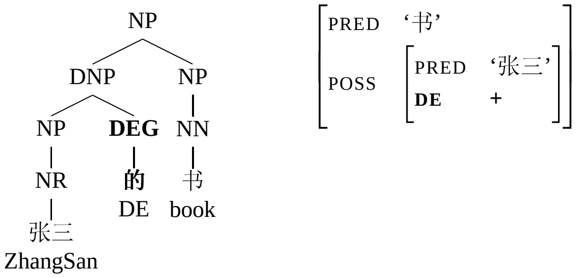
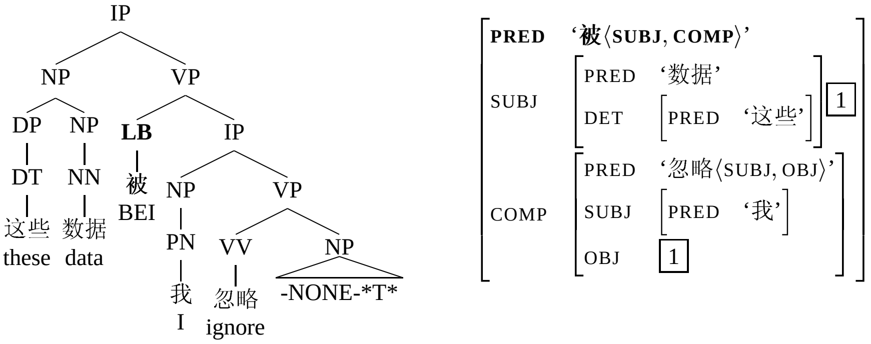


Figure 2: The CTB tree and our f-structure analysis of DE-phrase

feature, in that it also introduces the logic subject in long-BEI constructions as in (7), similar to the preposition “by” in the English passive construction. Furthermore, we do not analyse it as a subject marker, as short-BEI constructions as in (8) will be subjectless, where BEI marks nothing. And rather than treating it as a preposition, though the analysis can be argued for from a theoretical point of view, it does not always indicate passive voice, as in (9), where the embedded verb is intransitive. In line with Her (1991), we treat BEI as a verb. The advantage of this analysis is that it provides a unified account for embedded verbs, where verbs in BEI sentences have the same subcategorisation frames as those in their BEI-less corresponding sentences. Her (1991) treats BEI as a pivotal construction, where BEI requires an object and a non-finite VP complement. However, this is somewhat different from the CTB representation, where BEI takes a sentential complement. Both constructions are acceptable in Chinese without the presence of a complementiser. For practical purposes, we accept the tree representation in CTB and hence BEI requires a closed complement (COMP) in our f-structure, as exemplified in Figure 3.

(7) 这些 数据 被 我 忽略
these data BEI I ignore
'These data were ignored by me.'



(8) 他 被 授予 一等奖
he BEI award the top prize
'He was awarded the top prize.'

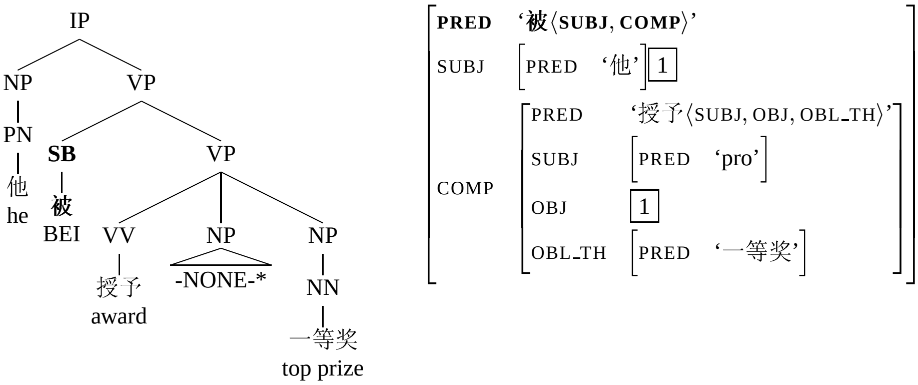


Figure 3: The CTB tree and our f-structure analysis of BEI-construction

(9) 猫 被 老鼠 跑 了
cat BEI mouse escape ASP
'The cat let the mouse escape.'

2.2 A New F-Structure Annotation Algorithm for CTB

The f-structure annotation method developed in Cahill et al. (2002) & Burke et al. (2004) builds on the CFG rule- and annotation-based architecture. By and large the algorithm works on local treebank subtrees of depth one (equivalent to a CFG rule).[3] In order to annotate the nodes in the tree, the algorithm partitions each sequence of daughters in the local subtree into three sections: left context, head and right context. Configurational information (left or right position relative to the head), category of mother and daughter nodes, and Penn treebank functional labels (if they exist) on daughter nodes are exploited to annotate nodes with f-structure functional equations. The annotation principles for Chinese in Burke et al. (2004) are fairly coarse-grained. However configurational and categorial information from local trees of depth one only is not always sufficient to determine the appropriate grammatical function (GF), as for example for DE-phrases (Figure 2). This means disambiguation of GFs for Chinese may require access to lexical information (common or proper noun in Figure 2) and more extensive contextual information beyond the local configurational and categorial structure.

In Cahill et al. (2002) & Burke et al. (2004), for each tree, the f-structure equations are collected after annotation and passed on to a constraint solver which produces an f-structure for the tree. Unfortunately, as explained in Cahill et al. (2002), the constraint solver's capability is limited: it can handle equality constraints, disjunction and simple set-valued feature constraints. However, it (i) fails to generate an f-structure (either complete or partial) in case of clashes between the automatically annotated features; and (ii) does not provide subsumption constraints to distribute distributive features into coordinate f-structures.

In order to avoid the limitations of the constraint solver, and in order to exploit more information for function annotation from a larger context than within the local tree, instead of indirectly generating the f-structure via functional equations annotated to c-structure trees, we adopt an alternative approach which transduces the treebank tree into an f-structure via an intermediate dependency structure, directly constructed from the original c-structure tree, as shown in Figures 4 & 5.

The basic idea is that the $\uparrow=\downarrow$ (or the equivalent $\phi(n_i)$=$\phi(n_j)$ equations in Figure 4) head projections in the classical LFG projection architecture allow us to collapse a c-structure tree into an intermediate, unlabelled dependency structure as in Figure 5. The intermediate unlabelled dependency structure is somewhat more abstract and normalised (compared to the original c-structure tree) and is used as input to an f-structure annotation algorithm, which is simpler and more general than the conventional f-structure algorithms (Cahill et al., 2002; Burke et al., 2004),

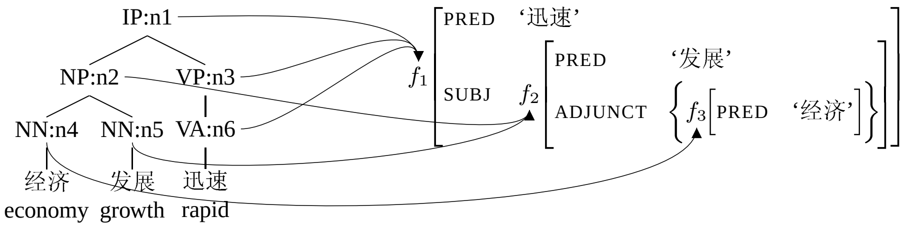
[3] Though it also uses some non-local information.

directly operating on the original, more complex and varied c-structure trees.

The new f-structure annotation architecture is illustrated in Figure 5, and includes two major steps:

I. First, we extract all predicates from the (local) c-structure tree, using head-finding rules similar to that used in Collins (1999), adapted to Chinese data and CTB5.1. Collapsing head-branches along the head-projection lines, the c-structure configuration is projected to an intermediate unlabelled dependency structure, augmented with CFG category and order information inherited from the c-structure.

II. Second, we use high-level annotation principles exploiting configurational, categorial, functional as well as lexical information from the intermediate unlabelled dependency structure to annotate grammatical function and other f-structure information (to create a labelled dependency structure, i.e. an LFG f-structure).

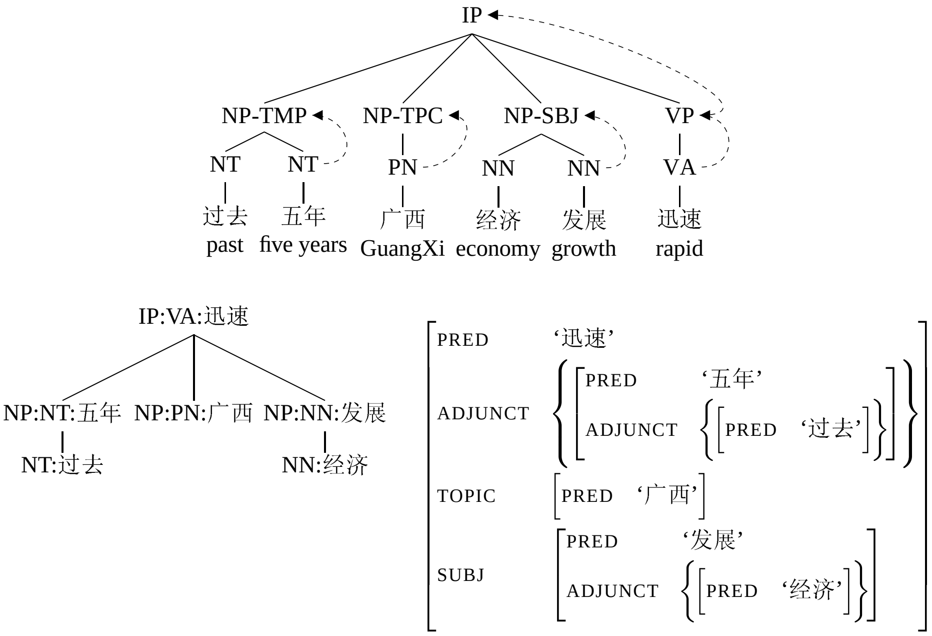


ϕ-correspondence:
ϕ(n1)=ϕ(n3)=ϕ(n6)=f_1
ϕ(n2)=ϕ(n5)=f_2
ϕ(n4)=f_3

f-structure
(f_1 PRED)=‘迅速’
(f_2 PRED)=‘发展’
(f_3 PRED)=‘经济’
(f_1 SUBJ)=f_2
(f_2 ADJUNCT)=f_3

Figure 4: ϕ-projection from c-structure to f-structure

By abstracting away from the ‘redundant’ c-structure nodes in our intermediate dependency representations, the annotation principles can apply to non-local sub-trees. This allows us to disambiguate different GFs in a larger context and resort to lexical information. As a more abstract dependency-like structure is used to mediate between the c- and f-structure, the algorithm always generates an f-structure, and there are no clashing functional equations causing the constraint solver to fail. Moreover, the intermediate dependency structure can easily handle distribution into coordinate structures by moving and duplicating the dependency branch associated with distributive functions. Furthermore, finite approximations of functional uncertainty equations resembling paths of non-local dependencies also can be computed on the intermediate dependency structure for the purpose of NLD recovery (this will be presented in section 3). Finally, in order to conform to the coherence condition and to produce a single connected f-structure for every CTB tree, a post-processing step is carried out to check duplications and to catch and add missing annotations.



(I) Predicate Extraction

(II) Function Annotation

Figure 5: The new f-structure annotation architecture for CTB

Our new annotation algorithm is somewhat similar in spirit to the conversion approach developed in Frank (2000), However in Frank (2000)'s algorithm the mapping of c-structure to f-structure is carried out in one step using a tree/graph rewriting system. Our method enforces a clear separation between the intermediate unlabelled dependency structure (predicate identification) and function annotation. Predicate identification maps c-structure into an unlabelled dependency representation, and is thus designed particularly for a specific type of treebank encoding and data-structures. In contrast, function annotation is accomplished on the dependency representation which is much more compact and normalised than the original c-structure representation, hence the function annotation rules are simpler and the architecture minimises the dependency of the annotation rules on the particular treebank encoding.

2.3 Experimental Evaluation

Similar to Cahill et al. (2002) & Burke et al. (2004), our new annotation algorithm is evaluated both quantitatively and qualitatively.

We apply the f-structure annotation algorithm to the whole CTB5.1 with 18,804 sentences. Unlike the CFG- and annotation-based predecessors (Cahill et al., 2002; Burke et al., 2004), the new algorithm guarantees that 100% of the treebank trees receive a single, connected f-structure.

For the purpose of qualitative evaluation, we selected 200 sentences from CTB-5.1 for which the f-structures are automatically produced by our annotation algorithm, and then manually corrected them to construct a gold-standard set in line with our Chinese LFG analyses presented in Section 2.1. Annotation quality is measured in terms of predicate-argument-adjunct (or dependency) relations. The relations are represented as triples $relation(predicate, argument/adjunct)$, following Crouch et al. (2002). The f-structure annotation algorithm is applied to two different sets of test data: (i) the original CTB trees, and (ii) trees output by Bikel's parser (Bikel and Chiang, 2000) trained on 80% of the CTB5.1 trees, exclusive of the 200 gold-standard sentences. Table 1 reports the results against the new 200-sentence set of gold-standard f-structures.

	CTB Trees			**Parser Output Trees**		
	Precision	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**
Preds Only	93.68	94.93	94.30	73.55	65.05	69.04
All GFs	95.25	96.75	96.00	84.00	71.77	77.40

Table 1: Quality of f-structure annotation

Table 1 shows that given high-quality input trees, the new algorithm produces high quality f-structures with f-scores of around 94%-96% for preds-only and all GFs, respectively. The corresponding scores drop by 20%-24% absolute on parser produced trees.

3 Recovery of Chinese Non-Local Dependencies for Parser Output

The drastic drop in the results on parser output trees is mainly due to labelled bracketing parser errors, but also because Bikel's parser (and most state-of-the-art treebank-based broad-coverage probabilistic parsers) does not capture non-local dependencies (or 'movement' phenomena).[4] As a result, the automatically generated f-structures produced from parser output trees are proto-f-structures, as they only represent purely local dependencies. In this section, we present a postprocessing approach to recover NLDs on the automatically generated proto-f-structures.

3.1 NLDs in Chinese

Non-local dependencies in CTB are represented in terms of empty categories (ECs) and (for some of them) coindexation with antecedents, as exemplified in Figure 6. Following previous work for English and the CTB annotation scheme (Xue and Xia, 2000), we use "non-local dependencies" as a cover term for all missing or

[4] The original parser does not produce CTB functional tags either, of which the f-structure annotation algorithm takes advantage (if they are present). To restore the CTB functional tags, we retrained the original parser to allow it to produce CTB functional tags as part of its output.

dislocated elements represented in the CTB as an empty category (with or without coindexation/antecedent), and our use of the term remains agnostic about fine-grained distinctions between non-local dependencies drawn in the theoretical linguistics literature.

Table 2 gives a breakdown of the most frequent types of empty categories and their antecedents. According to their different linguistic properties, we classify these empty nodes into three major types: null relative pronouns, locally mediated dependencies, and long-distance dependencies (LDDs).

	Antecedent	POS	Label	Count	Description
1	WHNP	NP	*T*	11670	WH traces (e.g. *OP*中国发射*T*的卫星)
2		WHNP	*OP*	11621	Empty relative pronouns (e.g. *OP*中国发射的卫星)
3		NP	*PRO*	10946	Control constructions (e.g. 这里不许*PRO*抽烟)
4		NP	*pro*	7481	Pro-drop situations (e.g. *pro*不曾遇到的问题)
5	IP	IP	*T*	575	Topicalisation (e.g. 我们能赢，他说*T*)
6	WHPP	PP	*T*	337	WH traces (e.g. *OP*人口*T*密集地区)
7		WHPP	*OP*	337	Empty relative pronouns (e.g. *OP*人口密集地区)
8	NP	NP	*	291	Raising & Bei constructions (e.g. 我们被排除*在外)
9	NP	NP	*RNR*	258	Coordinations (e.g. 鼓励*RNR*和支持投资)
10	CLP	CLP	*RNR*	182	Coordinations (e.g. 五*RNR*至十亿元)
11	NP	NP	*T*	93	Topicalisation (e.g. 薪水都用*T*来享乐)

Table 2: The distribution of the most frequent types of empty categories and their antecedents in CTB5.1.

Null Relative Pronouns (Table 2, rows 2 & 7) themselves are local dependencies, and thus are not coindexed with an antecedent. But they mediate non-local dependencies by functioning as antecedents for the dislocated constituent inside a relative clause.[5]

Locally Mediated Dependencies are non-local in that they are projected through a third lexical item (such as a control or raising verb) which involves a dependency between two adjacent levels and they are therefore bounded. This type encompasses: (Table 2, row 8) raising constructions, and short-bei constructions (passivisation); (row 3) control constructions, which include two different types: a generic *PRO* with an arbitrary reading (approximately equal to unexpressed subjects of *to*-infinitive and gerund verbs in English); and a *PRO* with definite reference (subject or object control).[6]

Long-Distance Dependencies differ from locally mediated dependencies, in that the path linking the antecedent and trace might be unbounded. LDDs include the following phenomena:

[5] Null relative pronouns in the CTB annotation are used to distinguish relative clauses in which an argument or adjunct of the embedded verb ‘moves’ to another position from complement (appositive) clauses which do not involve non-local dependencies.

[6] However in this case the CTB annotation does not coindex the locus (trace) with its controller (antecedent) as the *PRO* in Figure 6.

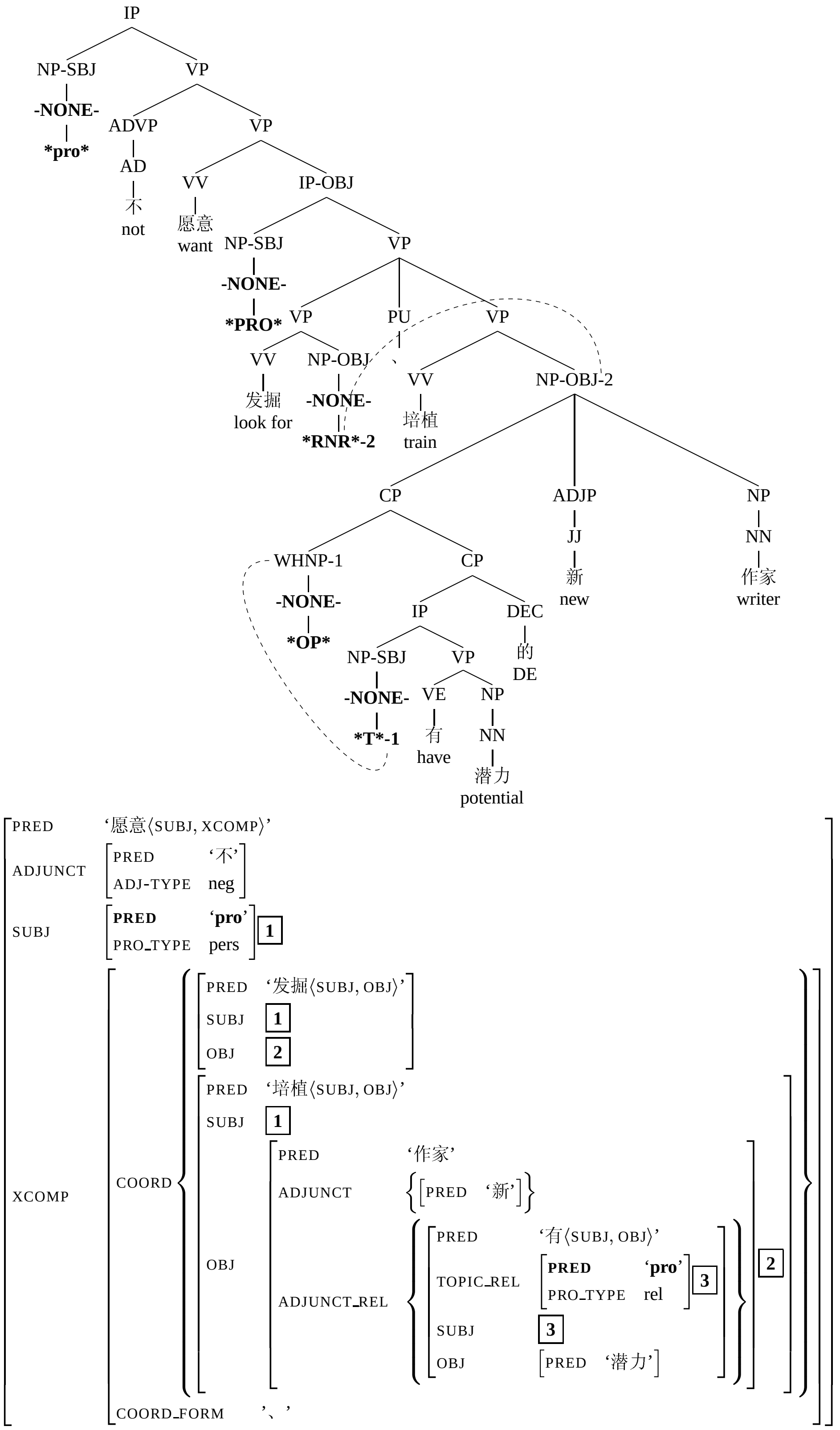


Figure 6: NLD example of sentence ‘*(People) don’t want to look for and train new writers who have potential.*’: the CTB tree and the corresponding f-structure.

Wh-traces in relative clauses, where an argument (Table 2, row 1) or adjunct (row 6) 'moves' and is coindexed with the 'extraction' site.

Topicalisation (Table 2, rows 5 & 11) is one of the typical LDDs in English, whereas in Chinese not all topics involve displacement, as shown in example (10).

(10) 北京 秋天 最 美
Beijing autumn most beautiful
'Autumn is the most beautiful in Beijing.'

Long-Bei construction as described above, takes a sentential complement which possibly involves long-distance dependencies, as in example (11).

(11) 约翰 被 玛丽 派 人 打 了
John BEI Mary send somebody hit ASP
'John was hit by somebody sent by Mary.'

Coordination is divided into two groups: right node raising of an NP phrase which is an argument shared by the coordinate predicates (Table 2, row 9); and the coordination of quantifier phrases (row 10) and verbal phrases as example (12), in which the antecedent and trace are both predicates and possibly take their own arguments or adjuncts.

(12) 我 和 他 分别 去 公司 和 *RNR* 医院
I and he respectively go to company and *RNR* hospital
'I went to the company and he went to the hospital respectively.'

Pro-drop cases (Table 2, row 4) are prominent in Chinese because subject and object functions are only semantically but not syntactically required. Nevertheless, we also treat pro-drop as a long-distance dependency as in principle the dropped subjects can be determined from the general (often inter-sentential)[7] context.

3.2 NLD Recovery Algorithm for CTB

Among these NLD types, LDDs cover various linguistic phenomena and are the most difficult to resolve. Inspired by Cahill et al. (2004), we recover long-distance dependencies at the level of f-structures, using automatically acquired subcategorisation frames and finite approximations of functional uncertainty equations describing LDD paths from the f-structure annotated CTB. Cahill et al. (2004)'s algorithm only resolves certain LDDs with known types of antecedents (TOPIC, TOPIC_REL and FOCUS). However as illustrated above, except for relative clauses, the antecedents in Chinese LDDs do not systematically correspond to types of grammatical function. Furthermore, more than half of all empty categories are not coindexed with an antecedent due to the high prevalence of pro-drop in Chinese.

[7] In this case, the 'pro' will be resolved by anaphora resolution in a later processing stage.

In order to resolve all Chinese LDDs represented in the CTB, we modify and substantially extend Cahill et al. (2004)'s algorithm as follows:

1. We extract LDD resolution paths p linking reentrances in f-structures automatically generated for the original CTB trees. To better account for all Chinese LDDs represented in the CTB, we calculate the probability of p conditioned on the GF associated with the trace t (instead of the antecedent as in Cahill et al. (2004)). The path probability $P(p|t)$ is estimated as Eq. 1 and some examples of LDD paths are listed in Table 3.

$$P(p|t) = \frac{count(p,t)}{\sum_{i=1}^{n} count(p_i,t)} \tag{1}$$

Trace (Path)	Prob.
ADJUNCT(↑TOPIC_REL)	0.9018
ADJUNCT(↑COORD TOPIC_REL)	0.0192
ADJUNCT(NULL)	0.0128
......	...
OBJ(↑TOPIC_REL)	0.7915
OBJ(↑COORD COORD OBJ)	0.1108
......	...
SUBJ(NULL)	0.3903
SUBJ(↑TOPIC_REL)	0.2092
......	...

Table 3: Examples of LDD paths

2. We extract the subcat frames s for each verbal form w from the automatically generated f-structures and calculate the probability of s conditioned on w. As Chinese has little inflectional morphology, we augment the word w with syntactic features including the POS of w, the GF of w, so as to disambiguate subcat frames and choose the appropriate one in a particular context. The lexical subcat frame probability $P(s|w, w\_feats)$ is estimated as Eq. 2 and some examples of subcat frames are listed in Table 4.

$$P(s|w, w\_feats) = \frac{count(s, w, w\_feats)}{\sum_{i=1}^{n} count(s_i, w, w\_feats)} \tag{2}$$

3. Given the set of subcat frames s for the word w, and the set of paths p for the trace t, the algorithm traverses the f-structure f to:

   - predict a dislocated argument t at a sub-f-structure h by comparing the local PRED:w to w's subcat frames s
   - t can be inserted at h if h together with t is complete and coherent relative to subcat frame s

Word:POS-GF(Subcat Frames)	Prob.
有:VE-adj_rel([subj, obj])	0.6769
有:VE-adj_rel([subj, comp])	0.1531
有:VE-adj_rel([subj])	0.0556
......	...
有:VE-comp([subj, obj])	0.4805
有:VE-comp([subj, comp])	0.2587
......	...
有:VE-top([subj, comp])	0.4397
有:VE-top([subj, obj])	0.3510
......	...

Table 4: Examples of subcat frames

- traverse f inside-out starting from t along the path p
- link t to its antecedent a if p's ending GF a exists in a sub-f-structure within f; or leave t without an antecedent if an empty path for t exists

4. Rank all resolution candidates according to the product of subcat frame and LDD path probabilities (Eq. 3).

$$P(s|w, w\_feat) \times \prod_{j=1}^{m} P(p|t_j) \tag{3}$$

As described in Section 3.1, besides LDDs, there are two other types of NLDs in the CTB5.1, and their different linguistic properties may require more fine-grained recovery strategies than the one described so far. Furthermore, as the LDD recovery method described above is triggered by dislocated subcategorisable grammatical functions, cases of LDDs in which the trace is not an argument in the f-structure, e.g. an ADJUNCT or TOPIC in relative clauses or a null PRED in verbal coordination, cannot be recovered by the algorithm. In order to recover all NLD types in the CTB5.1, we develop a hybrid methodology. The hybrid method involves four strategies (including the one described so far):

- Applying a few simple heuristic rules to insert the empty PRED for coordinations and null relative pronouns for relative constructions. The former is done by comparing the part-of-speech of the local predicates and their arguments in each coordinate; and the latter is triggered by GF ADJUNCT_REL in our system.
- Inserting an empty node with GF SUBJ for the short-bei construction and control & raising constructions, and relate it to the upper-level SUBJ or OBJ accordingly.
- Exploiting Cahill et al. (2004)'s algorithm, which conditions the probability of LDD path on the GF associated with the antecedent rather than the trace, to resolve the wh-trace in relativisation, including the ungovernable GFs TOPIC and ADJUNCT.

- Using our modified LDD resolution algorithm to resolve the remaining types.

3.3 Experimental Evaluation

For the experiments on NLD recovery, we use the first 760 articles of CTB5.1, from which 75 double-annotated files (1,046 sentences) are used as test data, 75 files (1,082 sentences) are held out as development data, while the other 610 files (8,256 sentences) are used as training data. Experiments are carried out on two different kinds of input: first on CTB gold standard trees stripped of all empty nodes and coindexation information; second, on the output trees of Bikel's parser.

We use the triple dependency relation encoding in the evaluation metric for NLD recovery. In the trace insertion evaluation, the trace is represented by the empty category, e.g. OBJ(发掘/look for, NONE) in Figure 6; and in the antecedent recovery evaluation, the trace is realised by the predicate of the antecedent, e.g. OBJ(发掘/look for, 作家/writer).

Table 5 shows the performance of the NLD recovery algorithm against (i) the CTB5.1 test set given the trees stripped of all empty nodes and coindexation and (ii) output trees by Bikel's parser. Table 6 gives the results of f-structure annotation for parser output after NLD resolution evaluated against the 200-sentence gold standard, which shows 2.3% and 2.6% improvement of pred-only measure and all-GFs measure respectively over the proto-f-structures (Table 1).

	CTB Trees			**Parser Output Trees**		
	Precision	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**
Insertion	92.86	91.45	92.15	67.29	62.33	64.71
Recovery	84.92	83.64	84.28	56.88	52.69	54.71

Table 5: Evaluation of NLD trace insertion and antecedent recovery

+NLD res.	**Precision**	**Recall**	**F-Score**
Preds Only	71.91	70.81	71.36
All GFs	80.41	79.61	80.01

Table 6: Evaluation of proper f-structures from NLD-resolved parser output

4 Conclusions and Future Work

We have reported on a project on inducing wide-coverage LFG approximations for Chinese from the CTB5.1. Our new two-stage annotation architecture provides an interface transducing c-structure trees to f-structures. The method avoids some of the limitations of the CFG rule- and annotation-based method. The more general annotation principles operating on intermediate unlabelled dependency representations allow us to scale the method to the whole Penn Chinese treebank and guarantee that every constituent-tree in the CTB5.1 can derive a complete f-structure. The separation of function annotation from the determination of the unlabelled dependency representations minimises the dependency of the functional annotation principles on the particular treebank encoding and data-structures. Our f-structure annotation algorithm is motivated by Chinese; however, in large parts it is less language-dependent than the CFG-rule- and annotation-based methods of Cahill et al. (2002) & Burke et al. (2004). As the method exploits information from a larger context, including non-local trees and lexical information, it may also benefit less configurational languages which exhibit relatively free word order, with morphology rather than phrasal position determining functional roles. Finally, the non-local dependency recovery method captures 'moved' constituents and produces a full-fledged f-structure from parser output.

Areas of current and future research include further extending the gold-standard and examining more kinds of constructions and linguistic phenomena particular to Chinese. We will also investigate ways of closing the gap between the performance of CTB trees and parser output trees, including improving parsing result for Chinese.

References

Bikel, Daniel M. and Chiang, David. 2000. Two Statistical Parsing Models Applied to the Chinese Treebank. In *Proceedings of the Second Chinese Language Processing Workshop*, pages 1–6, Hong Kong, China.

Burke, Michael, Lam, Olivia, Chan, Rowena, Cahill, Aoife, O'Donovan, Ruth, Bodomo, Adams, van Genabith, Josef and Way, Andy. 2004. Treebank-Based Acquisition of a Chinese Lexical-Functional Grammar. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation*, pages 161–172, Tokyo, Japan.

Cahill, Aoife, Burke, Michael, O'Donovan, Ruth, van Genabith, Josef and Way, Andy. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, pages 319–326, Barcelona, Spain.

Cahill, Aoife, McCarthy, Mairéad, van Genabith, Josef and Way, Andy. 2002. Automatic Annotation of the Penn Treebank with LFG F-Structure Information. In *Proceedings of the LREC Workshop on Linguistic Knowledge Acquisition and*

[†] We gratefully acknowledge support from Science Foundation Ireland grant 04/IN/I527 for the research reported in this paper.

Representation: Bootstrapping Annotated Language Data, pages 8–15, Las Palmas, Canary Islands, Spain.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph. D.thesis, Department of Computer & Information Science, University of Pennsylvania, Philadelphia, USA.

Crouch, Richard, Kaplan, Ronald M., King, Tracy H. and Riezler, Stefan. 2002. A comparison of evaluation metrics for a broad coverage parser. In *Proceedings of the LREC Workshop: Beyond PARSEVAL – Towards Improved Evaluation Measures for Parsing Systems*, pages 67–74, Las Palmas, Canary Islands, Spain.

Fang, Ji. 2006. *The Verb Copy Construction and the Post-Verbal Constraint in Chinese*. Ph. D.thesis, Department of Asian Languages, Stanford University, Standford, USA.

Frank, Anette. 2000. Automatic F-Structure Annotation of Treebank Trees. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the Fifth International Conference on Lexical Functional Grammar*, pages 226–243, Berkeley, CA, USA.

Her, One-soon. 1991. *Grammatical Functions and Verb Subcategorization in Mandarin Chinese*. Taipei: Crane Publishing.

Huang, Chu-Ren and Mangione, Louis. 1985. A Reanalysis of de: Adjuncts and Subordinate Clauses. In *Proceedings of West Cost Coast Conference on Formal Linguistics IV*, pages 80–91, University of California, Los Angeles, USA.

Kaplan, Ronald M. 1995. The formal architecture of lexical-functional grammar. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III and Annie Zaenen (eds.), *Formal Issues in Lexical-Functional Grammar*, pages 7–27, Standford, USA: CSLI Publications.

Sadler, Louisa, van Genabith, Josef. and Way, Andy. 2000. Automatic F-Structure Annotation from the AP Treebank. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the Fifth International Conference on Lexical Functional Grammar*, pages 226–243, Berkeley, CA, USA.

van Genabith, Josef., Sadler, Louisa and Way, Andy. 1999. Semi-automatic Generation of F-Structures from Treebanks. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the Fourth International Conference on Lexical Functional Grammar*, Manchester, UK.

Xue, Nianwen and Xia, Fei. 2000. The Bracketing Guidlines for the Penn Chinese Treebank (3.0). Technical Report 00-08, Institute for Research in Cognitive Science, University of Pennsylvania.

UNTANGLING THE RUSSIAN PREDICATE AGREEMENT KNOT

Hyun-Jong Hahm and Stephen Wechsler
University of Texas at Austin

Proceedings of the LFG07 Conference
Miriam Butt and Tracy Holloway King
(Editors)
2007
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

Russian predicates show a puzzling pattern of number agreement with their subjects. For example, the single-addressee use of the polite second person plural pronoun *vy* triggers *plural* number on Short Form predicate adjectives but *singular* on Long Form predicate adjectives. To solve this and other related puzzles, we draw upon several independently motivated assumptions: (i) the INDEX vs. CONCORD agreement distinction (Wechsler and Zlatic 2003; King and Dalrymple 2004); (ii) the analysis of singular target forms as marked both morphologically and semantically, with plurals filling in elsewhere (Wechsler 2004, 2005); and (iii) the nominal ellipsis analysis of Long Form predicate adjectives (Babby 1973; Siegel 1976; Bailyn 1994).

1 Introduction

Russian predicates exhibit a puzzling pattern of number agreement with their subjects, apparently conditioned in complex ways by both the type of agreement 'target' such as a finite verb or predicate adjective, and the semantics and form of the subject agreement 'trigger'. For example, like many other languages, Russian allows an honorific use of its second person plural pronoun *vy* to address a single person politely. Consider whether such a pronoun, used for a single addressee, triggers singular agreement (reflecting the meaning) or plural agreement (reflecting the form) on various targets. Russian predicate adjectives appear in two possible forms, the so-called Short Form (SF, see (1a)) and the Long Form (LF, see (1b) and (1c)). It turns out that a single-addressee use of *vy* triggers *plural* on SF adjectives (1a) but *singular* on LF adjectives (1b):

(1) a. Vy byli sčastlivy.
2PL be.past.PL happy.SF.PL
'You (one formal addresee or more than one addressee) were happy.'

b. Vy byli sčastlivyj.
2PL be.past.PL happy.LF.Nom.Masc.SG
'You (one formal male addressee) were happy.'

c. Vy byli sčastlivye.
2PL be.past.PL happy.LF.Nom.PL
'You (more than one addressee) were happy.'

In addition to showing the contrast between SF and LF adjectives, these data also illustrate *mixed agreement*, where a single subject triggers two different agreement values: in (1b) the finite verb is plural while the adjective is singular.

It is a complex but ultimately fairly straightforward matter to *describe* such puzzling agreement patterns by stipulating every allowable combination of trigger and target form. But one would like to go beyond a mere description and also *explain* facts like the ones illustrated in (1).

In this paper we offer such an explanation. To do so we draw upon several independently motivated assumptions: (i) the INDEX vs. CONCORD agreement distinction (Wechsler and Zlatic 2003; King and Dalrymple 2004); (ii) the analysis of singular targets as *marked* both morphologically and semantically, with plurals filling in elsewhere (Wechsler 2004, 2005); and (iii) the 'nominal ellipsis' analysis of Long Form predicate adjectives (Babby 1973; Siegel 1976; Bailyn 1994).

2 Polite Plurals

In many languages a second person plural pronoun can be used politely for a single person. Examples of such forms are French *vous*, Turkish *siz*, Persian *ʃomâ*, Romanian *Dumneavoastră*, and Russian *vy* and its cognates in other Slavic languages. Number agreement with such forms, when used of a single addressee, varies across languages even within Slavic (Comrie 1975; Corbett 1983). For example, Czech has mixed agreement with *vy*: finite verbs are plural while predicate adjectives are singular, as shown in (2c).

(2) Mixed agreement in Czech (Hahm 2006b)
a. Ty jsi čestný.
2SG be.2SG honest.Masc.SG
'You (one intimate male addressee) are honest.'
b. Vy jste čestní.
2PL be.2PL honest.Masc.PL
'You (multiple addressees) are honest.'
c. Vy jste čestný.
2PL be.2PL honest.Masc.SG
'You (one formal male addressee) are honest.'

Number on the predicate adjective varies depending on whether there is one or more than one addressee.

In contrast to the mixed agreement found in Czech, Serbo-Croatian has uniform agreement with *vi*: plural on both finite verbs and predicate adjectives, as shown in (3b).

(3) Uniform agreement in Serbo-Croatian (Wechsler 2004)

a. Ti   si   duhovit / duhovita.
2SG   AUX.2SG   funny.Masc.SG / funny.Fem.SG
'You (one informal male/female addressee) are funny.'

b. Vi   ste   duhoviti.
2PL   AUX.2PL   funny.Masc.PL
'You (one formal addressee or multiple addressees) are funny.'

Unlike (2b), sentence (3b), with plural on both agreement targets, can be used to address either a single person or more than one.

Turning now to Russian, as noted in the introduction, Russian number agreement on predicate adjectives depends on whether the adjective is a Short Form adjective (e.g. *krasiv* 'beautiful.SF') or a Long Form adjective (e.g. *krasivyj* 'beautiful.LF').[1] The polite, single-addressee use of *vy* triggers plural on SF adjectives but singular on LF adjectives. This contrast was illustrated in (1) above; a more complete paradigm appears here:

(4) Short Form adjectives

a. Ty   byl   sčastliv.
2SG   be.past.Masc.SG   happy.SF.Masc.SG
'You (one informal male addressee) were happy.'

b. Vy   byli / *byl   sčastlivy / *sčastliv.
2PL   be.past.PL/Masc.SG   happy.SF.PL/*SF.Masc.SG
'You (one formal or more than one addressee) were happy.'

(5) Long Form adjectives

a. Ty   byl   sčastlivyj.
2SG   be.past.Masc.SG   happy.LF.Nom.Masc.SG
'You (one intimate male addressee) were happy.'

b. Vy   byli / *byl   sčastlivyj.
2PL   be.past.PL/Masc.SG   happy.LF.Nom.Masc.SG
'You (one formal male addressee) were happy.'

c. Vy   byli   sčastlivye.
2PL   be.past.PL   happy.LF.Nom.PL
'You (more than one addressee) were happy.'

On the basis of (5b) and (5c) it looks like Russian LF adjectives show *semantic* rather than *grammatical* agreement, hence singular when the subject refers to one individual, plural for more than one. That this is not correct is

[1] Semantic differences between SF and LF adjectives are discussed below.

shown by agreement with pluralia tantum nouns, that is, nouns that are always morphologically plural but can refer to one or more than one entity, such as English *scissors* and *pants*. Regardless of whether they are used for singular or plural reference, Russian pluralia tantum nouns such as *očki* 'glasses' or *bryuki* 'pants' trigger *plural* agreement on both SF and LF adjectives, as shown here:

(6) SF adjectives
Èti otčki krasivy / *krasiv.
these glasses.PL beautiful.SF.PL / *SF.Masc.SG
'These glasses (one or more than one pair) are beautiful.'

(7) LF adjectives
Èti otčki krasivye / *krasivyj.
these glasses.PL beautiful.LF.Nom.PL/*LF.Nom.Masc.SG
'These glasses (one or more than one pair) are beautiful.'

Before tabulating these patterns we add one more type of predicate, namely predicate nominals. In keeping with a strong cross-linguistic tendency (Corbett 1983; Corbett 2000:194-5), Russian predicate nominals consistently show semantic agreement with both *vy* and pluralia tantum subjects (Hahm 2006a):[2]

(8) a. Ty byl geroem.
2SG be.past. Masc.SG hero.Inst.SG
'You (one informal male addressee) were a hero.'

b. Vy byli geroem.
2PL be.past.PL hero.Inst.SG
'You (one formal addressee) were a hero.'

c. Vy byli gerojami.
2PL be.past.PL hero.Inst.PL
'You (multiple addressees) were heroes.'

(9) a. Èti očki special'nyj instrument čtoby smotret' fil'm.
these glasses special.SG tool.SG so.that watch film
'These glasses (one pair) are a special tool to watch a (e.g. IMAX) movie.'

[2] In Russian, the present tense copula is null as shown in (9).

b. Èti očki special'nye instrumenty čtoby smotret' fil'm.
these glasses special.PL tool.PL so.that watch film
'These glasses (>1 pair) are special tools to watch a movie.'

Summarizing our findings on Russian predicate agreement, a normal singular subject triggers singular agreement on all targets, and a normal plural subject triggers plural agreement. When the subject is morphologically plural but semantically singular, we get the pattern shown in Table I.

subject trigger	**finite verbs**	**adjectives**		**predicate nominals**
		SF	**LF**	
vy (single addressee)	PL	PL	SG	SG
pluralia tantum	PL	PL	PL	SG

Table I. Russian predicate agreement with morphologically plural but semantically singular subjects.

A careful look at Table I should help the reader appreciate the difficulty of the problem. It will not do to stipulate either 'grammatical agreement' or 'semantic agreement' for LF adjectives since their behavior differs across the two types of trigger. And besides, as noted in the introduction, one would hope to explain rather than merely stipulate a solution. Our explanation involves three independently motivated factors, to which we turn next.

3 CONCORD and INDEX Agreement

Building on Pollard and Sag (1994) and Kathol (1999), Wechsler and Zlatic (2003) propose a theory of agreement based on the distinction between CONCORD and INDEX agreement (Wechsler and Zlatic 2000, 2003; King and Dalrymple 2004). An agreement trigger such as a noun or pronoun carries both CONCORD and INDEX agreement feature sets, which are understood as grammaticalizations of morphological and semantic properties, respectively (but not reducible to them). CONCORD is related to trigger morphology such as declension class and typically determines NP-internal agreement. The referential INDEX determines anaphoric agreement (e.g. between pronoun and antecedent), because anaphoric binding itself is modeled as INDEX-sharing (Pollard and Sag 1994). While CONCORD features reflect the morphological properties of the NP trigger, INDEX features tend to reflect the semantics of the NP trigger.

Normally the CONCORD and INDEX values for person, number, and gender simply match, but some mismatches exist. These mismatches are detectible by the phenomenon of mixed agreement. For example, Serbo-Croatian has a class of singularia tantum nouns like *deca* 'children' that trigger feminine singular on targets within the NP and neuter plural on pronouns (Corbett 1983; Wechsler and Zlatic 2003):

(10)

Posmatrali	smo	ovu	dobru	decu.
watched.1PL	AUX	this.Fem.SG	good.Fem.SG	children.Acc

Ona	su	se	lepo	igrala.
they.Neut.PL	AUX.3PL	REFL	nicely	played.Neut.PL

‘We watched those good children. They played well.’
(example from Wechsler and Zlatić 2003)

$$\textit{deca:}\ \begin{bmatrix} \text{CONCORD} & \text{fem.sg} \\ \text{INDEX} & \text{neut.pl} \end{bmatrix}$$

As noted above, NP-internal agreement tends toward CONCORD while anaphoric pronoun agreement is INDEX agreement. Predicate targets are mixed, as we will see below.

Returning next to Russian, we will ascertain the agreement features of the relevant agreement triggers, such as pronouns and pluralia tantum nouns, and the specifications for the various predicate targets.

4 Agreement triggers

Russian pronouns have the familiar paradigm formed by crossing three person values with two number values.

(11) a. $\left.\begin{array}{l} \textit{Ja}\ \text{'I'} \\ \textit{Ty}\ \text{'you (SG)'} \\ \textit{On}\ \text{'he'} \end{array}\right\}$ …*byl* 'be.past.Masc.SG' ...

b. $\left.\begin{array}{l} \textit{My}\ \text{'we'} \\ \textit{Vy}\ \text{'you (PL)'} \\ \textit{Oni}\ \text{'they'} \end{array}\right\}$ …*byli* 'be.past.PL' ...

The past tense verb forms shown in (11) confirm that Russian has a true number feature cutting across the person values and grouping together the pronouns as shown.[3]

[3] Cysouw (2003) argues that many languages lack a true number distinction in first and second person pronouns. Wechsler (2004, 2005) applies this idea to French mixed agreement, noting that French lacks target forms that cluster together the purported

Based on the agreement facts in Section 2 above, we propose the following lexical entries for polite pronoun *vy* and pluralia tantum nouns:

(12) a. *vy*: Pron (↑ PRED) = 'PRO'
(↑ PERS) = 2nd
(↑ CONC NUM) = PL
(↑ INDEX NUM) = ($\uparrow_{\sigma}$ NUM)

b. *bryuki:* N (↑ PRED) = 'PANTS'
(↑ CONC NUM) = PL
(↑ INDEX NUM) = PL

The pronoun *vy* is 'morphologically plural', hence its CONC(ord) value is PL(ural). But its INDEX number is tied to its semantic number, as encoded by the last equation in that entry (σ is the semantic projection function). In contrast, *bryuki* 'pants' is PL(ural) in both features, regardless of semantic cardinality. Before showing how these specifications work in our analysis, we present some independent evidence to support them.

Recall that the INDEX feature set resides on the referential index, hence it is tracked by anaphoric binding. The pronoun *vy* differs systematically from pluralia tantum nouns like *bryuki* 'pants' with respect to number agreement determined on anaphoric pronouns. As shown in (13a), a Russian *pluralia tantum* antecedent binds a plural pronoun, much like in English: *The trousers$_i$ are too tight; they$_i$ need to be altered.* It also takes a plural relative pronoun (see (13b)) ((13) and (14) are from Hahm 2006a):

(13) a.

Ja	kupil	eti	bryuki	včera.
1SG	bought.1SG	this.PL	pants	yesterday

Ja	lyublyu	ix / *ego.
1SG	love.1SG	they.Acc / it.Acc

'I bought a pair of pants yesterday. I love them.'

b.

Èti	bryuki,	kotorye /*kotoryj	dala
this.PL	pants.PL	rel-pron.PL/*SG	gave

mne	moya	babuška,	moi	lyubimye.
to.me	my	grandmother	my.PL	favorite.PL

'These pants, which my grandmother gave me, are my favorite.'

singular versus plural pronouns. The agreement shown in (11) shows that such an analysis is inappropriate for Russian.

The personal pronoun *ix* and relative pronoun *kotorye* are plural forms, supporting the plural INDEX number on *bryuki* ‘pants.’ However, with *vy* a singular relative pronoun is preferred when used for singular reference (i.e. with one addressee):

(14) Vy, kotoraja / kotoryj (>>kotorye) stol'ko
you rel-pron.Fem / Masc.SG (>> PL) so.much

čitaete, mnogo znaete.
read.2PL much know.2PL

‘You (one formal addressee), who read much, know much.’

Summarizing, *bryuki* ‘pants’ has a plural index while *vy* ‘you’ has singular or plural index depending on the meaning.

5 Agreement targets

Now let us turn our attention to the agreement targets, considering first the finite verbs and SF adjectives. Recall from Table I above that these targets consistently show ‘grammatical agreement’ across the different types of trigger, hence plural for the grammatically plural *vy* and pluralia tantum nouns. It would seem that the semantics of plurality can be safely ignored. As a first approximation, then, the lexical entries of singular verbs and SF adjectives would contain the equation (↑SUBJ CONC NUM) = SG , while their plural counterparts would have (↑SUBJ CONC NUM) = PL .

However, in Russian as in English, French, and perhaps all languages, agreement always retains some shadow of its semantic side, a semantic side that emerges in special contexts that block the grammatical feature. For example, the number value on predicates with a coordinate NP subject seems to reflect the semantic number of the subject, as in these examples:

(15) a. [Moj lučšij drug i redaktor moej
my.SG best.SG friend.SG and editor.SG my.Gen.Fem.SG

biografii] byl zdes' s vizitom.
autobiography.Gen.Fem.SG be.past.Masc.SG here with visit

‘My best friend and the editor of my autobiography (referring to one person) was here for a visit.’

b.

[Moj	lučšij	drug	i	redaktor	moej
my.SG	best.SG	friend.SG	and	editor.SG	my.Gen.Fem.SG

biografii]	byli	zdes'	s	vizitom.
autobiography.Gen.Fem.SG	be.past.PL	here	with	visit

'My best friend and the editor of my autobiography (referring to two different people) were here for a visit.'

With the singular verb in (15a) the subject is understood as singular (the speaker's best friend is her biographer), while the plural verb in (15b) brings with it a plural interpretation. The same applies to the English translations, incidentally (Farkas and Ojeda 1983; Farkas and Zec 1993; Farkas and Zec 1995).

We are faced by a paradox: these agreement targets seem to show grammatical agreement in some situations, apparently ignoring meaning (see Table I), but show semantic agreement in others.

To solve this paradox we follow Wechsler (2005) in positing that the singular target form is marked for singular both grammatically and semantically, as shown in (16a). Note that this lexical entry has two agreement equations, one for CONCORD and one for the semantic interpretation. The singular form is thus the marked form in the singular / plural opposition. The corresponding plural form is unmarked, exhibiting an 'elsewhere distribution', that is, applying whenever the conditions for the singular form are not met. Perhaps pending a more adequate formalization in a sophisticated theory of markedness such as Optimality Theory, we can capture this distribution with the disjunctive specification shown in (16b):

(16) a. *krasiv:* A (↑ PRED) = 'BEAUTIFUL<SUBJ>'
(↑ SUBJ CONC NUM) = SG
((↑ SUBJ)$_\sigma$ NUM) = SG

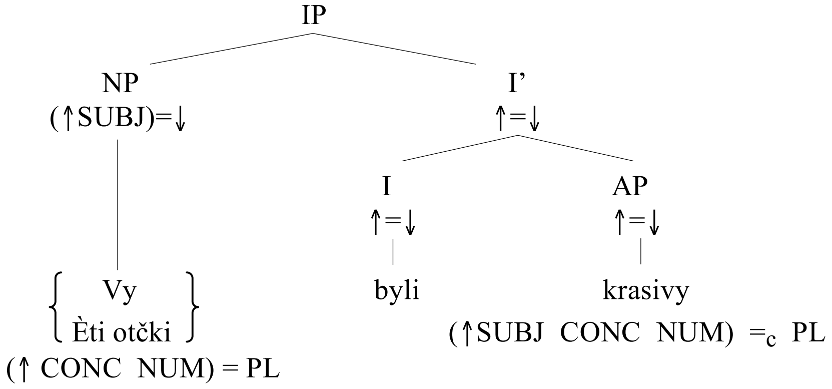
b. *krasivy:* A (↑ PRED) = 'BEAUTIFUL<SUBJ>'
{ (↑ SUBJ CONC NUM) =c PL |
((↑ SUBJ)$_\sigma$ NUM) = PL }

Given the constraining equation in its lexical entry, the plural target form *krasivy* (the SF adjective 'beautiful') effectively 'checks' the subject for morphological plurality, otherwise imposing plural semantics. That is, if the constraining equation is not satisfied, because the subject lacks a plural CONCORD feature, then the semantic equation must be active. In effect the plural target feature must be motivated either by morphology or semantics.

Let us assume that a coordinate NP as in (15) lacks a CONCORD feature entirely, perhaps because it is endocentric and CONCORD (or at least (CONCORD NUMBER)), is non-distributive (King and Dalrymple 2004). Then all the facts surveyed follow: *vy* and pluralia tantum nouns are morphologically plural, i.e. they have a plural CONCORD feature, so the verb or SF adjective cannot be singular, and the plural form does not impose plural semantics. But the coordinate NPs in (15) lack a CONCORD feature, so they allow either singular or plural, imposing semantic singularity or plurality, respectively.

C- and f-structures for representative examples where the subject is grammatically plural are shown in (17):

(17) a. c-structure for *Vy byli krasivy* and *Èti otčki byli krasivy*:



b. f-structure for *Vy byli krasivy /*krasiv*
‘You were beautiful.’
and *Èti otčki byli krasivy /*krasiv*
‘These glasses were beautiful.’:

$$\begin{bmatrix} \text{PRED} & \text{'BEAUTIFUL} < \text{SUBJ} > \text{'} \\ \text{TENSE} & \text{PAST} \\ & \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO' } \textit{or} \text{ 'glasses'} \\ \text{CONC} & [\,\text{NUM} \quad \text{PL}\,] \end{bmatrix} \end{bmatrix}$$

In contrast, a coordinate NP subject lacks the (↑ CONC NUM) = PL equation; hence the constraining equation on the adjective is not satisfied, so the adjective imposes semantic plurality.

6 Long form adjectives

Turning now to LF adjectives, recall that pluralia tantum nouns trigger plural on LF adjectives, but *vy* triggers semantic agreement:

(18) Èti otčki krasivye / *krasivyj.
these glasses.PL beautiful.LF.Nom.PL/*LF.Nom.Masc.SG
'These glasses (one or more than one pair) are beautiful.'

(19) a. Vy byli sčastlivyj.
2PL be.past.PL happy.LF.Nom.Masc.SG
'You (one formal male addressee) are happy.'

b. Vy byli sčastlivye.
2PL be.past.PL happy.LF.Nom.PL
'You (more than one addressee) were happy.'

To understand this fact, we adopt a longstanding proposal that an LF adjective in predicate position is really an attributive adjective modifying a null nominal head (Babby 1973; Siegel 1976; Baylin 1994). That is, (19a) can be paraphrased roughly as 'You are a happy one'. As in many languages, Russian expresses 'one-anaphora' by eliding the head noun from the NP. Within LFG we do not need literally to posit a null head, so we will express this idea differently, but the essential insight is taken from the works listed above.

Let us briefly review the evidence for this analysis of LF adjectives. First, in addition to serving as predicates, LF adjectives can also serve as nominal attributive modifiers, while SF adjectives cannot. This immediately predicts a systematic difference between the two, since we know independently that Russian allows noun ellipsis for one-anaphora. In diachronic perspective this evidence is even stronger: LF adjectives in Old Russian were *only* used as prenominal attributive modifiers and could not be predicative (Bailyn 1994). Also, LF adjectives inflect for case, a property typical of NP-internal items, while SF adjectives do not inflect for case.

There is also compelling semantic evidence. LF predicate adjectives have the partitive semantics typical of one-anaphora, as shown by the following contrasts (Siegel 1976):

(20) a. Prostrantsvo beskonečno (SF) / *beskonečnoe (LF).
'Space is infinite.' (cp. #Space is an infinite one.)

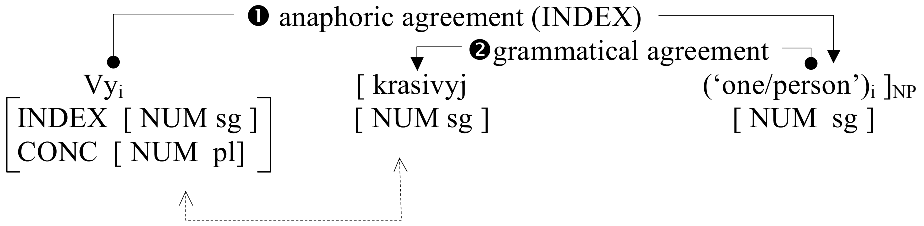
b. Vse jasno (SF) / *jasnoe (LF).
'Everything is clear.' (cp. #Everything is a clear one.)

c. Prixodit' domoj očen' prijatno (SF) / *prijatnoe (LF).
'To come home is very pleasant.'
(cp. #To come home is a very pleasant one.)

Compare the English translations. The LF adjectives suggest selection from some presupposed larger set.

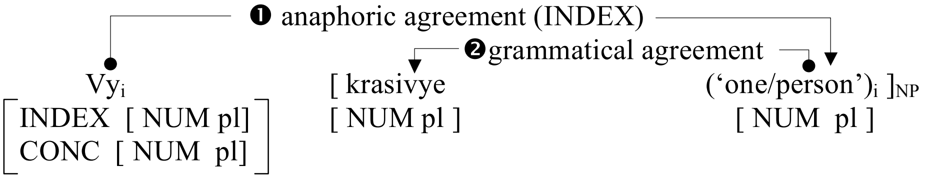
How is this relevant to agreement? We put forth the following proposal. The LF adjective actually shows grammatical (CONCORD) agreement with its null nominal head. That null head, being anaphoric, shows INDEX agreement with its antecedent, the subject. This gives the appearance of INDEX agreement.

(21) a. Vy krasivyj.
'You (one formal addressee) are beautiful'.

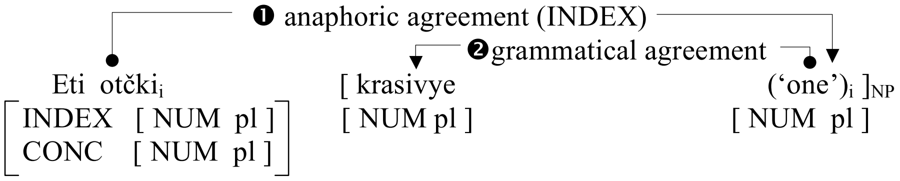


Result: appears to be INDEX agreement.

b. Vy krasivye.
'You (more than one addressee) are beautiful.'



c. Eti otčki krasivye.
'These glasses (one or more pairs) are beautiful.'



Recall from Section 4 above that *bryuki* 'pants' has a plural INDEX, while the INDEX on *vy* has a number value tied to its semantic number. The lexical entries in (12) are repeated here for convenience:

(22) a. *vy*: Pron (↑ PRED) = 'PRO'
(↑ PERS) = 2nd
(↑ CONC NUM) = PL
(↑ INDEX NUM) = ($\uparrow_{\sigma}$ NUM)

b. *bryuki:* N (↑ PRED) = 'PANTS'
(↑ CONC NUM) = PL
(↑ INDEX NUM) = PL

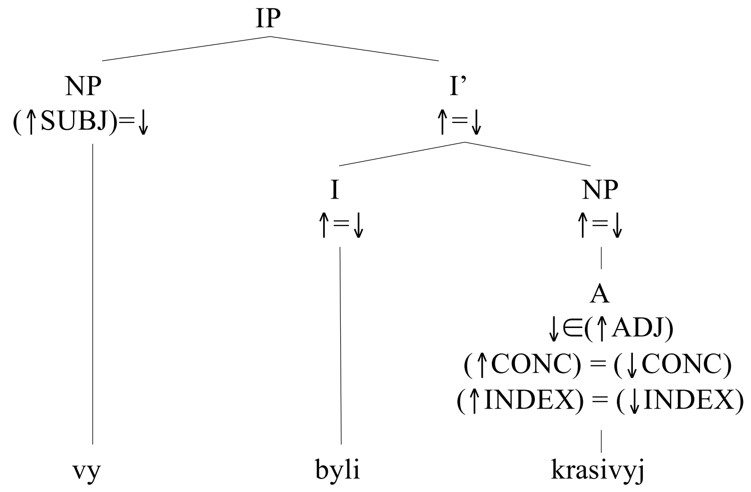
In Section 4 we supported these features on the basis of agreement in anaphoric binding. So our analysis of LF agreement in terms of one-anaphora effectively assimilates LF agreement to the other anaphoric agreement facts.

The null nominal head analysis can be expressed in LFG as follows. The LF adjective has an optional equation to introduce the implicit anaphoric PRED ('ONE <SUBJ>'). (The inside-out function application equation in the second line in (23a) places this PRED feature on the f-structure for the NP dominating the adjective. See the f-structure in (23c).) The variant including that optional equation is the predicative adjective, and the variant without it is a prenominal attributive modifier.

(23) Vy byli krasivyj$_{\text{[LF.M.SG].}}$
'You (one formal addressee) were beautiful.'

a. *krasivyj*: A (↑ PRED) = 'BEAUTIFUL'
(((ADJ ∈ ↑) PRED) = 'ONE <SUBJ>')
(↑CONC NUM) = SG
(↑CONC GEND) = MASC
(↑CONC CASE) = NOM
(↑INDEX NUM) = SG
($\uparrow_{\sigma}$ NUM) = SG

b. c-structure for *Vy byli krasivyj*:



c. f-structure for *Vy byli krasivyj:*

$$\begin{bmatrix} \text{PRED} & \text{'ONE} < \text{SUBJ} > \text{'} \\ \text{CONC} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{CASE} & \text{NOM} \end{bmatrix} \\ \text{INDEX} & [\text{NUM SG}]_{i} \\ \text{ADJ} & \{ [\ \text{PRED} \quad \text{'BEAUTIFUL'}\ ]\ \} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{CONC} & [\ \text{NUM} \quad \text{PL}\ ] \\ \text{INDEX} & [\ \text{NUM} \quad \text{SG}\ ]_{i} \end{bmatrix} \end{bmatrix}$$

Lastly, for completeness we will consider predicate nominals. Recall from Table I above that predicate nouns show semantic number agreement with their subjects. Hence a singular predicative noun has a lexical entry like the following.

(24) *instrument:* N (↑ PRED) = 'TOOL <SUBJ>'
(↑ CONC NUM) = SG
$((\uparrow \text{SUBJ})_{\sigma}$ NUM) = SG

This sort of pure semantic agreement is typical of predicate nominals across languages (Corbett 1983). It is probably explained by the fact that nominals can refer, so the number feature semantically modifies the predicate nominal itself. In that sense the correlation between the number of the predicate nominal and the subject may not be agreement at all, strictly speaking, but rather a consequence of semantic composition. On the latter view, in the equation above, the expression $((\uparrow \text{SUBJ})_{\sigma}\ \text{NUM})$ would be replaced with $(\uparrow_{\sigma}\ \text{NUM})$. Then the singular *instrument* denotes a property of a single tool, and the semantic effects on the subject are just a side effect of semantic composition.

7 Conclusion

The complex agreement patterns described in this paper can be understood as an interaction of independently motivated grammatical factors. First of all, we applied an earlier proposal by Wechsler (2005) that singular agreement targets are marked for both grammatical and semantic singularity, so that the plural counterpart, being distributionally unmarked, fills in the other options. In effect it is disjunctively specified for grammatical or semantic plurality: hence it checks the subject for morphological plurality, imposing semantic plurality if it fails to find that plural feature.

The main innovation of this paper is the idea that LF adjectives behave like anaphors with respect to agreement because they modify an implicit anaphor in the predicate position.

References

Babby, Leonard H. (1973). The Deep Structure of Adjectives and Participles in Russian. *Language* 49(2): 349-360.

Bailyn, John (1994). The Syntax and Semantics of Russian Long and Short Adjectives: An X'-Theoretic Account. *Annual Workshop on Formal Approaches to Slavic Linguistics: The Ann Arbor Meeting: Functional Categories in Slavic Syntax*. J. Toman. Ann Arbor, Michigan Slavic Publications**:** 1-30.

Comrie, Bernard (1975). Polite Plurals and Predicate Agreement. *Language* 51(2): 406-418.

Corbett, Greville (1983). *Hierarchies, Targets and Controllers: Agreement Patterns in Slavic*. London, Croom Helm.

Corbett, Greville G. (2000). *Number*, Cambridge University Press.

Cysouw, Michael (2003). *The Paradigmatic Structure of Person Marking*. Oxford University Press.

Farkas, Donka F. and Almerindo Ojeda (1983). Agreement and coordinate NPs. *Linguistics* 21: 659-673.

Farkas, Donka F. and Draga Zec (1993). *Agreement and Pronominal Reference*. Santa Cruz, CA, Linguistic Research Center, University of California.

Farkas, Donka F. and Draga Zec (1995). Agreement and Pronominal Reference. *Advances in Roumanian Linguistics*. G. Cinque and G. Giusti. Philadelphia, John Benjamins**:** 83-101.

Hahm, Hyun-Jong (2006a). Number Agreement in Russian Predicates. *Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, CSLI Publications, Stanford.

Hahm, Hyun-Jong (2006b). 'Uniform or Mixed agreement due to the Personal Pronouns.' Paper presented at Midwest Slavic Conference, Ohio State University.

Kathol, Andreas (1999). Agreement and the Syntax-Morphology Interface in HPSG. *Studies in Contemporary Phrase Structure Grammar*. R. Levine and G. Green. New York, Cambridge University Press.**:** 223--274.

King, Tracy H. and Mary Dalrymple (2004). Determiner agreement and noun conjunction. *Journal of Linguistics* 40(01): 69-104.

Pollard, Carl and Ivan Sag (1994). *Head Driven Phrase Structure Grammar*. Stanford and Chicago, CSLI Publications and University of Chicago Press.

Siegel, Muffy (1976). Capturing the Russian Adjective. *Montague Grammar*. B. H. Partee**:** 293-309.

Wechsler, Stephen (2004). Number as Person. *Empirical Issues in Syntax and Semantics 5*. O. Bonami and P. C. Hofherr**:** 255-274.

Wechsler, Stephen (2005). *Markedness and Meaning in Agreement*. LFG 2005, Bergen, Norway.

Wechsler, Stephen and Larisa Zlatic (2000). A Theory of Agreement and Its Application to Serbo-Croatian. *Language* 76(4): 799-832.

Wechsler, Stephen and Larisa Zlatic (2003). *The Many Faces of Agreement*. Stanford, California, CSLI Publications.

EXTENDING THE APPLICABILITY OF LEXICAL MAPPING THEORY

Anna Kibort

Surrey Morphology Group,
University of Surrey, UK

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

LFG grants syntactic functions a central role and has developed a theory of argument structure, Lexical Mapping Theory (LMT), which is independent of phrase-structure trees and thus able to account for morpholexical derivations. Yet some fundamental phenomena falling within the scope of morpholexical analysis – such as morphosemantic (meaning-altering) operations, phenomena referred to elsewhere as 'demotions', or subjectlessness – are currently denied satisfactory LMT accounts. This paper offers a way of extending LMT to phenomena which are awkward or impossible to handle with the current widely accepted versions of LMT.

While retaining the main component of LMT – the feature decomposition of syntactic functions – I suggest the following set of revisions: (1) restoring the early LFG distinction between argument positions and semantic roles; (2) allowing the semantic participants to change order and re-associate with different argument positions for non-default (morpho*semantically* altered) mappings; (3) fixing the order of (syntactic) argument positions; (4) reformulating the principles of argument-to-function mapping to make fuller use of the markedness hierarchy of syntactic functions and render the Subject Condition redundant; and (5) using a mechanism of increasing markedness to account for morpho*syntactic* operations referred to as 'demotions' in RG. I demonstrate that these revisions make LMT a cleaner formalism which is immediately applicable to some important phenomena that have so far escaped (good) analyses.[1]

1 Revision 1: Restoring the early LFG distinction between argument positions and semantic roles

LFG's argument structure is the locus of the mapping between semantic roles and grammatical functions. Because it maps from some kind of semantic or conceptual representation to a syntactic representation of grammatical functions, it is widely accepted that argument structure is a representation of the syntactic arguments of a predicate and that it contains some amount of semantic information, even though researchers still do not agree on how much. See Dalrymple (2001:197-200) for an overview of two major approaches to the content and representation of argument structure within LFG: Jackendovian and Dowtyian; and Butt (2006:Chapter 5) for a critical account of Jackendoff's and Dowty's linking theories and the way they have been combined with LMT.

Although the discussion of the semantic component of LMT has concentrated on the source and classification of the semantic content ascribed to the arguments (drawing from the Conceptual Semantics framework of Jackendoff 1983, 1990; or

[1] I gratefully acknowledge the current ESRC grant RES-051-27-0211. I also wish to thank Miriam Butt and Yehuda Falk for their very helpful comments and questions on this paper, some of which remain to be addressed in further work.

the Proto-Role classification of Dowty 1991), another relevant issue concerns the degree of association (i.e. either fusion or separation) of the semantic information and the syntactic argument positions. It is the second issue which falls under the scope of the proposed first revision.

Early LFG representations of argument structure implied a dissociation of argument positions and semantic roles, for example (Bresnan 1982:6):

(1) (SUBJ) (OBJ)
| | ← lexical assignment of grammatical functions
'LOVE (arg 1 , arg 2)' ← predicate argument structure
(agent) (theme)

Dalrymple (2001:198) attributes the following representation of the semantic form for *give* to Kaplan & Bresnan (1982):

(2) SUBJ OBJ OBL_{GOAL}
'give ⟨ __ , __ , __ ⟩'
AGENT THEME GOAL

and explains that the semantic form was thought of as 'expressing a kind of logical formula encoding aspects of the meaning of the sentence as well as the relation between thematic roles and their syntactic functions.'

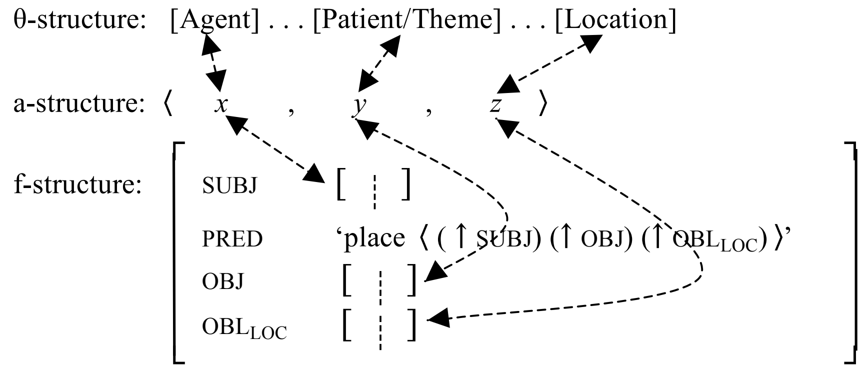
With the advent of LMT (Bresnan & Kanerva 1989; Bresnan & Zaenen 1990), which offered a substantive account of grammatical functions, argument positions and semantic roles became explicitly fused: 'the grammatically significant participant-role relations in the structure of events are represented by a-structures. An a-structure consists of a predicator with its argument roles, an ordering that represents the relative prominence of the roles, and a syntactic classification of each role indicated by a feature' (Bresnan & Zaenen 1990:48):

(3) *pound* ⟨ *ag* *pt* ⟩
[– o] [– r]

Although LMT currently exists in several variants, and there is no agreement about the substance of the participant roles, most researchers seem to adopt a model of argument structure corresponding to the representation in (3) and do not question the collapsed distinction between argument positions and semantic roles.

However, the need to separate these two tiers of representation has already had strong proponents such as Grimshaw (1988:1), T. Mohanan (1990/1994:15ff), Ackerman (1991:12; 1992:57ff), Joshi (1993), Alsina (1996:37), Ackerman & Moore (2001:40ff). In his LFG textbook, Falk (2001:105) offers the following representation of the mappings which are captured by LMT:

(4) *place*:


θ-structure: [Agent] . . . [Patient/Theme] . . . [Location]
a-structure: ⟨ x , y , z ⟩
f-structure:
SUBJ [ ]
PRED 'place ⟨ (↑ SUBJ) (↑ OBJ) (↑ OBL_LOC) ⟩'
OBJ [ ]
OBL_LOC [ ]

He emphasises that 'a-structure is a representation of the syntactic argument-taking properties of a lexical item' (2001:105); 'arguments fit empty positions in the meaning of a predicate' and 'can be identified by their role in the predicate's meaning' (2001:101). Hence, 'LMT maps between θ-structure and a-structure, and between a-structure and f-structure'; as a syntactic representation, a-structure 'only deals with syntactically relevant aspects of θ-structure and is the locus of constraints' (2001:105).

Four types of arguments can be put forward in support of the distinction between argument positions (corresponding to Falk's a-structure) and semantic roles (corresponding to Falk's θ-structure):

(i) The strongest evidence in support of this distinction comes from pairs of clauses that exhibit alternative assignments of grammatical functions to the semantic participants competing for the same argument status. Many different types of alternations have been identified where, holding constant both the predicate and the selected participants, there are two (and sometimes more than two) ways of matching the same set of grammatical functions with the participants which are available for mapping. I argue that the different options arise because the mapping is done indirectly, via an independent tier of representation: the argument structure positions (which correspond to Falk's a-structure). A common type of alternation involves two arguments within the verb phrase, either of which can be specified as an object (OBJ) or an oblique (OBL$_\theta$). An example is locative alternation, discussed in Ackerman (1991; 1992) and Ackerman & Moore (2001) (see also Levin 1993:49-55 for references):

(5) a. *The peasant loaded* *(the) hay* *onto the wagon.*
OBJ OBL$_\theta$

b. *The peasant loaded* *the wagon* *with (the) hay.*
OBJ OBL$_\theta$

Levin (1993:Chapter 2) gives the following examples of other alternations in English which involve arguments within a verb phrase: the material/product alternation (transitive) (e.g. *Martha carved a toy out of the piece of wood* ~ *Martha carved the piece of wood into a toy*), the fulfilling alternation (*The judge presented a prize to the winner* ~ *The judge presented the winner with a prize*), the image impression alternation (*The jeweller inscribed the name on the ring ~ The jeweller inscribed the ring with the name*), the *with/against* alternation (*Brian hit the stick against the fence ~ Brian hit the fence with the stick*), the *through/with* alternation (*Alison pierced the needle through the cloth ~ Alison pierced the cloth with a needle*), the *blame* alternation (*Mira blamed the accident on Terry ~ Mira blamed Terry for the accident*), the *search* alternations (*Ida hunted the woods for deer ~ Ida hunted deer in the woods ~ Ida hunted for deer in the woods*), the possessor and attribute alternation (*I admired him for his honesty ~ I admired the honesty in him*). Finally, Levin notes that the class of English verbs including *grow* participate in the intransitive material/product alternation where either of the participants can be specified as a subject (SUBJ) or an oblique (OBL_θ): *That acorn will grow into an oak tree ~ An oak tree will grow from that acorn.*

Although variants of the constructions involve the same predicates, participants, and even the same grammatical functions, there are some semantic differences associated with the variants (e.g. a holistic vs partitive effect of the locative alternation). However, crucially, neither is more basic than the other, or neither is derived from the other – in this respect, they have equal status. A simple way of capturing the fact that the same predicate may have two (or more) options of matching its participants with grammatical functions is to dissociate the tier of semantic participants from the tier of argument positions.

(ii) The distinction between semantic participants and argument positions is already implicit in standard LFG accounts of ‘empty’ (athematic) argument roles of raising verbs (Zaenen & Engdahl 1994:200, 203; Bresnan 2001:309, 317). The representations of a-structures of the subject-raising verb *seem* (as in *He seemed to me to be happy*) and the object-raising verb *believe* (as in *I believe him to be happy*) contain athematic arguments which are expressed as gaps in the argument list, because *He* is not a semantic subject of *seem*, and *him* is not a semantic object of *believe*. The following diagrams are from Bresnan (2001:309, 317; but see section 4 below for the alternative):

$$\begin{array}{lllll} (6) & \textit{seem} & \_\_ & \langle\ x & y\ \rangle \\ & & [-\mathrm{r}] & [-\mathrm{o}] & [-\mathrm{o}] \\ & & \mathrm{SUBJ} & \mathrm{OBL}_\theta & \mathrm{XCOMP} \end{array}$$

$$\begin{array}{lllll} (7) & \textit{believe} & \langle\ x & y\ \rangle & \_\_ \\ & & [-\mathrm{o}] & [-\mathrm{o}] & [-\mathrm{r}] \\ & & \mathrm{SUBJ} & \mathrm{XCOMP} & \mathrm{OBJ} \end{array}$$

Similarly, the non-raising version of *seem* has an athematic subject which in English has to be filled by an expletive (*It seemed to me that John was happy*):

$$\begin{array}{lllll} (8) & \textit{seem} & \_\_ & \langle\ x & y\ \rangle \\ & & [-\mathrm{r}] & [-\mathrm{o}] & [-\mathrm{o}] \\ & & \mathrm{SUBJ} & \mathrm{OBL}_\theta & \mathrm{COMP} \end{array}$$

The athematic arguments are represented outside the angled brackets, which indicates that they do not belong to the set of semantic participants of the event denoted by the predicate. Nevertheless, they do have a specific position in the argument structure relative to the other hierarchically ordered semantic participants, which gives them greater or lesser priority in the mapping of grammatical functions (Bresnan 2001:309). Having no semantic content, they receive the inherent syntactic classification of [– r]. Thus, athematic arguments imply the existence of a distinct level of argument positions separate from the semantic level, and the representations in (6)-(8) can be translated to the following notation:

(9)			*x*	*y*	
			|	|	
seem	⟨	arg	arg	arg	⟩
		[– r]	[– o]	[– o]	

(10)		*x*	*y*					*x*		*y*	
		|	|					|		|	
believe	⟨	arg	arg	arg	⟩	or	⟨	arg	arg	arg	⟩
		[– o]	[– o]	[– r]				[– o]	[– r]	[– o]	

(iii) The distinction between semantic participants and argument positions enables a better analysis of essentially *syntactic* phenomena (‘morphosyntactic operations’) such as passivisation and locative inversion (see sections 3 & 5 below) which are available to unergative versus unaccusative predicates, respectively. The notion of an ‘underlying slot which comes first and is, or is not, a subject’ is not easily expressible in thematic terms; in fact, it has been demonstrated that it is impossible to find a common semantic denominator for either the class of syntactically unaccusative, or unergative verbs (e.g. Rosen 1984; Wechsler 1995).

(iv) There is general intuition that LMT should be capable of handling morpholexical causativisation, though there is not yet a solution that is widely accepted and has been proven to be applicable to the full variety of causatives cross-linguistically. However, two of the most widely published LFG analyses of causatives, Falk (2001:114-119), who provides a brief account building on the work of Alsina (1996), and Ackerman & Moore (2001), who build on the work on T. Mohanan (1994), both appeal to the distinction between semantic participants and argument positions. Ackerman & Moore in particular argue that a model of argument structure which has an independent valency level predicts that there can be predicate formation processes which introduce semantic properties, but which may not lead to an increase in valency (2001:46). They examine several different instances of causativisation which do not involve an increase in valency and conclude that these data provide empirical motivation for the theoretical assumption that valency slots (i.e. argument positions) constitute an independent level of representation which is used to mediate the relation between semantic roles (understood by them as sets of semantic entailments of the predicate) and grammatical function assignment (2001:48ff).

2 Revision 2: Allowing semantic participants to change order and re-associate with different argument positions for non-default mappings

The most widely used versions of LMT have a fixed hierarchy of thematic roles which determines the ordering of argument positions. The following thematic hierarchies are from Bresnan (2001:307) and Falk (2001:104), respectively:

(11) agent > beneficiary > experiencer/goal > instrument > patient/theme
> location

(12) agent > patient/beneficiary > instrument > theme
> path/location/reference object

Most LFG researchers derive the content of their thematic hierarchy either from the Conceptual Semantics framework of Jackendoff (1983; 1990), or from the Proto-Role proposal of Dowty (1991) (see also Butt 2006:Ch.5 for discussion).

However, Levin & Rappaport Hovav (2005:Ch.6) show that it is impossible to formulate a thematic hierarchy that will capture all generalisations involving the realisation of arguments in terms of their semantic roles; Newmeyer (2002) cites 18 *distinct* thematic hierarchies on offer, none of which comes close to working. Ackerman & Moore (2001:27) cite Gawron (1983) as a good critique of the shortcomings associated with delimiting classes of verbs and identifying finite lists of discrete semantic roles. To overcome these shortcomings they assume, following Dowty, that an argument of a predicate is a set of the predicate entailments that is specific to a participant in the event denoted by the predicate; they propose that sets of proto-properties can be ordered from most proto-agentive to most proto-patientive, and they formulate a well-formedness condition on the linking of entailment sets to valency slots (2001:44-45).[2]

As a result of the shift of perspective on semantic participants – from classifying them into discrete roles to seeing them as sets of semantic entailments of the predicate – it is expected that the same semantic participants may align with the available argument positions in two (or more) different ways, as was exemplified in the previous section by locative and other alternations. Furthermore, it is also expected that the semantic participants may ‘change order’ and re-associate with different argument positions for derived, morpho*semantically* altered, predicates.

The following example from Polish shows a morphosemantically derived predicate in which a sentient causer of the event (normally interpreted as the agent) is portrayed as ‘unwilful’, i.e. not responsible for the action:

(13) *Wylała* *mi* *się* *zupa.*
spilt.3SG.FEM me.DAT REFL soup(FEM).NOM
‘The soup has spilt to me.’ (meaning: ‘I have spilt the soup unintentionally.’)

The resulting construction is the common Slavonic anticausative, in which the patient/theme is lexicalised as the subject and the ‘unwilful’ agent is lexicalised as a dative argument (secondary object).

Thus, even though there may be a default ordering of semantic participants,

[2] Note, however, that the first suggestion of integrating Dowty’s Proto-Role proposal into LMT came from Zaenen (1993). For an overview and discussion of her approach, see Butt (2006:135-138).

evidently it can be altered, the alteration being driven by the change in the interpretation of the predicate together with its sets of entailments. The most straightforward way to model this with LMT (see next section for examples) is to allow the same semantic participants to 'realign' and link to different argument positions for different types of clauses which may or may not differ in valency.

3 Revision 3: Fixing the order of (syntactic) argument positions

After separating the semantic information from the syntactic level of argument positions I argue, following Zaenen (1993:151) and Ackerman & Moore (2001:44ff), that priority should be given to the *syntactic* representation of the predicate's valency rather than the *semantic* representation of thematic roles with which argument positions are linked. Therefore, it is the ordering of argument positions holding at the valency level of argument structure that is fixed, with each position ('argument slot') coming with a particular (fixed) syntactic specification. The following represents the valency template available to a base (non-derived) predicate:

(14) < arg_1 arg_2 arg_3 arg_4 ... arg_n>
[–o/–r] [–r] [+o] [–o] [–o]

In the case when all the slots are used (i.e. none are bypassed), the argument in the first slot can be classified as either [–o] or [–r]; the argument in the second slot can only receive [–r] classification; and so on. This ordering corresponds to LFG's hierarchy of syntactic functions (proposed after Keenan & Comrie 1977), but it is based on LMT's atomic values [+/– r/o] instead of *final* grammatical functions. Since valency slots correspond to particular types of predicate entailments,[3] if the base predicate does not have a particular set of entailments, the slot corresponding to that set of entailments is not invoked. Thus, for a particular predicate, the angled brackets contain all and only the selected valency slots for the arguments associated with that predicate, both core and non-core.

As was outlined in section 2, semantic participants may be understood as having a certain default ordering, but their actual ordering is flexible, not fixed. This means that under certain conditions, the actual semantic participants of the event may map onto the argument positions listed in (14) in more than one way. For example, some semantic participants may compete for a certain argument slot (as in locative etc. alternations), or a semantic participant may map onto an unused (but syntactically pre-specified) argument slot (as in the linking of the unwilful agent to a dative in Slavonic, or in 'dative shift' in English – see below). In

[3] Note that many Dowtyian approaches, including Ackerman & Moore's (2001), adopt two proto-property sets: proto-agent and proto-patient. However, other researchers have suggested adding a third set: proto-recipient (see Primus 1999). For base predicates, the entailments set of the third argument slot proposed here (arg_3) corresponds to this proto-property set. I will refer to it as proto-beneficiary, since the term 'beneficiary' has been more common in thematic hierarchies. It has been noted that proto-beneficiary needs to be distinguished only for some, but not all, languages.

derived predicates, such re-alignments of participants result from a meaning-altering (i.e. morphosemantic) operation on the predicate's argument structure. The interpretation of the roles of the participants is altered due to the fact that, in the end, the participants bear a different grammatical function to the one they would be getting 'by default'.

Using different types of predicates I will now illustrate that revisions 1-3 do, on the whole, produce the same syntactic pre-specification as LFG's basic principles for determining the choice of syntactic features (patientlike roles are [− r], secondary patientlike roles are [+ o], and other semantic roles are [− o]) though, importantly, not in predicates with non-applied beneficiaries, which receive a much simpler analysis in the reformulated LMT than in standard LMT accounts.

First, I will deal with the anticausative exemplified in (13). The argument structure of the basic, non-derived causative variant in (15) is modelled in (16):

(15) *Wylałem* *zupę.*
spilt.1SG.MASC soup(FEM).ACC
'I have spilt the soup.'

(16)
$$\begin{array}{lcc} & x & y \\ & | & | \\ \textit{wylałem}\ \langle & \mathrm{arg}_1 & \mathrm{arg}_2\ \rangle \\ & [-\,\mathrm{o}] & [-\,\mathrm{r}] \end{array}$$

The formation of the anticausative predicate results from an operation which deletes the first argument from the argument structure frame of the base predicate, leaving behind an orphaned semantic role (*x*) (see Levin & Rappaport Hovav 1995 for a corresponding analysis of externally caused intransitive verbs in English which participate in the causative alternation). Since the predicate loses an argument, and hence its valency decreases, the operation may be referred to as 'lexical detransitivisation'.

The second argument, which can now bear the function of the subject, is interpreted as a 'pseudo-agent', but the event is still understood as requiring an external cause(r). Polish (unlike English) has a strategy of reintroducing the orphaned causer, interpreted as an unwilful agent, to syntax via the argument position of the secondary object (the dative) (for detailed discussion of the anticausative see Kibort 2004:Ch.3). I repeat example (13) as (17) and model it in (18):

(17) *Wylała* *mi* *się* *zupa.*
spilt.3SG.FEM me.DAT REFL soup(FEM).NOM
'The soup has spilt to me.' (meaning: 'I have spilt the soup unintentionally.')

(18)
$$\begin{array}{lcc} & y & \boldsymbol{x} \\ & | & | \\ \textit{wylała się}\ \ \langle & \mathrm{arg}_2 & \mathrm{arg}_3\ \rangle^{4} \\ & [-\,\mathrm{r}] & [+\,\mathrm{o}] \end{array}$$

Second, I will outline the mappings in constructions with beneficiaries. In

[4] I have left gaps in the representations of argument frames only for an easier reading of the diagrams. The gaps have no theoretical significance. Instead, theoretical significance is attributed to the rank of the particular argument position.

many familiar languages, including Polish, dative case marking distinguishes the beneficiary/maleficiary argument from the patient/theme. In Polish, the dative argument differs from obliques in that it cannot be multiplied, though like obliques as well as primary objects (in appropriate contexts) it can be omitted. It can be optionally added to any Polish clause: almost any clause can be expanded to include an optional beneficiary referring to 'self', marked for dative, regardless of the number and type of other dependants of the predicate, and without altering the semantic or syntactic mappings in the predicate's argument structure. Once a semantic participant is selected for the dative in the base predicate, it is not possible to either promote this argument to subject (as in passivisation) or change its status to object (as in 'dative shift') through any argument-structure alteration in the predicate. The dative fits well the LMT's description of the secondary object specified for [+ o]. In the revised version of LMT offered here, it is identified with the unique third argument position (arg_3). In Polish, this argument position is available to predicates both for non-derived mappings (of optional beneficiaries), as in:

(19) *Piotr* *dał* *monetę* *Jankowi.*
Peter(MASC).NOM gave.3SG.MASC coin(FEM).ACC John(MASC).DAT
'Peter gave a/the coin to John.'

(20)
$$\begin{array}{lccccl} & & x & y & b & \\ & & | & | & | & \\ \textit{dał} & \langle & \text{arg}_1 & \text{arg}_2 & \text{arg}_3 & \rangle \\ & & [-\text{o}] & [-\text{r}] & [+\text{o}] & \end{array}$$

and for morphosemantically altered mappings (e.g. of unwilful agents, as in (18)).

English ditransitives, which have been the subject of considerable debate in LFG, receive a much simpler account in the revised LMT. Modern English does not mark its beneficiaries for dative. Instead, an English beneficiary is expressed either adpositionally (headed by a preposition), like an oblique:

(19) a. *Peter handed a drink to John.*
b. *Both parents cooked supper for the children.*

(20) 'non-dative-shifted'
$$\begin{array}{lccccl} & & x & y & b & \\ & & | & | & | & \\ \textit{handed/cooked} & \langle & \text{arg}_1 & \text{arg}_2 & \text{arg}_4 & \rangle \\ & & [-\text{o}] & [-\text{r}] & [-\text{o}] & \end{array}$$

or in a syntactic argument which is not headed by a preposition, which occupies the surface position of the direct object and behaves like a direct object with respect to passivisation:

(21) a. *Peter handed John a drink.*
b. *Both parents cooked the children supper.* (Bresnan 2001:315-316)

(22) 'dative-shifted'
$$\begin{array}{lccccl} & & x & \boldsymbol{b} & \boldsymbol{y} & \\ & & | & | & | & \\ \textit{handed-to/cooked-for} & \langle & \text{arg}_1 & \text{arg}_2 & \text{arg}_3 & \rangle \\ & & [-\text{o}] & [-\text{r}] & [+\text{o}] & \end{array}$$

In the non-dative-shifted predicate, as in (20), the third argument position (arg_3) is not invoked in English. English has lost the morphological means to distinguish this argument from the primary object and hence base predicates treat beneficiaries as obliques. Note the lack of syntactically intransitive English clauses comprising only subjects and datives but no direct objects:[5] **Both parents cooked the children* meaning: 'Both parents cooked for the children'; and the ungrammaticality of the attempted dative in: *He sold three cars* (**John*), *He gave the book* (**me/him*). However, through dative shift, verbs of a certain class in English are capable of recovering their dative argument position: dative shift (or, dative alternation) in English is a morphosemantic operation on argument structure which alters the mapping of the semantic participants of the predicate onto argument positions by remapping the beneficiary onto the primary object position, and 'downgrading' the theme to the secondary object position.

The analysis sketched out above accounts for the passivisability patterns of the non-dative-shifted and dative-shifted predicates in English, and avoids invoking an additional constraint, the Asymmetric Object Parameter (which rules out argument structures with two unrestricted [– r] arguments for some languages), which was proposed specifically to handle languages with dative shift.[6] Further examples showing the redundancy of the AOP will be given below.

Finally, earlier in this section I outlined the mapping that occurs in the anticausative, see examples (17)-(18). This construction results from a type of morphosemantic operation, lexical detransitivisation, that targets directly the level of argument positions. I suggested that the anticausative operation deletes the first (core) argument from the valency frame of the base predicate:

(23) *I spilt the soup.*

$$\begin{array}{lccccc} & & x & y & \\ & & | & | & \\ \textit{spilt}_{\text{trans}} & \langle & \text{arg}_1 & \text{arg}_2 & \rangle \\ & & [-\text{o}] & [-\text{r}] & \end{array}$$

(24) *The soup spilt.*

$$\begin{array}{lccccc} & & x & y & \\ & & & | & \\ \textit{spilt}_{\text{intrans}} & \langle & & \text{arg}_2 & \rangle \\ & & & [-\text{r}] & \end{array}$$

(Recall also that the anticausative does not delete the semantic participant – typically, the event denoted by the verb does not cease to require an external causer. I demonstrated that some languages with anticausatives have a way of optionally retrieving the causer to project it to syntax through a different argument position.)

It is expected that an operation with the opposite effect to lexical

[5] A possible exception are clauses with the verb *give* which, for some speakers, has retained a fossilised structural dative (arg_3) position, as in (20), even in the base variant. Hence: ?*Peter gave John*, ?*A book was given John (by Peter).* See Kibort (2004:79-88) for examples, discussion and references.

[6] The Asymmetric Object Parameter is undesirable for one more reason: in the revised LMT, transitive unaccusatives (the class of verbs including *cost, last,* and *weigh*) are those predicates whose both arguments (arg_1 and arg_2) are pre-specified as [– r], hence their unavailability for passivisation (see section 5). The fact that the Parameter does not need to be invoked to account for dative shift leaves no reason to keep it. This, in turn, frees the revised LMT from a theory-internal solution.

detransitivisation can also be found. This is ‘lexical transitivisation’, which targets the same level of representation of the predicate as the anticausative, the level of argument positions, and adds to it a core argument. Dative shift, discussed above, is an example of a lexical transitiviser. Like the detransitivising anticausative in English, it is not expressed with any special verbal morphology. However, we acknowledge that the predicate and its set of entailments have been altered by indicating that the base verb’s lexical meaning has changed, e.g. from *handed* and *cooked*, to *handed-to* and *cooked-for*. The following non-dative-shifted examples are repeated from (19)-(20):

(25) a. *Peter handed a drink to John.*
b. *Both parents cooked supper for the children.*

(26)
$$\begin{array}{llll} & x & y & b \\ & | & | & | \\ \textit{handed/cooked} & \langle\ \text{arg}_1 & \text{arg}_2 & \text{arg}_4\ \rangle \\ & [-\text{ o}] & [-\text{ r}] & [-\text{ o}] \end{array}$$

and the following dative-shifted examples are repeated from (21)-(22):

(27) a. *Peter handed John a drink.*
b. *Both parents cooked the children supper.*

(28)
$$\begin{array}{llll} & x & \boldsymbol{b} & \boldsymbol{y} \\ & | & | & | \\ \textit{handed-to/cooked-for} & \langle\ \text{arg}_1 & \text{arg}_2 & \text{arg}_3\ \rangle \\ & [-\text{ o}] & [-\text{ r}] & [+\text{ o}] \end{array}$$

Dative shift increases the transitivity of the base mono-transitive predicate (*handed*, *cooked*) by adding an ‘objective’ [+o] argument to its valency frame. The arguments are ordered according to LMT’s atomic values [+/– r/o], and the new argument slot occupies a position that conforms to this ranking. The semantic participants map onto the new set of argument positions in a way that matches the sets of semantic entailments produced by the derived predicate (*handed-to*, *cooked-for*).

English does not express the addition of a new core argument with verbal morphology, and also has a different option of expressing the beneficiary: mapping it onto an oblique argument. However, many languages do not have the option of expressing the beneficiary as an oblique argument, and their strategy to bring beneficiaries and other peripheral participants into the verb’s lexical meaning is the transitivising applicative, a construction which is typically marked by special verbal morphology.

In the standard LMT account, the transitivising applicative is analysed as adding a new theta role to the theta structure of a verb, below the highest role (Alsina & Mchombo 1988, 1990, 1993; see also Bresnan & Moshi 1993). Bresnan & Moshi explain that ‘[t]his change in the argument structure is induced by an underlying change in the lexical semantic structure. (...) Informally, the action of the base verb *v* is applied to a new argument *x*, yielding a derived meaning paraphrasable as “do *v* for/to/with/at *x*”’ (1993:73, ft. 30).

In the revised LMT, the transitivising applicative is analysed as targeting the same level of representation of the predicate as the anticausative, the level of argument positions, and adding an argument pre-specified as [+ o] to the valency frame of the base predicate. In this respect, it is like dative shift (to which this construction has been likened in the literature), except that it is accompanied by dedicated verbal morphology. Applicative formation increases the transitivity of the base verb, and allows the semantic participants to map onto the new set of argument positions in a way that matches the entailment sets produced by the derived predicate (e.g. 'eat-for' when a beneficiary is added; 'eat-with' when an instrument is added; or 'eat-because-of' when a motive is added; etc.).

The 'applied' participant is typically mapped onto the second argument position of the primary object (which enables it to become a passive subject). However, for many predicates, the entailment sets corresponding to the two object positions ([– r] and [+ o]) allow the peripheral participant and the patient/theme to re-align and map in either way. Whichever participant maps onto the primary object position ([– r]) may become a passive subject. As can be expected, the argument in the primary object position ([– r]) is also privileged over the argument in the secondary object position ([+ o]) with respect to adjacency to the verb and availability for long-distance extraction (Bresnan & Moshi 1993:59-61).

Passivisation patterns in Kichaga (as described in Bresnan & Moshi 1993) reveal that several different mapping options are available for the base predicate which has been subjected to applicative transitivisation and has two participants competing for the primary object position. For illustration, I have schematised some options below, using thematic labels for the participants only for easier reading:

(29)		*agent*	*benef*	*pat/theme*	
		agent	*instr*	*pat/theme*	
		agent	*loc*	*pat/theme*	
		agent	*motive*	*pat/theme*	
		agent	*pat/theme*	*benef*	
		agent	*pat/theme*	*instr*	
		agent	*pat/theme*	*loc*	
		|	|	|	
	⟨	arg$_1$	arg$_2$	arg$_3$	⟩
		[– o]	[– r]	[+ o]	

Although the primary object argument is privileged (can become a passive subject, is adjacent to the verb and available for long-distance extraction), Kichaga treats both objects in the same way with respect to object marking on the verb.

Languages like Kichaga are referred to as 'symmetric'. In standard LMT accounts this is understood with reference to the Asymmetric Object Parameter. It is argued that the AOP, which regulates the occurrence of argument structures with two unrestricted [– r] arguments, is present in asymmetric languages like English and Chichewâ, but lacking in symmetric languages like Kichaga (Alsina & Mchombo 1988; Bresnan & Moshi 1993). In the revised LMT, there is no need to invoke the AOP, and symmetric languages can be defined as those which allow both their ‘applied’ participant and their patient/theme to be mapped onto either of the object argument positions ([– r] or [+ o]).[7]

The other type, ‘asymmetric’ languages, impose restrictions, or limitations on their secondary object position ([+ o]). In those languages, the [+ o] argument slot is not suitable for the mapping of the beneficiary participant (whether a ‘dative-shifted’ beneficiary as in English, or an ‘applied’ beneficiary as in Chichewâ); only the primary ([– r]) object is treated as an object with respect to object marking on the verb; and the secondary ([+ o]) object cannot be ‘dropped’ (left unspecified) in the transitivised predicate (this applies regardless of whether the predicate has undergone dative shift or applicative transitivisation).

Preventing the secondary object position from accepting beneficiaries results in fewer mapping options in asymmetric languages such as Chichewâ:

(30)	*agent*	*benef*	*pat/theme*	
	agent	*instr*	*pat/theme*	
	agent	*loc*	*pat/theme*	
	~~*agent*~~	~~*pat/theme*~~	~~*benef*~~	
	agent	*pat/theme*	*instr*	
	agent	*pat/theme*	*loc*	
	|	|	|	
⟨	arg$_1$	arg$_2$	arg$_3$	⟩
	[– o]	[– r]	[+ o]	

Thus, both the passivisation patterns in asymmetric languages, as well as different treatment of the two types of objects (primary and secondary) in asymmetric languages, can be accounted for by the revised LMT without having to invoke an additional parameter such as the AOP.

It has also been noted that, in some languages, the transitivising applicative can add more than one core argument – specifically, it has been found to add up to two core arguments, both in symmetric and asymmetric languages (Bresnan & Moshi 1993:52). In the revised LMT, the second applied argument position will also be pre-specified as [+ o], and the grammatical function mapped onto this argument will be OBJ$_\theta$. The two secondary objects will be distinguished by their subscripts.

4 Revision 4: Reformulating the principles of argument-to-function mapping to make fuller use of the markedness hierarchy of syntactic functions and render the Subject Condition redundant

[7] Note that, according to the standard LMT account, in symmetric languages a predicate has its third argument pre-specified as [+ o] for some clauses (e.g. an unaltered active) and as [– r] for other clauses (e.g. passive). Although LMT allows to interpret both pre-specifications as being appropriate for ‘patient-like’ arguments, the selection of either one or the other pre-specification for the same argument in the same predicate requires a non-monotonic change of information.

The features [+/– r], (thematically/semantically) (un)restricted, and [+/– o], (non)objective, group grammatical functions into natural classes (Bresnan & Kanerva 1989; Bresnan & Zaenen 1990; see also Bresnan 2001:308):

(31)

	[– r]	[+ r]
[– o]	SUBJ	OBL$_\theta$
[+ o]	OBJ	OBJ$_\theta$

where OBL$_\theta$ abbreviates multiple oblique functions, and OBJ$_\theta$ abbreviates secondary objects. Since the negatively specified features in diagram (31) indicate unmarked feature values, SUBJ is the least marked grammatical function and the restricted object (OBJ$_\theta$) is the most marked function, and the diagram can be read as a 'markedness hierarchy of syntactic functions' (Bresnan & Moshi 1993:71) or a 'partial ordering of basic argument functions' (Bresnan 2001:309):

(32) Markedness Hierarchy of Syntactic Functions
[–o]/[–r] SUBJ > [–r]/[+o] OBJ, [–o]/[+r] OBL$_\theta$ > [+o]/[+r] OBJ$_\theta$

In one of the most widely accepted versions of LMT, the Markedness Hierarchy feeds into the syntactic principles for mapping argument structures to surface grammatical functions, i.e. the Mapping Principles (Bresnan 1990; Bresnan 2001:311):

(33) (a) Subject roles:
(i) a [–o] argument is mapped onto SUBJ when initial in the argument structure;[8] otherwise:
(ii) a [–r] argument is mapped onto SUBJ.
(b) Other roles are mapped onto the lowest (i.e. most marked) compatible function on the markedness hierarchy.

However, the Mapping Principles in (33) do not make full use of the Markedness Hierarchy, even though it is possible to derive the principles of argument-to-function mapping directly from the Markedness Hierarchy, without

[8] The actual LFG formulation of this mapping principle is as follows: '$\underset{[-o]}{\hat{\theta}}$ is mapped onto SUBJ when initial in the a-structure' (Bresnan 2001:311), where $\underset{[-o]}{\hat{\theta}}$, referred to as the 'logical subject', is defined as 'the most prominent semantic role of a predicator' (p. 307). However, this formulation seems to contain superfluous information. Specifically, due to the Subject Condition, LFG excludes the formation of predicates with no core arguments; according to the principles of semantic classification of thematic roles for function, LFG allows only those thematic roles which will map onto 'subjective' (core) or oblique (non-core) functions to be classified as [–o]; and finally, due to the thematic hierarchy (and the Subject Condition), thematic roles which will map onto oblique functions can never be initial in the argument structure or higher than the 'subjective' role. It follows from this that a [–o] argument which is *initial* in the argument structure (i.e. has position adjacent to the left bracket; see also Falk 2001:108) can *only* be the most prominent thematic role, and it can never be an oblique participant. Thus, the formulation of the subject mapping principle in (33a)(i) is in fact just a more concise, but still faithful, version of the LFG principle.

building in the condition that the first *encountered* argument needs to be pre-specified as either [–o] or [–r]. Hence, I propose a reformulation of the Mapping Principles into the following, single Mapping Principle:

(34) MAPPING PRINCIPLE
The ordered arguments are mapped onto the highest (i.e. *least* marked) compatible function on the markedness hierarchy.

The reformulated Mapping Principle achieves correct mappings for various classes of predicates discussed in the literature (including unaccusatives and ditransitives – see below for examples), but avoids stipulating any specific principles where their result is already partially determined by the markedness hierarchy. In this way, it avoids redundancy both in the account of the mapping itself, as well as in the formulation of any conditions or constraints pertaining to the subject.

Thus, the Subject Condition ('Every predicator must have a subject'; e.g. Bresnan 2001:311) is now redundant, since the provision of the subject for any personal clause is ensured by the more general Mapping Principle. Note that the Subject Condition had been assumed incorrectly even when it was allowed to be parametrised, since it would rule out inherently impersonal predicates in a language which otherwise has to be analysed as having the parameter present. Without the Subject Condition, it is now possible to account for inherently impersonal predicates which have no subject at any level of analysis (a-structure, f-structure, or c-structure) (see Kibort 2006 for examples and discussion).

I will now give a few concise examples illustrating the correct mappings achieved by the revised Mapping Principle in (34):

(a) with unergative[9] transitive verbs such as *clean* in *I clean the floor*, the Mapping Principle ensures that the first ([–o]) argument is linked to SUBJ and the second ([–r]) argument is linked to OBJ;

(b) with unaccusative intransitive verbs such as *come* in *I come*, the Mapping Principle ensures that the first ([–r]) argument is linked to SUBJ because this is the grammatical function which is the highest compatible one on the markedness hierarchy in (32);

(c) with unaccusative transitive verbs such as *cost* in *The book cost £10*, the Mapping Principle ensures that the first ([–r]) argument is linked to SUBJ (the highest grammatical function compatible with this specification) and the second ([–r]) argument is linked to OBJ;

(d) the non-raising version of the verb *seem*, as in *It seems to me that John was happy*, selects three argument positions: the (athematic) subject position and two non-core argument positions for the expression of the experiencer and the proposition; the positions are pre-specified as [–r], [–o], and [–o], respectively; the Mapping Principle ensures that the first (athematic) ([–r]) argument is linked to SUBJ, the second (experiencer) argument is linked to OBL$_{\theta}$, and the third

[9] Following the RG tradition, I treat unergativity/unaccusativity as a primarily syntactic phenomenon, and as irrespective of transitivity (hence it is orthogonal to valency). For some discussion, see Kibort (2004:71-75, 357).

(propositional) argument is also linked to a type of oblique function, COMP;[10]

(e) the raising version of the verb *seem*, as in *He seemed to me to be happy*, selects three argument positions: the subject position and two non-core argument positions for the expression of the experiencer and the infinitival complement; the positions are pre-specified as [–r], [–o], and [–o], respectively (note that *seem* cannot have an unergative, i.e. passivisable subject argument in either version, non-raising or raising); the Mapping Principle ensures that the first ([–r]) argument is linked to SUBJ, the second (experiencer) argument is linked to OBL_θ, and the third (infinitival complement) argument is linked to a type of oblique function, XCOMP;

(f) the raising version of the verb *believe*, as in *I believe him to be happy*, selects three argument positions: the subject position, the primary object position, and a non-core position for the expression of the infinitival complement; the positions are pre-specified as [–o], [–r], and [–o], respectively; the Mapping Principle ensures that the first ([–o]) argument is linked to SUBJ, the second argument is linked to OBJ, and the third (infinitival complement) argument is linked to a type of oblique function, XCOMP;[11]

(g) non-derived predicates with a proto-beneficiary participant, as in (19)-(20), derived dative-shifted predicates, as in (21)-(22), and derived predicates with an 'applied' non-core participant, as in (29) and (30), receive straightforward argument-to-function mapping by the Mapping Principle: their first argument (whether [–o] or [–r]) is linked to SUBJ, their second ([–r]) argument is linked to OBJ, and their third ([+o]) argument is linked to OBJ_θ.

5 Revision 5: Using a mechanism of increasing markedness to account for morpho*syntactic* operations

Finally, using only a mechanism of increasing markedness, and retaining the principle of monotonicity, I will demonstrate how LMT can elegantly account for morpho*syntactic* phenomena that are referred to as 'demotions' in RG.

Morphosyntactic operations interfere with the 'default' argument-to-function mapping, but do not affect the lexical or semantic tiers of representation of the predicate (i.e. both the argument positions and the alignment of the participants with the argument positions remain unaffected). Hence, morphosyntactic operations do not affect the interpretation of the predicate together with its sets of semantic entailments, or the interpretation of the roles of the semantic participants. They affect only the final mapping of grammatical functions to arguments. Logically, this can be done only in one way: since the Markedness Hierarchy orders grammatical functions from the least restricted to the most restricted, and the Mapping Principle matches the ordered arguments with functions beginning from the least marked functions (i.e. the highest ones in the Markedness Hierarchy), the only way to disrupt this default mapping is by restricting the

[10] Zaenen & Engdahl (1994) analyse COMP and XCOMP as specialised type of OBL_θ.

[11] See Falk (2001:140), example (54), for a corresponding analysis.

unrestricted arguments before applying the Mapping Principle. I refer to this as a ‘mechanism of increasing markedness’: a morphosyntactic operation can only restrict an argument by adding a ‘marked’ specification ([+r] or [+o]) to its syntactic pre-specification. The principle of monotonicity ensures that the restriction of [+r] cannot be added to a [–r] argument, and the restriction of [+o] cannot be added to a [–o] argument, as these operations would involve a change of information in the argument structure.

Hence, the available morphosyntactic (i.e. restricting) operations are: adding the [+r] specification to a [–o] argument; adding the [+o] specification to a [–r] argument; and adding the [+r] specification to a [+o] argument. Each of these operations would not only change the mapping of the grammatical function onto the affected argument, but also, in consequence of that altered mapping, it may also affect the mapping of grammatical function(s) onto other argument(s) (if the predicate selects more than one argument).

In brief, the morphosyntactic operation which restricts the first, unergative, argument pre-specified as [–o] by adding to it the [+r] specification is passivisation. As a result of this restriction, the first argument receives an oblique grammatical function (OBL$_\theta$) (hence the RG term ‘demotion of subject to an oblique’), and the second (core) argument, if there is one, receives the SUBJ function by the Mapping Principle, as in (35). If there is no second core argument, the resulting construction is an impersonal passive, as in (36):[12]

(35)

$$\begin{array}{lccc} & x & y & \\ & | & | & \\ \textit{verb}_{\text{passive}}\ \langle & \text{arg}_1 & \text{arg}_2 & \rangle \\ & [-\text{o}] & [-\text{r}] & \\ & \mathbf{[+r]} & & \\ & \text{OBL}_\theta & \text{SUBJ} & \end{array}$$

(36)

$$\begin{array}{lcc} & x & \\ & | & \\ \textit{verb}_{\text{passive}}\ \langle & \text{arg}_1 & \rangle \\ & [-\text{o}] & \\ & \mathbf{[+r]} & \\ & \text{OBL}_\theta & \end{array}$$

The morphosyntactic operation which restricts the first, unaccusative, argument pre-specified as [–r] by adding to it the [+o] specification is locative inversion (Bresnan & Kanerva 1989). As a result of this restriction, the first argument receives the OBJ function (hence ‘demotion of subject to an object’). If the verb selects a non-core [–o] argument, by the Mapping Principle it will receive the SUBJ function, as in (37). If there is no [–o] argument, the resulting construction is inversion without the locative, as in (38):[13]

(37)

$$\begin{array}{lccc} & x & z & \\ & | & | & \\ \textit{verb}_{\text{loc.inv.}}\ \langle & \text{arg}_1 & \text{arg}_4 & \rangle \\ & [-\text{r}] & [-\text{o}] & \\ & \mathbf{[+o]} & & \\ & \text{OBJ} & \text{SUBJ} & \end{array}$$

(38)

$$\begin{array}{lcc} & x & \\ & | & \\ \textit{verb}_{\text{(loc).inv.}}\ \langle & \text{arg}_1 & \rangle \\ & [-\text{r}] & \\ & \mathbf{[+o]} & \\ & \text{OBJ} & \end{array}$$

[12] See Kibort (2001) and (2004) for detailed discussion of the passive, including arguments for the ‘demotional’, as opposed to ‘promotional’, analysis of the passive, and arguments against analysing the oblique agent as an adjunct (esp. 2004:360-363).

[13] For examples and discussion, see Kibort (2001) and (2004), esp. (2004:364-368).

The morphosyntactic operation which restricts the second, primary object argument pre-specified as [–r] by adding to it the [+o] specification can be called 'object preservation' (Kibort 2004:368-372). As a result of this restriction, in a situation where the second argument could receive the subject function by the Mapping Principle, it is prevented from doing so and is instead 'preserved' as an OBJ. This is observed, for example, in the common personal active with subject instrument that may not be conceptualised as an agent, as in the Polish equivalent of *The axe broke the slab*, represented in (39),[14] and in inherently impersonal predicates whose *only* argument is a 'primary patientlike' object, e.g. Polish *słychać ją* 'hear.[NON-PERSONAL] her.ACC' (Kibort 2006), represented in (40):

(39)		*x*	*y*	*z*		(40)		*y*	
			|	|				|	
verb$_{\text{obj.pres.}}$	⟨		arg$_2$	arg$_4$	⟩	*verb*$_{\text{NON-PERS}}$	⟨	arg$_2$	⟩
			[–r]	[–o]				[–r]	
			[+o]					**[+o]**	
			OBJ	SUBJ				OBJ	

Finally, the morphosyntactic operation which restricts the third, secondary object argument pre-specified as [+o] by adding to it the [+r] specification can be understood as 'secondary object preservation'. As a result of this restriction, in a situation where the second argument could receive the object function, it is prevented from doing so and is instead 'preserved' as an OBJ$_{\theta}$. This is occurs in the Polish anticausative, as in (17)-(18), where, after the removal of the first argument from the predicate's valency frame, the remaining core argument is mapped onto subject, but the retrieved causer participant (the 'unwilful agent') can only have the secondary object function, but not a primary object function in this construction.[15]

6 Conclusions

In the sections above I have outlined a revised Lexical Mapping Theory which has theoretical and practical (descriptive) advantages over the currently

[14] One of the semantic factors which determine the mapping of the instrument participant (i.e. a peripheral participant) onto a particular argument position is whether the entity behind the instrument participant can be conceptualised as the causer of the event. Intermediary instruments which may not be conceptualised as agents (unless they are personified), but which may be mapped onto active subjects, do not have to be re-mapped onto the first argument position to be assigned the function of the subject. I argue in Kibort (2004:127-129, 371) that this is the correct analysis for Polish.

[15] This could be due to the fact that in a non-derived, 'causative' predicate, there can always be a proto-beneficiary participant expressing the causer through a reflexive pronoun. More generally, while in Polish the two types of object preservation are obligatory in the constructions or predicates that I exemplified, there may be languages in which these operations occur as a result of optional choice, just like passivisation and locative inversion, with the two options (object preserved vs object non-preserved) having different discourse or other functions.

used, accepted versions of LMT. It combines the best insights about argument structure mappings from dispersed sources into a coherent model. I have demonstrated that it can handle a wide range of complex phenomena handled by the accepted LTM variants, as well as constructions that standard LMT does not or could not handle (e.g. morphosemantically altered predicates with participant-to-function mappings which do not conform to the preferred thematic hierarchy; the impersonal passive; the (locative) inversion without the locative argument; inherently impersonal predicates). The revised LMT enables an elegant account of dative shift and the transitivising applicative, without having to compromise the principle of monotonicity. It eschews some redundant or theory-internal solutions, and, as demonstrated by the precursors of revision 1, promises a fruitful approach to the analysis of causatives. The suggested theoretical revisions to LMT may, furthermore, enable it to apply more universally and account for participant-to-argument mappings in languages other than those whose relational clause structure can be uncontroversially described with the use of syntactic functions.

References

Ackerman, Farrell. 1991. Locative alternation vs. locative inversion. In: Halpern, Aaron L. (ed.) *The Proceedings of the Ninth West Coast Conference on Formal Linguistics 1990.* Stanford, CA: CSLI Publications. 1-13.

Ackerman, Farrell. 1992. Complex predicates and morpholexical relatedness: locative alternation in Hungarian. In: Sag, Ivan A. & Anna Szabolcsi (eds) *Lexical Matters.* Stanford, CA: CSLI Publications. 55-83.

Ackerman, Farrell & John Moore. 2001. *Proto-Properties and Grammatical Encoding. A Correspondence Theory of Argument Selection.* Stanford, CA: CSLI Publications.

Alsina, Alex. 1996. *The Role of Argument Structure in Grammar: Evidence from Romance.* Stanford, CA: CSLI Publications.

Alsina, Alex & Sam A. Mchombo. 1988. Lexical mapping in the Chichewâ applicative construction. Paper presented to the Summer Working Group on Argument Structure and Syntax, CSLI, Stanford University. Revised and expanded from a paper presented at the 19th Annual African Linguistics Conference, Boston University, April 14-17, 1988.

Alsina, Alex & Sam A. Mchombo. 1990. The syntax of applicatives in Chichewâ: problems for a theta theoretic asymmetry. *Natural Language and Linguistic Theory* 8:493-506.

Alsina, Alex & Sam A. Mchombo. 1993. Object asymmetries and the Chichewâ applicative construction. In: Mchombo, Sam A. (ed.) *Theoretical Aspects of Bantu Grammar.* Stanford, CA: CSLI. 17-45.

Bresnan, Joan. 1982. The passive in lexical theory. In: Bresnan, Joan (ed.) *The Mental Representation of Grammatical Relations.* Cambridge, MA: MIT Press. 3-86.

Bresnan, Joan & Lioba Moshi. 1993. Object asymmetries in comparative Bantu syntax. In: Mchombo, Sam A. (ed.) *Theoretical Aspects of Bantu Grammar.* Stanford, CA: CSLI. 47-91.

Bresnan, Joan & Jonni M. Kanerva. 1989. Locative inversion in Chichewâ: a case study of factorization in grammar. *Linguistic Inquiry* 20(1):1-50.

Bresnan, Joan & Annie Zaenen. 1990. Deep unaccusativity in LFG. In: Dziwirek, Katarzyna, Patrick Farrell & Errapel Mejías-Bikandi (eds) *Grammatical Relations. A Cross-Theoretical Perspective.* Stanford, CA: CSLI Publications. 45-57.

Butt, Miriam. 2006. *Theories of Case.* Cambridge University Press.

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

Dowty, David. 1991. Thematic Roles and Argument Selection. *Language* 67:547-619.

Falk, Yehuda. 2001. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax.* Stanford, CA: CSLI Publications.

Gawron, J. M. 1983. *Lexical Representations and the Semantics of Complementation.* PhD thesis, University of California, Berkeley.

Grimshaw, Jane. 1988. Adjuncts and argument structure. *Occasional Paper* 36. Center for Cognitive Science, MIT, Cambridge, MA.

Jackendoff, Ray. 1983. *Semantics and Cognition.* Cambridge, MA: MIT Press.

Jackendoff, Ray. 1990. *Semantic Structures.* Cambridge, MA: MIT Press.

Joshi, Smita. 1993. *Selection of Grammatical and Logical Functions in Marathi.* PhD thesis, Stanford University.

Kaplan, Ronald & Joan Bresnan. 1982. Lexical-Functional Grammar: a formal system for grammatical representation. In: Bresnan, Joan (ed.) *The Mental Representation of Grammatical Relations.* Cambridge, MA: MIT Press. 173-281.

Keenan, Edward & Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8(1):413-461.

Kibort, Anna. 2001. The Polish passive and impersonal in Lexical Mapping Theory. In: Butt, Miriam & Tracy Holloway King (eds) *Proceedings of the LFG01 Conference, University of Hong Kong, Hong Kong*. Stanford, CA: CSLI Publications. 163-183.

Kibort, Anna. 2004. *Passive and Passive-Like Constructions in English and Polish.* PhD thesis, University of Cambridge. Available on-line.

Kibort, Anna. 2006. On three different types of subjectlessness and how to model them in LFG. In: Butt, Miriam & Tracy Holloway King (eds) *Proceedings of the LFG06 Conference, University of Konstanz, Germany*. Stanford, CA: CSLI Publications. 289-309.

Levin, Beth. 1993. *English Verb Classes and Alternations.* Chicago, IL: University of Chicago Press.

Levin, Beth & Malka Rappaport Hovav. 2005. *Argument Realization.* Cambridge: Cambridge University Press.

Mohanan, Tara. 1990/1994. *Argument Structure in Hindi.* Stanford, CA: CSLI Publications.

Newmeyer, Frederick J. 2002. Optimality and functionality: a critique of functionally-based optimality-theoretic syntax. *Natural Language and Linguistic Theory* 20:43-80.

Primus, Beatrice. 1999. *Cases and Thematic Roles: Ergative, Accusative and Active.* Tübingen: Niemeyer.

Rosen, Carol. 1984. The interface between semantic roles and initial grammatical relations. In: Perlmutter, David & Carol Rosen (eds) *Studies in Relational Grammar 2*. Chicago, IL: University of Chicago Press. 38-77.

Wechsler, Stephen. 1995. *The Semantic Basis of Argument Structure.* Stanford, CA: CSLI Publications.

Zaenen, Annie. 1993. Unaccusativity in Dutch: integrating syntax and lexical semantics. In: Pustejovsky, J. (ed.) *Semantics and the Lexicon.* Dordrecht: Kluwer. 129-161.

Zaenen, Annie & Elisabet Engdahl. 1994. Descriptive and theoretical syntax in the lexicon. In: Atkins, B.T.S. & Antonio Zampolli (eds) *Computational Approaches to the Lexicon: Automating the Lexicon II.* Oxford: Oxford University Press. 181-212.

TOWARDS A MORE LEXICAL AND FUNCTIONAL TYPE-LOGICAL THEORY OF GRAMMAR

Miltiadis Kokkonidis
University of Oxford

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

Type-Logical Lexical Functional Grammar is a new, radically lexicalist, and formally parsimonious theory, in essence a re-incarnation of Lexical Functional Grammar (Kaplan and Bresnan, 1982) in a type-logical formal framework very similar in formal nature to that of Type-Logical Categorial Grammar (Morrill, 1994; Moortgat, 1997). It puts emphasis on having a simple logical foundation as its formal basis and no empirically unmotivated primitives, representations, and mappings between them. It differs from TLCG in basing syntactic analyses on functional rather than constituent structure, to both LFG and TLCG in that it rejects syntactic categories as primitives, and to LFG in that it rejects c-structure as a linguistically significant representation and in being radically lexicalist. The present paper presents TL-LFG, the sequence of developments that lead to it, and its key differences from LFG.

1 Introduction

Type-Logical Lexical Functional Grammar is a new radically lexicalist and formally parsimonious theory of grammar, deeply influenced by Lexical Functional Grammar (Kaplan and Bresnan, 1982), but similar in formal nature to Type-Logical Categorial Grammar (Morrill, 1994; Moortgat, 1997). Its very existence serves as a reminder that certain associations between theories and formal settings are not a necessary consequence of their respective nature. LFG is model-theoretical, but TL-LFG is not. Type-Logical Categorial Grammar, unfortunately, often goes by the name Type-Logical Grammar, but TL-LFG is a type-theoretical theory of grammar that does not have syntactic categories but grammatical functions as primitives.

TL-LFG is the outcome of a series of developments related mostly to LFG's Glue syntax-semantics interface theory (Dalrymple et al., 1993; Dalrymple, 1999, 2001). Developments in Glue have emphasised formal elegance and simplicity, and this line of development carries over to TL-LFG. TL-LFG, being in essence LFG encoded in Glue, inherits the formal simplicity and elegance of Glue. Contrasting its design with the design of LFG highlights the various redundancies and unnecessary layers in the latter.

The design of TL-LFG pushes forward the idea that a theory should only have primitives that are empirically motivated. What is immediately observable is the written or pronounced word sequence and the meaning it has. What lies in between

[†]I am thankful to Avery Andrews, Anders Soegaard, Ash Asudeh, Doug Arnold, and Ron Kaplan for commenting on ideas seen in this paper at LFG 2007 or on other occasions. Special thanks go to Mary Dalrymple, for her always encouraging and insightful comments and excellent advise. Special thanks also to the two editors, Miriam Butt and Tracy Holloway King, whose professionalism, efficiency, and kindness never seizes to amaze me. This work has been supported by AHRC grant 2005/118667.

is theory-internal and must be justified. TL-LFG, as presented here,[1] assumes the principles of the Montagovian programme for natural language semantics and the LFG functional structure primitives. These are its foundational stones and what anyone accepting the theory would have to also consider a solid basis for any further development. Given those two elements, syntactic trees and 'semantic' projections as separate levels of representation are considered redundant in TL-LFG.

LFG claims that it is a functional theory based on the fact that it has f-structure in addition to the tree structures shared by many other theories (c-structure); TL-LFG claims that it is more functional than TL-LFG because it only relies on f-structure representations. LFG claims that it is a lexical theory because it deals with certain phenomena in the lexicon rather than in terms of transformations; TL-LFG makes the claim that it is more lexical as it deals with the entire syntax in the lexicon, having no syntactic rules as a formal object as such. These claims do represent a real difference between the formalisms on some level, but at the same time the design of TL-LFG offers an opportunity for the consequences of these differences to be examined in a new light.

The beauty, from a formal perspective, of TL-LFG is that it is based on a simple logical formal framework. Given that Glue is but a small piece in the jigsaw puzzle that is the LFG formalism, it is interesting to see that its simplest version to date (Kokkonidis, 2006), appropriately used, can replace much of the formal machinery of LFG.

There are two ways in which this paper discusses how one arrives at TL-LFG. Section 2 explains how the recent developments on the Glue type-system lead to TL-LFG. Section 3 discusses the differences between TL-LFG and LFG, and how one can peel off layers of the LFG architecture to get to the core of the theory. Conclusions are drawn in Section 4.

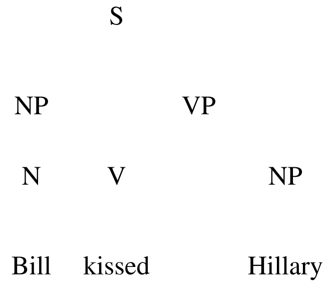
2 From Glue to TL-LFG

Lexical Functional Grammar was developed in the '70s by Ron Kaplan and Joan Bresnan (Dalrymple et al., 1995a). A remarkable fact about LFG is that in the three decades of use and development of the theory, its formal foundation has remained remarkably close to what Kaplan and Bresnan (1982) had proposed. While (sometimes significant) extensions and modifications to the theory have been proposed, the original architectural conception has by and large withstood the test of time. Moreover, the theory has been used by a diverse group of researchers that have found it particularly appealing for their line of work.

[1] In this paper the emphasis is on getting a basic LFG architecture in a type-logical setting. Argument structure, information structure, phonological structure, and morphological structure are not discussed, not because they are peripheral, nor because they are problematic, but because widening the scope of the discussion would not benefit making the basic points the paper intends to make. These are best made when a simpler LFG (closer to the original c-structure + f-structure proposal) is considered.

(1) Bill kissed Hillary.

(2)



$$f:\left[\begin{array}{ll} \textsc{pred} & p: \text{‘kiss}\langle\textsc{subj},\textsc{obj}\rangle\text{’} \\ \textsc{subj} & s: \left[\begin{array}{ll}\textsc{pred} & \text{‘bill’}\end{array}\right] \\ \textsc{obj} & o: \left[\begin{array}{ll}\textsc{pred} & \text{‘hillary’}\end{array}\right]\end{array}\right]$$

One area that had been problematic for some time for LFG was its syntax-semantics interface. A first difficulty lay in the fact that f-structure consists of unordered attribute-value pairs, whereas traditional compositional semantics was carried out on the nodes of trees with ordered children nodes. Although this issue had been addressed, one way or the other, early approaches did not deal in detail with the issues raised by interactions between scope and bound anaphora, or non-clausal quantification scopes arising from complex NPs and from intensional verbs with NP complements (Dalrymple, 1999). Glue was a particularly elegant proposal for the syntax-semantics interface in LFG. The developments outlined here brought enough encoding power to Glue to enable it to encode the information necessary for its purposes directly; TL-LFG is essentially Glue encoding f-structure information.

2.1 Early Glue (Dalrymple et al., 1993, 1995b)

Glue (Dalrymple, 1999, 2001) is a theory of the syntax-semantics interface based on linear logic (Girard, 1987). It was originally designed to solve the problem of f-structure-based compositional semantics for LFG. Developments in the following years were in various directions. One was the expansion of the fragment which the Glue syntax-semantics interface theory covers. Another was the expansion of the range of theories of grammar Glue was proposed as the syntax-semantics interface for: LTAG (Frank and van Genabith, 2001), HPSG (Asudeh and Crouch, 2002), CG (Asudeh and Crouch, 2001), and CFG (Asudeh and Crouch, 2001). A third direction was formal simplification. This was quite remarkable in light of the above two developments that one would assume would bring in additional requirements which in turn would necessitate enrichment after enrichment of the formalism with whatever added complexity such developments would come with.

As was the case with LFG, the foundational intuitions behind Glue have changed very little since its first appearance. In the case of LFG, changes and additions to the framework have not, overall, resulted in a simpler formal framework. Changes and additions in LFG, such as functional uncertainty, were motivated by the need to provide a means to enable the theory to deal with phenomena in a better way, so this statement is not meant as a criticism. But it is interesting to note that while

Glue analyses broadening its empirical coverage have been constantly appearing over the years, Glue followed a path constantly heading towards simplification.

The original Glue system (G0) of Dalrymple et al. (1993) was quickly superseded by the simpler system later introduced by Dalrymple et al. (1995b). The system presented there (G1) did away with G0-style rules and the use of the linear logic '!' ("of course!") modality.

The '!' modality was used in the G0 Glue system for specifying argument mapping principles such as the following:

(3)
$$\begin{array}{l} !(\forall f. \forall X. \forall Y. \\ \quad (((f \text{ SUBJ})_\sigma = X) \otimes ((f \text{ OBJ})_\sigma = Y)) \multimap \\ \quad (agent((f \text{ PRED})_\sigma, X) \otimes theme((f \text{ PRED})_\sigma, Y))) \end{array}$$

This would be used together with the semantic contributions from the three words in (1) to give its meaning.

(4)
['Bill'] :
$s_\sigma = bill,$

['kissed'] :
$\forall X. \forall Y. (agent(p_\sigma, X) \otimes theme(p_\sigma, Y) \multimap (f_\sigma = kiss(X, Y)),$

['Hillary'] :
$o_\sigma = hillary.$

While Dalrymple et al. (1993) use the '!' modality for their argument mapping principles, they also show, in a footnote, how this usage can be avoided. According to that analysis, adopted subsequently in G1, the lexical entries for the three words in (1) would simply make the following three semantic contributions in G0 with no need for argument mapping rules:

(5)
['Bill'] :
$s_\sigma = bill,$

['kissed'] :
$\forall X. \forall Y. (s_\sigma = X) \otimes (o_\sigma = Y) \multimap (f_\sigma = kiss(X, Y)),$

['Hillary'] :
$o_\sigma = hillary.$

Notice that given some f-structure f, in G0, its semantic projection f_σ is its meaning. This changes in G1. Compare (5) with (6). In G1, meanings are not assigned to semantic projections but associated with them through the '$\rightsquigarrow$' relation. Another difference was that as the mapping rules of Dalrymple et al. (1993) were abandoned by the time G1 was proposed, the '!' modality was not made part of the G1 logic. This simplified the Glue formalism considerably.

(6)

[‘Bill’] :
$s_\sigma \rightsquigarrow bill,$

[‘kissed’] :
$\forall X. \forall Y. (s_\sigma \rightsquigarrow X) \otimes (o_\sigma \rightsquigarrow Y) \multimap (f_\sigma \rightsquigarrow kiss(X, Y)),$

[‘Hillary’] :
$o_\sigma = hillary.$

In G0, semantic projections were meanings. In G1, ‘semantic projections’ were (not particularly interesting) feature structures. While in most cases, they have no internal structure and are simply empty, the ‘semantic projection’ of a noun phrases with a generalised quantifier would have a VAR and a RESTR attribute both having a feature structure that happens to be empty as their value. Given that two empty feature structures (functions from attributes to values) are always equal and two feature structures with only two empty feature structure valued features VAR and RESTR are also equal, there appears to have been a slight formal oversight in the move from G0 to G1 in this regard (Mary Dalrymple 2005, personal communication). One way to solve this problem would be to assume that for each f its semantic projection f_σ has an implicit copy of its PRED feature and that if f_σ has VAR and RESTR attributes their values are f-structures that also contain a copy of the PRED feature but also a feature VAR_OR_RESTR having the value ‘var’ and ‘rest’ respectively.[2]

The reason why G1 ‘semantic projections’ are feature structures is that Dalrymple et al. (1995b) wanted to be able to talk about a variable (entity) and a restrictor (truth value) associated with an f-structure, but did not want to introduce VAR and RESTR features in f-structure as they are semantic in nature, whereas f-structure is a syntactic structure.[3] In G0, there never was such a thing as a VAR or RESTR attribute, but there was no analysis for noun phrases containing determiners and common nouns either.[4]

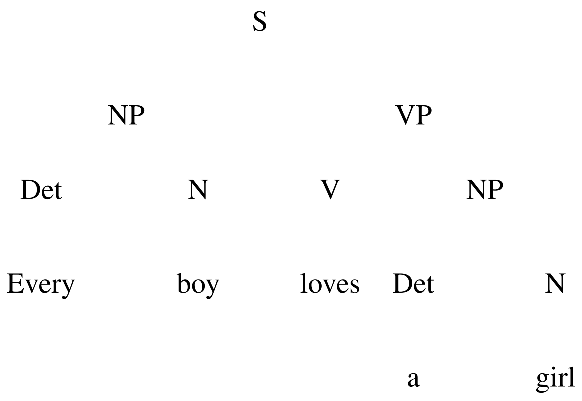
(7) Every boy loves a girl.

[2] A much simpler way of using these attributes but not G1/G2-style ‘semantic projections’ would be to have f-structures with a single VAR or RESTR feature outside their corresponding f-structure f with it as their value, instead of having VAR and RESTR as features of f. This solution was inspired by a combination of work on TL-LFG and one of the different solutions Kokkonidis (2007b) discusses for eliminating semantic projections.

[3] One could argue that the PRED features carrying an f-structure’s ‘semantic form’ also have a semantic flavour to them. Then the reason for not having VAR and RESTR attributes in the f-structure is that they are only needed by Glue; having ‘semantic projections’ as separate structures means that they only appear in Glue analyses and can be ignored by those working with other parts of LFG.

[4] Kokkonidis (2005, 2007b) demonstrates how such an analysis can be obtained without using those attributes; this analysis would have been expressible in G0 too, which would mean that the change from G0-style semantic projections to G1-style ‘semantic’ projections would not have been necessary.

(8)



$$\text{f:}\begin{bmatrix} \text{PRED} & \text{`love}\langle\text{SUBJ},\text{OBJ}\rangle\text{'} \\ \text{SUBJ} & \text{s:}\begin{bmatrix}\text{SPEC} & \text{`every'} \\ \text{PRED} & \text{`boy'}\end{bmatrix} \\ \text{OBJ} & \text{o:}\begin{bmatrix}\text{SPEC} & \text{`a'} \\ \text{PRED} & \text{`girl'}\end{bmatrix} \end{bmatrix}$$

$$f_\sigma : [\ \ ] \qquad s_\sigma : \begin{bmatrix} \text{VAR} & \\ \text{RESTR} & \end{bmatrix}[\ \ ]\Big] \qquad o_\sigma : \begin{bmatrix} \text{VAR} & \\ \text{RESTR} & \end{bmatrix}[\ \ ]\Big]$$

(9) ['every'] :
$\forall H.\forall R.\forall S.$
$(\forall X.((s_\sigma \text{VAR}) \rightsquigarrow_e X) \multimap ((s_\sigma \text{RESTR}) \rightsquigarrow_t R(X)))$
$\otimes$
$(\forall Y.((s_\sigma \rightsquigarrow_e Y)) \multimap (H \rightsquigarrow_t S(Y)))$
$\multimap$
$(H \rightsquigarrow_t \forall x. R(x) \rightarrow S(x)),$

['boy'] :
$\forall X.((s_\sigma \text{VAR}) \rightsquigarrow_e X) \multimap ((s_\sigma \text{RESTR}) \rightsquigarrow_t boy(X))$

['loves'] :
$\forall X.\forall Y.(s_\sigma \rightsquigarrow_e X) \otimes (o_\sigma \rightsquigarrow_e Y) \multimap (f_\sigma \rightsquigarrow_t love(X,Y)$

['a'] : $\forall H.\forall R.\forall S.$
$(\forall X.((o_\sigma \text{VAR}) \rightsquigarrow_e X) \multimap ((o_\sigma \text{RESTR}) \rightsquigarrow_t R(X)))$
$\otimes$
$(\forall Y.((o_\sigma \rightsquigarrow_e Y)) \multimap (H \rightsquigarrow_t S(Y)))$
$\multimap$
$(H \rightsquigarrow_t \exists y. R(y) \wedge S(y)),$

['girl'] :
$\forall Y.((o_\sigma \text{VAR}) \rightsquigarrow_e Y) \multimap ((o_\sigma \text{RESTR}) \rightsquigarrow_t girl(Y))$

2.2 G2: The First Type-logical Glue System (Dalrymple et al., 1997)

A further development was placing Glue on a type-logical setting with System F (Girard, 1989) as its basis (Dalrymple et al., 1997). The type-theoretic notation of this new system, G2, was neater, more concise and more readable than the notation of it predecessor, G1. Although, originally introduced in an effort to relate Glue to Categorial Grammar approaches, its popularity grew quickly to the point of replacing G1.

(10)

['every'] :
"$\lambda R.\, \lambda S.\, \forall x.\, S(x) \rightarrow R(x)$" : $\forall H.\, ((s_\sigma \text{VAR})_e \multimap (s_\sigma \text{RESTR})_t) \otimes (s_{\sigma e} \multimap H) \multimap H$

['boy'] :
"$\lambda x.\, boy(x)$" : $(s_\sigma \text{VAR})_e \multimap (s_\sigma \text{RESTR})_t$

['loves'] :
"$\lambda(x, y).\, loves(x, y)$" : $s_{\sigma e} \otimes o_{\sigma e} \multimap f_{\sigma t}$

['a'] :
"$\lambda R.\, \lambda S.\, \exists y.\, S(y) \wedge R(y)$" : $\forall H.\, ((o_\sigma \text{VAR})_e \multimap (o_\sigma \text{RESTR})_t) \otimes (o_{\sigma e} \multimap H) \multimap H$

['girl'] :
"$\lambda y.\, girl(y)$" : $(o_\sigma \text{VAR})_e \multimap (o_\sigma \text{RESTR})_t$

At the core of the type-logical approach to the syntax-semantics interface is the Curry-Hoard isomorphism (Howard, 1980) linking logics to the λ-calculus and type systems. The original Curry-Howard isomorphism was between proofs in intuitionistic logic and (well-typed) λ-terms of the simply-typed λ-calculus. The simply-typed λ-calculus has a type system that mirrors propositional intuitionistic logic. The G2 type system mirrors a higher-order logic with two sorts (e and t) but with various restrictions on quantification.

Using the terminology of the type-logical setting, the core idea behind the Glue theory of the syntax-semantics interface is that each atomic semantic contribution is assigned an appropriate syntax-semantics interface type. Given a word sequence, each typed atomic semantic contribution it makes is picked up and placed into Γ, the Glue typing context for that word sequence. A Glue implementation, in turn, finds all distinct (up to α-equivalence) normal-form terms M that have the target syntax-semantics type T for the word-sequence:

$$\Gamma \vdash M : T$$

(where Γ and T are given, and M is one of a number of possible compositions of type T of the atomic meanings in Γ).

2.3 First-Order Glue (Kokkonidis, 2007b)

Kokkonidis (2007b) proposed a first-order system (G3). The design of G3 does not rely on the ad-hoc restrictions and extensions Dalrymple et al. (1997) placed on System F to obtain G2; it is exactly what it appears to be: a first-order linear type system. While being formally simpler, G3 is also significantly more powerful than its predecessor due to its ability to encode arbitrarily complex hierarchical structures (using functions in the syntax for individuals).

The first step towards TL-LFG, however, comes from the alternative analyses of common nouns Kokkonidis (2007b) proposed that did not use VAR and RESTR attributes. These attributes were the most prominent example of some degree of structure in the so-called "semantic projections" that came with G1 and G2. These attributes could have been included in the f-structure but they were considered semantic in nature, therefore foreign to f-structure. Without internal structure, G1/G2-style "semantic projections" had no reason for existence. If f-structures were used directly, not only would the formalism be conceptually simpler, but the formal problem of non-uniqueness of 'semantic structures' mentioned earlier would have also been avoided. There would still be a formal complication as f-structures are complex formal objects that are not related directly to what the syntax of first-order logic individuals describes. This is why Kokkonidis (2007b) proposed a mapping from f-structures to simple atomic labels. These labels are the constants that can appear in expressions that can be arguments to base types in First-Order Glue.

However, this is not the only way things can be done. TL-LFG is based on the type system of First-Order Glue and encodes f-structures in it (Kokkonidis, 2007a). The basic idea is this: there is a finite number of attributes such as SUBJ, OBJ etc. In LFG, an f-structure is a (potentially partial) function from attributes to values. For every partial function f_p there is a corresponding total function f_t such that $f_t(x) = f_p(x)$ if f_p is defined for x and $f_t(x) = \bot$ otherwise, where $\bot$ is a special element of the range of f_t not in the range of f_p. This total function can be represented as a tuple whereby each position corresponds to a particular attribute and its value is the value of the attribute. In terms of first-order logic syntax for individuals, it can be represented as an N-ary function applied to its N-arguments. For an example, the f-structure value of a CASE feature for a noun that has either accusative or dative case in a language with cases NOM, ACC, GEN, DAT would look something like this:[5]

$$\mathit{fstr}(\bot, \bot, \ldots, \overbrace{-, \beta, -, \delta}^{casemarking}, \ldots, \bot, \bot).$$

Given the commutativity of linear logic (order-insensitivity with regards to the premises), and the importance of word order in natural languages, the question of how word-order constraints are captured arises. Inspiration for an answer can readily come either from the Prolog implementation of Definite Clause Grammars

[5] The analysis of Dalrymple et al. (2006) is used here.

(difference lists) or (the option taken here) from the basic setup of chart parsing (spans) (Kokkonidis, 2007a).

2.4 Instant Glue (Kokkonidis, 2006)

An ability to express word-order constraints in terms of simple features and an ability to encode arbitrarily nested feature structures brought First-Order Glue particularly close to being able to function as the basis for a grammar formalism, rather than as just the syntax-semantics interface. However, something was missing still: unification-based underspecification.

The Instant Glue implementation of Glue was based on a simple type system (G3$_i$) that only inhabited types with normal-form λ-calculus terms (Kokkonidis, 2006). That type system was chosen as the formal foundation for TL-LFG, both because of its normal-form property, but also because it is based on unification rather than quantification.

What this made possible is worth noting. Just like its predecessors,[6] the original First-Order Glue system, G3, as defined by Kokkonidis (2007b) only has universal quantification. But even if it did include existential quantification it would not be quite what one would want as the formal foundation for a type-logical LFG.

Let us first see what cannot be expressed without existential quantification.

(11) $_0$ Every $_1$ boy $_2$ loves $_3$ a $_4$ girl $_5$

The typing context for the above example would look similar to what one gets in First-Order Glue, except that instead of s, o, f, etc.[7] being labels for f-structures they would be the actual f-structures encoded in First-Order Glue. The question is then what is the target type. In Glue, it is t_f where f is the label of a prebuilt f-structure. In TL-LFG, f is not pre-constructed; it is meant to be built up as part of the concurrent syntactic analysis / semantic composition process. So there are no concrete values (except for the span and even for that in an incremental processing scenario the end point would be unknown). The natural solution would be to have existentially quantified variables as values for every attribute with an unknown value. But then, the actual value used would be subject to existential abstraction and therefore unavailable at the end of the derivation. So the entire functional syntactic analysis would just go to waste.

Unification provides a simple and elegant solution. In G3$_i$, all variables are free and equated with values, including other variables, on demand, using an assignment function that is updated throughout the course of the derivation. While the premises can be straightforwardly interpreted as having all the variables in their

[6] Existential quantification was considered as an option in the early days of Glue, and was even used in an analysis, but a dispreferred one.

[7] There would actually also be a number of intermediate structures, but that is a detail with respect to the present discussion.

$(\multimap Intro.)$

$$\frac{\Gamma, X : T \vdash E : T'}{\Gamma \vdash \lambda X.E : (T \multimap T')}$$

$(\multimap Elim.)$

$$\frac{\Gamma_1 \vdash A_1 : T'_1 \quad \ldots \quad \Gamma_N \vdash A_N : T'_N}{F : T_1 \multimap \ldots \multimap T_{N+1}, \Gamma_1, \ldots, \Gamma_N \vdash F\ A_1\ \ldots\ A_N : T_{N+1}{}_{[\sigma]}}$$

$[T_1{}_{[\sigma]}{=}T'_1{}_{[\sigma]}, \ldots, T_N{}_{[\sigma]}{=}T'_N{}_{[\sigma]}$, and T_{N+1} is a base type.]

where σ is some total function
from variables to individual denoting expressions
such that for any variable V, $\sigma(V) \neq V$.

Figure 1: TL-LFG (G3$_i$) Type-Inference Rules

types implicitly universally quantified, the interpretation of the variables in the target type is a bit more open ended. Both an interpretation assuming implicit universal quantification and another one assuming implicit existential quantification are possible, and both are useful. All uninstantiated[8] variables of the target type as originally specified can be thought of as universally quantified and all instantiated ones as existentially quantified.

(12)
[‘every’] :
“$\lambda R.\, \lambda S.\, \forall x.\, S(x) \rightarrow R(x)$” : $(e_s \multimap t_s) \multimap (e_s \multimap t_\alpha) \multimap t_\alpha$

[‘boy’] :
“$\lambda x.\, boy(x)$” : $e_s \multimap t_s$

[‘loves’] :
“$\lambda(x, y).\, loves(x, y)$” : $e_s \multimap e_o \multimap t_f$

[‘a’] :
“$\lambda R.\, \lambda S.\, \exists y.\, S(y) \wedge R(y)$” : $(e_o \multimap t_o) \multimap (e_o \multimap t_\beta) \multimap t_\beta$

[‘girl’] :
“$\lambda y.\, girl(y)$” : $e_o \multimap t_o$

Kokkonidis (2007b) investigated the two opposing trends with regards to having the ‘$\otimes$’ connective (tensor) in Glue, explained to what extend Glue analyses can avoid using it, but, targeting the second-order aspect of G2, chose to take a neutral stand with regards to whether the tensor should be included or excluded. Based on that discussion, I will assume the tensor to not be necessary for the purposes of either Glue or TL-LFG. This assumption leads to a simpler system. While a version of Instant Glue that includes the tensor exists, the version without it (Figure 1) is as simple as a first-order type system for Glue gets.

[8] A variable V is instantiated iff there is an X such that $(V, X) \in \sigma^*$ and X is a non-variable.

3 Differences with LFG

TL-LFG aims to be a simpler theory than LFG. That is a rather ambiguous statement. Formal simplicity does not necessarily come with ease of expressing linguistic facts and generalisations. It has been a priority for the LFG community to have intuitive representations and ways of expressing constraints. TL-LFG tries to build on this tradition, pushing even further both formal simplicity and ease of use.

3.1 From words to meanings in TL-LFG and LFG: An architectural comparison

LFG comes with a modular architectural design, based on separate representations (projections), linked through correspondence functions. While Figure 1 does not mention all the various different projections that have been assumed in the literature, it already gives a picture of the architectural complexity of LFG as described by Dalrymple (2001) (where f-structure was the only input to semantics in the analyses presented [9] as intended originally by Kaplan and Bresnan (1982)).

TL-LFG's architecture is a much more light-weight theory. In TL-LFG there is only one intermediate layer between a sequence of words and their meanings: atomic meanings with their syntax-semantics interface types.

LFG+Glue	TL-LFG
• input: word sequence • π: a mapping from strings to c-structure. • c-structure • ϕ: a mapping from c-structure to f-structure. • f-structure • σ: a mapping from f-structure to 'semantic' structures • meanings and Glue types • output: meanings	• input: word sequence • meanings and TL-LFG types • output: meanings

Table 1: Layers in LFG+Glue and TL-LFG

[9] Of course, Dalrymple (2001) does not fail to mention the understanding that other projections could be contributing to the semantic composition process. But the simplified picture the concrete examples of LFG syntax-semantics analyses in her book present is in line with the level of detail for the comparison between TL-LFG and LFG in the present paper.

3.2 No c-structure

Whether TL-LFG has phrase-structure rules and/or Immediate Dominance / Linear Precedence rules is an interesting question. The easy answer is to say that it does not; the unificational first-order Glue type-system that is its formal basis does not include such ID/LP rules. But this does not mean TL-LFG has no way of expressing the constraints such rules are used to express.

LFG as originally presented by Kaplan and Bresnan (1982) came with the following phrase structure rules for English:

(13)

$$\text{S} \rightarrow \begin{array}{c}\text{NP}\\ (\uparrow \textsc{subj}) = \downarrow\end{array} \quad \begin{array}{c}\text{VP}\\ \uparrow = \downarrow\end{array}$$

$$\text{VP} \rightarrow \text{V} \left(\begin{array}{c}\text{NP}\\ (\uparrow \textsc{obj}) = \downarrow\end{array}\right) \left(\begin{array}{c}\text{NP}\\ (\uparrow \textsc{obj2}) = \downarrow\end{array}\right) \left(\begin{array}{c}\text{PP}\\ (\uparrow (\downarrow \textsc{pcase})) = \downarrow\end{array}\right) \left(\begin{array}{c}\text{VP'}\\ (\uparrow \textsc{vcomp}) = \downarrow\end{array}\right)$$

$$\text{NP} \rightarrow \begin{array}{c}\text{Det}\\ \uparrow = \downarrow\end{array} \quad \begin{array}{c}\text{N}\\ \uparrow = \downarrow\end{array}$$

(14) John snores.

The relevant lexicon entries for (14) assuming this old c-structure analysis together with a standard (G2) Glue analysis (with first-order logic as the semantic representation language) are:[10]

‘John’	NP	$(\uparrow \textsc{pred}) = \text{‘JOHN’}$ $[\underline{john}] : e_{\uparrow_\sigma}$
‘snores’	V	$(\uparrow \textsc{pred}) = \text{‘SNORE (SUBJ)’}$ $[\lambda x.\, \underline{snore}(x)] : e_{(\uparrow \textsc{subj})_\sigma} \multimap t_{\uparrow_\sigma}$

The TL-LFG grammar that expresses this analysis consists of two lexical entries but no separate syntactic rules: all grammatical knowledge resides in the lexicon. The emphasis is on having few but effective primitives. Grammatical functions such as SUBJect and OBJect are primitives in TL-LFG and so are the semantic concepts of entity and truth value. ‘John’ makes a semantic contribution corresponding to a particular entity, $john$, and ‘snore’ one corresponding to a function taking an entity x and returning a truth value (true or false, depending on whether x is snoring or not).

[10] For the purpose of illustrating differences of the frameworks in practice, a simplistic view of syntax and semantics will be sufficient; any additional level of detail would complicate analyses at least equally for the two frameworks and, I claim, not more for TL-LFG than for LFG.

(15) $\underbrace{\text{‘John’}}_{j}$ $\quad$ ‘$john'$’ : e_j

(16) $\underbrace{s\ \text{‘snores’}}_{f}$ $\quad$ ‘$\lambda x.\, snore(x)$’ : $e_s \multimap t_f$ $\quad$ where $f = \left[\ \textsc{subj} \quad s\ \right]$

On the left-hand side one finds a schematic representation for the ORTHOgraphy and SPAN attributes. The one for ‘snores’ states that its SUBJect s is expected to precede it. Note that if we take the orthography, the SVO constraint, and the meaning with its type (function from entities to truth values) as observable facts, the only appearance of a theory-specific primitive is the SUBJ feature. This schematic representation for spans possibly augmented with explicit linear-precedence constraints corresponds to LFG's linear precedence constraints.

The other point to be made here is that the specification of word-order constraints used bears some resemblance to LFG's phrase structure rules. The word-order constraint for ‘snores’ is closely associated with the ‘S → NP VP’ rule of (13) as found in early LFG work (Kaplan and Bresnan, 1982). However, it lacks any mention of syntactic categories, only being concerned with the essential facts of word-order: the subject must precede the verb ‘snore’. The stipulation that the subject is an NP in LFG is redundant given the f-structure and semantic type information available, i.e. that the semantic type of the subject is e, and also its f-structure has $\perp$ as the value of its FORM feature meaning that it is not a prepositional phrase. In TL-LFG the concept of a noun phrase, just like that of a noun, is a concept definable in terms of its semantic and functional primitives.

One advantage of the TL-LFG approach is that it relieves the grammar writer from the burden of an additional layer of specification. It also provides a more abstract view of constituent structure that represents exactly what is necessary for determining the semantics. The LFG examples in this paper have been using a rather dated theory of c-structure. In more recent work Inflectional Phrases would be making their appearance, and in the works of some authors Determiner Phrases. The point is that if updating the theory of c-structure does not affect f-structure or semantic composition, c-structure is a redundant intermediate step from the word sequence input to semantics and vice versa. There are cases where updating the theory of c-structure will affect the syntax-semantics interface, namely when the grouping of words changes as this will normally mean that the semantics has to change, and it is exactly this fact that TL-LFG captures.

If the details of c-structure are not important for the syntax-semantics interface, they have no place in TL-LFG, which aims to be a minimalist theory of grammar. It has been one of the key ideas of LFG that a functional structure representation (a feature structure providing information about grammatical functions such as SUBJ and OBJ) is to be maintained in addition to a constituent structure one (a tree representing the phrase structure of the input string). It is easy to claim that TL-LFG is a more ‘functional’ theory than LFG because it only has f-structure as its syntactic representation.

There is substance to the claim. It was the original intent of Kaplan and Bresnan (1982) that f-structure be the sole input to semantics. This is true for TL-LFG, but not necessarily for LFG. If it were then LFG f-structure would encode all important syntactic relations Glue needs to have available. That is the case with relations such as SUBJ, OBJ, etc. but not necessarily, for instance, with modification relations. The LFG approach, of dumping adjuncts in a set feature helps keep the f-structure for a modified phrase very similar to that of the same phrase with the modification removed. The LFG approach has a positive impact with regard to complexity of grammar writing as the description of functional constraints that are not influenced by the presence of modifiers does not need to make special provisions in order to work when modifiers happen to be present. However, in LFG, convenience comes at a high price: f-structure does not encode syntactic relations relating to modification.[11] So, it encodes some grammatical relations but not all and can not be the sole input to semantics. It is able to capture the difference between 'John likes Mary' and 'Mary likes John', but not the difference between 'a fake golden gun' and 'a golden fake gun'. The current LFG view that this is acceptable is questionable, especially for a theory that claims to put emphasis on functional structure. TL-LFG is more 'functional' because its f-structure captures such syntactic relations.

Moreover, the TL-LFG analysis of scoping modification (Kokkonidis, 2007c) achieves having a sufficiently detailed f-structure representation without loosing the elegance and simplicity of LFG's f-structures. Trivial as it may seem, the key is using a more basic data-structure, lists, that unlike LFG sets, do not disregard the order in which modifiers are encountered in the input.

TL-LFG also comes with the claim that it is more 'lexical' than LFG because it does not have at its formal foundation phrase-structure rules or ID/LP rules. This is indeed true as one can see in (18), the lexical specification for 'snores' in raw TL-LFG lacking the syntactic sugar of the appealing presentation used in (16) and elsewhere in this paper. Indeed, that specification brings to mind theories such as Type-Logical Categorial Grammar where radical lexicalism reigns supreme and no phrase-structure rules as such exist. Yet can this also be said about the syntactically sugared version of TL-LFG used in (16)? Arguably, there is no separate syntactic rule as such. What is expressed on the left-hand side of the lexical entry is simply a constraint that applies to that particular lexical entry. Also the syntactic sugar for span specifications is only a way of expressing certain constraints in a more intuitive way; syntactically sugared TL-LFG is the same theory as TL-LFG,

[11] This is the case in prominent places of the LFG literature inviting criticism and solutions (Andrews and Manning; Andrews, 1993; 2004), but not a weakness of LFG as such. My impression is that when it comes to theory, there are strong voices supporting a simplified version of f-structure and more use of the inverse ϕ mapping, and when it comes to grammar engineering, LFG f-structure is much more detailed and autonomous. My criticism is directed towards LFG with insufficiently detailed f-structures. LFG-based grammar engineering (at least amongst the members of the ParGram community) tends to put the same kind of emphasis on f-structure that TL-LFG does and if one takes that version of LFG as the standard one then much of this criticism is inapplicable as such and should rather be seen as support for that approach to the role of f-structures in LFG.

much the same way as choosing not to display spans in a grammar when using a chart parser or difference lists in Prolog DCGs is a matter of presentation and convenience rather than of essence.

(17) $_0$ John $_1$ snores$_2$

(18) 'snores' '$\lambda x.\, snore(x)$' : $e_s \multimap t_f$

$$\text{where } f = \left[\begin{array}{ll} \text{SPAN} & \left[\begin{array}{ll} \text{START} & start \\ \text{END} & \nabla + 1 \end{array}\right] \\ \text{ORTH} & \text{`snores'} \end{array}\right],$$

$$s = \left[\begin{array}{ll} \text{SPAN} & \left[\begin{array}{ll} \text{START} & start \\ \text{END} & \nabla \end{array}\right] \\ \text{ORTH} & \\ \ldots & \end{array}\right], \text{and}$$

where ∇ is the current position in the word sequence.

The line of division between radical lexicalism and having phrase structure rules (or an equivalent) is pretty thin in TL-LFG. While TL-LFG follows, at the formal foundation level, the paradigm of radical lexicalism as found in, say, Type-Logical Categorial Grammar, the pretty straightforward (and familiar from chart parsing and/or Prolog DCGs) syntactic sugar that hides the underlying representation for the word sequence and positions within it allows for syntactic constraints to be expressed in a way that combines the best aspects of both CG and LFG approaches.

3.3 No 'semantic' forms and no 'semantic' projections

Developments within Glue have lead to the term 'semantic projection' being used (in G1 and G2) for rather uninteresting feature structures. Their intended use in G2 was just that they would distinguish between $e/1$ and $t/1$ base types of different f-structures. There is nothing semantic about them – different random numbers would do. Moreover, they fail to be unique as explained earlier. What so-called semantic projections were meant to do is import syntactic information of a very abstract nature (relations between distinct parts of an f-structure) into the Glue type system. In TL-LFG, there are no intermediaries; f-structures themselves are arguments to the syntax-semantics interface base-type constructors.

The incorporation of Glue into LFG meant also that the role of LFG's 'semantic forms' changed. In early LFG, 'semantic forms' had a clear syntax-semantics interface role. In current LFG with Glue, semantic constructors (the elements of the typing context with their corresponding meaning) have taken up the most essential roles of 'semantic constructions', not leaving much semantic substance to 'semantic forms'.

Investigating the role of these no-longer semantic 'semantic forms' reveals three facts: (i) they are used in relation to *syntactic* completeness and coherence;

(ii) they are used to make the f-structure containing them unique; (iii) they are used for presentation reasons. None of the above three roles has anything to do with semantics. In TL-LFG, it is clear where the semantics is specified; it is not in the functional *syntactic* structures but on the meaning side of semantic contributions (the left hand-side of the colon, the right-hand side being the syntax-semantics interface type).

Therefore it is not surprising that 'semantic forms', one of the most important concepts of LFG, is not part of TL-LFG. For presentation reasons having a feature such as ORTH seems more appropriate. The following section discusses completeness and coherence in TL-LFG and LFG; in TL-LFG the resource sensitivity of the formalism guarantees those principles without stipulation and the syntax-semantics interface types are instrumental in that. That leaves 'semantic forms' a single role in LFG, important for LFG to work, but not related to semantics: distinguishing between different f-structures. Again this is something different arbitrary numbers would achieve equally well.

TL-LFG was inspired by the elegance of the Glue syntax-semantics interface. Two pieces of formal machinery of LFG called 'semantic' ('semantic' projections and 'semantic' forms) are reducible to distinct but otherwise arbitrary numbers. The need for such formal hacks stems from the fact that f-structures and semantic projections have no direct connection to the word sequence they correspond to. In TL-LFG, f-structures have this direct connection in the form of the span feature (or an equivalent in terms of difference lists).

3.4 Completeness and Coherence

Completeness and Coherence are two very fundamental and important principles in LFG. However, these principles are not intrinsic to the formal framework. Nothing in the formal setup stops the syntactic rules of (13) from forming f-structures examples (19)–(21). There needs to be a piece of stipulation, the Completeness Principle, in order to mark example (20) as ungrammatical. There needs to be another piece of stipulation, the Coherence Principle, in order to mark example (21) as ungrammatical.

(19) John likes Mary.

(20) * John likes.

(21) John snores Mary.

While there is nothing objectionable about linguistic principles, the nature of Completeness and Coherence as additional pieces of stipulation shows that something was missing from the formal framework proper. In TL-LFG, (syntactic and semantic) Completeness and Coherence are automatically enforced due to the resource sensitivity of the Glue type system (Dalrymple et al., 1993). They are a consequence of the overall setup and type-logical formal foundation of the theory, rather than something that had to be added to it.

4 Conclusions

TL-LFG rejects the bulky formal (and theoretical) machinery LFG comes with, but not the importance LFG attaches to functional structure and functional constraints. Indeed it attaches more importance to f-structure than LFG and claims to be a more 'functional' theory as a result.

A very obvious argument in support of this claim is on the basis of TL-LFG having no c-structure representation. But if that is the case, the sceptic may wonder whether this is so simply because f-structure was turned into a kind of c-structure with features as is the case for HPSG.

If one concentrates on the structural organisation of a TL-LFG f-structure, it becomes obvious that this is not the case. TL-LFG f-structures have, in general, the same structural organisation as their corresponding LFG f-structures which in turn is quite different from that of their corresponding c-structure trees (in general f-structures are more flat).

As for what information goes into f-structures, the crucial addition to f-structure is the span attribute. Responding to a possible criticism that span information in the f-structure is a way of importing c-structure information into f-structure reveals interesting facts about both TL-LFG and LFG.

Spans relate to the word-sequence, not to any tree-structured analysis of it. They help relate the f-structure to the word-sequence in a simple and intuitive way. It is the span information that helps distinguish between the f-structures of the two occurrences of 'the' in a sentence like 'The coach praised the players'. If instead of encoding this relation in the f-structure itself, an LFG-style correspondence function was used for that purpose, there would need to be some other way of distinguishing between them. Indeed this could come from the semantics (uniqueness of 'semantic forms' for example), but it is not at all clear why this would be a better approach.

Moreover, the f-structure would have to contain this information. A correspondence function from f-structures to semantic structures would not help. This is why 'semantic forms' guaranteed to be non-equal even when their semantic content is the same are a part of f-structures in LFG. This discussion relates to why G1/G2 'semantic projections' fail to serve their purpose and why even if the obvious step of moving the 'semantic form' into the 'semantic projection' would not be a good idea.

TL-LFG, not only does not need to import c-structure information into f-structure in the guise of the SPAN feature,[12] it also keeps semantic information out of the f-structure. LFG distinguishes itself on the basis of using separate representations for linguistically different kinds of information, yet it had semantic information inside a syntactic structure. Moreover, had this not been the case, i.e. had semantic forms

[12] The main reason TL-LFG does not need to add c-structure information into its f-structures is that LFG f-structures tend to already contain enough information to distinguish between f-structures corresponding to different word sequences. This is due to the fact that LFG performs a number of checks at the level of f-structure, just like TL-LFG.

been placed inside semantic projections, in the most straightforward manner possible, the whole system would collapse because it would be unable to distinguish between f-structures that were meant to be different but in the absence of PRED features would be equal. This would be a direct consequence of doing the right thing, with respect to the projection architecture, and relying on a correspondence function rather than embedding the semantic information inside the f-structure using a feature (as is now the case).

While it was never the intent of this paper to challenge the projection architecture of LFG as such, it seems that suspiciously much depends on the PRED attributes and their 'semantic form' features inside f-structure to keep the LFG system together. One important role they play is in setting the subcategorisation requirements for Completeness and Coherence. In many ways, the f-structure PRED features have a syntax-semantics interface role. However, unlike the case with TL-LFG, in LFG there is nothing in the formal foundation of the theory that guarantees the Completeness and Coherence principles. In TL-LFG the resource-sensitive type-logical formal foundation of the theory does exactly that without further stipulation.

Returning to the claim that TL-LFG is more 'functional', the argument that this is so because c-structure disappears has a certain immediate appeal, but the essence is in examining the role and function of f-structure in the two theories. In TL-LFG it is there to capture any and all grammatical relations that would be important for the semantics; in LFG it captures some but not all of them.

Traditional phrase structure rules and the immediate dominance part of immediate dominance / linear precedence rule system are not a part of LFG and neither is the syntactic category system of LFG. Linear precedence rules are. In TL-LFG, all grammatical knowledge resides in the lexicon which makes it more 'lexical' than LFG. However, a bit of TL-LFG syntactic sugar hides low-level details of spans and gives a way of specifying spans and linear precedence constraints in an intuitive manner. To the extent that such constraints can be factored out of the lexicon TL-LFG could be seen as having rules and even constructional meaning. The point is that this is more a matter of presentation and convenience than theoretical essence.

Neither being more 'functional' nor being more 'lexical' mean much in themselves. It is TL-LFG's formal simplicity and parsimony combined with some of the best aspects of LFG that give these comparisons substance. Starting with the syntax-semantics interface and then building the details of the syntax based on a very successful theory lead to a re-incarnation of that theory in a different formal setting which was but a small fragment of the original theory's formal arsenal. Not only is the formal framework now simpler, but so is the conceptual framework: accounting for the facts involves fewer theory-internal concepts and representations, something achieved without complicating the part of the original theory preserved in the new theory. Finally, the new type-logical formal framework captures linguistic intuitions that the original framework left to stipulation.

References

Andrews, A. D. and Manning, C. D. 1993. Information-spreading and levels of representation in LFG. Technical Report CSLI-93-176, CSLI, Stanford University.

Andrews, Avery D. 2004. Glue Logic vs. Spreading Architecture in LFG. In Christo Mostovsky (ed.), *Proceedings of the 2003 Conference of the Australian Linguistics Society*.

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

Asudeh, Ash and Crouch, Richard. 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.

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

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. Syntax and Semantics Series, No. 42, Academic Press.

Dalrymple, Mary, Gupta, Vineet, Pereira, Fernando C.N. and Saraswat, Vijay. 1997. Relating Resource-based Semantics to Categorial Semantics. In *Proceedings of the Fifth Meeting on Mathematics of Language (MOL5)*, Schloss Dagstuhl, Saarbrücken, Germany, an updated version was printed in (Dalrymple, 1999).

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

Dalrymple, Mary, King, Tracy Holloway and Sadler, Louisa. 2006. Indeterminacy by underspecification. Poster presented at the LFG06 Conference.

Dalrymple, Mary, Lamping, John, Pereira, Fernado C.N. and Saraswat, Vijay. 1995b. A deductive account of quantification in LFG. In Kanazawa Makoto, Christopher J. Pinón and Henriette de Swart (eds.), *Quantifiers, Deduction and Context*, Center for the Study of Language and Information, Stanford, California.

Dalrymple, Mary, Lamping, John and Saraswat, Vijay. 1993. LFG semantics via constraints. In *Proceedings of the Sixth Meeting of the European ACL*, pages 97–105, European Chapter of the Association for Computational Linguistics, University of Utrecht.

Frank, Anette and van Genabith, Josef. 2001. GlueTag: Linear Logic 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.

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

Girard, Jean-Yves. 1989. *Proofs and Types*. Cambridge University Press.

Howard, William A. 1980. The formulae-as-types notion of construction. In J.R. Hindley and J.P. Selden (eds.), *To H.B. Curry: Essays on combinatory logic, lambda calculus and formalism*, Academic Press, conceived in 1969. Sometimes cited as Howard (1969).

Kaplan, Ronald M. and Bresnan, Joan. 1982. Lexical Functional Grammar: A formal system for grammatical representation. In Joan Bresnan (ed.), *The Mental Representation of Grammar Relations*, pages 173–281, MIT Press.

Kokkonidis, Miltiadis. 2005. Why glue your donkey to an f-structure when you can constrain and bind it instead? In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG05 Conference*, CSLI Publications.

Kokkonidis, Miltiadis. 2006. A Simple Linear First-Order System for Meaning Assembly. In *Proceedings of the Second International Congress on Tools for Teaching Logic*, Salamanca, Spain.

Kokkonidis, Miltiadis. 2007a. Encoding LFG f-structures in the TL-LFG type system. In *Proceedings of the Second International Workshop on Typed Feature Structure Grammars*, Tartu, Estonia.

Kokkonidis, Miltiadis. 2007b. First-Order Glue. *Journal of Logic, Language and Information* To appear in print. DOI: 10.1007/s10849-006-9031-0.

Kokkonidis, Miltiadis. 2007c. Scoping and Recursive Modification in Type-Logical Lexical Functional Grammar. In *Proceedings of the 12th Conference on Formal Grammar*, to appear.

Moortgat, Michael. 1997. Categorial Type Logics. In Johan van Benthem and Alice ter Meulen (eds.), *Handbook of Logic and Language*, Elsevier.

Morrill, Glyn. V. 1994. *Type Logical Grammar: Categorial Logic of Signs*. Dordrecht: Kluwer.

LFG AS A FRAMEWORK FOR DESCRIPTIVE GRAMMAR

Paul Kroeger

GIAL and SIL Intl.

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract:

LFG has a number of features that make it an attractive and useful framework for grammatical description, and for translation. These include the modular design of the system, the literal representation of word order and constituency in c-structure, a typologically realistic approach to universals (avoiding dogmatic assertions which make the descriptive task more difficult), and a tradition of taking grammatical details seriously.

> (This talk was presented as part of a panel entitled "Directions of LFG: Many Paths". Each of the six panel members was asked to "explain how LFG has related to their past work and to what they are doing now in their careers.")

Since leaving Stanford a lot of my work has been focused on training people to do field linguistics, so I have not actually been using the full LFG formalism in my daily work most of the time. But I have found that the conceptual structure of LFG provides a very good framework for grammatical description, and that is what I would like to talk about today.

A number of people have asked me at various times whether SIL still teaches Tagmemics to its field workers.[1] The short answer is "no"; there is a fair bit of variety from one place to another, but none of the major training programs currently use Tagmemics as their basic model. For those of you too young to remember Tagmemics, I might summarize it by saying it was Kenneth Pike's attempt to extend the methods of structural phonemic analysis to morphology and syntax. The phonologists among us may be eager to point out that structural phonemics was not a very satisfactory model of phonology either; but in its day it was considered a great triumph, the envy of the other social and behavioral sciences. And for all its limitations, Tagmemics was very successful in one important respect: using this framework, Pike was able to get people with fairly limited training in linguistics to study and describe languages that had never been studied before.

This kind of primary fieldwork on previously undescribed languages is a difficult thing to do, and I do not believe it has gotten nearly enough respect in American linguistics during the past 40 years or so; so I do not want to minimize in any way the contribution of Pike and his students and colleagues. But I would

[1] "SIL" originally stood for "Summer Institute of Linguistics", but now only the initials are used. For information about the organization, see www.sil.org.

have to admit that many Tagmemic grammars are frustrating and difficult to read. One problem is that concepts like “contrast” and “minimal pair” do not apply as neatly to phrase or sentence patterns as they do to phonemes. The more fundamental issue is that the goal of a Tagmemic grammar was to list the inventory of contrastive units of various types, just as a major goal of phonemic analysis was to state the inventory of consonants and vowels. So Tagmemic grammars tend to be essentially lists of clause patterns, sentence patterns, word patterns, etc. What is often lacking is any statement about the generalizations, i.e. the rules that constitute the grammar of the language.

Even when our goal is to write a purely descriptive grammar, I feel it is important to adopt a rule-based perspective. We may never express these rules in formal syntactic notation, but the rules of the grammar are an important part of what we are trying to describe.

Of course, the focus on grammar as a system of rules is the defining characteristic of generative linguistics. But my impression is that a lot of people who do field linguistics and descriptive grammar have given up on generative syntax. In fact, a fair number seem to have given up on syntax in general, either ignoring it or adopting the view that most apparent syntactic regularities can be reduced to semantic and/or pragmatic generalizations. I suspect that a major reason for this is that the models of formal syntax that they have been exposed to seem so far removed from observable reality and impractical for descriptive purposes.

I believe that LFG offers a much better framework for descriptive grammar than recent transformational models. The modular design of the system means that c-structure representations are a direct and literal representation of the word order of the sentence as it is actually pronounced (WYSIWYG), and of constituency in the classical sense. These are basic facts that any descriptive grammar needs to spell out. (My impression is that a lot of current work in formal syntax, and even some more functional approaches, largely ignore these issues.)

Moreover, the modular design of the system means that problems or novel solutions in one area of grammar do not necessarily lead to complications in other parts of the analysis. For example, I have been interested for some time in the problem of “symmetric(al) voice”, as it has become known in Austronesian syntax. (The term was first used, I believe, by Bill Foley at LFG98 in Brisbane.) Essentially, this means a voice alternation (change in the assignment of the SUBJ function) without demotion. In a language like Balinese, there are two different transitive clause types: one in which the agent is SUBJ, and another in which the patient is SUBJ. Both of these are fully transitive, meaning that agent and patient are both terms (direct core arguments); neither gets demoted to oblique or adjunct status, unlike familiar voice alternations such as passive and anti-passive. In a language like Tagalog, there are several additional voice options (dative, instrumental, locative) but again these are all fully transitive; in none of them is the agent demoted to oblique or adjunct status.

Now obviously this is a problem for the original, classical form of the Lexical Mapping Theory (as developed by Bresnan & Kannerva 1989, Alsina & Mchombo 1993, among others). When I arrived at Stanford in the late 1980s, there were several different theories of linking being developed, and these facts seemed to be a problem for all of them. But for LFG, they were a problem **ONLY** for the linking theory. Once you allow a non-canonical linking pattern for these languages, the rest of the system functions pretty normally. Wayan Arka (2003) later developed a model of LMT, adapted from Alex Alsina's model, that works for Balinese, and I am sure this could be further adapted for Tagalog. But having solved this problem does not force major changes to other aspects of the analysis.

I would like to contrast this with a very influential analysis of symmetric voice proposed within the Government and Binding framework by Guilfoyle, Hung and Travis (1992). They proposed that the agent is internal to VP at D-structure, specifically that it occupies [SPEC, VP]. Depending on the voice morphology of the verb, one argument will fail to get Case in its D-structure position and be forced to move into [SPEC, IP], the structural subject position. No theta roles get deleted or absorbed, so the agent is free to remain in [SPEC, VP] when it is not selected as subject. This is a very elegant model of non-demoting voice alternation. However, because the change of GFs is expressed in terms of phrase structure, it makes incorrect predictions about things like word order and long-distance extraction in Tagalog. When all information is represented in the same way, a change in any part of the system affects every part of the analysis. In contrast, the modular design of LFG allows us to address separately issues which are in fact independent of each other.

Wayan Arka (2003) also presented a fair bit of evidence for the claim that in Balinese Undergoer Voice clauses, in which the patient is the subject, the agent and verb form a very tight constituent which he labels V-bar. This is an interesting and typologically unusual claim. Similar claims have been made in other Western Malayo-Polynesian languages, e.g. Toba Batak (Schachter 1984) and West Coast Bajau (Miller 2007). In transformational theories that adopt the Uniform Theta Assignment Hypothesis (UTAH, Baker 1988), this hypothesis is virtually ruled out on theoretical grounds. Within LFG, however, because GFs and theta-roles are represented separately from constituent relations, it is simply an empirical issue: the analysis can follow the facts of the language.

Another thing that I appreciate about LFG is that there is a healthy respect for the degree to which languages may differ from each other. Universals that get proposed within LFG tend to be fairly well motivated typologically. Last October Peter Sells and I participated as “respondents” in a workshop on Austronesian syntax at UCSD. All of the papers were written within the Minimalist framework, and most of them adopted Kayne’s “anti-symmetry” hypothesis — essentially the claim that underlying phrase structure for all languages is strictly binary and right-branching — and Cinque’s recent proposals about universal D-structure positions for various types of adverbs. Whatever the merits of these proposals as theories of universal grammar, they impose an immense (and to my mind, intolerable) descriptive burden on languages whose surface word order is very different. Anti-symmetry seems like the most inconvenient possible analysis for a language like Malagasy.

Most of my students are not planning to do descriptive linguistics for its own sake. Many of them hope to use it to support their work in Bible translation, adult literacy, bilingual education, etc. The relevance of syntactic analysis to translation depends heavily on your philosophy of translation. Many years ago I read a review article about George Steiner’s book *After Babel.* I have not been able to find that article again, but as I recall the author was a Marxist literary critic who believed that a good translation was one that preserved the foreignness of the source text; the translation should feel almost as strange, difficult, and off-putting as the original would be for someone who does not speak the source language. For this type of translation, relatively little linguistic analysis is required since the form of the source text is largely preserved in a fairly literal way.

SIL has traditionally preferred a different model of translation. Local circumstances sometimes require a somewhat literal approach, but where possible the ideal is generally seen as a translation that is as faithful as possible to the meaning of the source text, but as natural as possible within the linguistic system of the target language. For this type of translation linguistic analysis is quite important. Any difference between the grammars of the source language and target language is a potential area where translators may be unconsciously influenced by the form of the source text, resulting in a loss of accuracy, clarity, and/or naturalness. Every detail of the grammar is important; the distinction between “core” and “periphery” is not too helpful in this context.

Ken Pike used to tell a story about a missionary that he met on one of his trips to Africa. This man had spent a year or so studying the local language and then got up to preach his first sermon. He told the people: “Jesus said, ‘I am the way, the truth and the life; no one comes to the Father but by me.’” The people replied, “Then we will worship you.” The man was horrified. He said, “No, wait, you don’t understand; ***Jesus*** said ‘I am the way, the truth and the life.’” The people answered, “If Jesus said it, we believe it; we will worship you.”

Obviously the people were interpreting a direct quote as indirect speech. I believe that the language in question was Bariba. Pike devotes several pages in his 1966 book on African languages to the factors which determine the choice between direct vs. indirect speech in Bariba, involving such things as the relative prominence of the speech act participants on the person-animacy hierarchy. He does not spell out what the grammatical difference between the two is, but in an obscure footnote he mentions that indirect speech is preferred and expected for reporting statements that are believed to be true.

Reported speech has been a source of difficulty in a number of other African languages as well. In some cases the use of the logophoric pronoun has caused problems. Other languages have a very productive and widely used category of "semi-direct" speech. Some translators have been reluctant to use this "semi-direct" form in their translations, even in contexts where they would always use it in natural speech, because they feel obligated to match the direct or indirect speech of the source text. The point is, every detail of grammar is potentially significant. The LFG implementation of large scale grammars for English, French, German, etc. has (I believe) instilled an ethos which takes the details seriously; they all have to be accounted for.

My own first attempt at translation came roughly two months after I began to study the Kimaragang language. I was asked to return to the state capital to help teach a seminar on translation principles. I was lecturing in the mornings, and in the afternoon the participants were practicing on short passages from the book of Acts, using a simplified Indonesian version as the source text. One of the participants turned out to be a Kimaragang; he spoke a dialect significantly different from the one I had been studying, but I could make out a fair bit of what he said.

One of the assigned passages was a somewhat bizarre story from a very early period in church history, when the small band of believers were practicing a kind of voluntary communism (or communalism), sharing their possessions with each other. A certain couple decided to sell a piece of land that they owned, keep part of the money and donate the rest; but to pretend that they were donating the full price. When the husband brought the money to St. Peter, Peter asked him: "Is this the full amount you got for that land?" The husband said "Yes." Peter said, "Why are you doing this? The land was yours, and after you sold it you were free to do whatever you like with the money. You are not lying to me, but to God." Immediately the husband fell down dead, and the young men wrapped up his body and carried him out to bury him. Some time later the wife came in, not knowing what had happened; the same conversation was repeated, with the same result.

As I looked over the first draft of the Kimaragang version of this story, I felt that the translator was doing a pretty good job until he got to the part where the husband falls down dead. At that point I realized that he had switched into a literal, nearly word-for-word rendering of the Indonesian. The two languages are very different in their structure, and the result seemed very unnatural and confusing to me. So I volunteered to fix it up, based on my two-months acquaintance with a related dialect. When the man read what I had written, he agreed: "Yes, that does sound better. Of course, if we say it that way people will think that the husband fell down dead, then he picked himself up and wrapped up his own body, carried himself out and buried himself."

Some years later, in the process of writing a dissertation about Tagalog, I think I finally figured out what went wrong. I had used pro-drop expecting it to signal subject continuity. But instead it was interpreted as indicating agent continuity. In these languages, as in most languages, agents tend to be highly topical in narrative. But because in Kimaragang (as in many Philippine-type languages) the agent is frequently not the grammatical subject, topic continuity is often not reflected by subject continuity.

Topic and focus are important issues in translation. I once heard a man named Roger van Otterloo talk about his initial attempts at translation in Kifuliru, a Bantu language of Zaire/Congo. They were beginning with one of the passages that talks about defending the rights of widows, orphans and foreigners. Roger proposed wording that simply said, "Don't steal from widows," and all the men with him began to laugh. He was afraid that he had perhaps gotten a tone wrong and said something improper, but he discovered that the problem was more interesting. Basic word order in Kifuliru, as in most Bantu languages, is SVO. But the immediate post-verbal position is also a structural focus position. So when he said "Don't steal from widows," the people heard "Don't steal from *widows*" (implying: "anyone else is fair game").

Nowadays everyone talks about topic and focus, but LFG was one of the first syntactic frameworks to integrate these pragmatic functions into the formal rule system. Joan Bresnan in particular was one of the pioneers in this area. I think this is a significant contribution to the field as a whole.

I once spent several months as an advisor to a committee of translators from one of the Land Dayak languages of western Borneo. They had been producing very literal renditions of the English Good News Bible, and I wanted to give them some sense of what a more natural style would look like; so I asked one of them to tell the story of Jonah and the whale in his own words. When he got to the point where the sailors ask Jonah how they can save themselves from the storm, he said (in Land Dayak): "Jonah told the sailor to throw himself into the sea." I was surprised and asked who exactly ended up in the water, but it turned out the meaning was correct; it was Jonah who went overboard. That was my first hint that the language allowed long-distance reflexives.

I told this story to K.P. Mohanan during one of my visits to Singapore, and Mo said something like, "Of course, what else would he say?" It seemed perfectly natural; just like Malayalam. But at that time I had not heard of long-distance reflexives in any closely related language, so I was not expecting it.

Now suppose that this had been a translation instead of a spontaneous story, and that the translator (following the English source text) had written: "Jonah told the sailor to throw him into the sea." I believe that this would have been interpreted in Land Dayak as meaning that some third person was to be thrown overboard, which of course is incorrect. But if I had not accidentally learned this fact about reflexive binding in Land Dayak, I would never have thought to check it, because it looks like a perfectly accurate and natural translation. This is an example of a "blind error", an error that no one would have caught without specific knowledge about that aspect of Land Dayak grammar.

Let me mention one final grammatical issue in translation. Malay/Indonesian shows a strong preference for the passive voice (specifically the *di-* passive) to encode main-line events in a narrative, especially where there is a series of actions by the same actor. Now in these contexts, it is clear that the actor is highly topical. In some functionalist approaches, such clauses cannot be analyzed as passives because the passive by definition is a construction that topicalizes patients. But syntactically the *di-* construction is clearly a passive: the patient has all the syntactic properties of a subject, and the agent has all the syntactic properties of an oblique argument. LFG takes both the syntactic relations and the pragmatic functions very seriously, but recognizes them as being distinct and logically independent of each other. Thus it is possible to ask, "What is the pragmatic function of the passive in language X" in a meaningful way, because the construction is not defined in terms of its pragmatic functions.

This is just one instance of a broader principle: languages can use the same syntactic constructions for quite different purposes. In a number of mainland Southeast Asian languages, the passive is used only for unfortunate events — the so-called ADVERSATIVE PASSIVE. In Biblical Hebrew, as in the Greek New Testament (probably due to Semitic influences), an agentless passive is often used as a way of describing God's actions without using any name to refer to God. Clearly the functions of the passive in Hebrew, Vietnamese, Malay and English are quite different from each other. I will not spell out the details here, but with a bit of imagination you can see that a translation from any one of these languages to any of the others which mechanically preserves active for active and passive for passive can lead to confusing (and sometimes hilarious) results.

In summary, I appreciate LFG's combination of precision and flexibility, its practical and realistic approach to syntactic analysis, and its attention to detail. All of these are important features of a good descriptive grammar.

References

Alsina, Alex and Sam A. Mchombo. 1993. Object asymmetries and the Chichewa applicative construction. In Sam Mchombo, ed., *Theoretical aspects of Bantu grammar*. Stanford, CA: CSLI Publications.

Arka, I Wayan, 2003. *Balinese morphosyntax: a Lexical-Functional approach*. Canberra: Pacific Linguistics.

Baker, Mark. 1988. *Incorporation: a theory of grammatical function changing*. University of Chicago Press.

Bresnan, Joan and Jonni M. Kanerva. 1989. Locative inversion in Chichewa: a case study of factorization in grammar. *Linguistic Inquiry* 20.1:1-50.

Cinque, Guglielmo. 1999. *Adverbs and Functional Heads: a Cross-linguistic Perspective*. Oxford: Oxford University Press.

Guilfoyle, Eithne, Hung, Henrietta, and Travis, Lisa. 1992. Spec of IP and Spec of VP: two subjects in Austronesian languages. *Natural Language and Linguistic Theory* 10:375-414.

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

Miller, Mark. 2007. A grammar of West Coast Bajau. PhD dissertation, University of Texas at Arlington.

Pike, Kenneth L. 1966. *Tagmemic and matrix linguistics applied to selected African languages*. Ann Arbor, MI: University of Michigan Center for Research on Language and Language Behavior; Office of Education, U.S. Dept. of Health, Education and Welfare.

Schachter, Paul. 1984. Semantic-role-based syntax in Toba Batak. In Schachter (ed.), *Studies in the structure of Toba Batak*. UCLA Occasional Papers in Linguistics No. 5. Los Angeles: Department of Linguistics, UCLA, pp. 122-149.

SINGLE CONJUNCT AGREEMENT AND THE FORMAL TREATMENT OF COORDINATION IN LFG

Jonas Kuhn and Louisa Sadler
University of Potsdam University of Essex

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

In cases of single conjunct agreement (SCA), the features of one conjunct within a coordinate structure control syntactic agreement between the coordinate NP and agreement targets external to that NP. This requires agreement processes to see inside the f-structure representation of the coordinate structure. Despite its intuitive simplicity, it has turned out to be surprisingly difficult to develop an approach to SCA in LFG, and existing approaches to SCA suffer from a range of technical inelegancies and/or empirical difficulties. We propose a novel approach to SCA which challenges the use of unordered sets for the representation of coordination at f-structure. Instead we propose a slightly more structured representation, which we call local f-structure sequences, for the representation of coordinate structures. Furthermore, we add more fine-grained subdistinctions to the standard LFG classification of features into distributive and non-distributive ones. This distinction controls the interpretation of feature path expressions as they interact with sets (of conjuncts) in f-structure, and the refined feature classification makes it possible to deal with SCA while keeping a simple, non-disjunctive formulation of the agreement constraints on the targets (such as verbs, which may combine with coordinate or simple NPs).

1 Introduction

In cases of single conjunct agreement (SCA), the features of one conjunct within a coordinate structure control syntactic agreement between the coordinate NP and agreement targets external to that NP (for example, a single conjunct like the feminine singular *kočka* (“cat”, FSG) in the Czech example (1) controls subject-verb agreement with the target verb *seděla* (“was sitting”, FSG)). This contrasts with the more familiar strategy of agreement with coordinate structures based on syntactic resolution, whereby the (resolved or calculated) features of the NP as a whole control agreement, i.e., an NP coordination of two or more singular NPs combines with a plural target verb etc.[1] In SCA it is usually the conjunct closest to the target which controls agreement (and SCA occurs much more frequently where the target precedes the controller as in (1)).

(1) Na rohožce seděla kočka a pes. Czech
on mat was.sitting.FSG [cat.FSG and dog.MSG]
“The cat and the dog were sitting on the mat.”

A crucial fact about SCA is that while a single conjunct controls certain syntactic agreement processes, other more semantically based agreement processes in the same language may be controlled by the resolved feature values of the coordinate structure as a whole. The Welsh example (2) demonstrates both strategies in a single sentence: subject-verb agreement with the target verb *gwelaist* (“saw”, 2SG) is an instance of SCA, while the anaphoric *eich hunain* (“yourselves”) agrees with the resolved features of the coordinate structure, 2nd person plural.

(2) Gwelaist ti a’th frawd eich hunain. Welsh
saw-2SG 2SG and-2SG brother 2PL self
2SG [2SG & 3SG] 2PL
“You and your brother saw yourselves.”

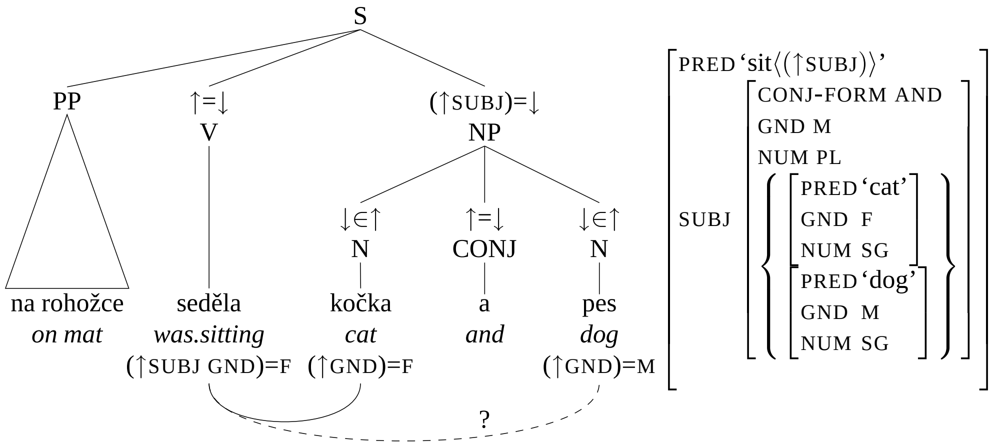
[*] Some of this work was carried out with the support of the AHRC under the research grant **Agreement and NP Coordination** (APN17606/AN10939) to Louisa Sadler (Essex) and Mary Dalrymple (Oxford), and with the support of the DFG under the Emmy Noether research grant **PTOLEMAIOS** to Jonas Kuhn (Potsdam).

[1] Single conjunct agreement is often referred to as partial agreement, terminology which is potentially confusing, since cases in which targets agree in person and gender but not in number are also referred to as partial agreement.

This observation indicates that both sets of features must be simultaneously available to control agreement in the syntax. As previous attempts at a precise formal account (which will be discussed in Section 3.2) show, it is an interesting challenge to accomodate these requirements while maintaining a general theory of agreement – i.e., without introducing a coordination-specific special case in the description of the agreement targets (e.g., the verb). It is, of course, desirable that agreement targets are blind to the status of their agreement controllers with respect to the presence or absence of coordination.

The (inconsistent) LFG analysis sketch for sentence (1) in (3) provides a visual characterisation of the issue: to maintain a principled account of agreement, the lexical annotation of the verb ((↑SUBJ GND)=F) should remain as it is. However, in the standard representation of a coordinate NP, the features available at the coordination level are only the resolved (masculine plural) features, and processes involving the conjunct-level features are only expected to work "across the board", i.e., assuming that all conjuncts would have to be marked identically as feminine singular. So, based on a standard conjunct-level process, (1) should be ungrammatical.

(3) *The challenge for an* LFG *analysis*



There are technical ways of making the features of the appropriate single conjunct available at the level of the coordinate structure (by introducing a distinction of different types of agreement features alongside each other, see Section 3.2.1), but to describe the full range of observable phenomena, the otherwise very elegant f-structure representation of coordinate structures has to be amended with a number of technical and construction-specific elements.

An alternative is to leave the f-structure representation unaltered, but introduce an explicit disjunction in the agreement constraints on the agreement target (the verb). This also has a very technical flavour and may run into problems with nested coordinations (Section 3.2.2).

The account we propose in this paper resolves the issue at a more abstract level. Previous attempts have left the status of unordered sets as the appropriate f-structure level representation for coordinate NPs unquestioned. It is one of the guiding principle of LFG that formal representations in particular components of the theory are chosen in such a way that they reflect the crucial empirical properties of the modelled phenomena: constituency is modelled by phrase structure trees (c-structure), which encode linear precedence and hierarchical structure in a natural way, while grammatical relations are modelled by feature structure representations (f-structure), which have no intrinsic concept of linear precedence. We appeal to this meta principle and argue that while plain, classical *sets* of f-structures model certain properties of coordinate NPs, some of the mathematical properties of sets turn out to be less adequate. If we use a slightly different formal device (which we call "local f-structure sequences", the descriptions used in the constraints describing a phenomenon like agreement may be interpreted in a way that caters more readily for the typological differences between languages we observe, including SCA).[2] By (slightly) altering the *interpretation* of the descriptional apparatus available for the formulation of constraints, we keep the case-by-case distinctions that would otherwise be required in agreement constraints, e.g. for verbs, "behind the scenes": no disjunctions are necessary in the actual lexical f-annotations or rule annotations. This is a move wholy in sympathy with the basic architectural philosophy of LFG.

The paper is organized as follows: after a discussion of the standard LFG analysis of coordination in Section 2, we review the relevant SCA data in some more detail in Section 3, as well as discussing previous LFG accounts. In Section 4, we present and discuss our approach, before concluding in Section 5.

2 Coordination in LFG

2.1 Set representation and distribution

The classical LFG analysis of coordination is based on two important assumptions (the former going back to Andrews (1983), the second due to Kaplan and Maxwell (1988/95)): First, the contribution of each conjunct is represented as an f-structure which forms an element of a *set* of f-structures. (A set of f-structures is defined as (a special case of) an f-structure.) The use of a set captures the fact that the number of conjuncts in a coordinate structure is in principle unbounded.

The second assumption combines this very intuitive *representational* idea with a clever notational short-hand, i.e., a way of allowing for simple, non-disjunctive *descriptions* in grammatical constraints. In terms of representation, no special feature is used to embed a coordinate structure's f-structure (set) within the larger f-structure that it contributes to: the set is inserted directly where the plain f-structure of a non-coordinated constituent would have gone (see (4), a slightly simplified f-structure for *I saw Bonnie and Clyde*). This captures the intuition that there is no difference between this sentence and *I saw the robbers* in terms of the depth of hierarchical embedding at the level of grammatical functions.

(4)
$$\begin{bmatrix} \text{PRED} & \text{'see}\langle(\uparrow\text{SUBJ})\ (\uparrow\text{OBJ})\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \end{bmatrix} \\ \text{OBJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'Bonnie'} \end{bmatrix} \begin{bmatrix} \text{PRED} & \text{'Clyde'} \end{bmatrix} \right\} \end{bmatrix}$$

The NP coordination rule is shown in (5): by virtue of the set membership annotations on the daughter NP nodes the entity referred to by ↑ is coerced into a set that will appear in the position of a non-coordinated NP's plain f-structure wherever the NP coordination rule is used.

(5)
$$\begin{array}{lllllllll} \text{NP} & \rightarrow & [& \text{NP} & \text{COMMA} &]^* & \text{NP} & \text{CONJ} & \text{NP} \\ & & & \downarrow\in\uparrow & & & \downarrow\in\uparrow & & \downarrow\in\uparrow \end{array}$$

It is important to note that this representational idea of coercing an f-structure into a set can only be made effective in the overall constraint-based system of LFG by a notational convention that Kaplan and Maxwell (1988/95) introduce: in the standard interpretation of f-structure path expressions, the application of a function (i.e., an f-structure feature like CASE, SPEC etc.) to a path denoting a set (like (↑OBJ) in (4), assuming that ↑ is referring to the outermost f-structure) is undefined.

[2] Effectively, we introduce a limited degree of sensitivity to string-level proximity to the resolution of f-descriptions in the cases where a description refers to an element of f-structure sequence (originating from a coordination).

Kaplan and Maxwell extend function application to sets of f-structures by providing a distributive interpretation for (otherwise undefined) path expressions referring to a set: whatever constraint such a path is used in, it will be applied to each of the set elements (a definition is shown in (6)). This extension provides a very elegant way of deriving the “across the board” effect that many constraints stated external to a coordinate structure show. In Figure 1, the rule annotation (↓CASE)=ACC is highlighted. Without the notational convention, this annotation (which would work correctly for a sentence like *I saw them*) would have no interpretation, since the f-structure referred to as ↓ is coerced into a set by the ↓∈↑ annotation further down the tree. With the notational convention, the desired effect follows directly with no further specification from the constraint as originally specified: all set elements have to satisfy the constraint (as **I saw her and he* is ungrammatical).

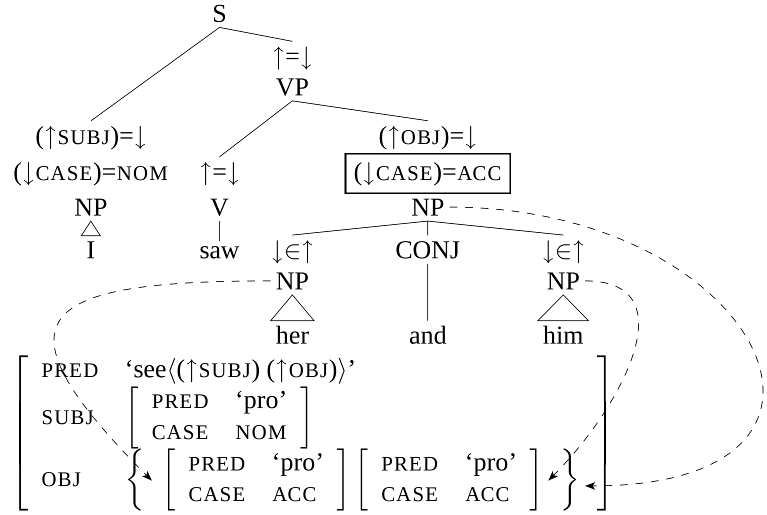


Figure 1: F-Descriptions referring to f-structure set

(6) *Notational convention (version 1)*

If some f-structure f is a set, then the value of an attribute a in f is v (that is, $(f\ a) = v$) iff for every $g \in f$, $(g\ a)$ includes the information in v.

(Kaplan and Maxwell, 1988/95), formulation from (Sells, 1985, 187)

There has been some debate (see e.g., (Maxwell and Manning, 1997)) about the formal relation implementing the intended distribution, with an earlier formulation in terms of generalization superceded by a formulation in terms of subsumption, but this is orthogonal to the current discussion and we do not discuss it further.

2.2 Distributive and non-distributive features

As the illustration in Figure 1 showed, morphosyntactic properties such as CASE distribute across conjuncts in a coordination, as do grammatical functions (even as part of functionally uncertain descriptions). On the other hand, some properties are coordination-level properties, that is they hold of the set itself, not of its members. This is true, for example, of the contribution of the conjunction, i.e., the CONJFORM feature.

Alongside CONJFORM, the agreement features PERS, NUM, GEND should be seen as properties of the coordination as a whole when a language applies a resolution strategy in agreement with coordinate NPs (as in English subject-verb agreement - *Bonnie and Clyde run/*runs*).

To capture this difference, a distinction is made between different classes of f-structure features (as discussed in Dalrymple and Kaplan (2000)): *distributive features* behave as just discussed, while *non-distributive features* are interpreted as properties of the f-structure set itself – i.e., constraints including such non-distributive features will not be distributed across set elements. Figure 2 provides an illustration; the convention is to show non-distributive features of a set inside an extra pair of square brackets surrounding the curly set brackets.

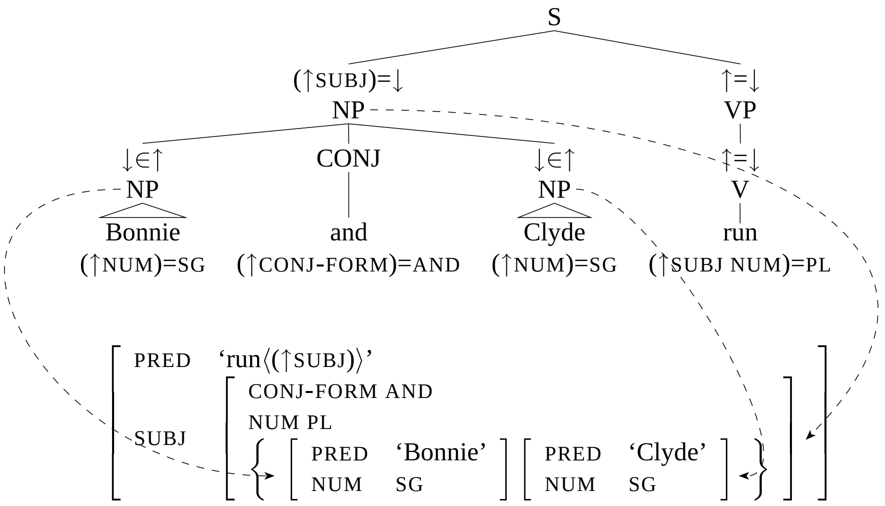


Figure 2: Non-distributive features

A formal LFG grammar must then include a declaration of the class that the features used belong to (typically, the distributive class is assumed to be the default, so just the non-distributive features have to be declared explicitly). Moreover the formulation of the notational convention (6) has to be replaced by (7):

(7) *Notational convention (version 2)*

Interpretation of $(f\ a) = v$:

- If (i) f is a plain f-structure,
  or (ii) f is a set and a is declared as a non-distributive feature,
  then $(f\ a)$ includes the information in v.
- Otherwise, if f is a set, then for every $g \in f$, $(g\ a)$ includes the information in v.

2.3 Semantic and morphosyntactic aspects of agreement

Before proceeding to a more detailed discussion of SCA in Section 3, we mention two further recent innovations which are relevant to coordination and motivated by agreement processes, but largely orthogonal to the specific issues of SCA.

King and Dalrymple (2004) (building on (Wechsler and Zlatić, 2000, 2003) and other work in HPSG) introduce a representational distinction between two sets of agreement features (compare (8)): (i) features under INDEX – for resolved features at the coordination level (which are semantically determined); and (ii) features under CONCORD – for morphosyntactic conjunct-level features. The former are generally declared as non-distributive, the latter as distributive. This representational distinction makes it possible to distinguish different agreement strategies within the same language by formulating constraints using the appropriate paths. Figure 3 illustrates the role of CONC(ORD) in English NP-internal conjunct-level agreement in number. English subject-verb agreement, on the other hand, involves (resolved) INDEX features, so a plural verb form would include the annotation (↑SUBJ INDEX NUM)=PL in agreement with a coordination of singular NPs. The use of this explicit representational distinction between two sets of agreement features provides for a fine-grained approach to agreement cross-linguistically, as demonstrated by King and Dalrymple (2004) and Dalrymple and Nikolaeva (2006).

(8)
$$\left[\begin{array}{ll} \text{INDEX} & \left[\begin{array}{l} \text{PER}\ldots \\ \text{NUM}\ldots \\ \text{GEN}\ldots \end{array}\right] \\ \text{CONC} & \left[\begin{array}{l} \text{PER}\ldots \\ \text{NUM}\ldots \\ \text{GEN}\ldots \end{array}\right] \end{array}\right]$$

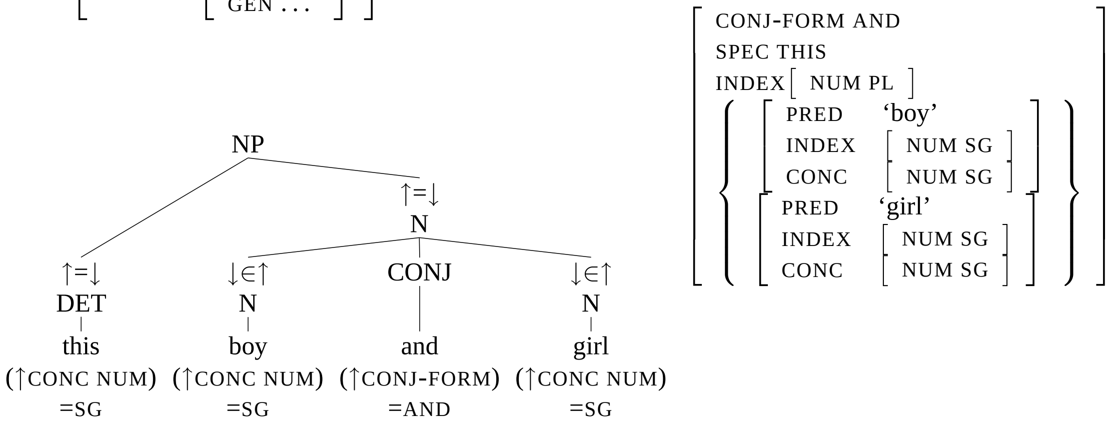


Figure 3: INDEX/CONCORD distinction following (King and Dalrymple, 2004)

Note that since all conjuncts are treated alike for the CONCORD features, the INDEX/CONCORD distinction does not itself provide a solution for SCA.

The second extension is due to Dalrymple and Kaplan (2000), who introduce the notion of (closed) marker sets as feature values for morphosyntactic features (instead of standard atomic values). For instance, in the Romance languages, the value of GEND may be either $\{\}$ (for feminine) or $\{\text{M}\}$ (for masculine). With such a representation, feature resolution can be simply modelled by set union over each conjuncts marker set – yielding $\{\}$ if all conjuncts are feminine, or $\{\text{M}\}$ if there is at least one masculine conjunct.

The elements making up the marker sets will differ depending on the distinctions a particular language makes. A possible encoding for PERS values is given in (9). The NP rule based on this analysis is shown in (10). (11) is an example f-structure for *you and John* (note that NUM is treated differently – it is assumed to be semantically rather than syntactically resolved).

(9)
- 1st person: $\{\text{S}\}$ [inclusive 1st person plural: $\{\text{S, H}\}$]
- 2nd person: $\{\text{H}\}$
- 3rd person: $\{\}$

(10)
$$\text{NP} \longrightarrow \begin{array}{ccc} \text{NP} & \text{CONJ} & \text{NP} \\ \downarrow\in\uparrow & & \downarrow\in\uparrow \\ (\downarrow \textsc{index pers}) \subseteq (\uparrow \textsc{index pers}) & & (\downarrow \textsc{index pers}) \subseteq (\uparrow \textsc{index pers}) \\ (\downarrow \textsc{index gen}) \subseteq (\uparrow \textsc{index gen}) & & (\downarrow \textsc{index gen}) \subseteq (\uparrow \textsc{index gen}) \end{array}$$

(11)
$$\begin{bmatrix} \textsc{index} \begin{bmatrix} \textsc{pers} & \{\textsc{h}\} \\ \textsc{num} & \textsc{pl} \end{bmatrix} \\ \left\{ \begin{bmatrix} \textsc{pred} & \text{'pro'} \\ \textsc{index} & \begin{bmatrix} \textsc{pers} & \{\textsc{h}\} \\ \textsc{num} & \textsc{sg} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \textsc{pred} & \text{'John'} \\ \textsc{index} & \begin{bmatrix} \textsc{pers} & \{\} \\ \textsc{num} & \textsc{sg} \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix}$$

3 Single Conjunct Agreement

3.1 The challenge posed by SCA data

In SCA, grammatical agreement with a coordinated NP is based on the features of just one of the conjuncts. This poses a number of challenges for an LFG analysis of agreement phenomena: the agreement features of the target are conceptually at the conjunct-level (like distributive features) but do not distribute across all conjuncts – so CONCORD is not the correct analysis. Moreover, as we saw in (2) evidence of contrast with other agreement processes (such as pronominal anaphora) may show that the controlling features in cases of SCA are also distinct from the resolved INDEX features of the coordinate NP as a whole. The distinction between distributive and non-distributive features does not, therefore, account for the the behaviour of the agreement features in cases of SCA.

SCA is not a marginal phenomenon but rather it is found in a broad range of languages, sometimes as an option alongside other patterns. The following brief overview is by no means exhaustive. Welsh, Irish and other Celtic languages show rightward SCA in predicate-argument agreement. In these languages, V, N and P heads preceding coordinated pronominal NPs agree with the initial conjunct (McCloskey, 1986; Rouveret, 1994; Sadler, 1999, 2003). Standard Arabic has the option of closest conjunct agreement in VS order, but uses resolved agreement in SV constructions (Aoun et al., 1994, 1999; Munn, 1999), and SCA is also found in Arabic vernaculars. It is described for Ndebele (Moosally, 1998) and Swahili (Marten, 2000, 2005). There is a variety of SCA data described for a number of Slavic languages, including some cases in Slovene of first conjunct agreement with target to the right ("furthest conjunct agreement") (Corbett, 1983, 1988). Portuguese has an option of both rightward and leftward closest conjunct agreement in head-modifer constructions (i.e. within NP) (Villavicencio et al., 2005) and other cases of SCA in Spanish and Portuguese are discussed in Camacho (2003) and Munn (1999)

Since there has been an extensive discussion of the data in the literature, we will here expand only on two particularly challenging observations: the option of "double edged" SCA in Portuguese, and the option of "furthest conjunct agreement" in Slovene.

3.1.1 Double edged single conjunct agreement

A particularly interesting pattern of SCA within coordinate NPs is found in Portuguese (de Almeida Torres, 1981; Villavicencio et al., 2005). Alongside the standard strategy of resolution illustrated in (12) (the resolution gender for MASC/FEM combinations is MASC), SCA is an option for *postnominal* adjectives, which then agree with the closest conjunct, as shown in (13).

(12) a. a parede e a janela vermelhas/*vermelhos
[the.FSG wall.FSG and the.FSG window.FSG] red.FPL/red.MPL
"the red wall and window"

b. a parede e o teto coloridos
[the.FSG wall.FSG and the.MSG ceiling.MSG] coloured.MPL
"the coloured wall and ceiling"

(13) a. estudos e profissão monástica
[studies.MSG and profession.FSG] monastic.FSG
"monastic studies and profession"

b. As maldições se cumpriam no povo e gente hebreia
The curses REFL fell [in the.MSG people.MSG and persons.FSG] Hebrew.FSG
"The curses fell on the Hebrew people."

c. O objectivo está claro: é perder, em pouco tempo, os quilos e as dobrinhas
the objective is clear: is to lose in little time, [the.MPL kilos.MPL and the fatty tissue.FPL]
acumuladas no inverno.
accumulated.FPL in the winter
"The objective is clear: to lose quickly the kilos and fat accumulated during the winter."

In situations where *prenominal* modifiers take scope over both conjuncts, SCA is highly preferred, and agreement is with the closest conjunct (14). However, some examples showing resolved number agreement (15) were found in the corpus study of Villavicencio et al. (2005).

(14) a. suas próprias reações ou julgamentos
his.FPL own.FPL [reactions.FPL or judgement.MPL]

b. pequenas partículas ou átomos
small.FPL [particles.FPL or atoms.MPL]

(15) a. os novos chefe e vice-chefe
the.MPL new.MPL [chief.MSG and vice-chief.MSG]

b. claras maioria e oposição
clear.FPL [majority.FSG and opposition.FSG]

Since modifiers at both the left and the right of a coordinated NP may show SCA, it follows that it may occur simultaneously *both prenominally and postnominally* with a single coordination.[3]

(16) a. Reconhecendo que a garantia de um tratamento igual para as mulheres e homens
recognising that the guarantee of a treatment equal for the.FPL [women.FPL and men.MPL]
refugiados pode exigir acções específicas a favor das mulheres.
refuge-.ADJ.MPL could demand actions specific to favour of the women
"To recognize that the guarantee of an equal treatment of the female and male refugees could make specific actions in favour of the women necessary."
http://www.cidadevirtual.pt/acnur/acn_lisboa/excom64.html

b. Os mitos e lendas brasileiras
the.MPL [myth.MPL and legend.FPL] Brazilian.FPL

[3] Of course, since MASC is the resolution gender, the occurrences of MASC agreement in these examples could be due to resolution operating alongside SCA within the same structure, but even under the most restrictive interpretation there is clear evidence of SCA both to the left and to the right in Portuguese NPs.

3.1.2 First Conjunct Agreement with rightward targets

A robust generalization about SCA seems to be that it occurs much more frequently in structures in which the agreement target precedes the agreement controller. Cases of SCA where the target is to the right of the nominal controller include the following from Slovene. These examples are also crosslinguistically unusual in that it is the furthest (rather than the closest) conjunct which is controlling agreement.

(17) a. Groza in strah je prevzela vso vas.
[horror.FEM.SG and fear.MASC.SG] has seized.FEM.SG the-whole village
"Horror and fear have seized the whole village." (Corbett 1983: 180)

b. knjige in peresa so se podražile.
[book.FEM.PL and pen.NEUT.PL] are selves got dear.FEM.PL
"Books and pens have become more expensive." (Corbett 1988: 26)

3.1.3 Data Summary

In summary, we see that the existence of patterns in which a single conjunct controls (some) agreement processes considerably complicates the array of possible strategies for syntactic agreement with (nominal) coordinate structures. In addition to (18-1) and (18-2) we must accommodate a number of further patterns.

(18)
1. Agreement with resolved coordination-level features
2. Grammatical concord at conjunct-level with *all* conjuncts (distributed)
3. Closest conjunct agreement (leftward or rightward)
4. Double edged closest conjunct agreement (both leftward and rightward)
5. First conjunct agreement with rightward targets [rare]

3.2 Previous approaches

Previous work on this phenomenon in LFG has explored a range of possible approaches to SCA using LFG's standard formal devices. We briefly outline some of this work in this section and suggest that it has a number of shortcomings before outlining a rather different approach in the following section. Existing approaches can be classified as *representation-based approaches* (Sadler, 1999, 2003; Villavicencio et al., 2005) or *description-based approaches* (Sadler, 1999, 2003; Falk, 2006); Asudeh (2005) can be characterised as a *"mixed" approach*.

3.2.1 Representation-based approaches

Representation-based approaches encode the agreement features of the appropriate single conjunct on the coordination structure as a whole, as the value of an additional feature alongside the INDEX feature which contains the resolved features of the coordinate structure as a whole. The appropriate annotations are specified on the coordination rule and agreement constraints are specified in terms of this additional feature. For example, in the following, AGR is the feature carrying the agreement features of the distinguished conjunct (Sadler, 1999, 2003).[4]

[4] The template call @ NP-CONJUNCT ensures that the standard f-annotations as in (10) are inserted for each conjunct.

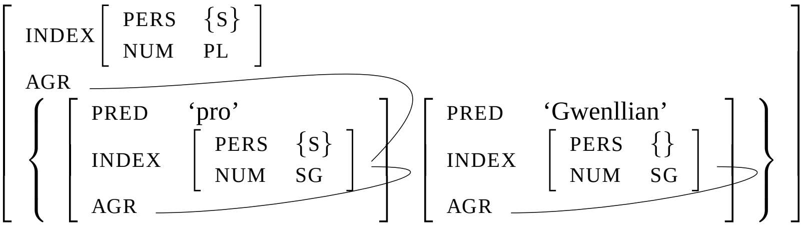
$$\text{(19)}\quad \text{NP} \longrightarrow \begin{array}{ccc} \text{NP} & \text{CONJ} & \text{NP} \\ @\ \text{NP-CONJUNCT} & & @\ \text{NP-CONJUNCT} \\ (\downarrow\text{IND}) = (\uparrow\ \text{AGR}) & & \end{array}$$

Predicate argument agreement is expressed as a constraint over the AGR feature and the constraint in (20) is intended to feed the appropriate value to AGR in non-coordinate structures, so that the agreeing predicate outside the coordination can simply constrain AGR features irrespective of whether or not the argument is itself a coordination. Other agreement processes, notably pronominal and reflexive anaphora, involve the INDEX feature rather than the AGR feature. The representation of the coordinate NP in an example such as (21) is given in (22).

(20) *Constraint on Nominal Lexemes:* (↑ INDEX) = (↑ AGR)

(21) Dw i a Gwenllian heb gael ein talu. Welsh
am.1S [1S and Gwenllian] without get 1PL pay
"Gwenllian and I have not been paid."

(22)



There are several drawbacks of this approach. It introduces a technically motivated feature-passing mechanism into the f-structure representation, something which LFG tries in general to avoid, and which is as problematic and inelegant as any other book-keeping feature. This non-distributive, coordination-level feature (AGR) is used to encode properties of an individual conjunct, blurring what is otherwise a clear conceptual separation between distributive (individual) and non-distributive (resolved) agreement features. Moreover double edged closest-conjunct agreement such as that which arises in Portuguese NPs can only be captured with a highly technical book-keeping representation using two sets of single-conjunct agreement features at the coordination level.

Postnominal ADJ agreement in Portuguese might be treated as follows on this approach:[5]

$$\text{(23)}\quad \text{NP} \longrightarrow \begin{array}{cc} \text{NP} & \text{A+} \\ \uparrow\ = \downarrow & \downarrow \in \uparrow \text{ADJ} \\ & (\downarrow\ \text{A-POSN}) =_c \text{POSTNOM} \end{array}$$

$$\text{(24)}\quad \text{NP} \longrightarrow \begin{array}{ccc} \text{NP} & \text{Conj} & \text{NP} \\ \downarrow \in \uparrow & \uparrow = \downarrow & \downarrow \in \uparrow \\ (\downarrow\ \text{IND GEN}) \subseteq (\uparrow\ \text{IND GEN}) & & (\downarrow\ \text{IND GEN}) \subseteq (\uparrow\ \text{IND GEN}) \\ & & (\downarrow\text{IND }) = (\uparrow\ \text{LAGR}) \end{array}$$

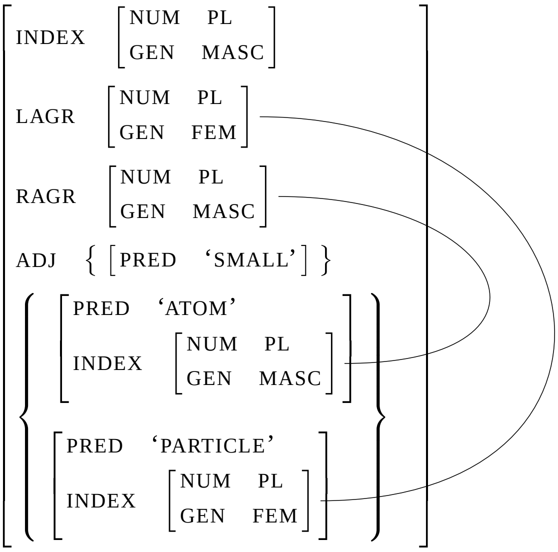
(25) *acumuladas* (↑ PRED) = 'ACCUMULATED'
(↑ A-POS) = POSTNOM
{ ((ADJ ∈ ↑) LAGR GEND) = FEM
((ADJ ∈ ↑) LAGR NUM) = PL
| ((ADJ ∈ ↑) IND GEND) = FEM
((ADJ ∈ ↑) IND NUM) = PL }

[5] Note that (25) includes inside-out designators for "leaving" the ADJUNCT set in which the adjective is introduced – these have nothing to do with the set we are dealing with for coordination.

To allow for percolation of features from both the rightmost and the leftmost conjunct you need in fact to distinguish both LAGR and RAGR at the level of the coordination (Villavicencio et al., 2005), leading to an f-structure along the following lines.

(26) pequenas partículas ou átomos
small.FPL [particles.FPL or atoms.MPL]

(27)



3.2.2 Description-based approaches

In contrast to a representation-based approach, a description-based approach would assume no additional representation either at the level of the coordinate structure or at the level of the individual conjunct. One possibility is to resort to a more complicated description on the agreement target itself to ensure that it is the agreement features of the distinguished conjunct that are picked up when the argument is a coordinate structure. The appropriate agreement controller might be picked out using f-precedence (Sadler, 1999; Falk, 2006).

(28) *F-precedence* (Kaplan and Zaenen, 1989/95)
f_1 f-precedes f_2 if and only if there are c_1 and c_2 such that c_1 is a rightmost element in $\phi^{-1}(f_1)$, c_2 is a rightmost element in $\phi^{-1}(f_2)$, and c_1 precedes c_2. (formulation of Bresnan (2001))

Description-based approaches formulate constraints in the lexical entry of agreement target (outside the coordination) that agree directly with the distinguished conjunct.[6] Thus in the case of Welsh, the V, N, or P expresses constraints over the linearly first member of the coordinate structure.[7] Consider for example the case of prepositional agreement in (29).

[6] A possible alternative is to use a tree-logic description of the SCA configuration (Kuhn, 2003) to avoid recourse to f-precedence. However this means using a very powerful tool to describe what is intuitively quite a simple relationship. Since agreement is clearly an f-structure phenomenon, any attempt to adress SCA by explicitly referring to c-structural aspects of the coordinate structure *in the lexical descriptions of the targets* does not seem to be quite appropriate.

[7] This account abstracts away from a number of subsidiary complications in the Celtic data involving the interaction with unbounded dependency constructions. Since these are orthogonal to our essentially formal point about types of approaches to SCA we do not discuss these intricacies further here.

(29) Roedd Wyn yn siarad amdanat ti a Siôn. Welsh
was.3S Wyn PROG speak about-2S [2S and Siôn]
"Wyn was talking about you and Siôn."

The relevant agreement constraint on the agreement target (the preposition, which agrees with its OBJ), would require something along the following lines:[8]

(30) *amdanat* (↑ PRED) = 'ABOUT⟨(↑ OBJ)⟩'
{ (↑ OBJ) = %A
| %A ∈ (↑ OBJ)
¬[(↑ OBJ ∈) $<_f$ %A] }
(%A PERS) = 2
(%A NUM) = SG

Although this sort of approach avoids the need to litter the representation with otherwise unmotivated features, a description-based approach along these lines necessitates a disjunctive formulation of all agreement constraints (on the target side), to allow for the presence or absence of coordination in the argument. Since the agreement contraints pick out the distinguished conjunct (from the set) directly, the account as outlined above does not generalize directly to cases of nested coordination, and such cases do exist, as shown in (31).

(31) Wyt ti a fi neu Peter a Mary yn mynd i ennill.
is.2SG [[2SG and 1SG] or [Peter and Mary]] PT go PT win
"Either you and I or Peter and Mary are going to win."

F-precedence itself is a very powerful tool and provides a very indirect way of referring to something that is intuitively quite simple, namely finding the "leftmost" or "rightmost" set element.

3.2.3 "Mixed" SCA account

Finally, it should be pointed out that elements of the description-based and the representation-based accounts can be combined, as they are in the proposal of Asudeh (2005). Asudeh proposes a representation in which the distinguished conjunct is both a member of the set (corresponding to the coordinate structure) and also the value of an additional attribute SEED within this same (hybrid) structure. He argues that this is independently motivated by its role as the "seed" in meaning construction for conjunction (Asudeh and Crouch, 2002). The representation of the coordinate NP in (32) is shown in (33).

(32) Daethost ti a Siôn.
came-2S [2S and Siôn]
"You and Siôn came."

[8] The use of ∈ in a feature path description such as "(↑ OBJ ∈)" in (30) allows one to pick up an arbitrary set element. The notation "%A" introduces a local variable for an f-structure, so the reference can be fixed across several f-equations. "¬[(↑ OBJ ∈) $<_f$ %A]" says that there cannot be any set elements f-preceding the f-structure fixed as %A – so it must be the leftmost element.

(33)
$$\left[\begin{array}{ll}
\text{CONJ AND} & \\
\text{INDEX} & \left[\begin{array}{ll}\text{NUM} & \text{PL}\\ \text{PERS} & 2\end{array}\right]\\
\text{SEED} & \\
\multicolumn{2}{l}{\left\{ \left[\begin{array}{ll}\text{PRED} & \text{‘pro’}\\ \text{INDEX} & \left[\begin{array}{ll}\text{NUM} & \text{SG}\\ \text{PERS} & 2\end{array}\right]\end{array}\right] \left[\begin{array}{ll}\text{PRED} & \text{‘Sion’}\\ \text{INDEX} & \left[\begin{array}{ll}\text{NUM} & \text{SG}\\ \text{PERS} & 3\end{array}\right]\end{array}\right] \right\}}
\end{array}\right]$$

The relevant conjunct is picked out as the seed conjunct by means of constraints using f-precedence in the annotation of the conjunction, shown in (34), and the agreement constraints on the target make reference to the SEED (in the case of coordinate structures).

(34) *and* $(\uparrow \text{CONJ}) = \text{and}$
$(\uparrow \text{SEED}) = (\uparrow \in)$
$\neg\,[\,(\uparrow \in) <_f (\uparrow \text{SEED}\,)\,]$

(35) *daethost* $(\uparrow\ \text{PRED}) = \text{‘COME}\langle\text{SUBJ}\rangle\text{’}$
$(\%\text{A PERS}) = 2$
$(\%\text{A NUM}) = \text{SG}$
$(\%\text{A PRED FN}) = \text{PRO}$
$((\ \%\text{A PRED}) = \text{‘PRO’})$
$\{\,(\uparrow \text{SUBJ SEED}) = \%\text{A} \mid (\uparrow \text{SUBJ}) = \%\text{A} \wedge\ \neg\,(\uparrow\ \text{SUBJ SEED})\,\}$

Note that on this account too, a disjunctive statement is needed for the agreement constraints (checking for the presence of a SEED feature in the last line of (35)). Moreover it seems that both nested coordinations and double edged SCA are problematic on this account.

3.2.4 Previous accounts: summary

We conclude that all previous accounts of SCA, which are based on the standard unordered set-based analysis of coordination, suffer from a range of technical inelegancies and/or empirical difficulties.

4 Proposal

As discussed in Section 1, the challenge posed by SCA is to keep the statement of agreement constraints maximally general (i.e., to avoid disjunctive, coordination-specific descriptions on the agreement targets such as verbs) while at the same time avoiding the augmentation of the f-structure representation for NP coordination with purely technical book-keeping features.

The solution we propose is a description-based approach in that it does not involve the addition of any additional features or embeddings in the f-structure representation. It does however change the formal character of the original set representation assumed for the coordinate structure. Sets are unordered, so there is no formal way of singling out particular elements by an operation that applies directly to the set. This makes it necessary to employ auxiliary constructs, i.e., either adding one or more explicit technical features at the coordination level, or indirectly falling back to c-structure (via f-precedence or possibly tree-logic descriptions) to recover linear precedence information that is missing from f-structure.

It is a meta principle of LFG to use formal representations displaying the desired modelling properties for the various parts of grammatical theory. We think that the technical issues posed by SCA for the standard set representation in coordination may be an indication that in this case, the formal properties of the representational device chosen are simply not fully adequate.[9] We propose the use of a slightly more structured representation for the collection of conjunct f-structures, what we call "local f-structure sequences" (lfsq's).

4.1 Singling out the contribution of a particular conjunct at f-structure

It is uncontroversial that agreement is a phenomenon that should be encoded at the level of f-structure. Since SCA facts show that agreement can be not only sensitive to (i) properties of the coordinate structure as a whole, or (ii) common properties of all the individual conjuncts – i.e., distributive properties, but also (iii) to properties of only the first or last conjunct, there should be a direct way *in f-structure* of picking up the relevant properties of the first or last conjunct.

One way of implementing this would be to define a new notation for deterministically picking up a particular element of an ordered set (maybe $(\uparrow \text{OBJ} \in_{\mathit{first}} \text{GEND})$ for picking up the first element's gender information).[10] This would however be a much more powerful tool than what the SCA issues seem to call for. At the same time, the notation would not resolve the issue that agreement constraints on the agreement target (the verb) should not be disjunctive, distinguishing a coordination and non-coordination case.

What we propose instead is to use the same notational "trick" that Kaplan and Maxwell (1988/95) used to account for the extended space of possible interpretation that a constraint like $(\uparrow \text{OBJ NUM}){=}\text{SG}$ can have. Ultimately, we will provide a third version of the notational convention (after (6) and (7)) that will make it possible to use this plain function application notation in the description of agreement targets like verbs, with the effect that they will be interpreted either (i) non-distributively at the coordination level, (ii) by classical conjunct-level distribution, or (iii) according to the SCA strategy, i.e., as applying to just one particular conjunct.

To trigger the various options, we assume a more fine-grained feature distinction affecting conjunct-level f-descriptions (the classical distributive feature descriptions): besides distributive features, there is a new type of "overlay" features, for which only the first or last element of the coordination representation is taken into account (the term "overlay" suggests that when looking at the coordination representation from the left or right, only the features of the peripheral element become

[9] Several of the properties that come with a set representation are questionable to a certain extent. For space reasons we only list them here briefly:

(i) No order among elements:
"I met Sue$_i$ and her$_i$ sister." vs. "I met her$_{*i}$ sister and Sue$_i$."
"Bill went to the city and rented a bike." vs. "Bill rented a bike and went to the city."

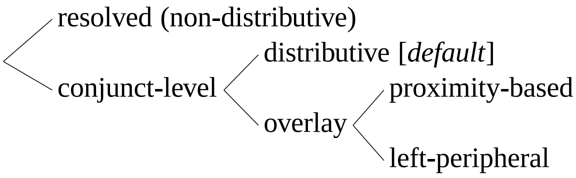
(ii) No duplicates
(The effect of this cannot normally be seen in LFG due to the instantiated interpretation of PRED values, but intuitively it is not clear why there should be the principled possibility of having two conjuncts that map to the same set element.)
"Our Wednesday schedule is Biology, Maths, Maths and French"

(iii) No reference to a specific element possible from "outside" (only an arbitrary member can be picked); this is the issue brought up by SCA.

[10] The ordering of the set would have to be defined as the elements are added by $\downarrow\in\uparrow$ constraints. For this, the notion of head precedence could be used, as implemented in XLE in order to model for instance the scoping order of adjectival modifiers in an ADJUNCT set. By specifying membership as "$\downarrow\in_{<_h<_s}\uparrow$" (in XLE notation: "`! $<h<s ^`"), scoping relations among the f-structure set elements are added, which follow the surface (head-precedence) order. An approach for SCA along these lines was suggested by Ron Kaplan in the discussion of the present paper at LFG07. Head precedence is computationally more manageable than full-fledged f-precedence, as the c-structural location at which the instantiated PRED value for an f-structure is introduced provides a clear, unique anchoring point.

visible, hiding the conjunct-level features of the other elements).[11] Features can be declared to be from one of the following classes:

(36) *Classification of features*



4.2 Aspects of locality

In principle, the intuitive "overlay" effect could be modelled in a standard set, augmented with some concept of precedence (either f-precedence, or head precedence, cf. Footnote 10): one could define an alternative to distribution that will pick out the first or last (in terms of the precedence relation) set element instead of distributing to all elements. (The decision of picking out the first vs. last element would have to be taken care of in the grammatical constraints describing agreement targets.)

However, this would not seem to model the effect that is really at play: Imagine a hypothetical c-structural configuration in which set elements are introduced not just within a single coordinate (NP) structure, but there are several, spatially separated exponents in c-structure, each of which contributes one or more elements to the same set in f-structure (a situation in which such an analysis would not be *entirely* unreasonable might be the case of extraposed "additions" to a coordinate structure like in "I saw an elephant and a zebra at the zoo yesterday, and a giraffe" – for our thought experiment, we would put the f-structures from *an elephant*, *a zebra* and *a giraffe* all in the same f-structure set under OBJ). If this phenomenon interacts with SCA and the agreement target sits between c-structural contributors to the set, would we expect that the single conjunct controlling agreement would always be the *globally* peripheral set element? This is what a general set-based approach would force us to assume. If a language has a strategy of SCA with targets preceding the controlling coordinate structure, we would rather expect that the next conjunct to the right of the target would be controlling agreement.

We believe that global aspects of precedence are not the driving force behind SCA. The fact that an overwhelming proportion of SCA phenomena are indeed *closest* conjunct agreement and the existence of double edged SCA point to a c-structural *proximity* effect, which requires a more local account.[12]

4.3 Local f-structure sequences

In order to be able to derive the locality effect, it is not sufficient to compare the set elements in terms of precedence. The relative position of the agreement target has to be taken into account too. We propose a technical solution that folds this check of proximity into the notational convention of function application in the presence of a coordination structure (a revised form of (6) and (7)).

In order to have a handle on the left and right edge of a coordination structure (from the point of view of f-structure), we assume a somewhat more structured and constrained representation structure

[11] A further distinction of overlay feature into proximity-based and peripheral features becomes necessary in order to be able to model the rare "furthest conjunct agreement" phenomenon.

[12] The rare cases of "furthest conjunct agreement" could be seen as evidence for a global effect. However, they would also be compatible with a salience-based, more local explanation: in these special coordinate structures, the speaker's attention may be attached to the first element and agreement – following speaker's attention – will "skip" the closer conjunct. For such a speaker's attention-based account (which would certainly require further elaboration), a locally confined representation would also seem more appropriate than a global representation.

than the classical set: what we call *local f-structure sequences* (lfsq's). Elements are ordered, and crucially, reference to the first and last element is possible: f_L, f_R (this will be used only in the notational convention however, so there seems to be no need to introduce new designators to LFG's functional description language). In order to exclude the puzzling hypothetical cases of multiple exponence for a single coordination, we posit that lfsq's have a unique anchoring point in c-structure, which has to be explicitly defined in the annotations of the coordination rule (see the *lfsq*($\uparrow$,$\mathcal{M}*$) annotation in (37), which defines the upper NP node (the mother of the conjunction) as the anchoring point for the lfsq referred to by $\uparrow$).[13]

(37) *Coordination rule*

$$\begin{array}{ccccc} \text{NP} & \longrightarrow & \text{NP} & \text{CONJ} & \text{NP} \\ & & \downarrow\in_{lfsq}\uparrow & lfsq(\uparrow,\mathcal{M}*) & \downarrow\in_{lfsq}\uparrow \\ & & & \uparrow=\downarrow & \end{array}$$

We are now in a position to formulate the refined version of the notational convention for the interpretation of function application ((38) below). Since the unique c-structural anchoring point of an lfsq is known (and can be stored during the process of f-structure constraint resolution), we can compare the anchoring point's string range (the word index of the first and last words it dominates) with the string range of any other node – in particular the nodes at which an f-description is introduced that includes a function application "entering" the coordinate structure representation, as it is introduced by the agreement target. For the proximity-based features, the interpretation of a path description will depend on the relation between the two string ranges: the left-most f-structure element of the lfsq will be picked in case the path description is to the left of the anchoring point; the right-most if it is to the right. The description may also originate from inside the coordinate structure, in which case a direct interpretation is chosen.

(38) *Notational convention (version 3)*

Interpretation of $(f\ a) = v$ in an f-annotation at a node with string range s_i-s_j:

- If (i) f is a plain f-structure, or (ii) f is a set or lfsq and a is declared as a **non-distributive** feature, then $(f\ a)$ includes the information in v.
- Otherwise, if f is a set or lfsq:
  - if a is declared as a **left-peripheral** feature, then $(f_L\ a)$ includes the information in v;
  - if a is declared as a **proximity-based** conjunct-level feature:
    - if f's c-structure anchor *precedes* string range s_i-s_j, then $(f_L\ a)$ includes the information in v;
    - if the anchor *follows* the string range, then $(f_R\ a)$ includes the information in v;
    - if the anchor falls into the string range: $(f\ a)$ includes the information in v
  - if a is **distributive**, then for every $g \in f$, $(g\ a)$ includes the information in v.

Since the technical distinctions are defined once and for all as a part of the general machinery (they are "behind the scenes" from the point of view of actual grammar specification), the descriptions needed to deal with SCA in lexical or rule annotations become surprisingly simple. Figures 4 and 5 show essentially the full set of required annotations if in the language under consideration, the agreement features are defined as proximity-based. The target's agreement description is propagated down to

[13] Alternatively, one could introduce the convention that the mother node for a $\downarrow\in_{lfsq}\uparrow$ annotation automatically becomes the anchoring point.

the appropriate set element, thanks to the more fine-grained notational convention. (What is ignored here is the INDEX/CONCORD distinction, which is compatible with the proposed modification of the formalism, and which is required to account for the simultaneous existence of non-distributive and conjunct-level features in one language.)

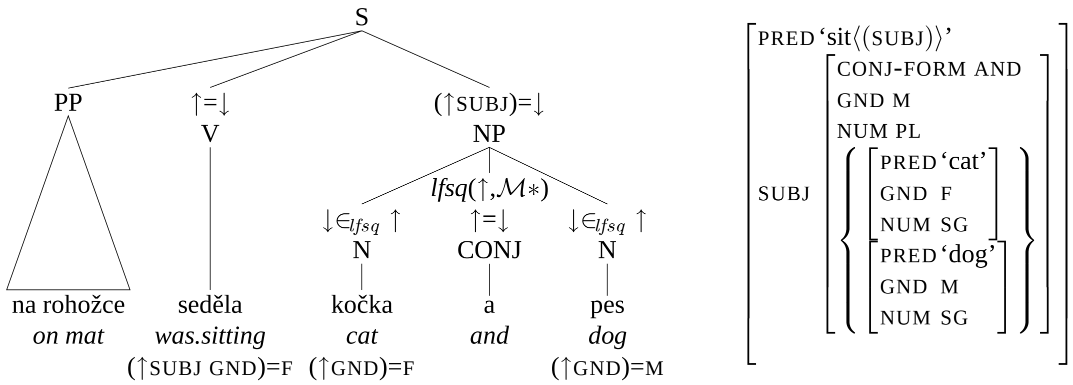


Figure 4: Representations and descriptions in SCA (example (1))

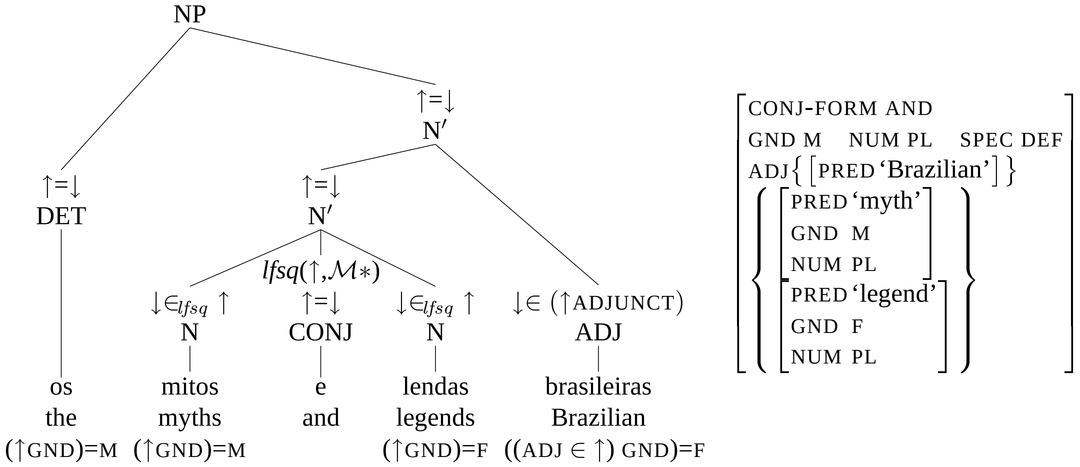


Figure 5: Representations and descriptions in double edged SCA (example (16b))

4.4 Discussion: changes to the LFG formalism

We think that the proposed account follows the original spirit of LFG – division of labour between representation and description, and the assumption of appropriate formal devices to represent the linguistic properties of the described entities. A limited degree of sensitivity to string-level proximity is introduced to the resolution of f-descriptions; this constitutes a considerable change in the character of f-structural constraints. However, the characteristics of the SCA phenomenon suggest that agreement *is* more sensitive to proximity than the classical division of labour between c-structure and f-structure allows the grammarian to express (other than in a rather round-about way). By introducing carefully controlled string precedence conditions in the notational apparatus, the original intuitive constraint formulation can be kept up for agreement in general – now extending to SCA.

The mechanism is less expressive than f-precedence (which is computationally problematic[14]), but more focused on the generalizations underlying the data.

5 Conclusion

We proposed an alternative way of looking at a long-standing issue in LFG and constraint-based theories of SCA more general. All attempts of formalizing SCA within the standard framework for coordination seem overly technical and unintuitive (at least when applied to double edged SCA like in Portuguese, or to nested coordinations).

By introducing a limited sensitivity to string proximity into the *interpretation* of f-descriptions as they are resolved in model construction, the original concise and intuitive descriptions for agreement constraints can be recovered.[15] Our analysis is essentially description-based, since the crucial effects are brought about without adding technical bookkeeping devices to the representation. In order to be able to make this change, we made small adjustments to the formal character of the representation structures assumed: we replaced sets in the f-structure representation of coordinate structures by local f-structure sequences (lfsq's). This is an example of taking LFG's meta principle seriously that underlying representation types and means of description should be chosen to match the needs from clear linguistic generalizations.

References

Andrews, Avery. 1983. Constituent coordination in LFG. Unpublished Ms., Australian National University, available from the LFG Archive (http://www.essex.ac.uk/linguistics/LFG/www-lfg.stanford.edu/archive/).

Aoun, Joseph, Elabbas Benmamoun, and Dominique Sportiche. 1994. Agreement and conjunction in some varieties of Arabic. *Linguistic Inquiry* 25:195–220.

Aoun, Joseph, Elabbas Benmamoun, and Dominique Sportiche. 1999. Further Remarks on First Conjunct Agreement. *Linguistic Inquiry* 30:669–681.

Asudeh, Ash. 2005. Semantic composition motivated first conjunct agreement. Unpublished LSA handout.

[14] In the XLE implementation of LFG, head precedence is provided as an approximation that relies just on the PRED-introduction c-structural element, not all nodes mapped to a given f-structure. In fact, XLE already goes a fairly long way towards ordered sets, as for instance scope statements can be added to set members (see Footnote 10). XLE also provides ways of influencing f-structure model building in the grammar code: the predefined template @COMPLETE can be used in an f-annotation to enforce "early" checking of constraining equations (compare the XLE online documentation at `www2.parc.com/isl/groups/nltt/xle/doc/`). While this mechanism cannot be used to implement our account, it indicates that there are other situations in which a strictly global view on the full sentence's f-structure is not desirable (at least computationally).

[15] Of course, this type of proposal comes with a certain danger: it is uncommon in linguistic LFG theorizing to propose modifications of the technical LFG devices that are "behind the scenes". This fact is probably quite healthy for the LFG community, as it ensures a clear and technically well-defined common ground shared in practically all LFG work. "Experimentation" is typically done by assuming extra projection levels. But since SCA is a phenomenon that has been tackled many times within the common ground system, it may be legitimate to take deeper modifications into consideration – in particular as they seem to fit into the overall system very smoothly. Of course, future work will have to show whether the approach can also resolve other empirical phenomena, as further justification for the adjustments.

Asudeh, Ash, and Richard Crouch. 2002. Coordination and parallelism in glue semantics: Integrating discourse cohesion and the Element Constraint. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG2002 Conference*. CSLI On-line Publications.

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

Camacho, José. 2003. *The Structure of Coordination: Conjunction and Agreement Phenomena in Spanish and Other Languages*. Dordrecht: Kluwer Academic Publishers.

Corbett, Greville G. 1983. Resolution rules: agreement in person, number and gender. In E. K. G. Gazdar and G. K. Pullum (eds.), *Order, Concord and Constituency*, pp. 175–214. Dordrecht: Foris.

Corbett, Greville G. 1988. Agreement: A partial specification based on slavonic data. In Michael Barlow and Charles A. Ferguson (eds.), *Agreement in Natural Language*, pp. 23–54. Stanford: CSLI.

Dalrymple, Mary, and Ronald M. Kaplan. 2000. Feature indeterminacy and feature resolution. *Language* 76:759–798.

Dalrymple, Mary, and Irina Nikolaeva. 2006. Syntax of natural and accidental coordination: Evidence from agreement. *Language* 82.

de Almeida Torres, Artur. 1981. *Moderna gramática expositiva da Língua Portuguesa*. Sao Paulo: Martins Fontes.

Falk, Yehuda. 2006. On the representation of case and agreement. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG2006 Conference*, Stanford, CA. CSLI On-line Publications.

Kaplan, Ronald, and John Maxwell. 1988/95. Constituent coordination in lexical-functional grammar. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell, and Annie Zaenen (eds.), *Formal Issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications. The first version of this paper appeared in COLING 88, Budapest, August 1988, pp. 303-305.

Kaplan, Ronald M., and Annie Zaenen. 1989/95. Long-distance dependencies, constituent structure, and functional uncertainty. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell, and Annie Zaenen (eds.), *Formal Issues in Lexical-Functional Grammar*, pp. 137–165. Stanford, CA: CSLI Publications. Originally appeared in *Alternative Conceptions of Phrase Structure*, ed. M. Baltin and A. Kroch (Chicago: Chicago University Press, 1989) 17–42.

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

Kuhn, Jonas. 2003. Generalizaed Tree Descriptions for LFG. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG2003 Conference*, Stanford, CA. CSLI On-line Publications.

Marten, Lutz. 2000. Agreement with Conjoined Noun Phrases in Swahili. *Afrikanistische Artbeitspapiere* 64, Swahili Forum VII:75–96.

Marten, Lutz. 2005. The dynamics of agreement and conjunction. *Lingua* 115:527–547.

Maxwell, John, and Christopher Manning. 1997. A theory of non-constituent coordination based on finite-state rules. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the First LFG Conference*, Grenoble. CSLI On-line Publications.

McCloskey, James. 1986. Inflection and Conjunction in Modern Irish. *Natural Language and Linguistic Theory* 4:245–82.

Moosally, Michelle. 1998. *Noun Phrase Coordination: Ndebele Agreement Patterns and Cross-Linguistic Variation*. PhD thesis, University of Texas at Austin.

Munn, Alan. 1999. First Conjunct Agreement: Against a Clausal Analysis. *Linguistic Inquiry* 75: 552–562.

Rouveret, Alain. 1994. *Le syntaxe du gallois*. Paris, France: Editions CNRS.

Sadler, Louisa. 1999. Non-distributive features and coordination in Welsh. In Miriam Butt and Tracy Holloway King (eds.), *On-line Proceedings of the LFG99 Conference*. CSLI On-line Publications.

Sadler, Louisa. 2003. Coordination and Asymmetric Agreement in Welsh. In Miriam Butt and Tracy Holloway King (ed.), *Nominals: Inside and Out*, pp. 85–118. Stanford, CA: CSLI.

Sells, Peter. 1985. *Lectures on Contemporary Syntactic Theories. An Introduction to Government-Binding Theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar*. Number 3 in CSLI Lecture Notes. Stanford, CA: CSLI Publications.

Villavicencio, Aline, Louisa Sadler, and Doug Arnold. 2005. An HPSG Account of Closest Conjunct Agreement in NP Coordination in Portuguese. In Stefan Müller (ed.), *Proceedings of the HPSG05 Conference*, Stanford, CA. CSLI Publications: http://www-csli.stanford.edu/publications.

Wechsler, Stephen, and Larisa Zlatić. 2000. A theory of agreement and its application to Serbo-Croatian. *Language* 76:759–798.

Wechsler, Stephen, and Larisa Zlatić. 2003. *The Many Faces of Agreement*. Stanford, CA: CSLI Publications.

ON ELLIPTICAL NOUN PHRASES IN HUNGARIAN

Tibor Laczkó

University of Debrecen

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King
(Editors)

2007

CSLI Publications

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

Abstract

In this paper I develop an LFG analysis of two noun phrase types in Hungarian that can be referred to as elliptical. In one of them, Type (A), the understood noun head is entirely missing from the construction and formally the head function is performed by the head of the final modifying constituent in the phrase. The major task here is to capture, in a lexicalist framework, the formal head properties of an adjective or a numeral. I employ an exocentric structure and introduce a pro noun head into f-structure by an appropriate functional annotation. In Type (B), which is always a (special) possessive construction, the noun head is represented by a pro-like morpheme attaching to the head of the possessor constituent. The fundamental challenge with respect to this construction type is that the morpheme appears to be phrasal in nature, and it can be recursively attached to the head, optionally in combination with the morpheme marking the plurality of the possessed noun. I assume that the morpheme in question is an argument taking predicate, and I capture scope relations and the possibility of recursion by means of a hierarchical sublexical representation.

1. Introduction

There are two very frequently used elliptical (or, depending on one's analysis, anaphoric) noun phrase types in Hungarian. Their shared property is that they lack an overt lexical head. Either the understood noun head is entirely missing from the construction, or it is represented by a pro-like morpheme attaching to the head of the possessor constituent. The goal of this paper is to develop an LFG analysis of these two construction types, which raise interesting questions related to the treatment of head-marking languages and "phrasal suffixes".

The paper has the following structure. In section 2, I characterize the two construction types to be analyzed. In section 3, I propose an LFG account of these phenomena. In section 4, I briefly show that a novel analysis of Hungarian possessive constructions can be naturally adopted in the approach I have worked out. This is followed by some concluding remarks in section 5.

2. The phenomena

In this section I present the relevant data. I describe the two construction types: the headless type (section 2.1) and the pro bound morpheme type (section 2.2). I also point out the challenges they pose for a lexicalist theory like LFG.

2.1. Type (A)

In the Type (A) constructions the noun head is missing from the expression entirely. The rightmost modifier in the “remainder” of the expression (whether an adjective or a numeral) functions formally as the head. This formal headness is manifested by the fact that all the nominal suffixes are attached to the head of this final constituent:

- plural markers (1c)
- case endings (1c-f)
- possessive agreement suffixes (1f)

Note that in nonelliptical noun phrases, numerals and adjectives take none of these suffixes, cf. (1a,b).

(1) a. a nyolc piros toll-at
the eight red pen-ACC
‘the eight red pens’

b. a tanár nyolc piros toll-á-t
the teacher.NOM eight red pen-3SG-ACC
‘the teacher’s eight red pens’

c. a piros-ak-at
the red-PL-ACC
‘the red ones’

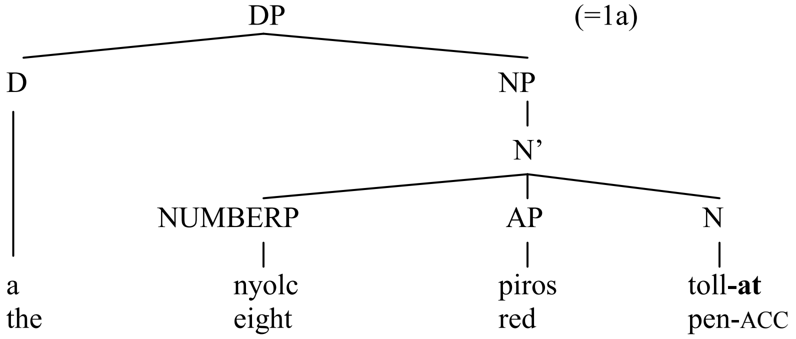
d. a nyolc piros-at
the eight red-ACC
‘the eight red ones’

e. a nyolc-at
the eight-ACC
‘the eight’

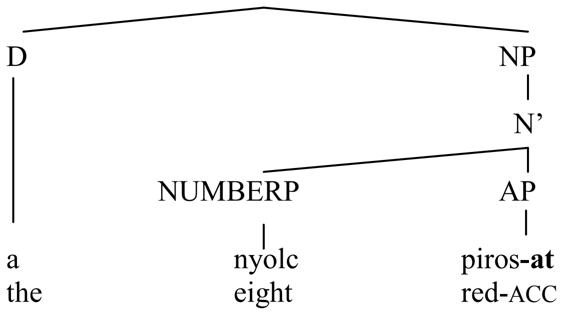
f. a tanár nyolc piros-á-t
the teacher.NOM eight red-3SG-ACC
‘the teacher’s eight red ones’

The c-structures for (1a-f) are shown in (2)-(6).

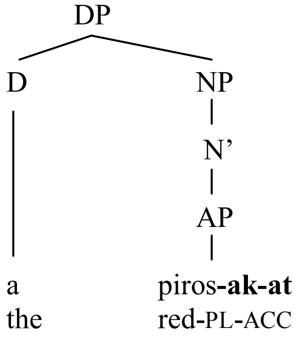
(2) (=1a)



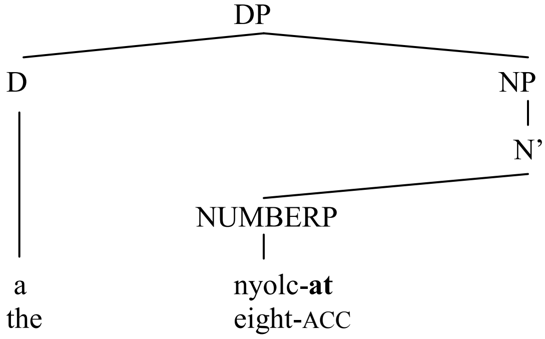
(3) a. (=1d)



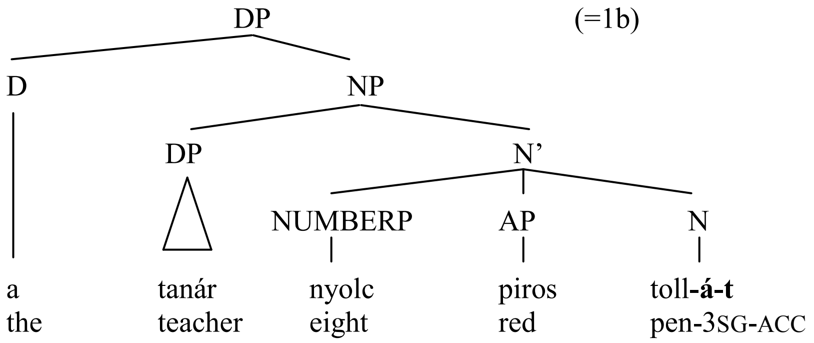
b. (=1c)



(4) (=1e)

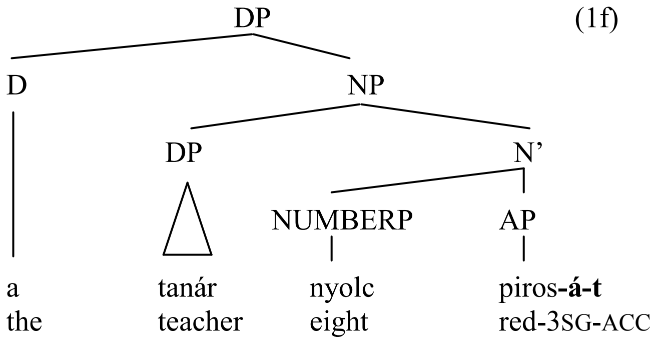


(5) (=1b)



Let me point out that in cases like (5), in theory the whole (definite) possessive construction as well as the definite possessor noun phrase should have their respective D positions filled. However, they would be adjacent, and, therefore, only one of them is phonetically realized. This issue does not concern us here, and for this reason I simply represent the D of the entire possessive DP without any justification and without any commitment to a possible LFG treatment of this phenomenon. (For detailed discussion and a GB analysis, see Szabolcsi (1994).)

(6) (1f)



1.2. Type (B)

Type (B) is a special possessive construction. Its special nature is due to the fact that the "phrasal" suffix *-é* attaching to the head of the possessor constituent stands for the possessed noun, cf. (7) and (5) vs. (8).

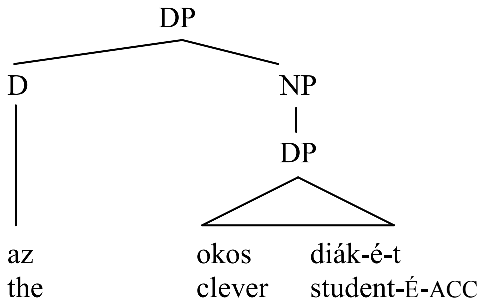
(7)	a	tanár	nyolc	piros	toll-á-t	és
	the	teacher.NOM	eight	red	pen-3SG-ACC	and
					az okos	diák-é-t
					the clever	student-É-ACC

'the teacher's eight red pens and the clever student's
((*i*) pens (*ii*) eight red pens)'

As the two versions in the English translation of (7) indicate, *-é* can stand for either the possessed head alone or a modifier + head sequence.

The c-structure for the elliptical DP in (7) is shown in (8).

(8)



The *-é* morpheme clearly has scope over the whole of the possessor phrase: e.g. on Bartos's (2000) MP account it assigns a Θ-role to this constituent, in addition to triggering the anaphoric interpretation of the missing possessee. This construction type has the following important additional properties, which are illustrated in (9).

(i) The *-é* constituent can be pluralized, and *-é* suffixation and pluralization are recursive.

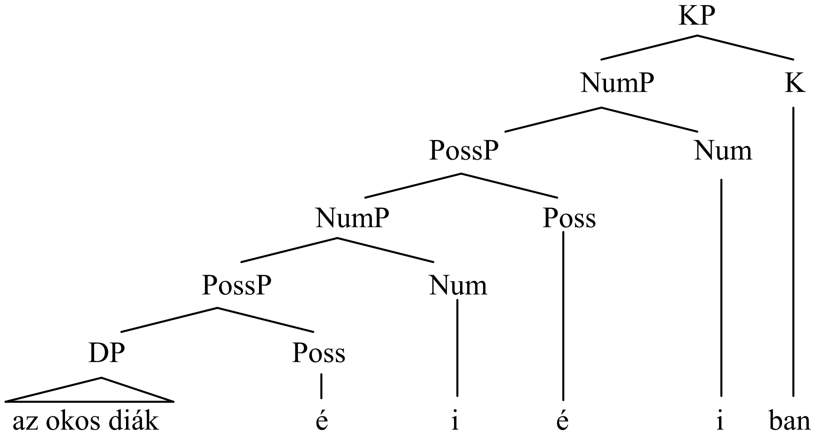
(ii) The entire *-é* phrase can be case-marked just like any other nominal expression.

(iii) The determiners and modifiers in the DP are, as a rule, interpreted as being associated with the most deeply embedded possessor, realized by the noun stem that *-é* attaches to.

(9) az okos diák-é-i-é-i-ban
the clever student-É-PL-É-PL-INE
ca. 'in those of those of the clever student'

Consider Bartos's (2000) syntactic (MP) analysis of (9).

(10)



There are two interrelated challenges for a lexicalist account of this construction type:

(i) the modelling of the recursion of a "phrasal" suffix to the effect that, on the face of it, several possessor DPs can be embedded within one another;
(ii) ensuring that, in the case of multiply embedded possessors, modification always applies to the deepest possessor.

2. An LFG analysis

2.1. Type (A)

Butt et al. (1999: 97-98), in their Parallel Grammar framework, outline an LFG analysis of basically similar German and English constructions, cf.:

(11) NPheadless → NPposs
(↑NUM)=sg
(↑PERS)=3
(↑PRED)='pro'
(↑PRON-TYPE)=null
(↑SPEC)=↓

(12) a. the dentist's

b.

$$\begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRON-TYPE} & \text{null} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \\ \text{SPEC} & \begin{bmatrix} \text{PRED} & \text{'dentist'} \\ \text{NTYPE} & \text{count} \\ \text{ANIM} & + \\ \text{CASE} & \text{gen} \\ \text{SPEC-TYPE} & \text{poss} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \\ \text{SPEC} & \begin{bmatrix} \text{SPEC-TYPE} & \text{def} \\ \text{SPEC-FORM} & \text{THE} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

In the spirit of this account, in my analysis of Type (A) constructions I postulate a special exocentric NP without a c-structure categorial head. A functional annotation associated with the final XP node provides an LFG-style ‘pro’ element, which serves as a basis for the appropriate anaphoric interpretation of the missing head in the given context.

The c-structure rules and their annotations also have to ensure that it is only in the case of elliptical noun phrases that an adjective or a numeral can be inflected and that it is always the final such element that is inflected, and in these cases the number and the case features of the XP provide the whole NP/DP with these features. Therefore, the following devices have to be applied.

(i) “Ordinary APs or NUMBERPs” must be negatively constrained for inflectional features. By “ordinary” I mean (a) APs or NUMBERPs in headed noun phrases and (b) APs or NUMBERPs in nonfinal positions in Type (A) elliptical noun phrases.

(ii) The final AP or NUMBERP must be associated with annotations that encode that the inflectional features of the A or NUMBER head are identical to those of the whole elliptical noun phrase. The following features are relevant in this connection: number, case, and, if a possessor is present in the construction, the agreement features of the possessor.

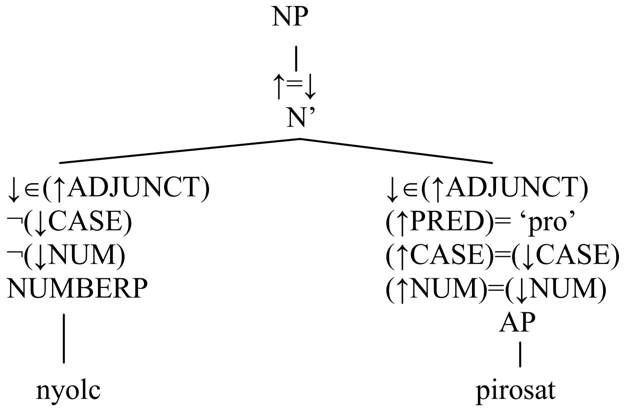
Consider the phrase structure rules in (13), whose functional annotations satisfy these requirements.

(13)	a. NP →	DP	N’
		(↑POSS)=↓	↑=↓
	b. N’ →	XP*	N
		↓∈(↑ADJUNCT)	↑=↓
		¬(↓CASE)	
		¬(↓NUM)	
		¬(↓POSS)	
	c. N’ →	XP*	{NUMBERP | AP}
		↓∈(↑ADJUNCT)	↓∈(↑ADJUNCT)
		¬(↓CASE)	(↑PRED)= ‘pro’
		¬(↓NUM)	(↑CASE)=(↓CASE)
			(↑NUM)=(↓NUM)

I provide the analysis of (1d), repeated here as (14a) for convenience, along these lines. (14b) shows the annotated c-structure representation of (1d) and I present the corresponding f-structure in (15).

(14) a. nyolc piros-at
eight red.SG-ACC
'eight red ones'

b.



(15)
$$\left[\begin{array}{ll} \text{PRED} & \text{'pro'} \\ \text{ADJUNCT} & \left\{\begin{array}{l} [\ \text{PRED} \quad \text{'eight'}] \\ \left[\begin{array}{ll}\text{PRED} & \text{'red'}\\ \text{NUM} & \text{sg}\\ \text{CASE} & \text{acc}\end{array}\right]\end{array}\right\} \\ \text{NUM} & \text{sg} \\ \text{CASE} & \text{acc}\end{array}\right]$$

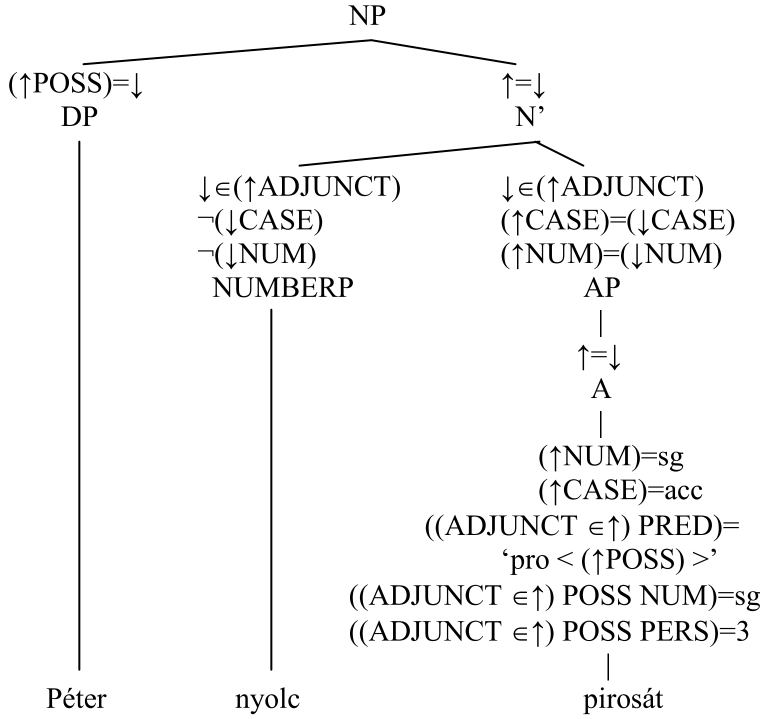
The predicate of the possession relationship and the featural information about the number and person of the possessor are also encoded by the possession morphology on the head of the final constituent in the form of inside-out function application.

Now consider the analysis of a Type (A) possessive construction, exemplified in (16a). I give the lexical form of the adjective used in this example in (16b), the annotated c-structure representation in (16c) and present the f-structure in (17).

(16) a. Péter nyolc piros-á-t
Peter.NOM eight red-3SG-ACC
'Peter's eight red ones'

b. pirosát, A ‘red’
(↑NUM)=sg
(↑CASE)=acc
((ADJUNCT ∈↑) PRED)= ‘pro < (↑POSS) >’
((ADJUNCT ∈↑) POSS NUM)=sg
((ADJUNCT ∈↑) POSS PERS)=3

c.



(17)

$$\left[\begin{array}{ll}
\text{PRED} & \text{‘pro} < (\uparrow\text{POSS}) >\text{’} \\
\text{POSS} & \left[\begin{array}{ll} \text{PRED} & \text{‘Peter’} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{array}\right] \\
\text{ADJUNCT} & \left\{\begin{array}{l} [\text{PRED} \quad \text{‘eight’}] \\ \left[\begin{array}{ll} \text{PRED} & \text{‘red’} \\ \text{NUM} & \text{sg} \\ \text{CASE} & \text{acc} \end{array}\right] \end{array}\right\} \\
\text{NUM} & \text{sg} \\
\text{CASE} & \text{acc}
\end{array}\right]$$

There is a strong motivation for the ‘pro’ analysis of Type (A) constructions: if there is no situational or linguistic context, the interpretation of the entire phrase is that it denotes people; that is, the ‘pro’ element I postulate has the [+human] feature, which is an instance of classical pro(arb), cf.:

(18) Én a gyors-ak-at kedvel-em.
I the fast-PL-ACC like.PRES-3SG
‘I like the fast ones (= people).’

(19) Tíz autó van az udvar-on. Én a gyors-ak-at
ten car is the yard-SUP I the fast-PL-ACC

kedvel-em.
like.PRES-3SG

‘There are ten cars in the yard. I like the fast ones (= cars).’

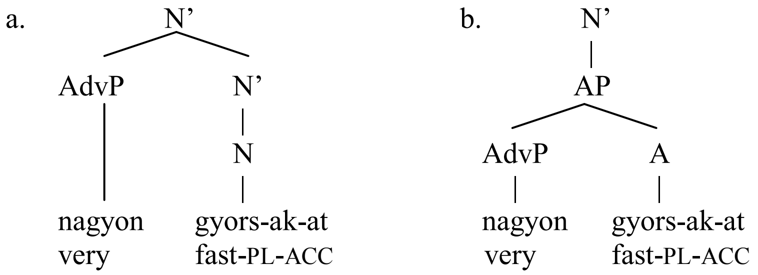
An alternative approach would be to assume that in Type (A) constructions, the head of the final constituent has undergone the lexical process of A → N conversion. There is, however, a very strong argument against such an account: in these elliptical constructions the final adjective or numeral has all the properties ordinary adjectives and numerals have. For instance, the adjective takes adverbial modification, it can be used in comparative and superlative forms, etc. Consider the following examples.

(20) Én a nagyon gyors-ak-at kedvel-em.
I the very fast-PL-ACC like.PRES-3SG
‘I like the very fast ones (= people).’

(21) Én a leg-gyors-abb-ak-at kedvel-em.
I the SUP-fast-COMP-PL-ACC like.PRES-3SG
‘I like the fastest ones (= people).’

The conversion approach would commit us to postulating that nouns, just like adjectives, can take adverbial modification, which would be rather counter-intuitive. Compare the undesirable, conversion-based representation in (22a) and the analysis based on my elliptical assumptions in (22b).

(22)



2.2. Type (B)

The most important assumptions and aspects of my analysis of Type (B) constructions are as follows.

1. I assume that the *-é* suffix is an LFG-style "pro" element.
2. It is the functional and semantic head of the whole nominal expression.
3. It is an argument-taking predicate with a (POSS) argument. The motivation for this assumption is that according to the majority of recent generative analyses of Hungarian possessive constructions the noun head and its possessive morphology make up a complex predicate (whether in the syntax or in the lexicon), and this complex predicate takes the possessor as its argument, cf. Szabolcsi (1994), Laczkó (2000), Bartos (2000), É. Kiss (2002), Chisarik and Payne (2003), etc. Given the fact that *-é* is most straightforwardly analyzable as a "pro possessive noun head" element, its argument taking capacity naturally follows.
4. When *-i*, the plural marker for possessed nouns attaches to *-é* immediately following it, I then take this plural suffix to be a functional co-head, pluralizing the nominal expression.
5. I employ articulated sublexical structures with functional annotations. The possible multiple attachment and the scope relations of the two morphemes, *-é* and *-i*, are modelled by a hierarchical organization of these sublexical structures.
6. The fact that determiners and modifiers are, as rule, associated with the most deeply embedded possessor, which is always realized by the noun head, is captured by the following mechanism. Admittedly, this is only one possible technical way of ensuring the correct interpretation of this construction type. On the issue of modification in the two elliptical phrase types, see section 4.
   - The functional annotations assigned to determiners and modifiers contain the (POSS+) function label. Here the + symbol means any number of (embedded) POSS functions, but at least one.
   - The (↑POSS+ PRED FN)~= pro equation states that the relevant possessor cannot be a 'pro'. No matter how many times *-é* is attached, there will always be only one non-'pro' possessor, the one realized by

the noun head to which the first -*é* suffix attaches, as required by the facts.

- It has to be ensured that the value of (POSS+) is the same in the relevant annotation pairs.

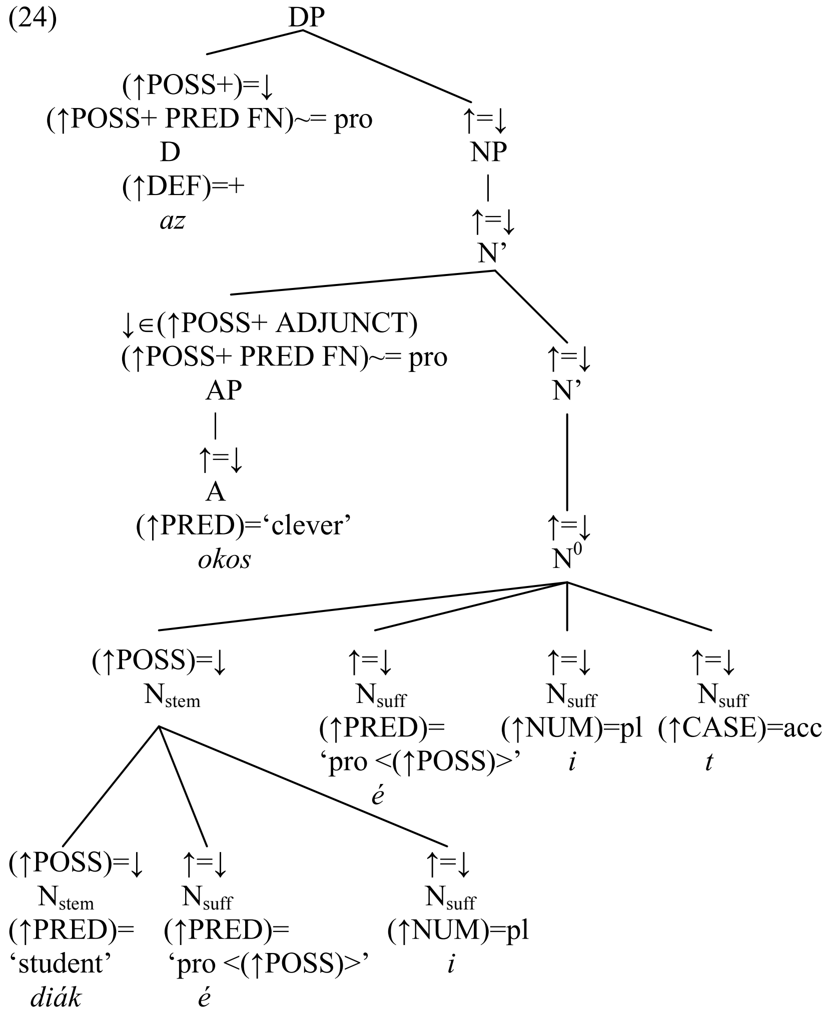
Let us see how this approach works through the analysis of the example in (23). I present the annotated c-structure in (24) and the f-structure in (27).

(23)

az	okos	diák-é-i-é-i-t
The	clever	student-É-PL-É-PL-ACC

ca. 'those of those of the clever student'

(24)



The use of hierarchical sublexical structures is not wide-spread in LFG, but it is not unprecedented either. See, for instance, Butt and King (2006) using a similar mechanism in their analysis of Urdu causatives. They in turn cite Kaplan et al. (2004) on the idea of sublexical rules, although these rules do not introduce hierarchical structures. (I am grateful to Tracy H. King for pointing out these facts to me in personal communication.) An alternative to this sublexical structural analysis may be to explore whether Wescoat's (2005) lexical sharing approach can be extended to the treatment of these Hungarian phenomena. I leave this to future research.

In the light of the points above describing the salient aspects of the analysis, most of the details of the representation should be straightforward. There is, however, an important technical problem that this representation does not address, and, consequently, does not solve. The problem is this. The current versions of the two members of the two pairs of functional annotations in (25) do not guarantee that the value of (POSS+) is the same in both members, which would be essential for the analysis to be adequate and not incorrectly overgenerate. In other words, if there are multiply embedded possessors then their numbers should match in the two members of each pair of functional equations. Otherwise we cannot ensure, among other things, that an adjunct should be represented in f-structure, and interpreted by our semantics, as modifying a non-pronominal possessor.

(25) a. (↑POSS+)=↓
(↑POSS+ PRED FN)~= pro

b. ↓∈(↑POSS+ ADJUNCT)
(↑POSS+ PRED FN)~= pro

One feasible solution, which I have developed in a Parallel Grammar framework, and which works efficiently, is as follows (for an overview of the Parallel Grammar Project, see Butt et al. (1999)). We can create a template for the relevant annotations in such a way that it contains disjunctive pairs of functional equations. The templates for (25a) and (25b) can be (26a) and (26b), respectively.

(26) a. { (↑POSS)=↓
(↑POSS PRED FN)~= pro
| (↑POSS POSS)=↓
(↑POSS POSS PRED FN)~= pro
| (↑POSS POSS POSS)=↓
(↑POSS POSS POSS PRED FN)~= pro }

b. { ↓∈(↑POSS ADJUNCT)
(↑POSS PRED FN)~= pro
| ↓∈(↑POSS POSS ADJUNCT)
(↑POSS POSS PRED FN)~= pro
| ↓∈(↑POSS POSS POSS ADJUNCT)
(↑POSS POSS POSS PRED FN)~= pro }

The disjunctive template in (26b), for instance, ensures that the ADJUNCT will precisely and exclusively be represented in f-structure as modifying the (only) non-pronominal possessor. Although in theory further embedding of possessors is possible, even ordinary possessive constructions hardly ever contain more than three possessors embedded within one another. As far as these *-é* “pronominal” constructions are concerned, not a single instance of more complex embedding has been attested. This is fundamentally due to human processing limitations, which are even stricter in these instances of multiple pronominal embedding. Naturally, the templates in (26) can always be augmented with further embedding if there is a justified need for this.

(27)

$$\left[\begin{array}{ll}
\text{PRED} & \text{‘pro} < (\uparrow\text{POSS}) >\text{’} \\
\text{NUM} & \text{pl} \\
\text{PERS} & 3 \\
\text{CASE} & \text{acc} \\
\text{POSS} & \left[\begin{array}{ll}
\text{PRED} & \text{‘pro} < (\uparrow\text{POSS}) >\text{’} \\
\text{NUM} & \text{pl} \\
\text{PERS} & 3 \\
\text{POSS} & \left[\begin{array}{ll}
\text{PRED} & \text{‘student’} \\
\text{NUM} & \text{sg} \\
\text{PERS} & 3 \\
\text{DEF} & + \\
\text{ADJUNCT} & \{[\text{PRED ‘clever’}]\}
\end{array}\right]
\end{array}\right]
\end{array}\right]$$

3. Adopting a new analysis of possessors

In Laczkó (2007) I develop a new account of Hungarian possessive constructions. It postulates that a lexical predication template converts an ordinary, nonrelational noun into a “raising” type predicate with an (XCOMP) propositional argument and a nonthematic (POSS) function. The possessiveness marker attaching to the noun head is the predicate (π) of the (XCOMP), and its open (POSS) argument is functionally controlled by the (POSS) of the raising predicate. In this section I briefly show that this new approach can be easily adopted by my analysis of the two elliptical constructions proposed in this paper.

(28a) is a Type (A) elliptical construction. (28b) shows the lexical form of the adjectival head as used in this structure according to my account presented in this paper, subscribing to the “traditional” view of possessive constructions. (28c) demonstrates the lexical form of the same adjective in the same construction on the basis of the new approach to possessive constructions as developed in Laczkó (2007).

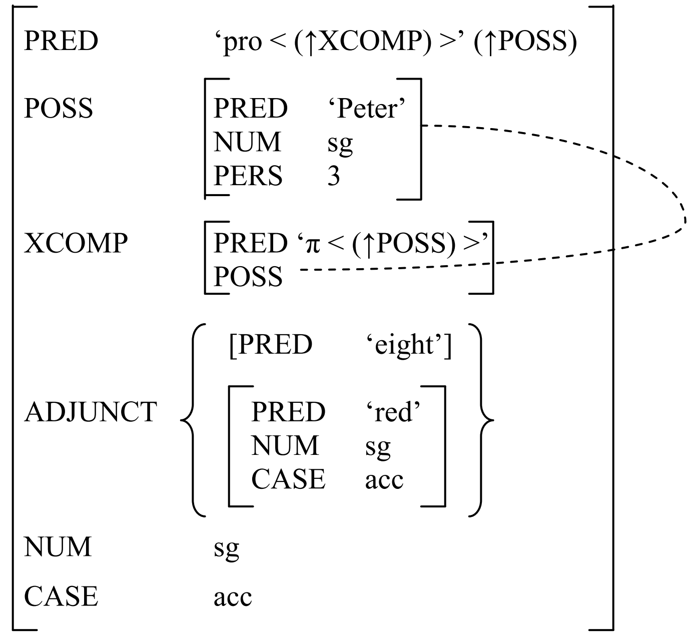
(28) a. Péter nyolc piros-á-t (=16a)
Peter.NOM eight red-3SG-ACC
‘Peter’s eight red ones’

b. pirosát, A ‘red’ (=16b)
(↑NUM)=sg
(↑CASE)=acc
((ADJUNCT ∈↑) PRED)= ‘pro < (↑POSS) >’
((ADJUNCT ∈↑) POSS NUM)=sg
((ADJUNCT ∈↑) POSS PERS)=3

c. pirosát, A ‘red’
(↑NUM)=sg
(↑CASE)=acc
((ADJUNCT ∈↑) PRED)=
‘pro < ((ADJUNCT ∈↑) XCOMP) >’ ((ADJUNCT ∈↑) POSS)
((ADJUNCT ∈↑) POSS)= ((ADJUNCT ∈↑) XCOMP POSS)
((ADJUNCT ∈↑) XCOMP PRED)=
‘π <((ADJUNCT ∈↑) POSS)>’
((ADJUNCT ∈↑) POSS NUM)=sg
((ADJUNCT ∈↑) POSS PERS)=3

The major difference is that the Poss morph attaching to the adjectival head does not introduce an ordinary possessive predicate. Instead, it introduces the functional annotational ingredients of the new, raising type analysis. For further details of the new possessive account, see Laczkó (2007). The f-structure of (28a) is as follows (compare it with (17)).

(29)



(23), repeated here as (30) for convenience, exemplifies Type (B) constructions.

(30) az okos diák-é-i-é-i-t
The clever student-É-PL-É-PL-ACC
ca. 'those of those of the clever student'

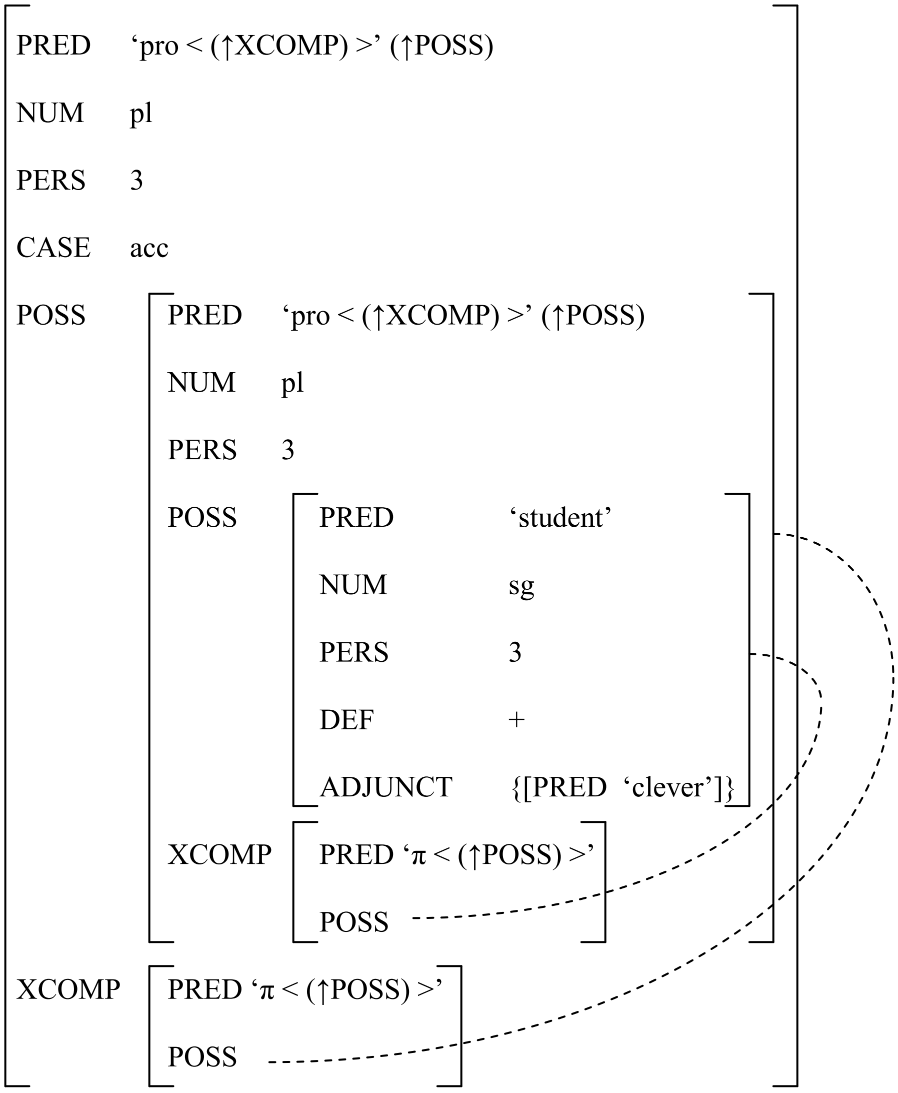
(31a) shows the lexical form of the *-é* morph in my analysis based on the "traditional" possessive view, while (31b) demonstrates the modified version capitalizing on the new perspective presented in Laczkó (2007).

(31) a. *-é*, N_{suff} [N__]$_N$ pro <(↑POSS)>'

b. *-é*, N_{suff} [N__]$_N$ pro <(↑XCOMP)>' (↑POSS)
(↑POSS)= (↑XCOMP POSS)

The f-structure of (30) on this new account is given (32). Compare it with (27).

(32)



4. Conclusion

In this paper I have developed an LFG analysis of two elliptical noun phrase types. In the case of Type (A), in which the understood noun head is entirely missing from the construction and formally the head function is performed by the head of the final modifying constituent in the phrase, I employ an exocentric structure and introduce a pro noun head into f-structure by an appropriate functional annotation. In the case of Type (B), which is always a (special) possessive construction and in which the noun head is represented by a pro-like morpheme attaching to the head of the possessor constituent, I assume that the morpheme in question is an argument taking predicate, and I capture the scope relations of this pro morpheme and the plural marker of the possessed noun as well as the possibility of recursion by a hierarchical sublexical representation.

Finally, let me make a short comment on modification in the constructions under investigation. In Type (A) the covert pro head must have a modifier, cf. (1c-f). In Type (B) the overt pro head, encoded by the *-é* suffix, must not have a modifier. This complementarity may be a part of the reason why Type (B) follows its special modification pattern, whose essence is that all the modifiers in the construction must alway be associated with the most deeply embedded, non-pronominal possessor. Another possible factor is that if in this type the modification of pro was possible, then this would inevitably lead to undesirable ambiguity.

References

Bartos, Huba 2000. Az inflexiós jelenségek szintaktikai háttere [The Syntactic Background of Inflectional Phenomena]. In: Kiefer, Ferenc (ed.) *Strukturális magyar nyelvtan 3. Morfológia* [Structural Hungarian Grammar 3. Morphology]. Budapest: Akadémiai Kiadó, 653–762.

Butt, Miriam – Tracy Holloway King – María-Eugenia Niño – Frédérique Segond 1999. *A Grammar Writer's Cookbook.* Stanford: CSLI Publications.

Butt, Miriam and Tracy H. King. 2006. Restriction for Morphological Valency Alternations: The Urdu Causative. *Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan*, ed. by Miriam Butt, Mary Dalrymple, and Tracy H. King. Stanford: CSLI Publications, 235–258.

Chisarik, Erika and John Payne. 2003. Modelling Possessor Constructions in LFG: English and Hungarian. *Nominals: Inside and Out*, ed. by Miriam Butt and Tracy H. King. Stanford: CSLI Publications, 181–199.

É. Kiss, Katalin. 2002. *The Syntax of Hungarian.* Cambridge: Cambridge University Press.

Kaplan, Ronald, J. T. Maxwell III, Tracy H. King, and R. Crouch. 2004. Integrating Finite-state Technology with Deep LFG Grammars. Proceedings of the Workshop on Combining Shallow and Deep Processing for NLP (ESSLLI). http://www2.parc.com/isl/groups/nltt/pargram/esslli04fst-xle.pdf.

Laczkó, Tibor. 2000. Derived Nominals, Possessors and Lexical Mapping Theory in Hungarian DPs. *Argument Realization*, ed. by Miriam Butt and Tracy H. King. Stanford: CSLI Publications, 189–227.

Laczkó, Tibor. 2007. Revisiting Possessors in Hungarian DPs: A New Perspective. *This volume.*

Szabolcsi, Anna. 1994. The Noun Phrase. *The Syntactic Structure of Hungarian. Syntax and Semantics 27*, ed. by Ferenc Kiefer and Katalin É. Kiss. New York: Academic Press, 179–274.

Wescoat, Michael T. 2005. English Nonsyllabic Auxiliary Contractions: An Analysis in LFG with Lexical Sharing. Butt, Miriam & King, Tracy H. eds. *Proceedings of the LFG '05 Conference*. Bergen: University of Bergen. (On-line publication: 2005 CSLI Publications, ISSN 1098-6782, http://www-csli.stanford.edu/publications/LFG2/lfg05.html)

REVISITING POSSESSORS IN HUNGARIAN DPs: A NEW PERSPECTIVE

Tibor Laczkó

University of Debrecen

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King
(Editors)

2007

CSLI Publications

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

Abstract

This paper offers a new LFG analysis of possessive DPs in Hungarian. This account is designed to overcome two difficulties that the majority of previous generative approaches had to face: *(a)* the problem of the (a)symmetrical cohead relationship between the noun head and the possessive marker *(b)* the problem of dual theta role assignment in GB (or its LFG equivalent) when the noun head is relational and the possessive marker is also considered an argument taking predicate. The new account postulates that a lexical predication template converts an ordinary, nonrelational noun into a "raising" type predicate with an (XCOMP) propositional argument and a nonthematic (POSS) function. The possessive marker is the predicate of the (XCOMP), and its open (POSS) argument is functionally controlled by the (POSS) of the raising predicate. The same lexical predication template is assumed to apply to relational nouns except that, as a result, they become "equi" predicates, that is, the (POSS) function introduced by the template is assigned to one of their arguments. This approach solves the above-mentioned problems, and it has several additional advantages.

1. Introduction

In the past two decades, several generative (GB, LFG, MP) analyses of Hungarian DPs in general and possessive DPs in particular have been developed. Most of them adopt or adapt some central insightful generalizations of Szabolcsi (1994). However, minimally there are two salient aspects of Szabolcsi's account, and those of other accounts sharing these aspects, that can be shown to be problematic (or at least these solutions can be taken to be rather marked in the given frameworks).

The first problem is that on these accounts the Poss predicate (that is, the possessive marker), realized by a morpheme attaching to the possessed noun, is minimally the central (co)head of the noun phrase; however, the possession relationship is DP/NP internal, not visible from "outside". This fact is only captured in a brute force manner.

The second problem is an instance of marked theta role assignment when both the Poss predicate and the relational (or deverbal) noun head are theta role assigners: the former is assumed to assign a formal theta role, and the latter is assumed to assign a contentful theta role.

The goal of this paper is to develop a more coherent (morpheme-based) LFG analysis that solves both these problems. I also show that it has additional favourable properties. The crucial aspect of the new solution is that a conversion process creates a "raising" predicate from an ordinary noun and an "equi" predicate from a relational/deverbal noun, and Poss attaching to this converted noun is the predicate of the propositional (XCOMP) argument of this raising or equi noun head.

The paper has the following structure. In section 2, I give a brief overview of some salient generative analyses of Hungarian possessive DPs and highlight two general problems most of them have to face, then I summarize Bresnan's (2001) account of English possessive constructions. In section 3, I present my new analysis. First, I analyze possessive DPs with ordinary noun heads (3.1). Second, I extend this approach to possessive constructions with relational/deverbal heads (3.2). Third, I discuss the most significant and most favourable aspects of the new account (3.3). In section 4, I reiterate the most important points of the paper.

2. Some previous generative accounts

In section 2.1 I discuss a basic assumption shared by several recent generative approaches to possessive DPs in Hungarian (2.1.1), and highlight those aspects of previous analyses that are directly relevant for our present purposes with special emphasis on the two central problems to be addressed (2.1.2-2.1.4). In section 2.2 I briefly show Bresnan's (2001) treatment of English possessive DPs, as it partially motivated my new account.

2.1. Hungarian possessive constructions

2.1.1. A basic assumption

It is a fairly generally accepted assumption that it is feasible to distinguish the Poss and the Agr morphemes (although they are often "fused"), cf. Szabolcsi (1994), Komlósy (1998, 2002), Bartos (2000), É. Kiss (2002). In my previous work, e.g. in Laczkó (1995, 2000), I did not adopt this assumption, neither did Chisarik and Payne (2003). On the basis of some new data, in my current analysis I also subscribe to this separation view. The standard argument for the separation is that in certain cases there is a morpheme intervening between the Poss morpheme and the Agr morpheme, cf. (1a), in which the intervening morpheme encodes the plurality of the possessed noun. In (1b) the Agr morph also encodes what under ordinary circumstances Poss expresses (in addition to the always unmarked singularity of the possessed noun). In (1c), by contrast, according to several analyses, the Poss morph also encodes 3SG (Agr) (again, in addition to the always unmarked singularity of the possessed noun).

(1) a. (az én) kalap-ja-i-m
the I hat-*Poss*-PL-*1SG*
'my hats'

b. (az én) kalap-om
the I hat-*Poss*.SG.*1SG*
'my hat'

c. (az ő)   kalap-ja
the he   hat-*Poss*.SG.*3SG*
'his hat'

In addition to the widely cited case in (1a), I have found a new and very strong argument for this separation: the use of emphatic pronouns in possessive constructions. The crucial point here is that when the possessor is an emphatic pronoun, the presence of the Agr morpheme on the possessed noun is strictly prohibited:

(2)   (emph.)pronoun-Agr   +   N-Poss(*-Agr)   /   N-Poss(*.Agr)

It is always the emphatic pronoun that is marked for agreement, and agreement marking on the noun head either by a separate morph or by fusion leads to ungrammaticality. In this construction type, of the two morphemes in question, it is only Poss that can (and must) be present.

(3)   a. a   mag-am   diák-ja   (*diák-om)
the   self-1SG   student-Poss   student-Poss.1SG
'my own student'

b. a   mag-atok   diák-ja   (*diák-otok)
the   self-2PL   student-Poss   student-Poss.2PL
'your own student'

Let me also point out in this connection that this construction type lends considerable support to the assumption that relational/deverbal nouns are also combined with the same Poss predicate in possessive constructions.

(4)   a. a   mag-am   elárul-ás-a   (*elárul-ás-om)
the   self-1SG   betray-DEV-Poss   betray-DEV-Poss.1SG
'my own betrayal'

b. a   mag-atok   elárul-ás-a   (*elárul-ás-otok)
the   self-2PL   betray-DEV-Poss   betray-DEV-Poss.2PL
'your own betrayal'

This calls for a uniform analysis of possessive constructions with ordinary noun heads, on the one hand, and relational/deverbal noun heads, on the other hand.

2.1.2. Szabolcsi (1994)

In Szabolcsi's (1994) classical GB analysis the noun head and the Poss make up a complex predicate in the lexicon. The theta role assigner is the Poss. According to Szabolcsi, this is formal theta role assignment.

		▼	Θ_F
(5)	az	ő	kalap-ja
	the	he.NOM	hat-Poss.3SG
	'his hat'		

Szabolcsi (1994: 197) provides the following interpretation for a possessive DP like (5).

(6) $\lambda x \lambda y[\mathrm{N}(x)\ \&\ \mathrm{R}(y,x)]$
'the set of pairs where *x* is a N and bears some relation R to *y*, and the range of *y* is restricted by the agr features'

As a rule, in these possessive constructions Poss equals R, and in the case of (5) N equals *hat*.

In my opinion the problem with the formula in (6) is that it does not satisfactorily capture the fact that the Poss, although it is assumed to be, and it is represented as, a copredicate in the complex lexical predicate, is subordinate to the noun head copredicate, cf.:

(7) a. *'the possession of the hat by him'

b. 'the hat that is possessed by him'

This is the first general problem I set out to solve here. It is characteristic of all Chomskyan approaches to Hungarian possessive DPs that adopt Szabolcsi's complex predicate analysis (whether this complex predicate is created in the lexicon, cf. Szabolcsi (1994), or in the syntax, cf. Bartos (2000) and É. Kiss (2002)) and it is equally characteristic of my previous LFG accounts, cf. Laczkó (1995, 2000).

The second general problem with Szabolcsi's and several other authors' approach (cf., again, Bartos (2000) and É. Kiss (2002), for instance) is related to possessive DPs with relational/deverbal noun heads. In these cases Szabolcsi postulates that both elements of the complex predicate assign their respective theta roles to one and the same constituent, the possessor. Again, the Poss predicate has its formal role to assign, while the argument-taking noun head is taken to assign a contentful role. In (8), for example, the deverbal head assigns the theme role it has inherited from the input verbal predicate.

```
           ▼          Θ_C      Θ_F
(8)   az   ő        megérkez-és-e
      the  he.NOM arrive-DEV-Poss.3SG
      ‘his arrival’
```

Obviously, this is a rather marked theta theoretical scenario. If the assignment of the formal role by itself can satisfy Theta Theory in the case of ordinary noun heads, cf. (7), then in the case of relational/deverbal heads the classical version of the Theta Criterion is inevitably violated. Naturally, the criterion can be, and it has been, loosened, cf. É. Kiss and Szabolcsi (1992), for instance, but the unquestionably marked aspect of this solution remains. My new analysis to be presented in this paper eliminates this problem as well.

It is also noteworthy that the semantic subordination of Poss to the noun head is even more surprising in the light of this dual theta role assignment mechanism. Szabolcsi claims that formal theta role assignment by Poss is instrumental in contentful theta role assignment by relational/deverbal nouns, that is, the latter is formally dependent on the former. In this respect, of the two copredicates one would expect Poss to be superior to, or at least equal to, but definitely not subordinate to, the relational/deverbal noun head.

2.1.3. Laczkó (2000)

In my LFG analysis in Laczkó (2000) I do not separate Poss and Agr and in this respect I assume complex possession morphology. In addition, I postulate that this possessive morphological complex does not behave in a uniform fashion.

(A) When it attaches to ordinary noun heads, it encodes agreement features, on the one hand, and it introduces an argument taking predicate (π) expressing extrinsic possession in the terminology of Barker (1995), or a general R-relation, cf., for instance, Szabolcsi (1994). The sole argument of this predicate has the “extrinsic possessor” (Π_e) semantic role, cf.:

(9) a. *kalap*, N PRED = ‘HAT’

b. *-(j)A$_1$*, [N__]$_N$ ‘π <Π_e>’
(POSS)
(↑POSS PERS) = 3
(↑POSS NUM) = SG
((↑POSS PRED) = ‘pro’)

c. *kalap-ja*, N PRED = ‘HAT-π <Π_e>’
(POSS)
(↑POSS PERS) = 3
(↑POSS NUM) = SG
((↑POSS PRED) = ‘pro’)

(B) When the possessive morphological complex attaches to a relational noun, which is an argument taking predicate, and it expresses intrinsic possession in the sense of Barker (1995), then this complex only encodes agreement features, and does not introduce a predicate. Compare (9b) and (10b).

(10) a. *apa*, N PRED = ‘FATHER < Π_i >’
(POSS)

b. *-(j)A$_2$*, [N__]$_N$
(↑POSS PERS) = 3
(↑POSS NUM) = SG
((↑POSS PRED) = ‘pro’)

c. *ap-ja*, N PRED = ‘FATHER <Π_i>’
(POSS)
(↑POSS PERS) = 3
(↑POSS NUM) = SG
((↑POSS PRED) = ‘pro’)

I treat possessive constructions with an argument taking deverbal noun head in a similar fashion.

This approach also faces the first general problem I pointed out in section 2.1.2. It can only capture the embedded nature of the possession relationship in a brute force way when the possessive morphological complex attaches to an ordinary noun head, because here, too, the two elements are taken to be on a par (“predicate composition” is assumed to take place).

The second general problem does not arise in this analysis, that is, there is no LFG style "dual" theta role assignment, because when the possessive morphological complex attaches to an argument taking noun head then it has no argument. Thus, the possessor is only an argument of the noun head. However, the cost of this is that this account cannot provide a uniform treatment of possessive constructions with either ordinary or relational/deverbal noun heads.

2.1.4. Komlósy (1998)

Komlósy's (1998) LFG analysis has the following major aspects to it. He separates Poss and Agr (of course, he also has to deal with cases of fusion). However, according to him Poss never encodes a predicate: in all its uses, it only introduces an existential constraint to the effect that there must be a (POSS) grammatical function in the construction. For the sake of easy comparison between Komlósy (1998) and Laczkó (2000), below I represent the version of the *-(j)A* morph in Komlósy's system that also carries agreement features.

(11) *-(j)A:* $[N\_\_]_N$
(↑POSS)
(↑POSS PERS)= 3
(↑POSS NUM)= SG
((↑POSS PRED)= 'pro')

Possessive constructions with relational/deverbal noun heads can be analyzed along these lines in a straightforward and principled manner. Poss introduces the (POSS) function, and this function is assigned to an argument of the relational/deverbal noun.

(12) *kocogás-a:* N, (↑PRED)= 'JOGGING < (↑POSS) >'
(↑POSS PERS)= 3
(↑POSS NUM)= SG
((↑POSS PRED)= 'pro')

It is easy to see that this solution, too, avoids the second general problem discussed in section 2.1.2: there is no "dual theta role assignment" here, either.

When Komlósy's Poss (without any argument structure) attaches to an ordinary noun head, the following lexical form is created. (For the sake of easy comparison, here, too, the version of the *-ja* morph is represented that also encodes 3SG agreement features.)

(13) *kalap-ja:* N, (↑PRED)= 'KALAP'
(↑POSS)
(↑POSS PERS)= 3
(↑POSS NUM)= SG
((↑POSS PRED)= 'pro')

On the face of it, it seems that this approach solves, in a trivial way, the first general problem discussed in section 2.1.2, the need for capturing the embedded nature of the possession relationship: such a semantic relationship is simply not introduced.

There are, however, three major problems with this analysis.

(A) In the case of ordinary noun heads coherence is violated. The (meaningful) possessor constituent in such constructions has the subcategorized (POSS) function; however, it is not a semantic argument of any predicate.

(B) Given that the (POSS) grammatical function is subcategorizable on Komlósy's account, too, but it is nonsemantic, cf. point (A), Komlósy is bound to state that the proper interpretation of the constituent receiving this function is adjunct-like in possessive constructions with ordinary noun heads. Thus, in addition to the coherence violation, the analysis has to face the problem of adjunct-like interpretation of the possessor in DPs with ordinary noun heads vs. argument-like interpretation of the possessor in DPs with relational/deverbal heads.

(C) This approach cannot capture the fact that (POSS) cannot be assigned to expletive elements in either Hungarian or English. This is especially significant in the case of deverbal noun heads. Poss introduces the (POSS) grammatical function, which on this account is always nonsemantic, and it is puzzling why a noun derived from a verb allowing an expletive subject or object is not compatible with an expletive possessor in either of these two languages.

2.2. Possessors in English: Bresnan (2001)

The essence of Bresnan's (2001) account of English possessive constructions is as follows. The lexical form of an ordinary noun without a predicate argument structure is augmented with a lexical predication template introducing a "subject" of predication, hence the term predication template.

(14) a. *hat$_1$*, N 'HAT < >' → b. *hat$_2$*, N 'HAT-OF <(↑POSS)>'

In effect, a lexical conversion process creates a relational noun from an ordinary, nonrelational one. The newly introduced argument is assumed to have subject-like properties.

In my new analysis to be proposed in the next section, I apply a conversion process similar to this in spirit; however, there are significant differences as far as the details are concerned.

3. The new approach

Contrary to my earlier view, in this new analysis I separate Poss and Agr, cf. the discussion in section 2.1.1. I set out to solve the two general problems of (i) modelling the "embedded" nature of the possession relationship and (ii) avoiding dual (parallel) theta role assignment. In addition, I aim at developing a uniform treatment of possessive constructions with either ordinary or argument taking (relational/derived) noun heads.

3.1. Ordinary nouns and possession

It is a generally held view that possession is a predicative relationship, cf. Szabolcsi (1994), den Dikken (1999), Laczkó (2000), Bresnan (2001), Chisarik and Payne (2003). In my new account I capture this relationship in an explicit morphosyntactic fashion that directly feeds semantics: the predicate of this possession relationship is Poss, which has one argument that receives the semantically unrestricted (POSS) grammatical function, cf.:

(15) *-(j)A*, [N__]$_N$ 'π | *x* is related to *y* | < (↑POSS) >'
y

This analysis is similar to that of attributive "relational" adjectives according to which these adjectives (and attributive adjectives in general) have no subject argument in their predicate argument structure, cf.:

(16) *proud*, A '*x* is proud of *y* < (↑OBL) >'
y

Let me point out in this connection that technically a subject argument approach could also be applied, compare (15) and (16), on the one hand, and (17) and (18), on the other hand.

(17) *proud*, A '*x* is proud of *y* < (↑SUBJ) (↑OBL) >'
x *y*
'pro'

(18) *-(j)A*, [N__]$_N$ 'π | *x* is related to *y* | < (↑SUBJ) (↑POSS) >'
x *y*
'pro'

For lack of space, here I cannot discuss the motivation for choosing the morphosyntactically simpler treatment shown in (15).

The two crucial aspects of the new account are the following.

(i) In Bresnan's (2001) predication template manner I postulate a lexical redundancy rule that converts an ordinary noun without an argument structure into a one-place nominal predicate.
(ii) This nominal predicate is a raising predicate, just like the verb *seem*.

Consider (19) illustrating this conversion process.

(19) a. *kalap$_1$*, N 'HAT < >' →

b. *kalap$_2$*, N 'HAT < (↑XCOMP) >' (↑POSS)
(↑POSS)= (↑XCOMP POSS)

Thus, the fundamental difference between Bresnan's (2001) template and mine is that the former simply introduces an argument structure with a possessor argument, while the latter introduces an argument structure with a propositional argument mapped onto (XCOMP), and also associated with a nonsemantic (POSS) function. The propositonal (XCOMP) requirement is, as a rule, satisfied by Poss attaching the the *kalap$_2$* type noun head.

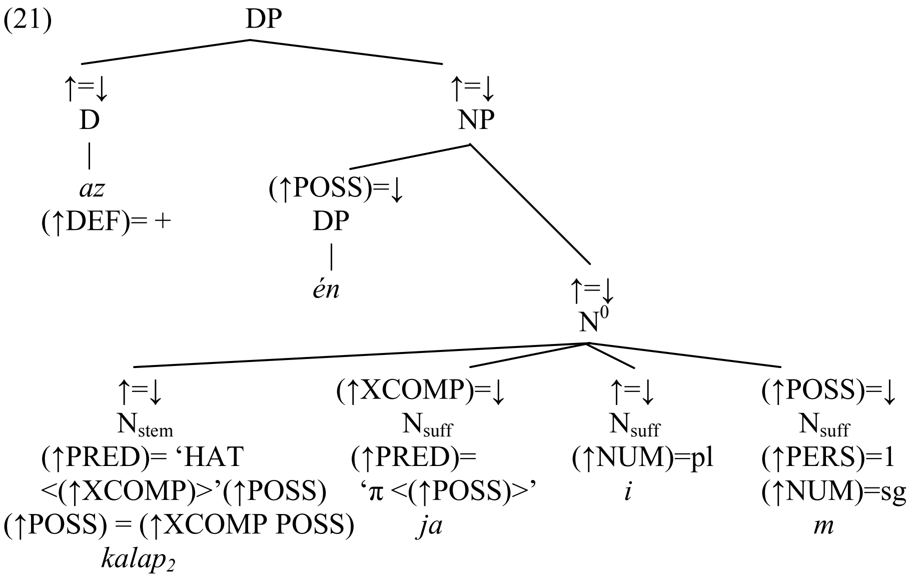
For the purposes of the presentation of my new analysis in this paper, I assume that in the nominal domain, in terms of semantically unrestricted functions, nominal predicates can only have (POSS) at their disposal, and this function is always introduced by the predication template. (This assumption is highly relevant in the case of possessive constructions with relational/deverbal noun heads.)

Let us now see the new account of a possessive construction like (1a), repeated here as (20) for convenience.

(20)	(az	én)	kalap-ja-i-m
	the	I	hat-*Poss*-PL-*1SG*
	'my hats'		

I show the c-structure of (20) in (21). For expository purposes I also include sublexical representation so that the f-structure contribution of each morph should be easily detectable. The lexical form of the complex noun head as used in this particular case is given in (22). I present the simplified f-structure of (20) in (23).

(21)



(22) kalapjaim, N (↑PRED)= 'HAT < (↑XCOMP) >' (↑POSS)
(↑POSS)= (↑XCOMP POSS)
(↑XCOMP PRED)= 'π <(↑POSS)>'
(↑NUM)=pl
(↑POSS PERS)=1
(↑POSS NUM)=sg

(23)

$$\left[\begin{array}{ll} \text{PRED} & \text{`HAT <(}\uparrow\text{XCOMP)>'(}\uparrow\text{POSS)} \\ \text{NUM} & \text{pl} \\ \text{DEF} & + \\ \text{CASE} & \text{nom} \\ \text{POSS} & \left[\begin{array}{lll} \text{PRED} & \text{`pro'} & \\ & \text{PERS} & 1 \\ & \text{NUM} & \text{sg} \\ & \text{CASE} & \text{nom} \end{array}\right] \\ \text{XCOMP} & \left[\begin{array}{ll} \text{PRED} & \text{`}\pi < (\uparrow\text{POSS}) >\text{'} \\ \text{POSS} & \text{-------} \end{array}\right] \end{array}\right]$$

As this f-structure representation shows, my new account provides a principled solution to the first general problem: modelling the embedded nature of the possessive relationship. The noun head selects this relationship as its propositional argument carrying the (XCOMP) function, in other words, possession is subordinate to the noun head.

3.2. Argument taking nouns in possessive constructions

The new account can be extended to argument taking nouns in a principled manner, and, thus, it makes a uniform analysis of possessive constructions with ordinary and argument taking predicates possible. I assume that an argument taking noun undergoes the same conversion process, that is, the same lexical predication template applies to it, introducing the propositional argument expressing the possessive relationship. The minimal contrast is that the (POSS) function introduced by this conversion is assigned to the argument (or one of the arguments) of the relational/deverbal noun, in other words: the conversion makes argument taking nouns equi (rather than raising) predicates.

(24) a.*húg*$_1$, N 'YOUNGER-SISTER-OF < Θ >'

b. *húg*$_2$, N 'YOUNGER-SISTER-OF < (↑POSS) (↑XCOMP) >'
(↑POSS) = (↑XCOMP POSS)

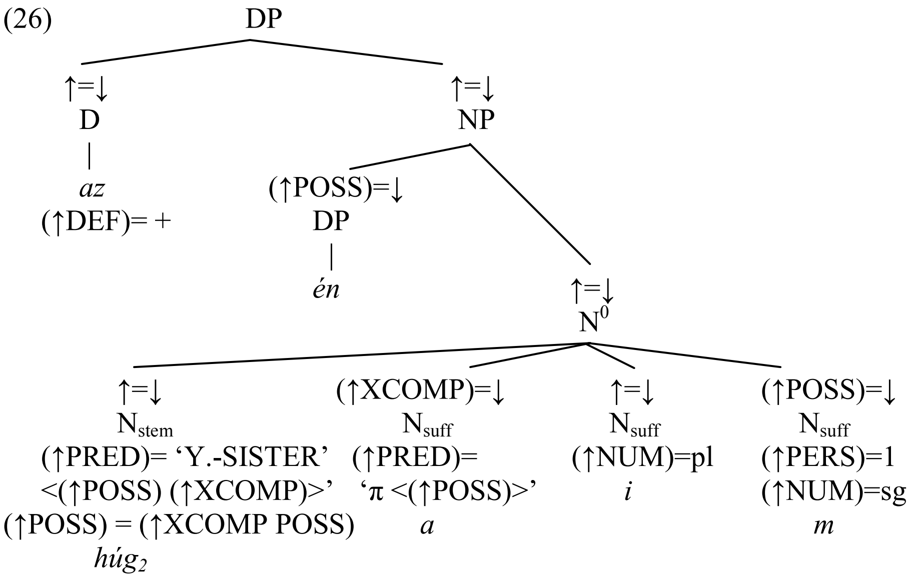
The following legitimate question could be raised in connection with (24a): Why does such a predicate need the lexical predication template? Why does it not assign the (POSS) function to its argument in a direct manner? My answer is this. I assume that nouns in general, whether ordinary or argument taking, are incapable of assigning this function. In the nominal domain mapping of an argument onto the (POSS) function is exclusively licensed by the predication template.

Let us now see the details of the new analysis of relational nouns through the example of (25).

(25) (az én) húg-a-i-m
the I younger.sister-*Poss*-PL-*1SG*
'my younger sisters'

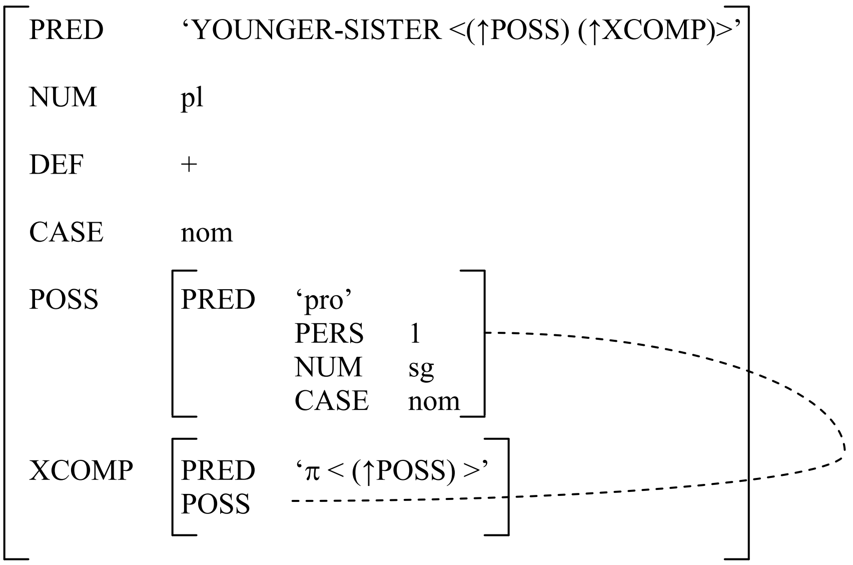
I show the c-structure of (25) in (26). For expository purposes I also include sublexical representation so that the f-structure contribution of each morph should be easy detectable. The lexical form of the complex noun head as used in this particular case is given in (27). I present the simplified f-structure of (25) in (28).

(26)



(27) húgaim, N (↑PRED)= 'YOUNGER-SISTER < (↑POSS) (↑XCOMP) >'
(↑POSS)= (↑XCOMP POSS)
(↑XCOMP PRED)= 'π <(↑POSS)>'
(↑NUM)=pl
(↑POSS PERS)=1
(↑POSS NUM)=sg

(28)



It is a major advantage of this new approach that it also solves the second general problem discussed in section 2.1.2: that of "dual theta role assignment". The relational/deverbal noun has a (POSS) argument and the Poss predicate (the PRED of XCOMP) also has an "open" (POSS) argument and the former functionally controls the latter (a typical LFG style equi scenario). That is, there are two arguments bearing the (POSS) function and there are two predicates each of which takes only one of the two arguments. This contrasts with the scenario in the majority of previous analyses, in which there are two predicates taking one and the same element as an argument.

As far as I can see, there is only one potential (and possibly apparent) problem with this new approach, which requires further investigation. In the case of deverbal nouns, if the input verb already has an (XCOMP) argument then as a result of the conversion process there will be two (XCOMP)s in the argument structure of such a deverbal noun. This seems to be a violation of biuniqueness. Consider the following example.

(29) a. a kerítés (Péter által-i) zöld-re fest-és-e
the fence.NOM Peter by-AFF green-SUBL paint-DEV-Poss.3SG
'the painting of the fence green'

b. festése, N (↑PRED)=
'PAINTING < ((↑OBL)$_{AG)}$, (↑POSS), (↑XCOMP$_1$) (↑XCOMP$_2$) >'
(↑POSS) = (↑XCOMP$_1$ SUBJ)
(↑POSS) = (↑XCOMP$_2$ POSS)
(↑XCOMP$_2$ PRED)= 'π < (↑POSS) >'
(↑NUM)=sg
(↑CASE)=nom
(↑POSS PERS)=3
(↑POSS NUM)=sg

It appears to me that there are at least the following two plausible avenues for exploring a solution to this problem.

(A) In classical LFG, too, there were several kinds of (XCOMP)s (on a fundamentally categorical basis): (VCOMP), (NCOMP), etc. It can be argued that predicates of different categories (V, N, etc.) denote (partially) different kinds of propositions: (XCOMP)$_\Theta$. Compare this with various versions of the oblique function: (OBL)$_\Theta$. It is also noteworthy in this connection that the status original (OBJ2) grammatical function in classical LFG has also been reconsidered, and now we have OBJ)$_\Theta$ instead. (Thanks to Tracy H. King for calling my attention to this additional factor.)

(B) Another possible solution could be based on the nature of the "open" grammatical function of the (XCOMP). It may be feasible or useful to distinguish between an (XCOMP) with the (SUBJ) open function and an

(XCOMP) with the (POSS) open function: $\text{(XCOMP)}_{\text{SUBJ}}$ vs. $\text{(XCOMP)}_{\text{POSS}}$. Naturally, for this distinction to be strongly motivated one would need some independent evidence. I leave this issue to future research.

3.3. On significant and favourable aspects of the new approach

3.3.1. Semantics is happy

The new account feeds semantics in an explicit and principled fashion. As I have already pointed out, the asymmetrical relationship between the noun head and Poss is plausibly captured by absolutely ordinary morphosyntactic means. The raising analysis ensures that the ordinary noun head assigns its nonthematic (POSS) function to the possessor constituent, which thus appears at the "highest" level in f-structure, formally linked to the noun head, while through the usual functional control mechanism this possessor constituent is the semantic argument of the "embedded" Poss predicate.

3.3.2. The subject-like nature of the possessor is retained

The classical "clause level subject in English — NP/DP level possessor in Hungarian" parallel as observed by Szabolcsi (1994) can be retained in a principled manner. The most important aspects of this parallelism are as follows.

(a) designated structural position:	[SPEC, IP] — [SPEC, NP]
(b) agreement:	subject~verb — possessor~possessed noun
(c) pro-drop:	subject pronoun — possessor pronoun
(d) extractability:	via [SPEC, CP] — [SPEC, DP]

All these generalizations can be kept, as the possessor occupies the usual, that is highest, c-structural (and f-structural) position, whether it is only formally linked to the noun head (in the case of ordinary noun heads) or it is also semantically linked to the head (in the case of relational/deverbal nouns). In Laczkó (1995, 2000) I adopt (a)-(c) in my LFG framework; (d) has to be treated diffferently in an LFG model: by using functional uncertainty. All these aspects of that analysis can be kept in the context of my new proposal.

3.3.3. Possession as a predication relationship

In section 3.1 I mentioned that it is a widely accepted view that possessive relationships are predicative in nature. In my opinion it is a further advantage of my novel analysis that it captures this generalization by ordinary morphosyntactic means. It employs the Poss predicate to introduce this relation into the possessive construction. Moreover, this account is a more explicit version of a Bresnan (2001) style predication template approach, too.

3.3.4. The (POSS) function is semantically unrestricted

The assumption that (POSS) is semantically unrestricted is essential for my new analysis. The reason for this is that it employs LFG style raising and equi devices, which involve functional control, and it is a basic generalization that only [–r] functions can participate in this kind of control relationship. However, this is not a problem at all. The [–r] status of (POSS) has been independently argued for: by Laczkó (1995, 2000), Komlósy (1998, 2002), and Chisarik and Payne (2003) for Hungarian; by Markantonatou (1995) for Modern Greek; by ∅rsnes (1995) for Danish; by Laczkó (1995) and Chisarik and Payne (2003) for English. Furthermore, although Bresnan (2001) does not elaborate on the status of (POSS) in English, her assumptions about the predication template (cf. "subject of predication"), on the one hand, and her treatment of verbal gerunds, according to which there is a functional control relation between the (POSS) of the DP and the (SUBJ) of the gerundive predicate, on the other hand, seem to suggest that she also needs to subscribe to the [–r] view of (POSS).

3.3.5. Explanation for the lack of expletive possessors

If the (POSS) grammatical function is considered semantically unrestricted, it has to be explained why it is incompatible with expletive elements, unlike (SUBJ) and (OBJ), which are readily compatible with expletives. The general factors and the essence of the explanation are as follows.

(a) The (POSS) function is assigned by two predicates: the noun head and Poss.
(b) For an ordinary (non-argument-taking) noun this (POSS) is nonthematic, so in theory it could be associated with an expletive element.
(c) However, the Poss predicate always assigns it to a semantic argument, and the two (POSS)'s are functionally identified by LFG's control mechanism.
(d) If (POSS) was assigned to an expletive element (by the noun head), this would inevitably lead to a violation of completeness, given that, as a result of functional control, the same expletive element would be required to satisfy the semantic argument need of the Poss predicate, which it would be unable to perform. Completeness would be violated.

3.3.6. The analysis can be extended to English

Although English is dependent marking (it has no Poss morpheme), at this stage of the investigation it seems worthwhile extending the new account to this language (and dependent marking languages in general). Technically this is very simple: Bresnan's (2001) predication template has to be replaced by my predication template.

(30) a. *hat*$_1$, N 'HAT < >' → b. *hat*$_2$, N 'HAT-OF <(↑POSS)>'

(31) *hat*$_2$, N 'HAT-OF <(↑XCOMP)>' (↑POSS)
(↑POSS) = (↑XCOMP POSS)
(↑XCOMP PRED)= 'π <(↑POSS)>'

The only difference between Hungarian and English is that in the latter the π predicate is always introduced by the template (and never by a morph(eme)), which is not at all an unusual or unprincipled solution in LFG.

I intend to explore the advantages and consequences of this extension in future work. Here I confine myself to briefly mentioning what advantages I envisage at this point.

(A) Just like in the case of Hungarian possessive DPs, this new analysis of English possessive DPs makes it possible to feed semantics more explicitly and in a more straightforward way.

(B) It can offer the same principled explanation for why (POSS) is incompatible with expletives in English possessive DPs, too. Consider Bresnan's (2001:293) examples.

(32) a. There appears to be a reindeer on the roof.

b. **There's* appearing to be a reindeer on the roof is an illusion.

c. It appears that there's a reindeer on the roof.

d. ??*Its* appearing that there's a reindeer on the roof is an illusion.

(C) In this new approach possessive constructions in head marking and dependent marking languages can be treated in a uniform way.

3.3.7. The new account can be adopted in GB/MP

The treatment of raising and equi constructions is commonplace in GB/MP; thus the analysis can, in theory, be easily translated into these Chomskyan models. The parallels between LFG and GB/MP with respect to these phenomena are straightforward.

A general remark is in order in this connection. If an account can be implemented in various frameworks, then this can often be regarded as a favourable aspect: it may suggest that correct theory-neutral generalizations have been made and a valid analysis has been developed.

4. Conclusion

Below I reiterate the most significant points of this paper.

1. I have proposed a new and more principled LFG treatment of Hungarian possessive constructions.
2. The gist of the analysis is that a lexical conversion process creates a raising predicate from an ordinary noun and an equi predicate from a relational/deverbal noun, and the Poss morpheme functions as the PRED of their (XCOMP) propositional argument.
3. This approach solves two classical problems: (i) modelling the "embedded" nature of the possessive relation and (ii) avoiding dual theta role assignment.
4. It makes a uniform analysis of all kinds of noun heads in Hungarian possessive constructions possible.
5. It can be extended to English; thus head-marking and dependent-marking languages can be treated in a uniform fashion.
6. It offers an explanation for why expletive elements cannot be possessors either in Hungarian or in English.
7. It can be translated into GB/MP in a principled manner.
8. There is one issue that requires further investigation: the nature of (XCOMP), or rather, exploring the possibility of postulating more than one (XCOMP) in the same argument structure: (XCOMP)$_{\Theta}$.

References

Barker, Chris. 1995. *Possessive Descriptions.* Stanford, CA: CSLI Publications.

Bartos, Huba. 2000. Az inflexiós jelenségek szintaktikai háttere [The Syntactic Background of Inflectional Phenomena]. *Strukturális magyar nyelvtan 3. Morfológia* [Structural Hungarian Grammar 3, Morphology], ed. by Ferenc Kiefer. Budapest: Akadémiai Kiadó, 653–762.

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

Chisarik, Erika and John Payne. 2003. Modelling Possessor Constructions in LFG: English and Hungarian. *Nominals: Inside and Out*, ed. by Miriam Butt and Tracy H. King. Stanford: CSLI Publications, 181–199.

Dikken, Marcel den. 1999. On the structural representation of possession and agreement. The case of (anti-)agreement in Hungarian possessed nominal phrases. *Crossing Boundaries: Theoretical Advances in Central and Eastern European Languages*, ed. by István Kenesei. Amsterdam: John Benjamins, 137–78.

É. Kiss, Katalin. 2002. *The Syntax of Hungarian.* Cambridge: Cambridge University Press.

É. Kiss, Katalin and Anna Szabolcsi. 1992. Grammatikaelméleti bevezető [Introduction to Grammatical Theory]. *Strukturális magyar nyelvtan 1. Mondattan* [Structural Hungarian Grammar 1, Syntax], ed. by Ferenc Kiefer. Budapest: Akadémiai Kiadó, 21–77.

Komlósy, András. 1998. *A nomen actionis argumentumainak szintaktikai funkcióiról* [On the Syntactic Functions of Nomen Actionis].Ms. Budapest: az MTA Nyelvtudományi Intézete.

Komlósy, András. 2002. *Another Look at Action Nominalizations in Hungarian.* Talk. Düsseldorf: 6th International Conference on the Structure of Hungarian.

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

Laczkó, Tibor. 2000. Derived Nominals, Possessors and Lexical Mapping Theory in Hungarian DPs. *Argument Realization*, ed. by Miriam Butt and Tracy H. King. Stanford: CSLI Publications, 189–227.

Markantonatou, Stella. 1995. Modern Greek deverbal nominals: an LMT approach. *Journal of Linguistics 31*, 267-299.

∅rsnes, Bjarne. 1995. *The Derivation and Compounding of Complex Event Nominals in Modern Danish – An HPSG Approach with an Implementation in Prolog*. Ph.D. dissertation. University of Copenhagen.

Szabolcsi, Anna. 1994. The Noun Phrase. *The Syntactic Structure of Hungarian. Syntax and Semantics 27*, ed. by Ferenc Kiefer and Katalin É. Kiss. New York: Academic Press, 179–274.

THE APPLICATIVE AFFIX
AND MORPHEME ORDERING IN CHICHEWA

Olivia S.-C. Lam

University of Oxford

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

Applicativization is highly productive in a language like Chichewa. The applicative affix augments the a(rgument)-structure of a verb by bringing in an additional semantic role, which is most frequently a benefactive, instrument or locative role. We show in this paper that the structure of the word can be represented in the form of a morphology-syntax interface tree, which makes it possible to refer to not only parts of the word but also the levels of representation that are associated with each morpheme. The a-structure is of particular interest, as this is the structure that the applicative and passive affixes alter. More importantly, these morphemes alter the existing a-structure, one that is the result of the interaction between the verb root and any other a-structure-changing morpheme that precedes in morphological form the morpheme in question. With the interface tree, it is possible to make reference to an intermediate a-structure, one that is associated with a particular morpheme on the tree. Morpheme order can thus be accounted for more straight-forwardly.

1. The Applicative Affix in Chichewa[1]

The applicative affix introduces a non-agentive phrase/clause that is not directly associated with the SUBJ function (contra the causative affix, for instance) (Mchombo 2004). It is an argument-structure-augmenting verbal affix, and most frequently introduces a benefactive, instrument or locative role into the a-structure. In Chichewa, this affix has two allomorphs: *-il-* and *-el-*. Which allomorph is selected and affixed to the verb is constrained by rules of vowel harmony. Consider the following examples:

[1] I wish to thank the participants of the LFG'07 conference for a number of very useful comments. A special thanks goes to Ron Kaplan for an illuminating discussion of parts of this paper, and to Mary Dalrymple for providing constructive feedback on earlier drafts of the paper.

(1) a. With the underived verb root *-pika* "cook":

mkango u-ku-phik-a nyemba
lion(3) 3SM-pres-cook-fv beans(10)[2]
'The lion cooked beans.'

b. A-structure: *-phika* < Ag, Pt >

(2) a. Benefactive role introduced by the applicative affix

Mkango u-ku-phik-**il**-a **ana** nyemba
lion(3) 3SM-pres-cook-appl-fv children(2) beans(10)
'The lion cooked the children beans.'

b. A-structure: *-phik-**il**-a* < Ag, **Ben**, Pt >

(3) a. Instrument role introduced by the applicative affix (Mchombo 2004:87, ex. 48b)

Kalulu a-ku-phik-**il**-a **mkondo** maungu
hare(1) 1SM-pres-cook-appl-fv spear(3) pumpkins(6)
'The hare is cooking pumpkins with (using) a spear.'

b. A-structure: *-phik-**il**-a* < Ag, **Instr**, Pt >

(4) a. Locative role introduced by the applicative affix (Mchombo 2004:87, ex. 49b)

Kalulu a-ku-phik-**il**-a **pa** **chulu** maungu
hare(1) 1SM-pres-cook-appl-fv on(16) anthill(7) pumpkins(6)
'The hare is cooking the pumpkins on the anthill.'

b. A-structure: *-phik-**il**-a* < Ag, Pt, **Loc** >

In (1a), the verb root is in its most basic form, without any a-structure-changing morpheme affixed to it. The verb root *-phika* "cook" is transitive, and subcategorizes for one object. The a-structure of the verb *-phika* is shown in (1b). Examples (2) to (4) show that an extra argument is licensed by the affixation of the applicative morpheme. In each of these cases, with the applicative affix *-il-* attached to the verb root *-phika*, the applied verb form becomes *-phik-il-a*, which subcategorizes for two objects. In (2), a

[2] Symbols and abbreviations used:
Acc = accusative case; Appl = applicative affix; Ben = benefactive role; fv = final vowel; fut = future tense; Instr = instrument role; Loc = locative role/locative case; OM = object marker; pres = present tense; Prop = proprietive case; pst = past tense; SM = subject marker; Th = theme role.
The number in the parentheses after a glossed noun shows the noun class of that noun.

benefactive argument *ana* “children” is introduced. In (3), an instrument argument *mkondo* “spear” is added, while in (4), the additional argument that is licensed is a locative argument *pa chulu* “on anthill”.

It is quite often the case that there is more than one a-structure-changing morpheme affixed to the verb root. Besides the applicative affix, other a-structure-changing morphemes include the passive, the causative and the reciprocal affixes. When there is more than one such affix on the verb, it is usually possible to have the morphemes affixed in more than one order. The difference in morpheme order results in a difference in meaning:

(5) *a-na-meny-an-its-a* (Alsina 1999:7, ex. 3)

Alenje	a-na-meny-**an-its**-a	mbuzi
Hunters(2)	2SM-pst-hit-**rcp-caus**-fv	goats(10)

‘The hunters made the goats hit each other.’

(6) *a-na-meny-ets-an-a* (Alsina 1999:7, ex. 4)

Alenje	a-na-meny-**ets-an**-a	mbuzi
Hunters(2)	2SM-pst-hit-**caus-rcp**-fv	goats(10)

‘The hunters made each other hit the goats.’

Since the order of morphemes has such an important role to play in the interpretation of a construction, there must be a way to accurately predict morpheme order and to correctly account for the effects that the morphemes have on the a-structure of the verb. We will first look at one such account proposed in Alsina (1999) and the problems that Alsina’s proposal faces in section 2. Section 3 provides an alternative way to account for morpheme order and the corresponding a-structure-altering effects, building on Sadler and Nordlinger’s (2004) analysis of case-stacking. Section 4 concludes the paper.

2. Alsina’s (1999) Instantiation of the Mirror Principle

It is generally accepted in the literature that morpheme order bears some relation to the order of processes triggered by these morphemes. To capture the relation between morphological changes and the corresponding syntactic effects induced by these morphemes, Baker (1985) proposes the Mirror Principle:

(7) The Mirror Principle (Baker 1985:375)
Morphological derivations must directly reflect syntactic derivations (and vice versa).

In the transformational theory that Baker assumes, this is achieved by allowing (bound) morphemes to appear under terminal nodes. A syntactic derivation (ex. movement) picks up the morpheme by moving into the position on the tree that is occupied by that morpheme. For instance, a causative derivation involves the movement of the verb root into the position that is occupied by the causative morpheme, which then attaches to the verb root to create the verb form V-CAUS. A single movement operation will give rise to a morphological derivation and a syntactic derivation.

Alsina (1999) suggests how the Mirror Principle can be captured in a non-transformational theory like LFG. The Mirror Principle is not a result of a sequence of transformations, but is a consequence of the order of morphological affixations and the order of their corresponding morpholexical operations during mapping from a-structure to f-structure. The operation associated with the morpheme that is closer to the verb root is applied first, and so the linear order of the morphemes reflects the order of the operations.

Argument-structure changing morphemes, such as the causative, applicative, passive and reciprocal morphemes, all have their own lexical entries, in which the change in argument structure to be effected by this morpheme is specified. Crucial to Alsina's proposal is the assumption that the a-structure of the verb root is altered in the way specified in the lexical entry of the morpheme upon affixation of that morpheme to the verb root in the lexicon. The Mirror Principle then follows as a consequence of the "morphological change and the a-structure change associated with the same morpholexical operation […] tak[ing] place at the same time" (Alsina 1999:24).

Take the applicative affix for example. The lexical entry of the applicative morpheme is given as follows:

(8) Lexical entry for the applicative affix (Alsina 1999:26)
[ir] $]_{V}$__ $< \; < \; \theta \ldots \underbrace{\theta \ldots > \mathrm{pt}} \; >$

The notation "$]_{V}$__" means that the item cannot be an independent form and must attach to the right edge of the verb stem. The a-structure alternation caused by this affix is such that the "theme is fused with the thematic role introduced" (Alsina 1999:24).

There are, however, some serious problems with Alsina's proposal. First, the notion of "fusion" of thematic roles is never clearly defined. In the case of the applicative affix, it is not at all clear what semantic basis there could be for making the claim that the thematic role introduced, whether it is a benefactive, instrument or locative role, had "fused" with another theme[3] role. Besides, the fusion does not seem to be constrained in any way. Can any two roles just fuse together?[4]

Another even more serious problem with Alsina's mapping analysis concerns cases in which there is more than one a-structure-changing morpheme on the verb. As an illustration, assume that there are two such morphemes on a verb root: V-Aff$_1$-Aff$_2$. Each of these morphemes makes one change to the a-structure. Aff$_1$ makes a change to the a-structure of the verb root, but Aff$_2$ alters the verb root's *modified* a-structure by Aff$_1$. In order to formalize this, there must be a way to talk about not only the "end point" a-structure, but also the intermediate a-structure.

Alsina attempted to do so by postulating that "morphological change and the a-structure change associated with the same morpholexical operation […] take place at the same time" (Alsina 1999:24). While this assumption is valid, his formalization faces a serious problem of creating new and temporary lexical items – the lexical entry of the affix interacts with that of the verb root, intrinsic classifications are assigned to the resulting roles, and the intermediate lexical item serves as the starting point of the morphological and morpholexical operation that follows:

> "The basic assumption is that *the assignment of intrinsic classifications and morphological composition interact in a cyclic manner*: intrinsic classifications apply to the underived a-structure and, successively, after any morphological process which alters its thematic content." (p. 29; author's emphasis in italics)

Each intermediate a-structure is thus accompanied by a partially derived word form, which also exists temporarily.

[3] In this paper, a theme role and a patient role are treated identically.

[4] The one constraint on the fusion of thematic roles is that, in an applicative operation, 'the role that is fused with the theme […] cannot be the highest thematic role' (Alsina 1999:26).

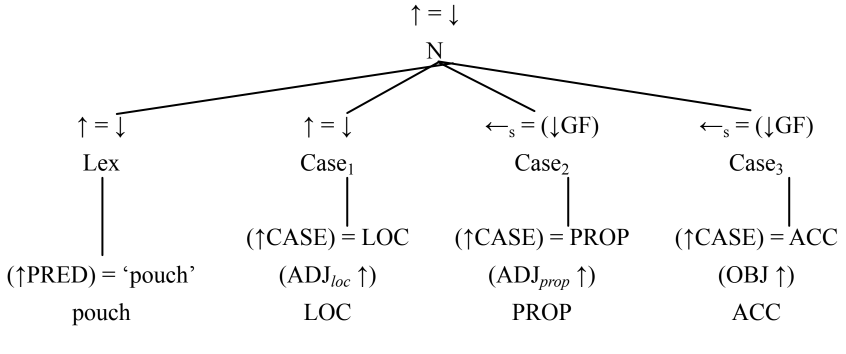
In the next section, we shall see how it becomes possible to make reference to different parts of word and the level(s) of representation associated with each of them by drawing insights from Sadler and Nordlinger's (2004) representation of morphological structures in the form of morphology-syntax interface trees.

3. An Alternative Proposal

3.1 Morphology-Syntax Interface Trees

In order to account for case-stacking phenomena in Australian languages, Sadler and Nordlinger (2004) adopt the Principle of Morphological Composition (PMC), originally proposed in Nordlinger (1998). Case-stacking is when more than one case affix is found on a nominal, and each of them contributes functional information to the f-structure that is defined by its following case morpheme. To achieve morphological composition more straight-forwardly, Sadler and Nordlinger (2004) assume that the morphological structure is represented by a flat interface tree between morphology and syntax[5]. The embedding relation between a case affix and its following case affix is represented by assigning the functional equation $\leftarrow_s = (\downarrow\text{GF})$ to the nodes dominating the non-initial case affixes:

(9) a. Morphological structure of the nominal *thara-ngka-marta-a* (pouch-LOC-PROP-ACC) represented as a morphology-syntax interface tree (This is a combination of the partial trees in Sadler and Nordlinger 2004:176-177, ex. 33-35.)



[5] Doug Arnold, Ron Kaplan and Louisa Sadler all pointed out that such an interface tree does not have to be flat in nature. The possibility of having a more hierarchical tree to represent the morphological structure of the verb, and therefore different functional annotations on the nodes, will be explored in future work.

b. F-structure for the nominal *thara-ngka-marta-a* (pouch-LOC-PROP-ACC) (Sadler and Nordlinger 2004:163, ex. 5)

$$\left[\begin{array}{ll}\text{OBJ} & \left[\begin{array}{ll}\text{CASE} & \text{ACC}\\ \text{ADJ}_{prop} & \left[\begin{array}{lll}\text{CASE} & \text{PROP} & \\ \text{ADJ}_{loc} & \begin{array}{l}\text{PRED}\\ \text{CASE}\end{array} & \left[\begin{array}{l}\text{`pouch'}\\ \text{LOC}\end{array}\right]\end{array}\right]\end{array}\right]\end{array}\right]$$

The functional annotation $\leftarrow_s = (\downarrow\text{GF})$[6] says "the f-structure defined by my sister to the left is a GF in my f-structure". Take, for instance, the nodes Case_1 and Case_2. As part of the lexical information specified by the case value LOC, some GF labelled ADJ_{loc} is required to exist at some level of the f-structure. The annotation on its sister node to the right, Case_2, indicates where this GF has to be – it has to be in the f-structure associated with that node, namely the f-structure called ADJ_{prop}. This gives the desired f-structure embedding, with the ADJ_{loc} function inside the ADJ_{prop} function.

3.2 The Proposal

The case-stacking phenomenon is similar to the morpheme ordering problem at hand in three respects: (i) there may be more than one affix on the stem; (ii) the order of affixes is significant; and (iii) any specification or change to a particular structure takes places sequentially. While Alsina (1999:24) assumes that "morphological change and the a-structure change associated with the same morpholexical operation […] take place at the same time", we assume that morphological composition motivates a-structure alternations.

3.2.1 The facts

We will, once again, work with the applicative affix and show how Sadler and Nordlinger's analysis can be extended to account for the order of the a-structure-changing morphemes in Chichewa. In order to show that the

[6] The arrow $\leftarrow_s$ refers to the immediately preceding sister node. Following Sadler and Norlinger (2004), this symbol is contrasted with the ← symbol (without the subscript s) that is found in off-path constraints (Sadler and Norlinger 2004:176).

order of the morphemes has an important role to play, a passive affix is also included on the verb root, together with the applicative affix. With two affixes, two morpheme orders are possible, but only one is acceptable. Consider the following examples:

(10) a. Example from (Alsina 1999:9, ex. (8b))
Mtsogoleri a-na-tumiz-**il-idw**-a zipatso (ndi ana)
leader(1) 1.sg-pst-send-**appl-pass**-fv fruit(8) by children(2)
'The leader was sent fruit (by the children).'

b. Example from (Alsina 1999:9, ex. (8b))
*Mtsogoleri a-na-tumiz-**idw-il**-a zipatso (ndi ana)
leader(1) 1.sg-pst-send-**pass-appl**-fv fruit(8) by children(2)
'The leader was sent fruit (by the children).'

Examples (10a) and (10b) have the same word order, but (10b) is ungrammatical while (10a) is grammatical. The only difference between the two examples lies in the order of the a-structure-changing morphemes on the verb root. In (10b), the passive morpheme precedes the applicative affix on the verb, whereas in (10a), the applicative affix precedes the passive morpheme.

3.2.2 The Analysis – An Interface Tree for *-tumiz-il-idw-* (send-pass-appl)

Assuming the Mirror Principle is at work, the grammaticality of (10a) and the ungrammaticality of (10b) lead to the conclusion that the applicative operation must take place before the passive operation (for a benefactive applied argument). For ease of discussion, we will focus on the following morphological fragments of the two verbs:

(11)	a.	-tumiz-il-idw-	send-appl-pass
	b.	*-tumiz-idw-il-	send-pass-appl

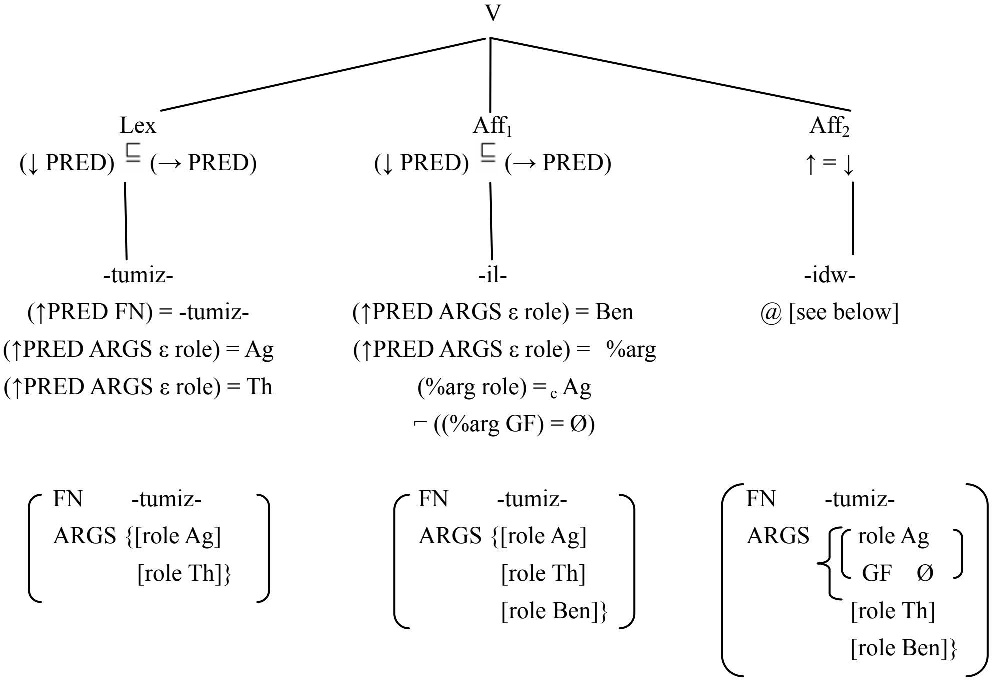
Let us take these morphological fragments and assign a morphological representation to each of them in the form of a partial interface tree. To obtain this interface tree, we need the annotation principle in (12):

(12) Annotation principle:

If there is/are a-structure-changing affix(es) on the verb, annotate the last a-structure-changing affix with ↑ = ↓. Annotate the verb root and any other a-structure-changing affix with the subsumption equation (↓ PRED)⊑(→ PRED).

The interface tree for (10a) is shown in (13):

(13) Interface tree for the well-formed verb form -tumiz-il-idw- (send-appl-pass)



@ { (↑PRED ARGS ε) = %arg [suppress agent]
(%arg role) = Ag
(%arg GF) = Ø

| ¬ (↑PRED ARGS role) = Ag [suppress benefactive]
(↑PRED ARGS ε) = %arg
(%arg role) = Ben
(%arg GF) = Ø

| ¬ { (↑PRED ARGS role) = Ag | (↑PRED ARGS role) = Ben }   [suppress recipient/
(↑PRED ARGS ε) = %arg   experiencer]
(%arg role) = Rpt/Exp
(%arg GF) = Ø

| ¬ { (↑PRED ARGS role) = Ag | (↑PRED ARGS role) = Ben | [suppress instrument]
(↑PRED ARGS role) = Rpt/Exp }
(↑PRED ARGS ε) = %arg
(%arg role) = Instr
(%arg GF) = Ø

| ¬ { (↑PRED ARGS role) = Ag | (↑PRED ARGS role) = Ben | [suppress theme]
(↑PRED ARGS role) = Rpt/Exp | (↑PRED ARGS role) = Instr }
(↑PRED ARGS ε) = %arg
(%arg role) = Th
(%arg GF) = Ø }

This interface tree has three nodes. The first one dominates the verb root, which is labelled Lex. The second one dominates the applicative affix and the last one dominates the passive affix. The tree shows the linear order of the morphemes on the verb, and therefore the order of any morpholexical process that each may be associated with.

The node labelled Lex is annotated with the equation (↓ PRED) ⊑ (→ PRED). The f-structure of Lex subsumes that of its right-sister node, which is Aff_1. Subsumption is necessary because the f-structure of the following node may contain more information than the f-structure of the current node (i.e. an additional semantic role licensed by an applicative affix). Besides, an equality equation cannot be assigned to this node because ultimately, the a-structure of the mother node V will be altered by the morpholexical operations triggered by the applicative and passive suffixes, and this f-structure should not be identical with that of the Lex node. Subsumption is defined as "a relation that holds between two f-structures *f* and *g* if *g* is compatible with but perhaps has more structure than *f*".[7]

The lexical entry of the verb *-tumiz-* "send" shows the number of arguments subcategorized by the verb and its semantic roles. It states that in its set of arguments in the PRED, there is an agent role, and there is a theme role.

Consider the lexical entry of the applicative affix:

[7] See Dalrymple (2001:161) for a formal definition of subsumption.

(14) Lexical entry of the applicative affix (benefactive)[8]

-il$_{\text{Ben}}$-	Aff	(↑PRED ARGS ε role) = Ben
		(↑PRED ARGS ε role) = %arg
		(%arg role) $=_c$ Ag
		⌐ [(%arg GF) = Ø]

The lexical entry in (14) states that the morpheme *-il*$_{\text{Ben}}$- is an affix, and that in the set of arguments of its PRED, there must be a benefactive role. Since (↑PRED ARGS ε role) = BEN is a defining constraint, it has the effect of introducing an additional role to the existing a-structure. This, of course, is licensed by the applicative affix.

The equations ((↑PRED ARGS ε role) = %arg), ((%arg role) $=_c$ Ag) and (⌐ ((%arg GF) = Ø)) together ensure that in the existing a-structure, there must be an agent role and that this agent role must be one that is not suppressed. These constraints capture the observation in Alsina (1999) that the applied argument cannot bear the most prominent semantic role. Ensuring that there is an agent role is sufficient for this affix, as the only more prominent semantic role on the thematic hierarchy than the benefactive role is the agent role.[9]

The relative prominence of semantic roles is also important in the formulation of the lexical entry for the passive affix, which is shown below:

[8] We assume that each type of applied argument is licensed by a different applicative affix, each of which has its own lexical entry, although in form all of them are the same. Support for this comes from Kinyarwanda, another Bantu language, in which there are different forms of applicative affixes. The form of the applicative affix is related to the role of the applied argument - benefactive: -ir/-er; instrument: -ish/-esh; and locative: -ho/-mo (Simango 1995:8).

[9] Lexical entry for the instrumental applicative affix:

-il$_{\text{Instr}}$-	Aff	(↑PRED ARGS ε role) = Instr
		(↑PRED ARGS ε) = %arg
		(%arg role) $=_c$ { Ag | Ben | Rpt/Exp}
		⌐ [(%arg GF) = Ø]

Lexical entry for the locative applicative affix:

-il$_{\text{Loc}}$-	Aff	(↑PRED ARGS ε role) = Loc
		(↑PRED ARGS ε) = %arg
		(%arg role) $=_c$ { Ag | Ben | Rpt/Exp | Instr }
		⌐[(%arg GF) = Ø]

(15) Lexical entry of the passive affix

-idw- Aff { (↑PRED ARGS ε) = %arg
(%arg role) = Ag
(%arg GF) = Ø

| ⌐ (↑PRED ARGS role) = Ag
(↑PRED ARGS ε) = %arg
(%arg role) = Ben
(%arg GF) = Ø

| ⌐ { (↑PRED ARGS role) = Ag | (↑PRED ARGS role) = Ben }
(↑PRED ARGS ε) = %arg
(%arg role) = Rpt/Exp
(%arg GF) = Ø

| ⌐ { (↑PRED ARGS role) = Ag | (↑PRED ARGS role) = Ben |
(↑PRED ARGS role) = Rpt/Exp }
(↑PRED ARGS ε) = %arg
(%arg role) = Instr
(%arg GF) = Ø

| ⌐ { (↑PRED ARGS role) = Ag | (↑PRED ARGS role) = Ben |
(↑PRED ARGS role) = Rpt/Exp |
(↑PRED ARGS role) = Instr }
(↑PRED ARGS ε) = %arg
(%arg role) = Th
(%arg GF) = Ø }

The lexical entry in (15) shows that the passive morpheme *-idw-* is an affix. Passivization involves the suppression of the highest semantic role. Here, the lexical entry of the passive affix ensures that the highest semantic role links to a null grammatical function. This highest role will no longer be available for linking. Moreover, passivization does not suppress the highest semantic role at any point in time, but it suppresses the highest semantic role at a particular point in the altering a-structure. There has to be a way to make reference to both the existing a-structure and the thematic hierarchy at the point of passivization. The constraint in the lexical entry in (15) does exactly this. The thematic hierarchy is built into the disjuncts. The constraint will always start by suppressing the agent, the highest semantic role on the thematic hierarchy, if there is an agent in the existing a-structure. If there is no agent, the next highest semantic role, the benefactive role, will be suppressed. The same logic applies for the other roles on the thematic hierarchy. As a summary:

(16) To suppress the highest thematic role in an existing a-structure,

   i. suppress agent.
   ii. If an agent does not exist, suppress benefactive.
   iii. If an agent and a benefactive do not exist, suppress recipient/experiencer.
   iv. If an agent and a benefactive and a recipient/experiencer do not exist, suppress instrument.
   v. If an agent and a benefactive and a recipient/experiencer and an instrument do not exist, suppress theme.

The last semantic role that can possibly be suppressed is a theme role. If the highest semantic role is a locative role, this means this is also the only role in the a-structure. Suppressing it will give an a-structure with no semantic roles in it. Besides, if it is the only role, it should be linked to the SUBJ function even without passivization, and there seems to be no reason for passivization to apply.

Let us revisit the interface tree in (13) and explain how the f-structure of the root V comes about. At the node Lex, the verb root *-tumiz-* "send", in its most basic form, subcategorizes for two arguments, an agent and a theme. This information comes from the lexical entry of the verb. The annotation (↓ PRED) ⊑ (→ PRED) on Lex passes the f-structure information of the PRED of Lex to the f-structure of PRED in its right-sister node, which is Aff$_1$. In Aff$_1$, there is an applicative affix, the lexical entry of which says that (i) the affix *-il*$_{\text{Ben}}$- licenses an extra benefactive role in the a-structure; and (ii) this applied role must not be the highest thematic role and that there must be an agent role, which is higher than the benefactive on the thematic hierarchy, in the a-structure. A modified f-structure results, which, according to the functional annotation on Aff$_1$ (↓ PRED) ⊑ (→ PRED), is passed to the f-structure of the PRED in its right-sister node, Aff$_2$. A passive affix is in Aff$_2$, and the lexical entry of the passive affix ensures that (i) a change to the a-structure of PRED will be brought about by the *-idw-* passive affix; and (ii) the most prominent semantic role in the a-structure of PRED is suppressed, meaning it is linked to a null GF. The a-structure of PRED, shown in (17), will have all the necessary a-structure modifications made to it after the sequential application of the applicative and passive operations on the verb root:

(17)

$$\begin{pmatrix} \text{FN} & \text{-tumiz-} \\ \text{ARGS} & \left\{ \begin{matrix} \begin{bmatrix} \text{role Ag} \\ \text{GF} \quad \varnothing \end{bmatrix} \\ [\text{role Th}] \\ [\text{role Ben}] \end{matrix} \right\} \end{pmatrix}$$

It is this a-structure that will be passed up to the root V according to the functional annotation ↑ = ↓ on Aff_2. The semantic roles will be linked to grammatical functions. The mapping is shown below:

(18)

-tumiz-il-idw-	<	Ag	Ben	Th >
		Ø		
AOP[10]			[-r]	[+o]
Defaults			[-r]	
			S/O	O
Well-Formedness Conditions			S	O

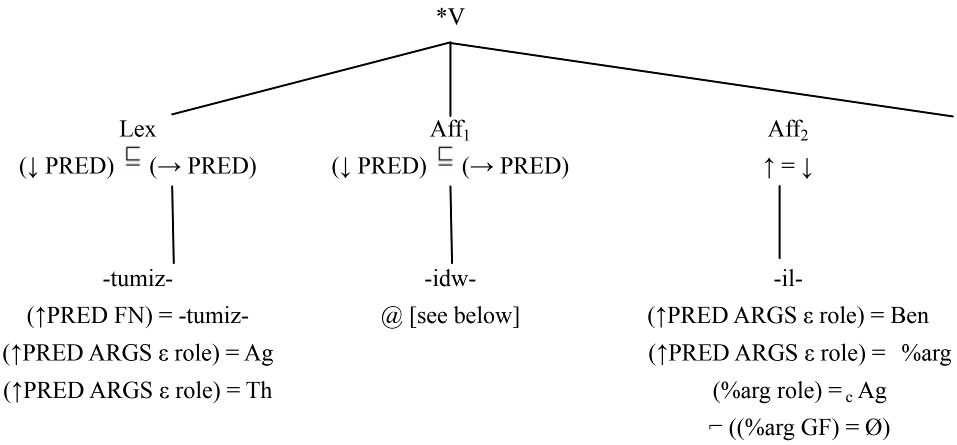
This accounts for the grammatical function realization in (10a).

3.2.3 Accouting for **-tumiz-idw-il-* (send-pass-appl)

The ungrammaticality of (10b), with the partial verb form **-tumiz-idw-il-* (send-pass-appl), can be easily accounted for. Here is the interface tree for (10b):

[10] AOP stands for ‘Asymmetric Object Parameter’. The AOP states that only one role can be intrinsically classified unrestricted [-r] (Bresnan and Moshi 1990:172). The AOP holds in Chichewa (Alsina and Mchombo 1989; Bresnan and Moshi 1990), thus, the theme role must be classified [+o] but not [-r] as the benefactive role has been classified [-r].

(19) Interface tree for the ill-formed verb form *-*tumiz-idw-il-* (send-pass-appl)



$$\left[\begin{array}{ll} \text{FN} & \text{-tumiz-} \\ \text{ARGS} & \{[\text{role Ag}] \\ & \;\;[\text{role Th}]\} \end{array}\right] \qquad \left[\begin{array}{ll} \text{FN} & \text{-tumiz-} \\ \text{ARGS} & \left\{\begin{array}{l} \left[\begin{array}{ll} \text{role} & \text{Ag} \\ \text{GF} & \text{Ø} \end{array}\right] \\ [\text{role Th}] \\ [\text{role Ben}]\} \end{array}\right. \end{array}\right]$$

@ { (↑PRED ARGS ε) = %arg
(%arg role) = Ag
(%arg GF) = Ø

| ¬ (↑PRED ARGS role) = Ag
(↑PRED ARGS ε) = %arg
(%arg role) = Ben
(%arg GF) = Ø

| ¬ { (↑PRED ARGS role) = Ag | (↑PRED ARGS role) = Ben }
(↑PRED ARGS ε) = %arg
(%arg role) = Rpt/Exp
(%arg GF) = Ø

| ¬ { (↑PRED ARGS role) = Ag | (↑PRED ARGS role) = Ben |
(↑PRED ARGS role) = Rpt/Exp }
(↑PRED ARGS ε) = %arg
(%arg role) = Instr
(%arg GF) = Ø

| ⌐ { (↑PRED ARGS role) = Ag | (↑PRED ARGS role) = Ben |
(↑PRED ARGS role) = Rpt/Exp | (↑PRED ARGS role) = Instr }
(↑PRED ARGS ε) = %arg
(%arg role) = Th
(%arg GF) = Ø }

The passive affix *-idw-* is under Aff_1, which immediately precedes the applicative affix in Aff_2. This order is a reflection of the passive operation being applied before the applicative operation. The a-structure information is passed from Lex to Aff_1. At Aff_1, the passive affix suppresses the highest semantic role such that it is linked to a null GF. This information is in turn passed on to the following morpheme Aff_2, where applicativization takes place. As specified by the lexical entry of the applicative affix, an additional benefactive role is introduced into the a-structure. The last two constraints for the applicative affix, however, cannot be satisfied. In the current a-structure, there is no agent role which is not at the same time linked to a null GF. The a-structure becomes ill-formed, and hence the ungrammaticality of (10b).

The verb form is morphologically licensed, i.e. in principle the verb can be derived. But this verb form does not have a well-formed a-structure, as not all constraints imposed by the applicative affix can be satisfied. The verb form, even if it could be formed at m-structure, cannot receive any grammatical function realization. As a result, a constructed example like (13b), even with nominals in the ordinary GF positions (c/f (10a)), is ungrammatical.

3.3 Advantages over Alsina's (1999) Treatment of Morpheme Ordering

The present analysis has a number of advantages over Alsina's treatment of morpheme ordering. These include: (i) the possibility of referring to intermediate, changing a-structures with the help of a morphology-syntax interface tree, without creating temporary, unwanted word forms; and (ii) a- to f-structure mapping will only take place once, from the "completed" a-structure after all the relevant morpholexical processes have taken place. We shall look at each of these in more detail.

In the present approach, the internal structure of the word formed via applicativization and passivization is represented in the form of an interface tree between morphology and syntax. It is here that any relevant morpholexical operation is represented. The word is parsed into its component stem and affixes, and an a-structure change can be thought of as taking place right there and then – “at the level of the information lexically associated with the affixes and not at the level of the derived word” (Sadler and Nordlinger 2004:171). Each relevant affix causes a change in a-structure in a particular way. This alternation targets the a-structure associated with the preceding morpheme(s). That the order of morpholexical operations is reflected by the order of morphemes is captured.

In this approach and unlike in Alsina’s proposal, we do not assume that intermediate morphological forms are created after each affixation of a morpheme. Alsina (1999:34) explicitly states that a new lexical item is created upon the affixation of an a-structure-changing morpheme, and that yet another such morpheme can be attached to this new lexical item. Intermediate morphological forms seem unnecessary and unmotivated, other than for the need in Alsina’s analysis to keep track of the order of morphemes and therefore the order of morpholexical operations. It also seems that such forms cannot be avoided – if a new a-structure is assumed to be associated with some word form, new intermediate lexical items are bound to appear.

No intermediate lexical items are created in the present analysis. By representing a fully derived lexical item as a morphology-syntax interface tree, it is possible to refer to intermediate a-structures without assuming intermediate word forms. The interface tree makes it possible to make reference to a particular level of representation (a-structure in this case) associated with a particular morpheme.

Once all the alternations to a-structure are completed, a- to f-structure mapping is performed. Only the arguments of well-formed a-structures will have GF realizations at f-structure. Ill-formed a-structures simply cannot serve as the input for a- to f-structure mapping. That the a- to f-structure mapping principles will only be applied once and that no intermediate lexical items are assumed make that present analysis a more elegant one.

4. Conclusions

Applicativization is highly productive in a language like Chichewa. The applicative affix augments the a-structure of a verb by bringing in an additional semantic role, which is most frequently a benefactive, instrument or locative role. It is not uncommon to find cases where there is more than one a-structure changing morpheme on the verb. In this paper, we have looked at one such verb form – a verb root is affixed with an applicative affix and a passive affix.

We have also shown in this paper that the structure of the word, with the verb root, applicative affix and passive suffix, can be represented in the form of a morphology-syntax interface tree, which makes it possible to refer to not only parts of the word but also the levels of representation that are associated with each morpheme. We were particularly interested in the a-structure, as this is the structure that the applicative and passive affixes alter. More importantly, these morphemes alter the existing a-structure, one that is the result of the interaction between the verb root and any other a-structure-changing morpheme that precedes in morphological form the morpheme in question. With the interface tree, it is possible to make reference to an intermediate a-structure, one that is associated with a particular morpheme on the tree, without having to assume intermediate lexical items as in Alsina's analysis.

References

Alsina, Alex. 1999. Where's the Mirror Principle? *The Linguistic Review*, 16(1):1-42.

Alsina, Alex and Sam A. Mchombo. 1989. Object asymmetries in the Chichewa applicative construction. Ms. Department of Linguistics, Stanford University, Stanford, CA. and the University of California, Berkeley.

Alsina, Alex and Sam A. Mchombo. 1990. The syntax of applicatives in Chichewa: problems for a theta-theoretic asymmetry. *Natural Language and Linguistic Theory*, 8(4): 493-506.

Alsina, Alex and Sam A. Mchombo. 1993. Object asymmetries and the applicative construction in Chichewa. In *Theoretical Aspects of Bantu Grammar*, ed. Sam A. Mchombo. Stanford, CA.: CSLI Publications.

Baker, Mark. 1985. The Mirror Principle and morphosyntactic explanation. *Linguistic Inquiry*, 16: 373-416.

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

Bresnan, Joan and Lioba Moshi. 1990. Object asymmetries in comparative Bantu syntax. *Linguistic Inquiry*, 21: 147-185.

Bresnan, Joan and Jonni Kanerva. 1989. Locative inversion in Chichewa: a case study of factorization in grammar. *Linguistic Inquiry*, 20(1): 1-50.

Dalrymple, Mary. 2001. *Syntax and Semantics Vol. 34, Lexical Functional Grammar*. San Diego, CA.: Academic Press.

Falk, Yehuda. 2001. *Lexical-Functional Grammar: an introduction to parallel constraint-based syntax*. Stanford, CA.: CSLI Publications.

Mchombo, Sam A. 2004. *The Syntax of Chichewa*. Cambridge: Cambridge University Press.

Nordlinger, Rachel. 1998. *Constructive Case: Evidence from Australian Languages*. Stanford, CA.: CSLI Publications.

Sadler, Louisa and Rachel Nordlinger. 2004. Relating morphology to syntax. *Projecting Morphology*. Stanford, CA.: CSLI Publications.

USING F-STRUCTURES IN MACHINE TRANSLATION EVALUATION

Karolina Owczarzak Yvette Graham
Josef van Genabith

National Centre for Language Technology
School of Computing
Dublin City University

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King
(Editors)

2007

CSLI Publications

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

Abstract

Despite a growing interest in automatic evaluation methods for Machine Translation (MT) quality, most existing automatic metrics are still limited to surface comparison of translation and reference strings. In this paper we show how Lexical-Functional Grammar (LFG) labelled dependencies obtained from an automatic parse can be used to assess the quality of MT on a deeper linguistic level, giving as a result higher correlations with human judgements.

1 Introduction

The use of automatic evaluation metrics became quite widespread in the Machine Translation (MT) community, mainly because such metrics provide an inexpensive and fast way to assess translation quality. It would be highly impractical to employ humans every time MT developers wished to test whether the changes in their system are reflected in the quality of the translations, so the appearance of string-based evaluation metrics such as BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) have been a great boost to the field. Both BLEU and NIST score a candidate translation on the basis of the number of *n*-grams shared with one or more reference translations, with NIST additionally using frequency information to weigh certain *n*-grams more than others. The metrics are fast to apply and intuitively easy to understand; however, these advantages come at a price. An automatic comparison of *n*-grams measures only the surface string similarity of the candidate translation to one or more reference strings, and will penalize any (even admissible and well-motivated) divergence from them. In effect, a candidate translation expressing the source meaning accurately and fluently will be given a low score if the lexical and syntactic choices it contains, even though perfectly legitimate, are not present in at least one of the references. Necessarily, this score would differ from a much more favourable human judgement that such a translation would receive.

The adequacy of string-based comparison methods has been questioned repeatedly within the MT community, with strong criticism for insensitivity to perfectly legitimate syntactic and lexical variation which can occur between the candidate and reference. However, almost all attempts at creating better metrics have been limited to the incorporation of local paraphrasing and/or surface reordering of elements, while ignoring structural levels of representation.

In this paper, we present a novel method that automatically evaluates the quality of translation based on the labelled dependency structure of the sentence, rather than its surface form. Dependencies abstract away from some of the particulars of the surface string (and CFG (Context-Free Grammar) tree) realization and provide a more "normalized" representation of (some) syntactic variants of a given sentence. The translation and reference files are analyzed by a treebank-based, probabilistic Lexical-Functional Grammar

(LFG) parser (Cahill et al., 2004), which produces a set of labelled dependency triples for each input. The translation set is compared to the reference set, and the number of matches is calculated, giving the precision, recall, and f-score for each particular translation.

In an experiment on 5,007 sentences of Chinese-English newswire text with associated segment-level human evaluation from the Linguistic Data Consortium's (LDC) Multiple Translation project,[1] we compare the LFG-based evaluation method with other popular metrics like BLEU, NIST, General Text Matcher (GTM) (Turian et al., 2003), Translation Error Rate (TER) (Snover et al., 2006),[2] and METEOR (Banerjee and Lavie, 2005), and we show that our labelled dependency representations lead to a more accurate evaluation that correlates better with human judgment. Although evaluated on a different test set, our method also outperforms the correlation with human scores reported for an earlier unlabelled dependency-based method presented in Liu and Gildea (2005).

The remainder of this paper is organized as follows: Section 2 gives a basic introduction to LFG; Section 3 describes related work; Section 4 describes our method; Section 5 gives results of two experiments on 5,007 sentences of Chinese-English newswire text from the Multiple Translation project; Section 6 discusses ongoing work; Section 7 concludes.

2 Lexical-Functional Grammar

In Lexical-Functional Grammar (Kaplan and Bresnan, 1982; Bresnan, 2001), sentence structure is represented in terms of c(onstituent)-structure and f(unctional)-structure. C-structure represents the word order of the surface string and the hierarchical organisation of phrases in terms of CFG trees. F-structures are recursive feature (or attribute-value) structures, representing abstract grammatical relations, such as SUBJ(ect), OBJ(ect), OBL(ique), ADJ(unct), etc., approximating to predicate-argument structure or simple logical forms. C-structure and f-structure are related in terms of functional annotations (attribute-value structure equations) which describe f-structures and are placed on c-structure trees.

While c-structure is sensitive to surface rearrangement of constituents, f-structure abstracts away from some of the particulars of the surface realization. The sentences *John resigned yesterday* and *Yesterday, John resigned* will receive different tree representations, but identical f-structures, shown in (1).

[1] http://www.ldc.upenn.edu/

[2] We omit HTER (Human-Targeted Translation Error Rate), as it is not fully automatic and requires human input.

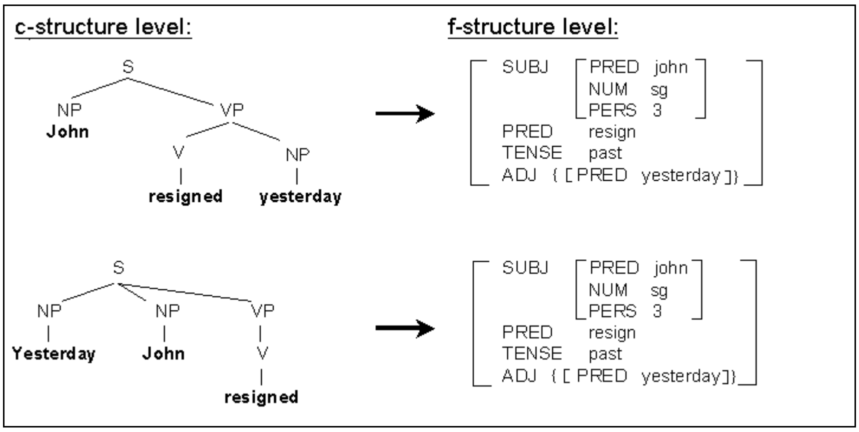


Figure 1. C-structure and f-structure

Note that if these sentences were a translation-reference pair, they would receive a less-than-appropriate score from string-based metrics. For example, BLEU with add-one smoothing[3] gives this pair a score of 0.76. This is because, although all three unigrams from the "translation" (*John*; *resigned*; *yesterday*) are present in the reference (*Yesterday*; *John*; *resigned*), the "translation" contains only one bigram (*John resigned*) that matches the "reference" (*Yesterday John*; *John resigned*), and no matching trigrams.

The f-structure can also be described in terms of a flat set of triples. In triples format, the f-structure in (1) is represented as shown in (2). The representation in (2) is simplified in that it omits index numbers which are carried by the words (this keeps track of multiple tokens of the same lexical item in a single sentence).

SUBJ(resign, john)
PERS(john, 3)
NUM(john, sg)
TENSE(resign, past)
ADJ(resign, yesterday)
PERS(yesterday, 3)
NUM(yesterday, sg)

Figure 2. A set of dependencies in the triples format

[3] We use smoothing because the original BLEU metric gives zero points to translations with fewer than one four-gram in common with the reference. We note also that BLEU is not intended for use at the segment level, but show this example for illustration only. In this example, we also ignore the punctuation in the segments to simplify things.

Cahill et al. (2004) presents a set of Penn-II Treebank-based LFG parsing resources. Their approach distinguishes 32 types of dependencies, including grammatical functions and morphological information. This set can be divided into two major groups: a group of predicate-only dependencies and a group of non-predicate (atomic) dependencies. Predicate-only dependencies are those whose path ends in a predicate-value pair, describing grammatical relations. For example, for the f-structure in (1), predicate-only dependencies would include: {SUBJ(resign, john), ADJ(resign, yesterday)}. Other predicate-only dependencies include: *apposition*, *complement*, *open complement*, *coordination*, *determiner*, *object*, *second object*, *oblique*, *second oblique*, *oblique agent*, *possessive*, *quantifier*, *relative clause*, *topic*, and *relative clause pronoun*. The remaining non-predicate dependencies are: *adjectival degree*, *coordination surface form*, *focus*, complementizer forms: *if*, *whether*, and *that*, *modal*, *number*, *verbal particle*, *participle*, *passive*, *person*, *pronoun surface form*, *tense*, and *infinitival clause*.

Such dependencies are often the basis of parser evaluation, where the quality of the f-structures produced automatically can be checked against a set of gold standard sentences annotated with f-structures by a linguist. The evaluation is conducted by calculating the precision and recall between the set of dependencies produced by the parser and the set of dependencies derived from the human-created f-structure. Usually, two versions of f-score are calculated: one for all the dependencies for a given input and a separate one for the subset of predicate-only dependencies.

In the experiments reported in this paper, we use the LFG parser developed by Cahill et al. (2004), which automatically annotates input text with c-structure trees and f-structure dependencies, obtaining high precision and recall rates.[4]

3 Related Research

3.1 String-Based Metrics

The insensitivity of BLEU and NIST to perfectly legitimate syntactic and lexical variation has been raised, among others, in Callison-Burch et al. (2006), but the criticism is widespread. Even the creators of BLEU point out that it may not correlate particularly well with human judgment at the sentence level (Papineni et al., 2002).

Recently a number of attempts to remedy these shortcomings have led to the development of other automatic MT evaluation metrics. Some of them concentrate mainly on allowing greater differences in word order between the translation and the reference, like General Text Matcher (Turian et al., 2003), which calculates precision and recall for translation-reference

[4] A demo of the parser can be found here:
http://lfg-demo.computing.dcu.ie/lfgparser.html

pairs, weighting contiguous string matches more than non-sequential matches, or Translation Error Rate (Snover et al., 2006), which computes the number of substitutions, insertions, deletions, and shifts necessary to transform the translation text to match the reference. Others try to accommodate both syntactic and lexical differences between the candidate translation and the reference, like CDER (Leusch et al., 2006), which employs a version of edit distance for word substitution and reordering; or METEOR (Banerjee and Lavie, 2005), which uses stemming and WordNet[5]-based synonymy. Kauchak and Barzilay (2006) and Owczarzak et al. (2006) use paraphrases in conjunction with BLEU and NIST evaluation to increase the number of matches between the translation and the reference; the paraphrases are either taken from WordNet (Kauchak and Barzilay, 2006) or derived from the test set itself through automatic word and phrase alignment (Owczarzak et al., 2006). Another metric making use of synonyms is the linear regression model developed by Russo-Lassner et al. (2005), which makes use of stemming, WordNet synonymy, verb class synonymy, matching noun phrase heads, and proper name matching. Kulesza and Shieber (2004), on the other hand, train a Support Vector Machine using features such as proportion of *n*-gram matches and word error rate to judge a given translation's distance from human-level quality.

3.2 Dependency-Based Metrics

The metrics described in Section 3.1 use only string-based comparisons, even while taking into consideration reordering. By contrast, Liu and Gildea (2005) present three metrics that use syntactic and unlabelled dependency information. Two of these metrics are based on matching syntactic subtrees between the translation and the reference, and one is based on matching headword chains, i.e. sequences of words that correspond to a path in the *unlabelled* dependency tree of the sentence. Dependency trees are created by extracting a headword for each node of the syntactic tree, according to the rules used by the parser of Collins (1999), where every subtree represents the modifier information for its root headword. The dependency trees for the translation and the reference are converted into flat headword chains, and the number of overlapping *n*-grams between the translation and the reference chains is calculated. Our method, by contrast, uses *labelled* LFG dependencies, partial matching, and *n*-best parses, allowing us to considerably outperform Liu and Gildea's (2005) highest correlations with human judgement (they report 0.144 for the correlation with human fluency judgement and 0.202 for the correlation with human overall judgement), although it has to be kept in mind that such comparison is only tentative, as their correlation results are calculated on a different test set.

[5] http://wordnet.princeton.edu/

4 LFG F-structure in MT Evaluation

As for parsing, the process underlying the evaluation of f-structure quality against a gold standard can be used in automatic MT evaluation as well: we parse the translation and the reference, and then, for each sentence, we check the set of translation dependencies against the set of reference dependencies, counting the number of matches. As a result, we obtain the precision and recall scores for the translation, and we calculate the f-score for the given pair. Because we are comparing two outputs that were produced automatically, there is a possibility that the result will not be noise-free.

To assess the amount of noise that the parser may introduce, we conducted an experiment where 100 English sentences were modified by hand in such a way that the position of adjuncts was changed, but the sentence remained grammatical and the meaning was not changed, as shown in (1).

(1) a. We must change this system, Commissioner.
b. Commissioner, we must change this system.

This way, an ideal parser should give both the source and the modified sentence the same f-structure, similarly to the case presented in (1). The modified sentences were treated like a translation file, and the original sentences played the part of the reference. Each set was run through the parser. We evaluated the dependency triples obtained from the "translation" against the dependency triples for the "reference", calculating the f-score, and applied other metrics (TER, METEOR, BLEU, NIST, and GTM) to the set in order to compare scores. The results, including the distinction between f-scores for all dependencies and predicate-only dependencies, are given in Table 1.

	upper bound	**modified**
TER	0.0	6.417
METEOR	1.0	0.9970
BLEU	1.0	0.8725
NIST	11.5232	11.1704 (96.94%)
GTM	100	99.18
dep f-score	100	96.56
dep_preds f-score	100	94.13

Table 1. Scores for sentences with reordered adjuncts

The baseline column shows the upper bound for a given metric: the score which a perfect translation, word-for-word identical to the reference, would obtain.[6] In the other column we list the scores that the metrics gave to the "translation" containing reordered adjuncts. As can be seen, the dependency and predicate-only dependency scores are lower than the perfect 100, reflecting the noise introduced by the parser.

To show the difference between the scoring based on LFG dependencies and other metrics in an ideal situation, we created another set of a hundred sentences with reordered adjuncts, but this time selecting only those reordered sentences that were given the same set of dependencies by the parser (in other words, we simulated having the ideal parser). As can be seen in Table 2, other metrics are still unable to tolerate legitimate variation in the position of adjuncts, because the sentence surface form differs from the reference; however, it is not treated as an error by the parser.

	upper bound	**modified**
TER	0.0	7.841
METEOR	1.0	0.9956
BLEU	1.0	0.8485
NIST	11.1690	10.7422 (96.18%)
GTM	100	99.35
dep f-score	100	100
dep_preds f-score	100	100

Table 2. Scores for sentences with reordered adjuncts in an ideal situation

5 Correlations with Human Judgement - MultiTrans

5.1 Experimental Design

To evaluate the correlation with human assessment, we used the data from the Linguistic Data Consortium Multiple Translation Chinese (MTC) Parts 2 and 4, which consists of multiple translations of Chinese newswire text, four human-produced references, and segment-level human evaluation scores for a subset of the translation-reference pairs. Although a single translated segment was always evaluated by more than one judge, the judges used a different reference every time, which is why we treated each translation-reference-human score triple as a separate segment. In effect, the test set created from this data contained 16,800 segments. We randomly selected 5,007 segments as our test set, while the remaining segments served as a training corpus for those versions of our test method that required the training

[6] Two things have to be noted here: (1) in case of NIST the perfect score differs from text to text, which is why we provide the percentage points as well, and (2) in case of TER the lower the score, the better the translation, so the perfect translation will receive 0, and there is no bound on the score, which makes this particular metric extremely difficult to directly compare with others.

of weights. As in the previous experiment, the translation was scored using BLEU, NIST, GTM, TER, METEOR, and our labelled dependency-based method.

5.2 Labelled Dependency-Based Method

The results, presented in Table 3, show that although the basic labelled dependency-based evaluation method achieves a high correlation with human scores for translation fluency, it is only average in its correlation with human judgement of translation accuracy, falling short of some string-based metrics. This suggests that the dependency f-score, at least as calculated in the evaluation method used for parsing, might not be the ideal reflection of the true quality of the translation. This could be due to the dependency triple f-score assigning equal weight to each dependency triple. For parser evaluation this is appropriate, but for MT evaluation it may not be. Since the task of automatic MT evaluation attempts to replicate human judgments of a given candidate translation for adequacy and fluency, the type of relation that the dependency encodes may influence its importance in the evaluation.

H_FL		H_AC		H_AV	
GTM	0.172	**METEOR**	0.278	**METEOR**	0.242
dep	0.161	**NIST**	0.273	**NIST**	0.238
BLEU	0.155	**dep**	0.256	**dep**	0.235
METEOR	0.149	**dep_preds**	0.240	**dep_preds**	0.216
NIST	0.146	**GTM**	0.203	**GTM**	0.208
dep_preds	0.143	**BLEU**	0.199	**BLEU**	0.197
TER	0.133	**TER**	0.192	**TER**	0.182

Table 3: Pearson's correlation between human scores and evaluation metrics. Legend: dep = dependency-based method, _preds = predicate-only, M = METEOR, H_FL = human fluency score, H_AC = human accuracy score, H_AV = human average score.

For example, predicate-only dependencies (like SUBJ, OBJ, ADJ, etc.) encode a specific relation between two items, and only when both of these items happen to occur in that specific labelled dependency relation is the dependency counted as a match against the reference. This proves problematic when using dependencies to evaluate MT output, since we might encounter lexical variation: in a candidate-reference pair *John quit yesterday* and *John resigned yesterday* none of the predicate-only dependencies will match, e.g. candidate: {SUBJ(quit, John), ADJ(quit, yesterday)}, reference: {SUBJ(resign, John), ADJ(resign, yesterday)}. The predicate-only score would therefore be zero. However, if we allow partial matches for predicate-only dependencies, this should accommodate cases where an object might find itself in the correct relation, but with an incorrect partner. This modified method would give us an f-score of 0.5 (candidate: {SUBJ(quit,_), SUBJ(_,John), ADJ(quit,_), ADJ(_,yesterday)}; reference: {SUBJ(resign,_), SUBJ(_,John), ADJ(resign,_), ADJ(_, yesterday)}).

Another problem stemming from the equal treatment of all dependencies is that lexical items and their resulting grammatical categories naturally differ with respect to how many atomic (non-predicate) dependencies they generate. For example, a noun phrase like *the chairman* generates three atomic dependencies from its atomic features PERS, NUM and DET, whereas a verb like *resign* might generate only a single atomic dependency for its TENSE feature. As a result, the f-score for the overall dependency triples match implicitly weights the words in the sentence by the number of atomic features the word receives at f-structure level. For example, if an MT system incorrectly translates the noun *chairman*, it affects the final score three times as much as an incorrect translation of the word *resign*. Individual lexical items can easily be given an even influence on the final score by assigning each an equal weight in the overall score, irrespective of the number of dependency relations they generate. This means that a partial f-score is calculated at the lexical item level from all the dependencies relating to this item, and then all the partial f-scores are averaged at the segment level to give the final f-score for the segment.

In addition to this, the information encoded in predicate-only dependencies and atomic feature-value pairs could relate to human judgments of translation quality differently. We investigated this by calculating a score for the atomic features only and a separate score for the predicate-only triples and combining the two scores using automatically optimized weights.

We implemented a number of ways in which predicate and atomic dependencies combine in order to arrive at the final sentence-level f-score, and we calculated the correlation between each of these combinations and human assessment of translation quality. The results of these modifications are presented in Table 4. Interestingly, all the improved f-score calculations raise the correlation with human MT evaluation scores over the values displayed by the original f-score calculation; the only scores showing lower correlation than the traditional method are partial f-scores for predicates-only and atomic-features-only. It is also important to note that this increase in correlation, even if not enough to outperform the highest-ranking string-based metrics in the areas of human fluency and accuracy judgement (GTM and METEOR, respectively), is nevertheless enough to place one of the dependency-based f-score calculations (partial match for predicate dependencies plus all non-grouped atomic dependencies) at the top of the ranking when it comes to the general correlation with the average human score (which combines fluency and accuracy).

Method	**H_FL**	**Method**	**H_AC**	**Method**	**H_AV**
p+a(g)	0.1653	**pm+a**	0.2666	**pm+a**	0.2431
pm+a	0.1648	**w_pm+w_a(g)**	0.2648	**w_pm+w_a(g)**	0.2415
pm+a(g)	0.1648	**pm+a(g)**	0.2631	**pm+a(g)**	0.2409
w_pm+w_a(g)	0.1641	**w_p+w_a(g)**	0.2560	**p+a(g)**	0.2360
w_p+w_a(g)	0.1631	**a(g)**	0.2560	**w_p+w_a(g)**	0.2352
original	0.1613	**original**	0.2557	**a(g)**	0.2348
a(g)	0.1610	**p+a(g)**	0.2547	**original**	0.2347
pm	0.1579	**pm**	0.2479	**pm**	0.2283
p	0.1427	**p**	0.2405	**p**	0.2165

Table 4: Pearson's correlation between human scores and variations of f-score dependency scores. Types of dependencies: p = predicate, pm = partial match for predicate, a = atomic, a(g) = atomic grouped by predicate, w_ = optimally weighted, original = basic f-score, H_FL = human fluency score, H_AC = human accuracy score, H_AV = human average score.

Note also that almost all versions of our method show higher correlations than the results reported in Liu and Gildea (2005): 0.144 for the correlation with human fluency judgement, 0.202 for the correlation with human overall judgement, with the proviso that the correlations are calculated on a different test set.

6 Current and Future Work

Fluency and accuracy are two very different aspects of translation quality, each with its own set of conditions along which the input is evaluated. Therefore, it seems unfair to expect a single automatic metric to correlate highly with human judgements of both fluency and accuracy at the same time. This pattern is very noticeable in Table 3: if a metric is (relatively) good at correlating with fluency, its accuracy correlation suffers (GTM might serve as an example here), and the opposite holds as well (see METEOR's scores). It does not mean that any improvement that increases the method's correlation with one aspect will result in a decrease in the correlation with the other aspect; but it does suggest that a possible direction for development would be to target these correlations separately, if we want our automated metrics to reflect human scores better. At the same time, string-based metrics might have already exhausted their potential when it comes to increasing their correlation with human evaluation; as has been pointed out before, these metrics can only tell us that two strings differ, but they cannot distinguish legitimate grammatical variance from ungrammatical variance. As the quality of MT improves, the community will need metrics that are more sensitive in this respect. After all, the true quality of MT depends on producing grammatical output which describes the same concepts (or proposition) as the source utterance, and the string identity with a reference is only a very arbitrary approximation of this goal.

In order to maximize the correlation with human scores of fluency, we plan to look more closely at the parser output, and implement some basic transformations which would allow an even deeper logical analysis of input (e.g. passive to active voice transformation).

As to the correlations with human judgments of accuracy, we found that adding WordNet synonyms to the matching process increases the scores. The use of synonyms in matching allows us to account for legitimate lexical variation that can occur between the translation and the reference. For example, if our "translation" in Figure 1 *John resigned yesterday* contained the verb *quit* instead of *resign*, the number of matches would decrease even though a human judge would be able to recognize the equivalent meaning; however, if we automatically search WordNet synonym sets and find that *quit* and *resign* are in fact synonyms, we can still count the match. Results of these experiments are presented in Owczarzak et al. (2007a,b).

7 Conclusions

In this paper we present a novel way of evaluating MT output. So far, most metrics have relied on comparing translation and reference on a string level. Even given reordering, stemming, and synonyms for individual words, current methods are still far from reaching human ability to assess the quality of translation. Our method compares the sentences on the level of their grammatical structure, as exemplified by their f-structure labelled dependency triples produced by an LFG parser. The labelled dependency-based method can be further augmented by allowing partial matching for predicate dependencies or WordNet synonyms. In our experiments we showed that one version of the dependency-based method correlates higher than any other metric with the average human score. The use of labelled dependencies in MT evaluation is a rather new idea and requires more research to improve it, but the method shows potential to become an accurate, yet automatic, evaluation metric.

Acknowledgements

This work was partly funded by Microsoft Ireland PhD studentship 2006-8 for the first author of the paper and Science Foundation Ireland studentship P07077-60101 for the second author. We would also like to thank our reviewers for their insightful comments. All remaining errors are our own.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation*

Measures for MT and/or Summarization at the Association for Computational Linguistics Conference 2005: 65-73. Ann Arbor, Michigan.

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

Aoife Cahill, Michael Burke, Ruth O'Donovan, Josef van Genabith, and Andy Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations, In *Proceedings of Association for Computational Linguistics 2004*: 320-327. Barcelona, Spain.

Chris Callison-Burch, Miles Osborne and Philipp Koehn. 2006. Re-evaluating the role of BLEU in Machine Translation Research. *Proceedings of the European Chapter of the Association for Computational Linguistics 2006*: 249-256. Oslo, Norway.

Michael J. Collins. 1999. Head-driven Statistical Models for Natural Language Parsing. Ph.D. thesis, University of Pennsylvania, Philadelphia.

George Doddington. 2002. Automatic Evaluation of MT Quality using N-gram Co-occurrence Statistics. *Proceedings of Human Language Technology Conference 2002*: 138-145. San Diego, California.

Ronald M. Kaplan and Joan Bresnan. 1982. *Lexical-functional Grammar: A Formal System for Grammatical Representation.* In J. Bresnan (ed.), The Mental Representation of Grammatical Relations. MIT Press, Cambridge.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. *Proceedings of Human Language Technology – North American Chapter of the Association for Computational Linguistics Conference 2006*: 45-462. New York, New York.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of* the *Conference on Theoretical and Methodological Issues in Machine Translation 2004*: 75-84. Baltimore, Maryland.

Gregor Leusch, Nicola Ueffing and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. *Proceedings of European Chapter of the Association for Computational Linguistics Conference 2006*: 241-248. Trento, Italy.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization at the Association for Computational Linguistics Conference 2005.* Ann Arbor, Michigan.

Karolina Owczarzak, Declan Groves, Josef van Genabith, and Andy Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT

Evaluation. *Proceedings of the Workshop on Statistical Machine Translation at the Human Language Technology – North American Chapter of the Association for Computational Linguistics Conference 2006*: 86-93. New York, New York.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007a. Dependency-Based Automatic Evaluation for Machine Translation. *Proceedings of the HLT-NAACL 2007 Workshop on Syntax and Structure in Statistical Machine Translation*: 86-93. Rochester, New York.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007b. Labelled Dependencies in Machine Translation Evaluation. *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation*: 104-111. Prague, Czech Republic.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of Association for Computational Linguistics Conference 2002*: 311-318. Philadelphia, Pennsylvania.

Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik. 2005. A Paraphrase-based Approach to Machine Translation Evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland, College Park, Maryland.

Mathew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciula. 2006. A Study of Translation Error Rate with Targeted Human Annotation. *Proceedings of the Association for Machine Translation in the Americas Conference 2006*: 223-231. Boston, Massachusetts.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and Its Evaluation. *Proceedings of MT Summit 2003*: 386-393. New Orleans, Luisiana.

DESIGNING AND IMPLEMENTING DISCRIMINANTS FOR LFG GRAMMARS

Victoria Rosén, Paul Meurer and Koenraad de Smedt
University of Bergen and Unifob AKSIS

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

We extend discriminant-based disambiguation techniques to LFG grammars. We present the design and implementation of lexical, morphological, c-structure and f-structure discriminants for an LFG-based parser. Chief considerations in the computation of discriminants are capturing all distinctions between analyses and relating linguistic properties to words in the string. Our work is mostly tested on Norwegian, but our approach is independent of the language and grammar.

1 Introduction

The use of linguistically motivated handwritten grammars in realistic applications is dependent on the capacity to automatically resolve ambiguities produced by the grammar. Statistical techniques for disambiguation by parse ranking require training of the parser on a previously analyzed and disambiguated corpus—a treebank. Quality controlled treebanks that can serve as gold standards cannot be constructed without considerable manual effort towards ambiguity resolution. Intelligent ways of minimizing these efforts have been the subject of earlier research in the context of different tasks and formalisms (Carter, 1997; Van der Beek et al., 2002; Oepen et al., 2004). In our work on treebanking by automatically parsing a corpus with an LFG grammar, we have employed and further developed such techniques.

In this paper we explain in depth how discriminants can be extended to LFG grammars and how we have implemented them. The paper is structured as follows. First we present previous work on discriminants. Then we describe our design of various types of discriminants for LFG grammars. These will be illustrated and motivated with examples parsed with the Norwegian grammar developed at the University of Bergen within the Parallel Grammar project (Butt et al., 2002). Furthermore, we describe their implementation, i.e. the computation of discriminants from linguistic structures. Finally, we discuss the presentation and use of discriminants. The LFG Parsebanker, a toolkit developed at the University of Bergen in the TREPIL[1] and LOGON[2] projects, implements the computation and presentation of LFG discriminants.

2 Previous Work on Discriminants

Discriminant-based disambiguation was first presented by Carter (1997) as a time-saving method for treebanking. Carter's aim was to train a linguistic analyzer for several domains and tasks, each one requiring a separate analyzed and disambiguated corpus. In this context, it is clearly desirable to optimize the efficiency

[†] This work was supported in part by a grant from the Research Council of Norway. We would like to thank John Maxwell at PARC for his help with implementation issues.

[1] http://gandalf.aksis.uib.no/trepil

[2] http://www.emmtee.net/

of manual disambiguation. Inspecting full analyses proved to be “a tedious and time-consuming task”. In contrast, a few lexical or structural properties are often sufficient to distinguish the one intended analysis from many other analyses. Examples of properties that involve relatively simple choices are PP attachment, word senses, and the arity of predicates. Calling such distinguishing properties *discriminants,* Carter implemented their identification and presentation in his TreeBanker tool. He designed various discriminants, including constituents, semantic triples, word senses, sentence types, and grammar rules used.

In the TreeBanker’s graphical interface, the user can label discriminants as either good or bad, or can leave them undecided. Carter defined a set of inferencing rules based on these decisions. If a discriminant is marked by the user as bad, then all analyses that contain this property are rejected, whereas if a discriminant is marked by the user as good, then only analyses that contain it are kept. Thus, the set of analyses is narrowed down, until only one analysis remains. Furthermore, a discriminant that is true only of analyses that have already been rejected must be bad. Conversely, a discriminant that is true of all the still undecided analyses must be good (assuming there is at least one good analysis). In the cases where a discriminant is inferred to be either good or bad for all analyses, it loses its discriminatory power, i.e. it is trivial, and hence it need not be presented to the user, who can thus concentrate on more relevant choices.

Carter pays special attention to “user-friendly” discriminants which are easy for humans to judge and are prominent in the display. The efficiency of this method, as compared to presenting all the full analyses to the user, can be appreciated from the fact that a combination of a small number of local ambiguities can result in a large number of analyses. Carter mentions an example with 154 analyses, for which 318 discriminants are computed, yet only two discriminant choices are necessary to select the correct analysis.

Discriminants have also been used in at least two other projects, both HPSG-based. In the context of the Alpino project (Van der Beek et al., 2002), a large treebank was built using manual disambiguation based on Carter’s principles but with a different design. Lexical discriminants, representing ambiguities that result from lexical analysis, are always presented to the annotator first, because it is claimed that lexical decisions are easy to make. Furthermore, constituent discriminants represent alternative groupings of words in constituents, and dependency triples represent alternative paths in a dependency tree. These can be compared to our c-structure and f-structure discriminants, which will be presented in sections 3.2 and 3.3 respectively.

The LinGO Redwoods project (Oepen et al., 2004) was aimed at building a dynamic treebank as a testbed for grammar development. Since grammar development presupposes frequent automatic reparsing of a corpus, automatic redisambiguation is highly desirable. This was achieved by storing the annotator’s discriminant choices and reapplying them when reparsing. To our knowledge, LinGO Redwoods was the first project to closely integrate treebanking and grammar development in this way. Properties related to constituents (i.e. use of a grammar rule over a specific substring), lexical items (part of speech), semantics (primary predicate) and node labeling were used as discriminants. With the help of a suitable tool for identifying and presenting these discriminants, an annotator performance of about 2000 sentences per week was achieved.

In all work with discriminants, Carter's rules for narrowing down the set of analyses based on the annotator's choice of discriminants, as well as his rules for narrowing down the set of discriminants so that only the nontrivial ones are kept, are essential. But even though, by means of Carter's rules, enough discriminant choices will eventually lead to a single analysis, this analysis is not necessarily the correct one. There may be no correct analysis among the ones that the parser produced, or a wrong discriminant choice could have eliminated the correct analysis. To assist the annotator in making the right choices, a sophisticated, user-friendly tool that identifies and presents discriminants together with specific analyses is indispensable. Both the TreeBanker and the discriminant tools used in Alpino and LinGO Redwoods aim to provide such assistance in the context of the grammars and parsers they operate with. However, the types of discriminants, their computation, and even their presentation are not universal, but depend on the grammar formalism, the parser, and on user-oriented and system-oriented design choices. To our knowledge, there has been no previous work on designing discriminants for LFG grammars and implementing them for an LFG-based parser such as the Xerox Linguistic Environment (XLE) (Maxwell and Kaplan, 1993).

3 Designing Discriminants for LFG

The number of analyses of realistic sentences provided by a grammar may run into the thousands. In such cases, disambiguation by the sequential inspection of individual structures is prohibitively time consuming. XLE provides packed c- and f-structures which are compact representations of all the information in all analyses. In XLE's native interface it is possible to disambiguate interactively by choosing between alternatives indicated in the packed structures. While an important property of packed structures is that they are concise from a computing standpoint (Maxwell and Kaplan, 1993), this property is nevertheless of little help towards efficient manual disambiguation, since for sentences with multiple ambiguities, packed structures may become too unwieldy for a human to cope with. Disambiguation with discriminants does not suffer from the complexity issue that packed structures have, since each discriminant is local and may be chosen independently of all others.

There are often a large number of elementary properties that are not shared by all analyses, such as local c-structure node configurations and labels or f-structure attributes and values. Any such elementary property is a candidate for being a discriminant, for all such properties actually discriminate between analyses. However, in many cases it is impossible for a human disambiguator to pick out such elementary properties in isolation. In order for them to be reliably recognizable as properties of the intended analysis, they must be related to words in the string. This is a crucial point in the design of discriminants.

The present work on discriminants is focused on how they may be defined and used in an optimal way for LFG grammars. Discriminants should be designed so as to automatically identify all possible distinctions between analyses and make these recognizable to the annotator. It is important that the discriminants contain enough information to make it possible to uniquely identify them, but little enough information that they remain elementary local properties. The graphs representing the c-structure and f-structure must be fully traversed to find all possible distinctions between structures. We have defined four major types of discriminants for LFG grammars: lexical discriminants, morphological discriminants, c-structure discriminants and f-structure discriminants.

3.1 Lexical and Morphological Discriminants

We agree with Van der Beek et al. (2002) that lexical ambiguities are often the easiest to resolve. Two types of discriminants are meant to aid in resolving lexical ambiguities: lexical discriminants and morphological discriminants.

A *lexical discriminant* is a word form with its lexical category. Consider the Norwegian sentence in example (1) and its two c-structures in figure 1.

(1) *Glade fisker svømmer.*
Glad fish swim/swimmer
"Glad fish swim." / "Glad ones fish a swimmer."

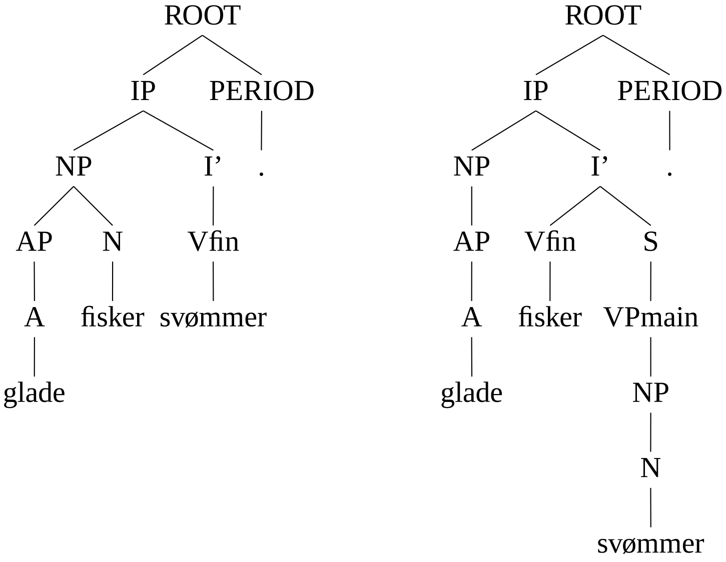


Figure 1: Two analyses for example (1), the left one corresponding to 'Glad fish swim', the right one to 'Glad ones fish a swimmer'

In this example, both *fisker* and *svømmer* may be either a noun or a verb, and because of this there are two quite different c-structures. The entire c-structures need not be examined, however, since determining the lexical category of either of these words is enough to determine which c-structure is the intended one. The relevant subtrees containing preterminal and terminal nodes for example (1) are shown in figure 2. Table 1 illustrates the representation of lexical discriminants for this example.

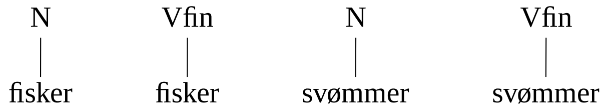


Figure 2: Subtrees defining lexical discriminants for example (1)

Table 1: Representation of lexical discriminants for example (1)

‘fisker’: N
‘fisker’: Vfin
‘svømmer’: N
‘svømmer’: Vfin

The lexical category specified in the discriminant is sometimes simply the traditional part of speech (e.g. N), sometimes a more fine-grained category (e.g. Vfin). Whatever preterminal node label occurs in the subtree will be the category in the discriminant.

Sometimes a word form may be ambiguous between different lexemes or between different forms of one lexeme within the same part of speech. This is the case in the present example. Even after the category N has been chosen by selecting the first discriminant in table 1, the word form *fisker* may still be an inflected form of the noun *fisk* “fish” or of the noun *fiske* “fishing”. Since lexical discriminants are not sufficient for the disambiguation of lexical ambiguities, we also define morphological discriminants. A *morphological discriminant* is a word with the tags it receives from morphological preprocessing. The two morphological analyses for the noun *fisker* are illustrated in figure 3, which shows a simplified version of the sublexical trees not usually displayed by XLE. The morphological discriminants for this example are represented as in table 2.

Table 2: Morphological discriminants for *fisker* in example (1)

fisk+Noun+Masc+Indef+Pl
fiske+Noun+Neut+Indef+Pl

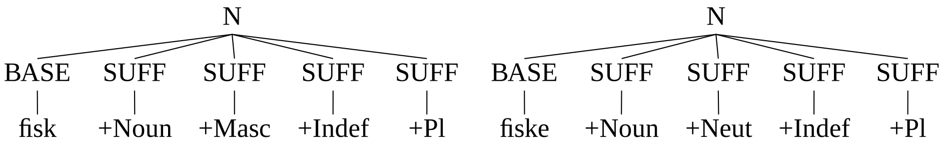


Figure 3: Two morphological analyses for *fisker*

Neither lexical nor morphological discriminants alone are sufficient for the full disambiguation of lexical ambiguities. As shown in the above examples, lexical discriminants cannot distinguish between different word forms that have different features and/or base forms but the same lexical category. In this example, full lexical disambiguation could have been achieved through selecting only morphological discriminants. This is not always the case, however, since not all words go through morphological preprocessing. For some words, morphosyntactic features may be directly encoded in the lexical entry. Therefore both lexical and morphological discriminants are necessary for lexical disambiguation. There are also cases where lexical ambiguities remain after all lexical and morphological discriminants have been chosen; we will return to these in section 3.3.

3.2 C-structure Discriminants

C-structure discriminants are important for the disambiguation of syntactic ambiguities. Their design aims at selecting an elementary local property of a tree. They are therefore based on minimal subtrees, a minimal subtree being defined as a mother node and her daughters. Since identical subtrees may occur more than once in the same analysis, these need to be related to the substring that they dominate. Example (2) involves two different PP attachment choices, as shown in the c-structure trees in figure 4. The substring which is relevant for the disambiguation of this example is shown with its bracketing in example (3). The simple breakdown of the substring into its immediate constituents is shown in the unlabeled bracketing in (3a), and the two different PP attachments are shown in the labeled bracketings in (3b) and (3c).

(2) *Vi fanget fisk med stang.*
We caught fish with fishing-rod
“We caught fish with a fishing rod.”

(3) (a) [[fisk] [med stang]]
(b) [$_{VPmain}$ [$_{NP}$ fisk] [$_{PP}$ med stang]]
(c) [$_{NP}$ [$_{N}$ fisk] [$_{PP}$ med stang]]

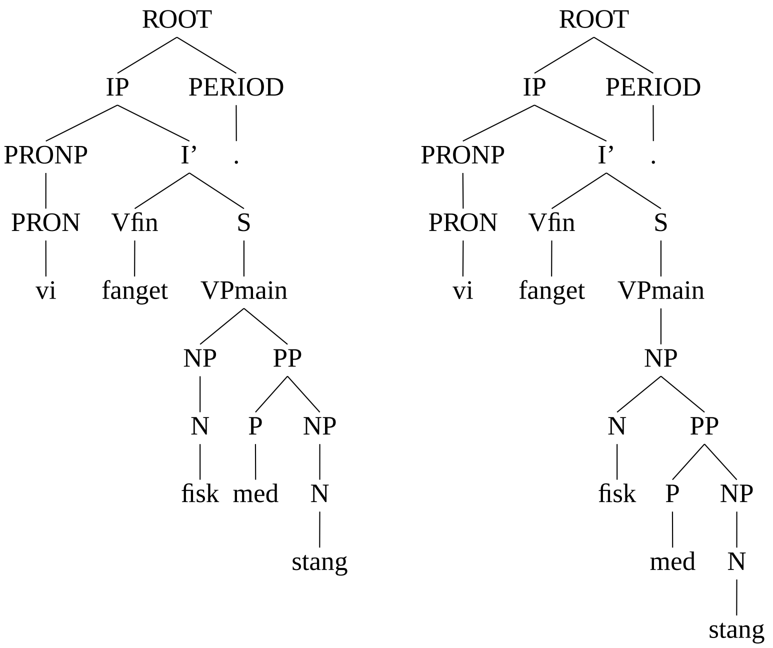


Figure 4: Two PP attachments for example (2)

C-structure discriminants are of two subtypes. An unlabeled top-level bracketing of a constituent substring is a *constituent discriminant*. A top-level bracketing of a constituent substring labeled by the rule which induces that bracketing is a *rule discriminant*. The c-structure discriminants for this example are shown in table 3.

Table 3: C-structure discriminants for the PP attachments in figure 4

fisk || med stang
VPmain → NP PP
NP → N PP

The top row in this table shows the representation of the constituent discriminant corresponding to the bracketed string in (3a). Instead of indicating the bracketing by enclosing the constituents in square brackets, the constituents are separated by two vertical bars. The second and third rows of the table illustrate the representation of rule discriminants, with the second row corresponding to (3b) and the third row corresponding to (3c). The representation of rule discriminants is simply expressed as a grammar rule, but this rule must be interpreted as the labeled bracketing of the string in question. Since rule discriminants are always displayed in a table cell underneath the corresponding constituent discriminant, it is always clear which substring the rule applies to.

Both types of c-structure discriminants can be useful: sometimes it is possible for an annotator to decide on the labeling as well as the bracketing, while in other cases one may wish to commit to a bracketing but not to a certain labeling. In the case in table 3, however, the constituent discriminant is actually trivial. Since both analyses share this constituent structure, the bracketing [[fisk] [med stang]] does not discriminate between analyses.

3.3 F-structure Discriminants

A c-structure may project more than one f-structure. In example (4), the constituent *hver time* may function as either OBJ or ADJUNCT.

(4) *Vi spiser hver time.*
we eat every hour
"We eat every hour."

F-structure discriminants are based on partial paths through f-structures. For f-structures it is not so apparent as for c-structures how to make local properties easily identifiable in discriminants, since the string is not represented in the f-structure. Therefore, the design of f-structure discriminants crucially exploits PRED values, which typically provide the most direct connection to words in the string. An f-structure discriminant is a minimal path through the f-structure from a PRED value to another PRED value or to an atomic value, a minimal path being one that does not cross any intermediate PRED values and does not contain cycles.

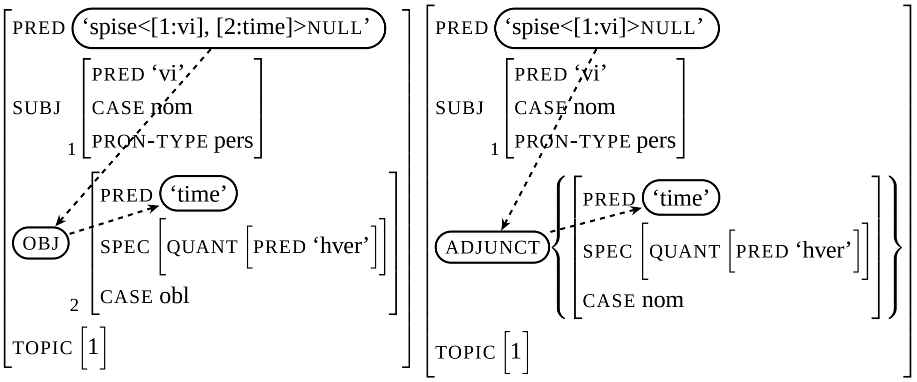


Figure 5: Simplified f-structures for example (4)

Table 4 represents some relevant f-structure discriminants for the example in figure 5. The empty brackets in the PRED values show the arity of the predicate. The first discriminant may thus be read: *the two-place predicate ‘spise’ has an* *object whose* PRED *value is 'time'*, while the second discriminant may be read: *the one-place predicate 'spise' has a set of adjuncts, one of which has the* PRED *value 'time'*. The PRED attributes themselves are omitted in the discriminants for brevity.

Table 4: Some f-structure discriminants for example (4)

'spise<[],[]>NULL' OBJ 'time'
'spise<[]>NULL' ADJUNCT > 'time'

The path in an f-structure discriminant is, however, not always from PRED value to PRED value. The word *barn* in example (5) is ambiguous between singular and plural, and the morphology tells us that not by assigning different morphological subtrees but by assigning a single tag *SP* representing both singular and plural. For this single tag, the rules in the grammar assign two different values, as shown in the packed f-structure in figure 6, where parentheses surround the alternate values for the number attribute. Since there are neither lexical nor morphological discriminants in cases like this, we must let f-structure discriminants describe paths from PRED values to atomic values, as shown in table 5.

(5) *Vi* *liker* *barn.*
We like child-SG/PL
"We like child/children."

$$\begin{bmatrix} \text{PRED} & \text{'like<[3:vi], [5:barn]>NULL'} \\ \text{SUBJ} & {}_{3}\begin{bmatrix} \text{PRED} & \text{'vi'} \\ \text{CASE} & \text{nom} \\ \text{PRON-TYPE} & \text{pers} \end{bmatrix} \\ \text{OBJ} & {}_{5}\begin{bmatrix} \text{PRED} & \text{'barn'} \\ \text{NUM} & \begin{pmatrix} a_1 \text{ pl} \\ a_2 \text{ sg} \end{pmatrix} \end{bmatrix} \\ \text{TOPIC} & \begin{bmatrix} 3 \end{bmatrix} \end{bmatrix}$$

Figure 6: Simplified packed f-structure for example (5)

Table 5: F-structure discriminants with atomic values for *barn*

'barn' NUM sg
'barn' NUM pl

Moreover, as mentioned earlier, not all words go through morphological preprocessing. Some words receive multiple features directly through a disjunction in the lexicon. An example is *den*, which can either be a demonstrative meaning “that” or an article meaning “the”. A simplified partial lexical entry for this word is shown in example (6).

(6) den D {(↑SPEC DET DET-TYPE) = demon
| (↑SPEC DET DET-TYPE) = article}

Since both have the category D (determiner) there are no lexical discriminants, and since this word does not go through morphological preprocessing, there are no morphological discriminants either. This ambiguity can therefore only be resolved in the f-structure. Two of the f-structure discriminants for *den* are shown in table 6.

Table 6: F-structure discriminants with atomic values for *den*

‘den’ DET-TYPE demon
‘den’ DET-TYPE article

The previous two cases have shown the necessity of allowing f-structure discriminants based on a minimal path from a PRED value to an atomic value. Since in general we do not know what atomic values will provide the only means of resolving an ambiguity for any grammar and any language, we have to allow every path from a PRED value to an atomic value to be a discriminant candidate. This gives rise to a very large number of discriminants with a high degree of redundancy. Nevertheless, the disadvantage of the large number of discriminants is outweighed by the assurance of having discriminants for all possible distinctions.[3] Furthermore, the number of redundant discriminants quickly diminishes as discriminant choices are made.

3.4 Discriminant Anchors

Each type of discriminant is designed so that it relates linguistic properties to words in the string in order to make it easy to recognize the desired properties. However,

[3] There are marginal cases where two differing c-structures or f-structures will have no discriminants, but these cases are very unlikely to occur with a real grammar. In concrete terms, the two (sub-)c-structures A → B → A → X and A → B → A → B → A → X are different but cannot be distinguished by discriminants; it is easy to see that all other cases are extensions of this example. A simple example of two differing f-structures that cannot be distinguished by (our) discriminants is given by the following pair: $\begin{bmatrix} \text{A} & {}_{1}x \\ \text{B} & \begin{bmatrix}1\end{bmatrix} \end{bmatrix}$ and $\begin{bmatrix} \text{A} & x \\ \text{B} & x \end{bmatrix}$. All other non-cyclic examples have in common with the given minimal one that the tree expansions of both f-structures are identical, that is, the f-structures only differ in whether two attributes share their values or have (distinct) values with identical expansions. The situation with f-structures containing cycles is somewhat more complicated, but comparable.

the same word or substring may occur more than once in the same string. In order to allow the correct identification, and hence, disambiguation, of identical substrings, discriminants are *anchored* to their string positions in terms of character count (which for technical reasons is the least problematic to calculate).

Consider the repeated word *fisker* in example (7). If we did not take string position into account, these two occurrences of the same word form would result in identical discriminants. By anchoring the discriminants in string positions as illustrated in table 7, identical substrings can always be disambiguated correctly. The anchors *10* and *31* refer to the position of the first character in the word *fisker* in its two occurrences.

(7)

De	*store*	*fisker*	*spiser*	*de*
you/they/the/that	big	fish.N/fishing.N/fish.V	eater.N/eat.V	the/that
små	*fisker.*			
small	fish.N/fishing.N/fish.V			

"The big fish eat the small fish."/ "The small fish, the big fish eat."/ "Those big fish eat the small fish."/etc.

Table 7: Anchored morphology discriminants for the word *fisker* in example (7)

10	'fisker': N
10	'fisker': Vfin
31	'fisker': N
31	'fisker': Vfin

In some cases, a single anchor is not sufficient to ensure that discriminants that should be distinct actually are distinct. Consider again example (7), and assume that the noun discriminants have been chosen for both occurrences of *fisker*. Since Norwegian is a V2 language, we are still left with an ambiguity as to which NP is the SUBJ and which is the OBJ. The f-structure discriminants shown in table 8 have two anchors. The first anchor refers to the position of the verb *spiser* which projects the PRED value 'spise<[],[]>NULL'. The second anchor refers to the position of the noun *fisker* which projects the PRED value 'fisk'. Doubly anchored discriminants are those which are paths from PRED value to PRED value.

Table 8: Doubly anchored f-structure discriminants for example (7)

17:10	'spise<[],[]>NULL' SUBJ 'fisk'
17:31	'spise<[],[]>NULL' SUBJ 'fisk'
17:10	'spise<[],[]>NULL' OBJ 'fisk'
17:31	'spise<[],[]>NULL' OBJ 'fisk'

4 Calculation of Discriminants

Discriminants are calculated on the basis of packed c- and f-structures, which are internally represented as directed (not necessarily acyclic) graphs, where each node is labeled with the context (the set of solutions) for which it is valid. This means that the c- and f-structures for a given solution may be recovered by discarding all nodes whose context does not contain that solution. It is, however, crucial to note that neither the discriminants themselves nor the algorithm that computes them depends on the solutions being packed; the algorithm uses packed solutions solely for efficiency reasons and could easily be modified to operate on unpacked c- and f-structures.

In XLE a context is represented as a set of (compatible) choices. The choices corresponding to a packed structure are organized in AND/OR graphs, and each solution corresponds to a maximal selection of compatible choices. A maximal selection can be characterized as a choice of a maximal path in each AND branch of the choice tree. Non-maximal selections correspond to sets of solutions.[4] A typical choice graph looks like figure 7, where $\wedge$ = AND and $\vee$ = OR, and a possible selection corresponding to a single solution is given by (a_2, c_1, e_1, b_2), where c_1 is redundant. The graph encodes 12 solutions.

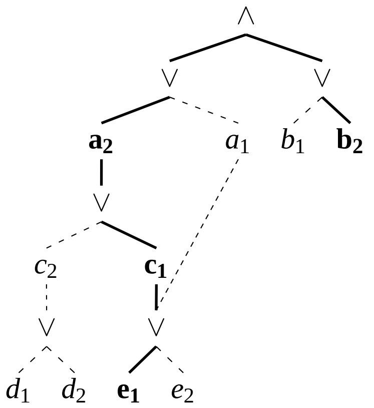


Figure 7: Choice tree with a highlighted path to a single solution

The solutions encoded in a choice graph can easily be enumerated (and thus ordered) using a depth-first multi-traversal of the graph, and a given context can thus be mapped to a bit vector that encodes solutions that are contained in the context by ones and solutions not contained in the context by zeros. For easier processing, all node contexts in the packed structures are converted to solution bit vectors.

As a first step in the calculation of discriminants, the packed graphs are traversed, and all relevant local properties are computed, each of them being associ-

[4] Note that not every solution set can be represented by a selection of compatible choices.

ated with the context in which it is valid. These local properties (together with their contexts) are called *discriminant candidates*.

Since we want to keep apart discriminant candidates with identical local properties that are related to different positions in the source string, we also record the string position from which the local property originates (the anchor of the discriminant). This is straightforward for c-structure discriminants calculated on the basis of c-structure graphs (lexical, morphological and c-structure discriminants) since c-structure lexical nodes are directly associated with string positions. In the case of f-structure discriminants, however, one first has to identify the c-structure node the semantic form of the predicate was projected from. The discriminant anchor is then given by the string position associated with the leftmost lexical node below that node.

If discriminant candidates originating from different parts of the graph have identical patterns (i.e. express the same local properties) and have the same anchors, we combine them into one new discriminant candidate whose context is the union of the original candidates' contexts. Many of the calculated discriminant candidates may be trivial, as their local properties might be valid for all solutions. Removing these trivial discriminant candidates yields the set of proper discriminants.

Let us consider a simple example. In figure 8, we see the packed c-structure and the choice tree for the four-way ambiguous string in example (8).

(8) *Det* *regnet.*
that.D/it.PRON/it.PRONexpl rain/rained/calculated
"That rain." / "It calculated." / "That (one) calculated." / "It rained."

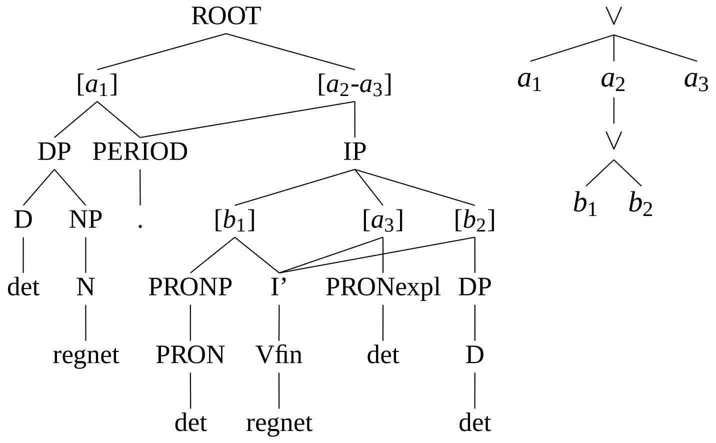


Figure 8: Packed c-structure and choice tree for example (8)

Following the algorithm outlined above, we obtain the c-structure rule and constituent discriminant candidates in table 9. All the rule discriminant candidates are different, thus each of them is a proper discriminant. The associated constituent discriminants, however, are trivial, since the contexts of identical candidates add up to the context containing all solutions. A grouping of the resulting discriminants by identical constituents is presented in table 10. Note that despite the name *rule discriminant*, this kind of discriminant is computed exclusively on the basis of the structures, while access to the grammar rules that assigned those structures is not required.

Table 9: C-structure rule and constituent discriminant candidates for example (8)

anchor	**labeled bracketing**	**substring bracketing**	**context**	**solution vector**
1	ROOT → DP PERIOD	det regnet | .	a_1	1000
1	ROOT → IP PERIOD	det regnet | .	$a_2 - a_3$	0111
1	DP → D NP	det | regnet	a_1	1000
1	IP → PRONP I’	det | regnet	b_1	0100
1	IP → DP I’	det | regnet	b_2	0010
1	IP → PRONexpl I’	det | regnet	a_3	0001

Table 10: Grouping of c-structure discriminants for example (8)

anchor	**discriminant**	**# of solutions**	**solution vector**
1	det regnet | .	(4)	1111
1	ROOT → DP PERIOD	1	1000
1	ROOT → IP PERIOD	3	0111
1	det | regnet	(4)	1111
1	DP → D NP	1	1000
1	IP → PRONP I’	1	0100
1	IP → DP I’	1	0010
1	IP → PRONexpl I’	1	0001

The lexical and morphological discriminants are also computed from the c-structure. In table 11 the lexical discriminant candidates for example (8) are shown. Two of the discriminant candidates (those in boldface) have identical patterns and anchors, so they must be combined to give a proper discriminant.

The computation of morphological discriminants, too, is based on the packed c-structures; this time, however, the sublexical subtrees of the c-structures are considered. Each morphological feature (including the base form) of an analyzed word gives rise to a branch of a sublexical subtree. A candidate for a morphological discriminant is then the concatenation of the base form and all features that can be read off of the sublexical nodes for a given word (or, equivalently, for a given anchor position) and solution.

Table 11: Lexical discriminant candidates for example (8)

anchor	lexical rule	context	solution vector
1	**‘det’ : D**	a_1	1000
1	‘det’ : PRON	b_1	0100
1	**‘det’ : D**	b_2	0010
1	‘det’ : PRONexpl	a_3	0001
5	‘regnet’ : N	a_1	1000
5	‘regnet’ : Vfin	$a_2 - a_3$	0111

The word *fisker* in example (9) is ambiguous between a verb and a noun. Thus, the word *fisker* has two morphological analyses, which surface in the sublexical subtrees in figure 9. We can read off the two discriminant candidates in table 12.

(9) *Jeg fisker.*
I fisherman.N/fish.V
“I am fishing.” / “I, (a) fisherman.”

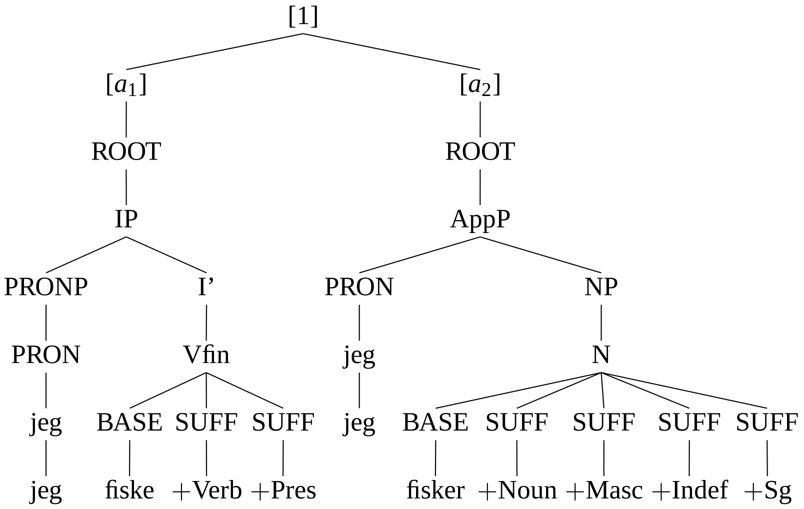


Figure 9: Packed c-structure including sublexical nodes for example (9)

It is important to bear in mind that only those words that are assigned morphological features via XLE’s morphology module will have nontrivial sublexical subtrees and thus potentially give rise to morphological discriminants. Readings of ambiguous words which are directly listed in the LFG lexicon can still be disambiguated using lexical discriminants if their lexical categories are different.

Table 12: Morphological discriminant candidates for example (9)

anchor	**morphology**	**context**	**solution vector**
5	fiske+Verb+Pres	a_1	10
5	fisker+Noun+Masc+Indef+Sg	a_2	01

To exemplify the computation of f-structure discriminants, we consider the sentence in example (4) and its f-structures in figure 5. The relevant parts of the packed f-structure are shown in figure 10. In the packed f-structure, attribute values are annotated with the choices for which they are valid. This sentence is ambiguous, as apparent from choices a_1 and a_2, the ambiguity being manifest solely in the f-structure. An attribute in a packed structure may have more than one possible value, but the choices for those values have to be mutually exclusive, such that only one value or no value remains for each single solution. In such cases, for example the alternative PRED values indexed by a_1 and a_2 in figure 10, the set of values is enclosed in parentheses.

$$\left[\begin{array}{ll}
\textsc{pred} & \left(\begin{array}{l} a_1 \text{ `spise<[1:vi],[2:time]>}\textsc{null}\text{'} \\ a_2 \text{ `spise<[1:vi]>}\textsc{null}\text{'} \end{array}\right) \\
\textsc{subj} & {}_{1}\!\left[\begin{array}{ll} \textsc{pred} & \text{`vi'} \\ \textsc{pron-type} & \text{pers} \end{array}\right] \\
\textsc{obj}\ a_1 & {}_{2}\!\left[\begin{array}{ll} \textsc{pred} & \text{`time'} \\ \textsc{spec} & \left[\textsc{quant}\ \left[\textsc{pred}\ \text{`hver'}\right]\right] \end{array}\right] \\
\textsc{adjunct} & \left\{ a_2\ \left[2\right] \right\} \\
\textsc{topic} & \left[1\right]
\end{array}\right]$$

Figure 10: Simplified partial f-structure for example (4)

Applying the algorithm for f-structure discriminants, we obtain the candidates in table 13, which are all proper discriminants.

5 Display and Use of Discriminants

As mentioned above, a large number of discriminants may be computed for a sentence. This guarantees that there will be enough discriminants for virtually every distinction between structures, so that full disambiguation can always be achieved.

Table 13: F-structure candidates for example (4)

anchor	f-structure path	context	solution vector
0	_TOP ‘spise<[],[]>NULL’	a_1	10
0	_TOP ‘spise<[]>NULL’	a_2	01
4	‘spise<[],[]>NULL’ SUBJ ‘vi’	a_1	10
4	‘spise<[],[]>NULL’ TOPIC ‘vi’	a_1	10
4	‘spise<[],[]>NULL’ OBJ ‘time’	a_1	10
4	‘spise<[]>NULL’ SUBJ ‘vi’	a_2	01
4	‘spise<[]>NULL’ TOPIC ‘vi’	a_2	01
4	‘spise<[]>NULL’ ADJUNCT ‘time’	a_2	01

By considering every node in the c-structure and f-structure and filtering out those that are the same for every analysis, one essentially obtains all discriminants. If, in spite of computing all discriminants, several analyses are left but no discriminants, then, disregarding marginal cases like those discussed in footnote 3, there must be a spurious ambiguity in the grammar and the analyses must be identical.

However, the annotator usually does not need to use all discriminants in the disambiguation process. In fact, in many cases just a few discriminant choices are needed to select the correct analysis amongst many. There is often considerable redundancy, because many discriminants are not independent of others. In order to make the annotator’s choices easier, it is therefore interesting to at least rank and perhaps also filter the discriminants that are presented to the annotator. Annotators will choose those that are the easiest and most useful to them. Our system keeps track of which discriminants are chosen. With this information, the display can be optimized so that, for instance, discriminants which are often chosen can be displayed first, and those that are not needed can be hidden from the display. Much work is still to be done in this area since it must be based on considerable testing in actual practice.

We have developed a toolkit that computes all discriminants and which is a testbed for optimizing their display. XLE-Web is a web-based interface to XLE with packed c- and f-structures and discriminants. The LFG Parsebanker is like XLE-Web, but also stores analyses and discriminant choices, and supports search in the stored analyses. For further details on this work, we refer to earlier publications (Rosén, Meurer, and De Smedt, 2005; Rosén et al., 2005; Rosén, De Smedt, and Meurer, 2006).

We currently display lexical and morphological discriminants first for several reasons. It has been pointed out that lexical ambiguities are often easier to decide on than others (Van der Beek et al., 2002; Oepen et al., 2004). Annotator decisions on lexical ambiguities also tend to be very reliable decisions, since they require little knowledge of the grammar. Decisions on lexical ambiguities are likely to be safer than decisions on syntactic ambiguities because lexical and morphological discriminants contain such a small amount of information. Furthermore, decisions on lexical ambiguities are highly likely to be reapplicable on reparsing with a new version of the grammar, since part of speech changes and changes in morphological analysis will be rare.

With respect to syntactic ambiguities, different branchings are very intuitive (at least for linguists) and require little knowledge of the grammar. In many cases, branchings are quite independent of the grammatical theory used. For these reasons, we present both constituent and rule discriminants to the annotator.

Although not every discriminant is equally easy to decide on, the human disambiguator usually has enough choices of where to begin disambiguation that this does not really matter. Even though discriminant choices can be made independently, the discriminants themselves are not always independent. Choices also normally cause the resolution of other, dependent local ambiguities, making the disambiguation process even more efficient. Furthermore, a discriminant's applicability does not depend on the grammar, but only on the structures, so that discriminants can often be reused in an incremental parsebanking approach.

Discriminants can be exploited in various ways. The first and foremost application is in efficient manual disambigation to supplement the automatic parsing of a corpus, an approach also known as parsebanking. Parsebanking offers quality benefits over the manual construction of a treebank, including the avoidance of formal errors, consistency within the treebank and consistency with a grammar.

Another use of discriminants is in stochastic parse disambiguation. This approach uses properties of c- and f-structures as feature functions to train a stochastic parse ranking model (Riezler et al., 2002). XLE has property templates that can be used for this purpose. We have done experiments using our discriminants instead of the property templates. Preliminary testing of these two approaches has provided results that are better for discriminants than for property templates (Oepen et al., 2007).

6 Conclusion

In creating discriminants for LFG grammars, we have been guided by two important design principles. One principle is that enough discriminants must be computed to distinguish between all analyses. This means that all nodes in both c-structures and f-structures must be examined for possible discriminant candidates. The other main principle is that all distinctions must be represented in such a way that an annotator can easily relate them to words in the string. This ensures that disambiguation can be achieved quickly and efficiently.

Another important consideration has been our objective of making our methodology language and grammar independent. Our independence from particular languages and grammars follows from our approach which only builds on formal properties of representations. It would be possible to extend our design of LFG discriminants to other projections. Although the Norwegian grammar has an MRS projection, and discriminants could be calculated on MRS properties, we have chosen not to do so. Since all LFG grammars have both c-structures and f-structures, complete disambiguation on these levels will be possible for any grammar and language.

One consequence of computing discriminants for all distinctions between representations is the large number of resulting discriminants. Often, however, individual structural differences are not independent of other differences. Rather than trying to eliminate some redundant discriminants by exploiting language specific interdependencies in their computation, we prefer to handle redundancy in their presentation. We have begun work on discriminant presentation in the context of the LFG Parsebanker, but this will be the focus of future research on how annotators use the tool. With the help of the LFG Parsebanker, the discriminants make it feasible to create large parsebanks for languages that have a broad coverage LFG grammar, something that until now has been impossible in practice because of the difficulty of disambiguating.

References

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan*.

Carter, David. 1997. The TreeBanker: A tool for supervised training of parsed corpora. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 598–603, Providence, Rhode Island.

Maxwell, John and Ronald M. Kaplan. 1993. The interface between phrasal and functional constraints. *Computational Linguistics*, 19(4):571–589.

Oepen, Stephan, Dan Flickinger, Kristina Toutanova, and Christopher D. Manning. 2004. LinGO Redwoods, a rich and dynamic treebank for HPSG. *Research on Language & Computation*, 2(4):575–596, December.

Oepen, Stephan, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. 2007. Towards hybrid quality-oriented machine translation. On linguistics and probabilities in MT. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, Studies in Informatics. University of Skövde.

Riezler, Stefan, Tracy Holloway King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*.

Rosén, Victoria, Koenraad De Smedt, Helge Dyvik, and Paul Meurer. 2005. TREPIL: Developing methods and tools for multilevel treebank construction. In Montserrat Civit, Sandra Kübler, and Ma. Antònia Martí, editors, *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 161–172.

Rosén, Victoria, Koenraad De Smedt, and Paul Meurer. 2006. Towards a toolkit linking treebanking to grammar development. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, pages 55–66.

Rosén, Victoria, Paul Meurer, and Koenraad De Smedt. 2005. Constructing a parsed corpus with a large LFG grammar. In *Proceedings of LFG'05*, pages 371–387. CSLI Publications.

Van der Beek, Leonoor, Gosse Bouma, Robert Malouf, and Gertjan Van Noord. 2002. The Alpino dependency treebank. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University.

C-STRUCTURES AND F-STRUCTURES FOR THE BRITISH NATIONAL CORPUS

Joachim Wagner[†], Djamé Seddah[‡], Jennifer Foster[†] and
Josef van Genabith[†]

[†] National Centre for Language Technology (NCLT)
School of Computing
Dublin City University

[‡] Laboratoire Langages, Logiques, Informatique,
Cognition (LaLIC)
Université Paris-Sorbonne

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

We describe how the British National Corpus (BNC), a one hundred million word balanced corpus of British English, was parsed into Lexical Functional Grammar (LFG) c-structures and f-structures, using a treebank-based parsing architecture. The parsing architecture uses a state-of-the-art statistical parser and reranker trained on the Penn Treebank to produce context-free phrase structure trees, and an annotation algorithm to automatically annotate these trees into LFG f-structures. We describe the pre-processing steps which were taken to accommodate the differences between the Penn Treebank and the BNC. Some of the issues encountered in applying the parsing architecture on such a large scale are discussed. The process of annotating a gold standard set of 1,000 parse trees is described. We present evaluation results obtained by evaluating the c-structures produced by the statistical parser against the c-structure gold standard. We also present the results obtained by evaluating the f-structures produced by the annotation algorithm against an automatically constructed f-structure gold standard. The c-structures achieve an f-score of 83.7% and the f-structures an f-score of 91.2%.

1 Introduction

We describe a parsing experiment involving the British National Corpus (BNC) (Burnard, 2000) and a treebank-based parsing architecture for Lexical-Functional Grammar (LFG) (Kaplan and Bresnan, 1982) that reuses a lexicalised, history-based, generative, probabilistic parser (Charniak, 2000), a discriminative reranker (Charniak and Johnson, 2005) and an f-structure annotation algorithm (Cahill et al., 2002, 2004; Burke, 2006) in a pipeline process: the parser and reranker produce c-structures from which f-structures are produced via the annotation algorithm. We show how this technology can be scaled to parse the 100 million word BNC into both c-structure and f-structure representations. We investigate the effect on performance when moving from the domain upon which the LFG parsing resources have been trained (financial newspaper text) to the more varied text of the BNC.

The paper is structured as follows: In Section 2 we describe the LFG parsing architecture. In Section 3 we review related work. Section 4 provides a brief introduction to the BNC. The f-structures constructed in our pipeline processing architecture can only be as good as the c-structures produced by the parsers in the first phase of our parsing architecture. Section 5 presents the c-structure parsing experiments, including BNC preprocessing, parsing issues, gold standard tree construction and evaluation against the gold standard trees. Section 6 presents the f-structure annotation experiments including coverage, the automatic construction of an f-structure gold standard and evaluation against the f-structure gold standard. Section 7 summarises and outlines ongoing and future research.

[†] We are grateful to IRCSET (basic research grant SC/02/298 and postdoctoral fellowship P/04/232) and Science Foundation Ireland (04/IN/I527) for funding this research. We wish to acknowledge the SFI/HEA Irish Centre for High-End Computing for the provision of computational facilities and support. We would like to thank Aoife Cahill and Gzregorz Chrupała for their help.

2 Treebank-Based LFG Parsing

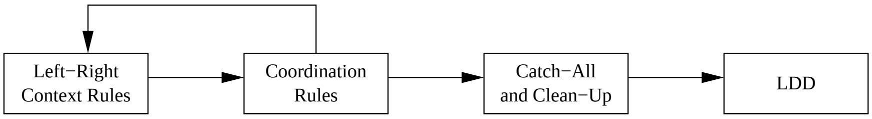


Figure 1: F-Structure Annotation Algorithm Modules

Cahill et al. (2002, 2004) present an LFG f-structure annotation algorithm that annotates Penn-II-style treebank trees (Marcus et al., 1994) with f-structure information (equations), from which a constraint solver can produce an f-structure. The algorithm is modular with four components (Figure 1) taking Penn-II trees as input and automatically adding LFG f-structure equations to each node in the tree. The annotation algorithm is described at length in Burke (2006). Here we will be brief. First, head finding rules are used to head-lexicalise the treebank trees. This partitions the daughters in local subtrees of depth one into a left context, followed by the head, followed by a right context. Left-right annotation matrices state generalisations over nodes occurring in these contexts: e.g. DET nodes in the left context of local subtrees rooted in an NP receive the LFG f-structure annotation $\uparrow$ SPEC:DET = $\downarrow$. Heads are annotated $\uparrow=\downarrow$. The leftmost sister NP to a V head rooted in a VP is annotated $\uparrow$ OBJ = $\downarrow$ etc. In order to keep annotation matrices perspicuous and concise, coordination is treated by a separate component. A Catch-All and Clean-Up component corrects overgeneralisations of the previous modules and provides default annotations for nodes which did not receive an annotation. The LDD component translates traces and coindexation representing long-distance dependencies in Penn-II treebank trees into reentrancies in f-structure. Lexical information is provided automatically in terms of macros associated with Penn-II part-of-speech (POS) tags.

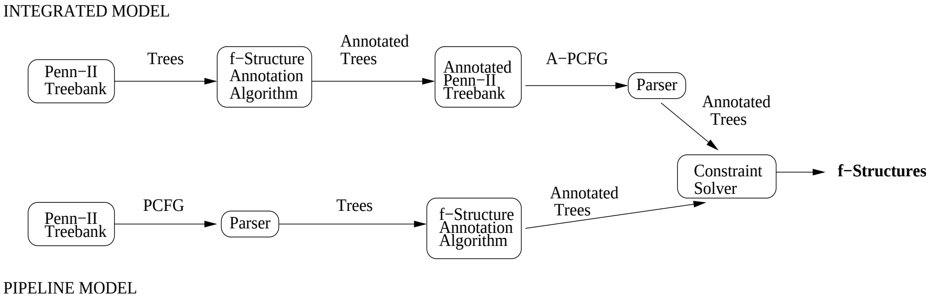


Figure 2: Treebank-based LFG Parsing: Parsing Architectures

Cahill et al. (2002, 2004) provide two parsing architectures exploiting the f-structure annotation algorithm: in the *pipeline* architecture a probabilistic context-free grammar (PCFG) or a lexicalised history-based Markov grammar (Collins, 1999; Charniak, 2000) is extracted from the Penn-II treebank (WSJ sections 02-21), new text is parsed into trees, these trees are passed to the f-structure annotation algorithm and f-structures are produced. In the *integrated* architecture, the Penn-II treebank is first annotated with f-structure equations using the f-structure annotation algorithm and an annotated PCFG (an A-PCFG) is extracted (from WSJ sections 02-21). An A-PCFG carries f-structure equations associated with CFG categories in the rules. The A-PCFG is used to parse new text into trees with f-structure annotations from which f-structures are produced. The two parsing architectures are shown in Figure 2.

Because c-structure parser output in general does not produce traces and coindexation in parse trees to represent long distance dependencies, in both parsing architectures long distance dependencies are resolved at the level of f-structure using finite approximations of LFG functional uncertainty equations and subcategorisation frames automatically learned from the f-structure annotated Penn-II treebank (Cahill et al., 2004; O'Donovan et al., 2004). In this paper, the pipeline parsing architecture is employed with Charniak and Johnson's reranking parser (Charniak and Johnson, 2005) to parse the BNC.

3 Related Work

In the field of information extraction, there have been attempts to obtain predicate-argument structures from the output of statistical parsers. Pasca and Harabagiu (2001) use head finding rules to "read off" binary dependencies from the output of Collins' parser (Collins, 1999) to help analyse questions into corresponding answer types in a high-performance QA system. Surdeanu et al. (2003) present a Propbank (Kingsbury et al., 2002) trained role labeling system to label the output of Collins' parser for use in an information extraction system. These systems do not resolve long distance dependencies, although Surdeanu et al. (2003) integrate an anaphora resolution component into their IE processing pipeline.

The LFG annotation algorithm takes as input a c-structure and produces an f-structure from which dependency relationships between words can be obtained. Thus, it is related to the constituent-to-dependency conversion methods used to prepare English training material for dependency parsing (Yamada and Matsumoto, 2003; Johansson and Nugues, 2007). The dependencies produced by the conversion method of Yamada and Matsumoto (2003), which is based on the head-finding rules of Collins (1999), do not attempt to capture long-distance dependencies. They are handled by the conversion method of Johansson and Nugues (2007).

Deep linguistic grammars have been automatically acquired from treebanks for Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) and Combinatory Categorial Grammar (CCG) (Steedman, 2000). Hockenmaier and Steedman (2002) and Clark and Curran (2004) show how wide-coverage probabilistic CCGs can be extracted from the Penn-II treebank and how these resources can be used in a number of probabilistic parsing models, resolving long-distance dependencies. Their CCG-based parsing system has been used to analyse the one billion word Gigaword corpus (Curran et al., 2007). Miyao and Tsujii (2002, 2004) show how wide-coverage, HPSG resources can be acquired from the Penn-II treebank and how the feature forest model can be used for efficient parsing and parameter estimation. The HPSG resources resolve long distance dependencies. Miyao et al. (2006) show how these resources can be adapted to the medical domain.

To our knowledge, we are the first to parse the BNC in its entirety *and* to evaluate the parsing on a hand-annotated subset. Briscoe and Carroll (2002) report that almost all of the written section of the BNC (90 million words, see Section 4) has been parsed using the RASP parser, which combines a hand-crafted grammar with a probabilistic parse selection model. Baldwin et al. (2004) describe how the Lingo ERG (Copestake and Flickinger, 2000), a hand-crafted HPSG English grammar, was tested by using it to parse BNC sentences. The aim of their work was not to parse the BNC but to test the grammar's coverage using corpus data, and only a small subsection of the written section of the BNC was parsed.

4 British National Corpus

The BNC is a one hundred million word corpus of written and spoken English from a variety of sources. It is a balanced corpus and is designed to be a representative sample of British English from the late twentieth century. Written text comprises 90% of the BNC: 75% of this is non-fiction. The written text is taken from newspapers, published and unpublished letters, school and university essays, academic journals and novels. The spoken component of the BNC consists of transcriptions of spontaneous unscripted dialogue with participants of various ages, regions and social classes, and transcriptions of more formal speech, e.g. business meetings, speeches or radio shows. The BNC is automatically tagged for part-of-speech using the CLAWS4 tagger (Garside et al., 1987), with an estimated tagging accuracy of 97.5%.[1] A two million word subset has been manually tagged using a richer tagset. It is encoded in SGML, with metadata expressed at the document (e.g. document source, genre, id) and sentence (e.g. sentence id) level.

5 C-Structure Parsing

In this section we describe how the BNC was parsed using Charniak and Johnson's two-stagereranking parser (Charniak and Johnson, 2005). The first stage of the parser is a generative, history-based, lexicalised parser (Charniak, 2000) which outputs a list of *n*-best parse trees. The second stage is a discriminative reranker which re-orders the *n*-best list produced by the first stage parser. This parser achieves a

[1] This figure was obtained from `http://www.natcorp.ox.ac.uk/docs/bnc2error.htm`.

labelled Parseval f-score of 91.3% on Section 23 of the Wall Street Journal corpus, and 85.2% on the Brown corpus standard test set (McClosky et al., 2006b).

Section 5.1 details the preprocessing steps which were applied to the BNC sentences to facilitate parsing. Section 5.2 discusses some of the issues involved in actually parsing the BNC sentences, Section 5.3 describes how a gold standard set of 1,000 parse trees was constructed, and, finally, Section 5.4 presents the results of evaluating the parses produced by the parser against the gold standard trees.

5.1 Preprocessing Steps

> "*Cleaning is a low-level, unglamorous task, yet crucial: The better it is done, the better the outcomes. All further layers of linguistic processing depend on the cleaniness of the data.*"
> (Kilgarriff, 2007, p.149)

In our approach we use a Penn-II American English trained parser to parse British English as represented by the BNC. In order to achieve optimal results, we carry out a number of preprocessing steps to adapt BNC coding conventions to what is expected by a Penn-II trained parser. This includes SGML entities, soft hyphens, quotes, currency symbols and spelling differences between American and British English. Adaptations carried out to more closely match the Penn Treebank (PTB) encoding conventions can be expected to improve the parse results because the number of unknown tokens for the parser is reduced. This section describes the preprocessing steps we applied to the BNC in order to obtain cleaner input for the c-structure parsing described in Section 5.2.

5.1.1 Extraction of Sentences

In the original BNC, the start but not the end of each sentence is marked. Usually, the end of a sentence is indicated by the start of the next sentence or the end of the document. Very occasionally (eighteen cases), other tags such as paragraph markers were used to detect the end of a sentence. While processing the BNC SGML files, various tags were exploited to annotate the sentences with additional information, for example whether they belong to headers, list items, spoken utterances, poems, etc. A tag that needs special attention is the `<gap>` tag. It marks omissions due to anonymisation and replaces various material including formulae and figures. To facilitate parsing, we automatically re-inserted text for gaps according to Table 1. The gap substitutions are recorded and are recoverable. In total, 51,827 gap substitutions were performed in 38,452 sentences (0.617 %).

5.1.2 UTF-8 Encoding of SGML Entities

The BNC uses a large number of SGML entities to represent special characters, symbols, fractions, etc. A mapping to UTF-8 was manually created based on the description in the file `bncents.dtd` included in the BNC distribution and Unicode

Gap Description	Substitution String
last or full name	Jon1234es
list of names	Jon1234es , Smi1234th and Mur1234phy
date	29/12/1970
list of dates	29/12/1970 , 30/12/1970 and 31/12/1970
list of countries	Germ1234any , Ire1234land and Spa1234in
address	11234 Sun1234set Avenue
name and address	Mur1234phy , 11234 Sun1234set Avenue
telephone number	0123/4561234
number	1231234
formula	1231234

Table 1: Gap substitutions: 1234 is replaced by a random number drawn from an exponential distribution.

character code charts[2] and other web resources. UTF-8 is not used in the PTB which only contains ASCII characters. Nevertheless, the conversion is useful as it also affects ASCII characters that are represented by an SGML entity in the BNC, for example, the dollar sign. For other characters, UTF-8 serves more as an intermediate format that allows us to keep as much information as possible and at the same time to visualise the intended symbols in normal text editors. 1,255,316 (20.156 %) BNC sentences contain UTF-8 characters. Quote and currency conversions (see below) reduce this number to 45,828 sentences (0.736%).

5.1.3 Disambiguation of Soft Hyphens

Inspection of the frequency tables of special characters revealed that soft hyphens occur in the BNC. They are supposed to mark hyphens inserted by pagination processes at the end of a line. Often, they are also used to mark possible hyphenation points.[3] As the PTB does not contain soft hyphens at all, we decided to replace them with the following simple strategy. We create three candidate substitutions (deletion, space, normal hyphen) and vote based on the frequency of the respective tokens and bigrams in the BNC. Manual evaluation of this strategy on 100 randomly extracted instances showed 6 clear errors and 12 unclear cases. Only 4,190 BNC sentences (0.067%) contain soft hyphens.

5.1.4 Normalisation of Quotes

The PTB uses and the PTB-trained parsers expect sequences of two single left or right quotes to represent left and right quotes. In most cases, distinct quotes in the

[2] `http://www.unicode.org/charts/` accessed during 2005 and 2006

[3] The correct usage is controversial – compare for instance the Wikipedia article on hyphens and the detailed discussion on the web-page `http://www.cs.tut.fi/~jkorpela/shy.html`

BNC can be easily converted to PTB-style. However, some sections of the BNC use neutral quotes. Very rarely, single quotes are used as well. In order to achieve optimal results, a more elaborate conversion is necessary. We disambiguate neutral quotes by replacing them with alternating left and right quotes. Existing unambiguous quotes are respected, so that a neutral quote after a left quote becomes a right quote. Single quotes are not changed as there would be a high risk of accidently damaging apostrophes. The total number of BNC sentences containing ambiguous neutral double quotes is 68,020 (1.092%).

5.1.5 Currency and Other Normalisations

The PTB uses individual tokens for currency and number, for example `US$` 2,000, while the BNC amalgamates them into a single token. Furthermore, the pound sign is the dominant currency symbol in the BNC while the PTB does not provide (much) training data for it.[4] Therefore, we map pound, yen and euro symbols to the dollar sign and, in a second step, insert a token boundary after each dollar sign to separate a possibly attached amount. The currency symbols are restored after parsing. A total of 69,459 BNC sentences (1.115%) contain currency symbols.

Additionally, dashes are replaced by PTB-style sequences of minus signs. Horizontal ellipsis is replaced by three full stops. Many fractions are represented by single entities in the BNC, and consequently mapped to single characters in Unicode (if possible), e.g. frac23 and U+2154 for two-thirds. The common fractions 1/4, 1/2, and 3/4 are re-written with normal numbers and a forward slash. Prime and double prime are encoded as single and double (neutral) quotes. The multiplication sign is replaced by ‘x’. The bullet and micro signs that are quite frequent in the BNC are not replaced because we could not find suitable examples in the PTB.

5.1.6 Translation to American English

The varcon package (`http://wordlist.sf.net`) was used to translate the BNC to American English. According to the source code and vocabulary file, the varcon translation process is only a matter of different spelling and words substitutions. Word order and tokenisation are not changed. If required, the original British English tokens can be written into the leaf nodes of parse trees. The varcon tool has been modified to not change “For” to “Four” because this substitution would also apply to the preposition “for” at the start of a sentence or in headings. The total number of BNC sentences that are changed by varcon is 333,745 (5.359%).

[4] Recently, we found out that the pound sign is represented by the # sign in the PTB, see `http://www.ldc.upenn.edu/Catalog/docs/treebank2/cl93.html`. Still, a substitution with the dollar sign can be justified by the larger amount of instances that provide more reliable statistics for the parser.

5.2 Parsing the BNC sentences

5.2.1 N-Best Parsing and Reranking

To carry out the BNC c-structure parsing, we used Charniak and Johnson's reranking parser (Charniak and Johnson, 2005) taken off the shelf from the June 2006 package.[5] We implemented a wrapper that supervises the parsing process because the earlier August 2005 version we used during development sometimes aborts parsing, leaving some sentences unparsed, or fails to terminate. The order of sentences is randomised to spread the effect of bugs evenly over the corpus and to make the duration of each task-farming package more similar (see Section 5.2.2)

The parsing and reranking phases were carried out separately. Charniak's parser allows for an ID string to be present in the `<s>` tag, but this string cannot contain any whitespace. Therefore, the SGML attribute-value list expressing the annotation was encoded with underscores instead of spaces and restored after parsing. The first-stage parser was instructed to respect the input tokenization and to output the 50-best parse trees. Of the 6,228,111 BNC input sentences, 6,218,384 (99.844%) were parsed successfully. After parsing, we applied the reranker to the output of the first stage. The reranker succeeded in reranking all sentences that had been parsed by the first-stage parser.

5.2.2 Time and Space Requirements

Parsing the preprocessed BNC is not a trivial task as it would require several months of computation time on a single PC and substantial disk space for the results. We decided to avail ourselves of a high-end computing facility to carry out this task. On 31 CPUs (AMD Opteron 250, 2.4 GHz single core), parsing the 6,228,111 sentences took 79.5 hours walltime, i.e. roughly 2,500 CPU hours or 1.425 seconds per sentence. x Table 2 shows the parsing time for 50-best parsing (and also for 2-best parsing that we considered initially), this time on a 2.8 GHz Pentium 4 and with an older version of the parser. There is a huge variance in parsing time and (unsurprisingly) sentence length is an important factor. Because of this, the observed parsing speed may not translate well to other corpora. The re-ranking process is faster than the parsing process – sentences were re-ranked at the rate of 0.15 seconds per sentence.

The space required for the output of parsing tends to get big compared to the raw sentences, especially for n-best parsing. However, the output can be compressed with very high compression ratios as the set of categories is small and n-best parses are similar to each other. We measured a compression ratio of 27.5 for GZip and 38.0 for BZip2 on the 8,000 sentences used in Table 2 for time measurements. The actual size of the 50-best parses including our SGML markup is 3.4 GB after compression with BZip2.

[5] reranking-parserJune06.tar with SHA1 832e63ce87196d5d0e54b6414020d1c786217936 downloaded from `ftp://ftp.cs.brown.edu/pub/nlparser/`

Length	$n = 2$	$n = 50$	**Increase** 2 → 50
00–04	0.407	0.414	1.72%
05–09	0.687	0.696	1.31%
10–14	1.153	1.169	1.39%
15–19	1.914	1.946	1.67%
20–24	2.559	2.577	0.70%
25–29	3.594	3.630	1.00%
30–34	4.683	4.664	-0.41%
35–39	6.116	6.139	0.38%

Table 2: Parsing time for n-best parsing in seconds per sentence, measured with 1,000 random BNC sentences in each length range

5.3 Constructing BNC Gold Standard C-Structures

A gold standard set of 1,000 BNC sentences was constructed by one annotator who manually corrected the output of the first stage parser of Charniak and Johnson's reranking parser. The sentences included in the gold standard were chosen at random from the BNC, subject to the condition that they contain a token which occurs as a verb in the BNC but not in the training sections of the WSJ section of the PTB. A decision was made to select sentences for the gold standard set which differ from the sentences in the WSJ training sections, and one way of finding different sentences is to focus on verbs which are not attested in the WSJ Sections 2-21. The gold standard sentences serve a dual purpose: as a set of test sentences (and it is in this role that they are being used in the research described here) and as a potential set of training sentences (for future research). Because they contain verbs which do not occur in the parser's training set, they are likely to represent a hard test for WSJ-trained parsers.

The following steps were carried out to obtain the gold standard sentences:

1. Using the BNC tag set, a list of all verbs occurring in the BNC was generated. A frequency count was associated with each verb.
2. All verbs in the BNC verb list were converted to their root form, duplicates were merged and frequency counts adjusted.[6]
3. The BNC verb root forms were converted to American English using the varcon tool (see Section 5.1.6).
4. Using the PTB tag set, a list of all verbs occurring in Sections 2-21 of the WSJ corpus was generated.
5. All verbs in the WSJ verb list were converted to their root form, and duplicates merged.

[6] Lemmatisation was carried out using the Xerox XLE xfst tool (Maxwell and Kaplan, 1996).

	LP	**LR**	**F-Score**	**%100 Match**
All Sentences	83.8	83.7	83.7	25.2
Less than 41 words	86.4	86.2	86.3	30.3

Table 3: BNC C-Structure Evaluation: Labelled Parseval

6. A list of all verb root forms in the BNC verb root form list which are not in the WSJ verb root form list was compiled (BNC-WSJ). This resulted in 25,874 root forms. Of these, 537 occurred more than one hundred times within the BNC, and 14,787 occurred only once.[7]

7. 1,000 forms were selected from the BNC-WSJ verb root form list, respecting the frequency of the verb forms so that root forms which occur frequently within the BNC were favoured over root forms occurring only once.

8. For each of the 1,000 verb roots, a sentence containing a conjugated form of this token was randomly selected from the BNC.

The PTB bracketing guidelines (Bies et al., 1995) and the PTB itself were used as references by the BNC annotator. Functional tags and traces were not annotated. The annotator noticed that the PTB parse trees sometimes violate the PTB annotator guidelines, and in these cases, the annotator chose the analysis set out in the guidelines. An example is the noun phrase *almost certain death* which occurred in a sentence in the BNC gold standard. According to Bies et al. (1995, p.179), this should be analysed as *(NP (ADJP almost certain) death)*, but a search for the word *almost* in the PTB yielded a similar example (in WSJ Section 9, 0946.prd) *almost unimaginable speed*, which was parsed as *(NP almost unimaginable speed)*. The BNC annotator analysed the phrase *almost certain death* as *(NP (ADJP almost certain) death)*, according to the guidelines. If a structure was encountered which was not mentioned in the PTB bracketing guidelines and no example of which could be found in the PTB, the annotator decided how it should be analysed and documented this decision. An example is the phrase *day in day out* which was analysed as a flat adverbial phrase. It took approximately sixty hours to construct the gold standard.

5.4 C-Structure Evaluation

Table 3 shows the results of evaluating the parses produced by Charniak and Johnson's parser against the gold standard parses described in Section 5.3 using the Parseval labelled precison/recall measures (Black et al., 1991). The results were calculated using the *evalb* software and the *new.prm* parameter file, which are distributed with the parser. The precision figure represents the number of correct constituents divided by the total number of constituents produced by the parser.

[7] The most frequently occurring verb lemma in the BNC which does not appear (as a verb) in WSJ2-21 is *mutter*, which occurs 1,871 times.

	LP	LR	F-Score	%100 Match
All Sentences	85.5	85.4	85.4	27.9
Less than 41 words	88.2	88.0	88.1	33.5

Table 4: BNC C-Structure Evaluation: Unlabelled Parseval

Constituent Type	Precision	Recall	F-Score
NP	86.8	88.4	87.6
VP	81.6	81.8	81.7
S	80.0	81.8	80.9
PP	80.2	82.1	81.1
SBAR	75.8	77.6	76.7
ADVP	80.3	77.4	78.8
ADJP	67.2	69.5	68.3
WHNP	91.9	96.8	94.3
PRT	61.4	84.3	71.1
WHADVP	97.3	95.5	96.4

Table 5: BNC C-Structure Evaluation: 10 Most Frequent Constituents

The recall figure represents the number of correct constituents divided by the total number of constituents in the gold standard set. The f-score is the harmonic mean of precision and recall. A constituent in a test parse tree is considered to be correct if it spans the same sequence of words and has the same label as a constituent in the corresponding gold standard parse tree. Unlabelled precision/recall figures are shown in Table 4: these figures reflect a more relaxed notion of correctness, whereby a constituent in a test parse is considered correct if a constituent spanning the same sequence of words in the corresponding gold tree can be found, regardless of whether the labels on both constituents match. Tables 3 and 4 both also contain the percentage of sentences which achieve an f-score of 100%.

The results in Tables 3 and 4 show that Charniak and Johnson's parser performs quite well on BNC data with a labelled f-score of 83.7%, considering that it achieves an f-score of 91.3% on Section 23 of the Wall Street Journal corpus and 85.2% on the Brown Corpus (McClosky et al., 2006a). This is quite encouraging because the BNC sentences represent a different domain to the Wall Street Journal and the sentences in the gold standard contain verbs which do not occur as verbs in the parser's training data. The quality of the c-structure trees means that the f-structures which are generated from them are more likely to be reliable: previous research (Judge et al., 2006) has shown that, given good CFG trees, the f-structure annotation algorithm can produce good f-structures. Table 5 shows the labelled precision and recall figures for the ten most frequent constituents in the BNC test set (in descending order of their frequency). Among the frequently occurring constituents, ADJP, SBAR, ADVP and PRT are the categories with the most room for improvement for Charniak and Johnson's parser. The parser performs well on NP and WH constituents.

6 F-Structure Annotation

In Section 6.1 we describe how the LFG Annotation Algorithm (see Section 2) was applied to the BNC c-structures produced by Charniak and Johnson's reranking parser (see Section 5) to yield BNC f-structures. In Section 6.2 we describe how the BNC f-structures were evaluated, we present evaluation results and, in an error analysis, discuss some low-scoring examples.

6.1 Annotation Process

The same high-end computing facility which was used to perform the first stage of the c-structure parsing (see Section 5.2.2) was employed to apply the annotation algorithm to the c-structures. The output of the reranking parser is a list of 50 parse trees, and the highest ranked of these is passed as input to the annotation algorithm. The annotation is faster than the c-structure parsing, with an annotation rate of approximately 16 sentences per second. The annotation algorithm fails for just one sentence:

> *And , and , you know , they 've got paid youth officer 's working in Harlow , now they are , there are , they 're over they 're over stretched it 's true and , but we , I mean what were doing here is actually supplementing there service and were not meeting all , we would n't of erm meeting all the demands , but the important thing I think is that were continuing to erm , you know , were trying to do something about it , and one of the things that were trying to do as officer 's in the Local Government Unit is work with Leisure Services and get them to put more resources into doing things for young people .* (BNC D95.477)

6.2 F-Structure Evaluation

6.2.1 F-Structure Evaluation Procedure

In order to evaluate the f-structures generated by the annotation algorithm, it is necessary to have a set of reference or gold standard f-structures. Unfortunately, due to time constraints, we do not have a hand-corrected set of gold standard f-structures. Instead we followed the established procedure (Hockenmaier and Steedman, 2002; Miyao and Tsujii, 2004) of automatically constructing a reference set of f-structures by applying the annotation algorithm to the manually corrected gold standard c-structures (see Section 5.3). We then evaluate the f-structures produced by applying the annotation algorithm to the c-structure parser output against

Attribute	Precision	Recall	F-Score	WSJ
OVERALL	91.1	91.4	**91.2**	94.3
PRED-ONLY	86.5	86.1	**86.3**	91.1

Table 6: BNC F-Structure Evaluation

the reference f-structures by computing precision and recall on the f-structures as sets of term descriptions (following Crouch et al. (2002)).

6.2.2 F-Structure Evaluation Results

Attribute	Precision	Recall	F-Score	WSJ
adjunct	83.3	83.6	**83.4**	89
num	96.6	97.3	**96.9**	97
pers	97.2	97.9	**97.5**	98
obj	90.1	90.4	**90.2**	94
subj	89.6	87.4	**88.5**	93
tense	97.4	96.3	**96.8**	95
det	96.5	96.4	**96.4**	98
pron_form	98.5	99.0	**98.7**	99
coord	84.0	82.1	**83.0**	89
xcomp	87.7	85.6	**86.6**	91

Table 7: BNC F-Structure Evaluation: Ten Most Frequent Attributes

The f-structure evaluation results for the 1,000 BNC test sentences are shown in Table 6. Evaluated in this way, the BNC sentences receive an f-score of 91.2%. When attributes with atomic values (e.g. `num`, `tense` and `pers`) are omitted, the f-score goes down to 86.3%. The fourth column in Table 6 shows the results of performing the same evaluation on Section 23 of the WSJ. The annotation algorithm is applied to Section 23 gold standard parse trees (stripped of functional tags and traces) and the resulting f-structures are compared to those produced by applying the annotation algorithm to Charniak and Johnson parser output.

Table 7 shows the individual evaluation results for the ten most frequently occurring attributes, and Table 8 for the remaining less frequent pred-only attributes. Again, the WSJ Section 23 results are shown for comparison. It is clear from Table 7 that atomic-valued attributes such as `tense`, `num` and `pers` attributes suffer little when moving from the WSJ to the BNC domain, unlike the arguably more important attributes of `subj`, `adjunct`, `obj` and `xcomp`. Table 8 shows that there is significant room for improvement for the attributes `relmod`, `topicrel`, `comp`, `quant`, `app`, `obj2` and `topic`.

Attribute	Precision	Recall	F-Score	WSJ
poss	95.7	94.7	**95.2**	96
comp	72.6	73.7	**73.1**	87
topicrel	77.3	79.8	**78.5**	88
relmod	64.5	69.6	**67.0**	80
quant	87.6	82.7	**85.1**	95
obl	69.0	61.9	**65.3**	71
obl_ag	91.5	91.5	**91.5**	96
app	42.9	43.8	**43.3**	86
obj2	47.1	55.8	**51.1**	71
topic	50.0	75.0	**60.0**	87
focus	100.0	75.0	**85.7**	59
obl2	50.0	33.3	**40.0**	22

Table 8: BNC F-Structure Evaluation: Other Pred-Only Attributes

6.2.3 Two Low-Scoring Examples

Consider the BNC examples (1) and (2):

(1) *They've been digging coal in small private mines in the area for centuries* (BNC K1J.1996)

(2) *Grey-haired, stooping, shabbily dressed* (BNC GVT.0268)

The top part of Figure 3 shows the gold standard parse tree and Charniak and Johnson parser output tree for example (1). Notice that the parser has incorrectly tagged *digging* as a noun and has incorrectly adjoined the prepositional phrase *for centuries* to the noun *area*. The dependency triples in Figure 3 show how mistakes in c-structure are propagated to f-structure. The mistagging of *digging* leads to the misanalysis of *coal* as a verb. The identification of *for* as an adjunct of *area* is a straightforward consequence of the c-structure mistake. Notice also that the "gold" f-structure in Figure 3 is not completely accurate: it incorrectly analyses the sentence as being passive. This is a consequence of the *unsupervised* creation of the reference f-structure set. Figure 4 shows the reference and test c-structures and f-structures for example (2). This example shows again that mistakes at the c-structure level are propagated to the f-structure level, and that mistakes can be introduced by the annotation algorithm, resulting in less than perfect reference f-structures.

7 Conclusions and Future Work

In this paper we have described in some detail how Lexical-Functional Grammar c-structures and f-structures were extracted from the one-hundred million word

Gold
(S (NP They) (VP 've (VP been (VP (VBG digging) (NP coal) (PP in (NP (NP small private mines) (PP in (NP the area)))) (PP for (NP centuries))))))
Test
(S (NP They) (VP 've (VP been (VP ***(NN digging)*** *(NP coal) (PP in (NP (NP small private mines) (PP in* ***(NP (NP the area) (PP for (NP centuries)))****)))))))*

Gold	Test
adjunct(dig,in)	**adjunct(coal,in)**
subj(dig,pro)	**subj(coal,pro)**
subj(be,pro)	subj(be,pro)
adjunct(mine,in)	adjunct(mine,in)
passive(be,+)	passive(be,+)
obj(in,area)	**adjunct(area,for)**
tense(be,past)	tense(be,past)
adjunct(dig,for)	**num(digging,sg)**
obj(dig,coal)	**pers(digging,3)**
xcomp(be,dig)	**xcomp(be,coal)**
participle(dig,pres)	**obj(coal,digging)**

Figure 3: Gold and Test C-Structures and F-Structures for Example (1)

British National Corpus using a treebank-based parsing architecture. Two steps were involved in this process: the first step was the parsing of the sentences into context-free trees or LFG c-structure; the second step was the annotation of these trees to produce LFG f-structures, from which semantic dependencies can be extracted. A thorough description of the preprocessing required to parse a corpus as large as the BNC was provided. This paper also described a c-structure gold standard for the BNC, presented Parseval results for Charniak and Johnson's reranking parser on this gold standard, and provided an evaluation of the f-structures against an automatically generated set of gold standard f-structures. The c-structures achieve an f-score of 83.7% and the f-structures an f-score of 91.2%. Our research demonstrates that it is feasible to provide a reasonably accurate LFG analysis of a very large body of sentences in a robust, non-labour-intensive way.

Our next goal is to investigate to what extent parsing accuracy can be improved by performing self-training experiments. We have already begun work on this: following McClosky et al. (2006a), we have re-trained Charniak and Johnson's first-stage parser on BNC parse trees produced by the two-stage reranking parser, and have obtained a statistically significant f-score increase of 1.7% (Foster et al., 2007). When the annotation algorithm is applied to the output of this self-trained parser, accuracy goes up from 91.2% to 91.7%. We aim to build on these small improvements by applying more sophisticated domain adaptation methods.

The f-structure annotation algorithm is sensitive to the presence of Penn-II

Gold

(FRAG (ADJP (ADJP Grey-haired), (ADJP stooping), (ADJP (ADVP shabbily) dressed)))

Test

*(FRAG **(ADJP Grey-haired , stooping)** , (VP (ADVP shabbily) dressed))*

Gold	Test

`tense(dress,past)`
`adjunct(dress,shabbily)`
`adjunct(dress,grey-haired)`
`adjunct(dress,stooping)`
`participle(stooping,pres)`
	`tense(dress,past)`
	`adjunct(dress,shabbily)`
	`adjunct(stooping,grey-haired)`

Figure 4: Gold and Test C-Structures and F-Structures for Example (2)

functional tags and will use such information to assign f-structure functional equations, backing off to simpler categorical and configurational information if functional tags are not present. Here we use the annotation algorithm on raw parser output trees with no functional tags. In the future we aim to apply a post-parsing Penn-II functional tag labeller, e.g. Chrupała et al. (2007), to raw parser output prior to the application of the annotation algorithm. Other aims include the manual correction of our reference f-structures so that more confidence can be placed in the f-structure evaluation results and the application of our pipeline parsing architecture to the BNC using other c-structure parsers, e.g. Bikel (2004).

References

Baldwin, Timothy, Bender, Emily M., Flickinger, Dan, Kim, Ara and Oepen, Stephan. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-04)*, volume Six, pages 2047–2050, Lisbon, Portugal.

Bies, Ann, Ferguson, Mark, Katz, Karen and MacIntyre, Robert. 1995. Bracketing Guidelines for Treebank II Style, Penn Treebank Project. Technical Report Tech Report MS-CIS-95-06, University of Pennsylvania, Philadelphia, PA.

Bikel, Daniel. 2004. Intricacies of Collins Parsing Model. *Computational Linguistics* 30(4), 479–511.

Black, Ezra, Abney, Steve, Flickinger, Dan, Gdaniec, Claudia, Grishman, Robert, Harrison, Philip, Hindle, Donald, Ingria, Robert, Jelinek, Fred, Klavans, Judith, Liberman, Mark, Marcus, Mitchell, Roukos, Salim, Santorini, Beatrice and Strzalkowski, Tomek. 1991. A Procedure for Quantitatively Comparing the

Syntactic Coverage of English Grammars. In *Proceedings of the 1991 DARPA Speech and Natural Language Workshop*, pages 306–311.

Briscoe, Ted and Carroll, John. 2002. Robust Accurate Statistical Annotation of General Text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-02)*, pages 1499–1504, Las Palmas, Gran Canaria.

Burke, Michael. 2006. *Automatic Treebank Annotation for the Acquisition of LFG Resources*. Ph. D.thesis, School of Computing, Dublin City University, Ireland.

Burnard, Lou. 2000. User Reference Guide for the British National Corpus. Technical Report, Oxford University Computing Services.

Cahill, Aoife, Burke, Michael, O'Donovan, Ruth, van Genabith, Josef and Way, Andy. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 320–327, Barcelona, Spain.

Cahill, Aoife, McCarthy, Mairéad, van Genabith, Josef and Way, Andy. 2002. Parsing with PCFGs and Automatic F-Structure Annotation. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the Seventh International Conference on LFG*, pages 76–95, Stanford, CA: CSLI Publications.

Charniak, Eugene. 2000. A Maximum Entropy Inspired Parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*, pages 132–139, Seattle, WA.

Charniak, Eugene and Johnson, Mark. 2005. Course-to-fine n-best-parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the ACL (ACL-05)*, pages 173–180, Ann Arbor, Michigan.

Chrupała, Grzegorz, Stroppa, Nicolas, van Genabith, Josef and Dinu, Georgiana. 2007. Better Training for Function Labeling. In *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-07)*, pages 133–138, Borovets, Bulgaria.

Clark, Stephen and Curran, James. 2004. Parsing the WSJ using CCG and Log-Linear Models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 104–111, Barcelona, Spain.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph. D. thesis, University of Pennsylvania.

Copestake, Ann and Flickinger, Dan. 2000. An Open-source Grammar Development Environment and Broad-coverage English Grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-00)*, Athens, Greece.

Crouch, Richard, Kaplan, Ron, King, Tracy Holloway and Riezler, Stefan. 2002. A comparison of evaluation metrics for a broad coverage parser. In *Proceedings of the LREC Workshop: Beyond PARSEVAL – Towards Improved Evaluation Measures for Parsing Systems*, pages 67–74, Las Palmas, Gran Canaria.

Curran, James R., Clark, Stephen and Bos, Johan. 2007. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the Demonstrations Session of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 29–32, Prague, Czech Republic.

Foster, Jennifer, Wagner, Joachim, Seddah, Djamé and van Genabith, Josef. 2007. Adapting WSJ-Trained Parsers to the British National Corpus using In-domain Self-training. In *Proceedings of the Tenth International Workshop on Parsing Technologies (IWPT-07)*, pages 33–35, Prague, Czech Republic.

Garside, Roger, Leech, Geoffrey and Sampson, Geoffrey (eds.). 1987. *The Computational Analysis of English: a Corpus-Based Approach*. Longman, London.

Hockenmaier, Julia and Steedman, Mark. 2002. Generative Models for Statistical Parsing with Combinatory Categorial Grammar. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 335–342, Philadelphia.

Johansson, Richard and Nugues, Pierre. 2007. Extended Constituent-to-Dependency Conversion for English. In Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (eds.), *Proceedings of NODALIDA 2007*, pages 105–112, Tartu, Estonia.

Judge, John, Cahill, Aoife and van Genabith, Josef. 2006. QuestionBank: Creating a Corpus of Parse-Annotated Questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING-ACL-06)*, pages 497–504, Sydney, Australia.

Kaplan, Ron and Bresnan, Joan. 1982. Lexical Functional Grammar, a Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 173–281, Cambridge, MA: MIT Press.

Kilgarriff, Adam. 2007. Googleology is Bad Science. *Computational Linguistics* 33(1), 147–151.

Kingsbury, Paul, Palmer, Martha and Marcus, Mitch. 2002. Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference (HLT-02)*, San Diego, CA.

Marcus, Mitchell, Kim, Grace, Marcinkiewicz, Mary Ann, MacIntyre, Robert, Bies, Ann, Ferguson, Mark, Katz, Karen and Schasberger, Britta. 1994. The

Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 110–115, Princeton, NJ.

Maxwell, John and Kaplan, Ron. 1996. An Efficient Parser for LFG. In *Proceedings of the First LFG Conference*, Grenoble, France.

McClosky, David, Charniak, Eugene and Johnson, Mark. 2006a. Effective Self-Training for Parsing. In *Proceedings of the Human Language Technology Conference and North American chapter of the ACL annual meeting (HLT-NAACL-06)*, pages 152–159, New York.

McClosky, David, Charniak, Eugene and Johnson, Mark. 2006b. Reranking and Self-Training for Parser Adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING-ACL-06)*, pages 337–344, Sydney, Australia.

Miyao, Yusuke, Ohta, Tomoko, Masuda, Katsuya, Tsuruoka, Yoshimasa, Yoshida, Kazuhiro, Ninomiya, Takashi and Tsujii, Jun'ichi. 2006. Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING-ACL-06)*, pages 1017–1024, Sydney, Australia.

Miyao, Yusuke and Tsujii, Jun'ichi. 2002. Maximum Entropy Estimation for Feature Forests. In *Proceedings of the Human Language Technology Conference (HLT-02)*, San Diego, CA.

Miyao, Yusuke and Tsujii, Jun'ichi. 2004. Deep Linguistic Analysis for the Accurate Identification of Predicate-Argument Relations. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-04)*, pages 1392–1397, Geneva, Switzerland.

O'Donovan, Ruth, Burke, Michael, Cahill, Aoife, van Genabith, Josef and Way, Andy. 2004. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II Treebank. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 368–375, Barcelona, Spain.

Pasca, Marius A. and Harabagiu, Sandra M. 2001. High Performance Question/Answering. In *The 25th Annual International ACM SIGIR Conference (SIGIR-01)*, pages 366–374, New Orleans, Louisiana.

Pollard, Carl and Sag, Ivan A. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications.

Steedman, Mark. 2000. *The Syntactic Process*. MIT Press.

Surdeanu, Mihai, Harabagiu, Sandra, Williams, John and Aarseth, Paul. 2003. Using Predicate-Argument Structure for Information Extraction. In *Proceedings of the 41st Annual Meeting of the ACL (ACL-03)*, pages 8–15, Sapporo, Japan.

Yamada, Hiroyasu and Matsumoto, Yuji. 2003. Statistical Dependency Analysis with Support Vector Machines. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT-03)*, pages 195–206, Nancy, France.

PREPOSITION-DETERMINER CONTRACTIONS: AN ANALYSIS IN OPTIMALITY-THEORETIC LEXICAL-FUNCTIONAL GRAMMAR WITH LEXICAL SHARING

Michael T. Wescoat

University of California, Davis

Proceedings of the LFG07 Conference

Miriam Butt and Tracy Holloway King (Editors)

2007

CSLI Publications

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

Abstract

I consider preposition-determiner contractions in various European languages and offer a uniform approach using optimality-theoretic Lexical-Functional Grammar with lexical sharing. Lexical sharing allows a preposition and a determiner to share a contraction, which is a single word; this predicts that contractions are disallowed if anything intervenes between the preposition and determiner. Constraints on the syntactic relationship between the preposition and determiner are handled with the mechanisms of Lexical-Functional Grammar. Optimality-theoretic constraints predict when contractions are required in place of independent prepositions and determiners and when they are not. I suggest that this combination of mechanisms may be useful in tackling phenomena beyond preposition-determiner contractions.

1 The problem

Various European languages exhibit *preposition-determiner* (P-D) *contractions*, in which a P and a D appear to coalesce, with idiosyncratic changes in shape. For instance, in the Italian paradigm in (1), mutations of P include *con* → *co*, *di* → *de*, and *in* → *ne*; changes to D include *il* → *l* and gemination of initial *l*. The German forms in (2) show elision of *d*V from D, feeding the loss of P-final *n* before *m*.

(1)

'the'	il	lo	l'	i	gli	la	le
a 'to'	al	allo	all'	ai	agli	alla	alle
con 'with'	col	collo	coll'	coi	cogli	colla	colle
da 'from'	dal	dallo	dall'	dai	dagli	dalla	dalle
di 'of'	del	dello	dell'	dei	degli	della	delle
in 'in'	nel	nello	nell'	nei	negli	nella	nelle
su 'on'	sul	sullo	sull'	sui	sugli	sulla	sulle

(2)

'the'	das	dem	der
an 'at'	ans	am	
auf 'on, onto'	aufs		
bei 'at, with'		beim	
für 'for'	fürs		
in 'in, into'	ins	im	
von 'from, of'		vom	
zu 'to'		zum	zur

Beyond the obvious morphological issues, P-D contractions pose significant challenges for syntax. I take the position that there is enough commonality among P-D contraction phenomena across languages to warrant a uniform syntactic analysis with minor variants. I focus on three issues that such an approach must address.

1.1 Adjacency

Riemsdijk (1998:651–667) argues that P-D contractions arise where a P and a D may occur side-by-side, as illustrated by the Italian data in (3). In (3a), the P-D contraction *nel* 'in the' occupies the same position as the P-D sequence *per il* 'for the.' Italian nominal syntax places the quantifier *tutto* 'all' between P and D, as in

(3b). When *tutto* is present, the P-D contraction *nel* becomes impossible, as (3c) and (3d) show. The Greek data in (4) display the same pattern.

(3) a. **nel** gruppo ∼ **per il** gruppo ‘in / for the group’
b. **in** *tutto* **il** gruppo ∼ **per** *tutto* **il** gruppo ‘in / for all the group’
c. ***nel** *tutto* gruppo
d. **tutto* **nel** gruppo

(4) a. **sta** pedhiá ∼ **ghia ta** pedhiá ‘to / for the children’
b. **se** *ola* **ta** pedhiá ∼ **ghia** *ola* **ta** pedhiá ‘to / for all the children’
c. ***sta** *ola* pedhiá
d. **ola* **sta** pedhiá (Riemsdijk 1998:664 and C. Condoravdi, p.c.)

For German, Riemsdijk appeals to the idiom in (5a). Note that the uninflected adverb *gut* ‘well’ cannot occur between *das* ‘the’ and *Drittel* ‘third.’ Adding *auf* ‘on’ to this idiom yields (5b), where *gut* separates P and D. As a result, the P-D contraction *aufs* ‘on the’ is impossible, as shown in (5c) and (5d). The adjacency condition on P-D contraction seems quite general and thus must be captured.

(5) a. *gut* **das** Drittel (***das** *gut* Drittel) ‘a little more than one third’
b. **auf** *gut* **das** Drittel ‘on a little more than one third’
c. ***aufs** *gut* Drittel
d. **gut* **aufs** Drittel (Riemsdijk 1998:663)

1.2 Syntactic relationships between P and D

At times P-D contractions arise in contexts where P and D are not in the ‘canonical’ syntactic relationship, in which D is head of P’s object. German exhibits contrasts according to the relationship between P and D. In (6), where D resides in an *adjunct* of the object of P, the P-D contraction *vom* ‘of the’ is unacceptable. In contrast, given a colloquial dative *possessor*, as in (7a), the P-D contraction *vom* may be used as in (7b); P takes the whole possessed phrase as its object (note that it governs the dative case of *seinem* ‘his’), and D lies in the object’s possessor. A grammar must distinguish the acceptable syntactic relationships between P and D.

(6) a. **von** [[**dem** König treu ergebenen] Dienern]
of the king faithfully devoted servants
‘of servants faithfully devoted to the King’
b. ***vom** König treu ergebenen Dienern

(7) a. [[dem Bürgermeister] sein Gehalt]
the mayor his salary
‘the mayor’s salary’
b. **vom** Bürgermeister seinem Gehalt (Riemsdijk 1998:655, 658)

1.3 When to use P-D contractions

Some languages allow P-D contractions and P-D sequences as stylistic alternates, e.g. in German, *ins Kino* ∼ *in das Kino* ‘to the cinema.’ Other languages favor P-D contractions over independent P and D, e.g. in Italian, *nel gruppo* ∼ **in il gruppo* 'in the group,' and in Greek, *sta pedhiá* ∼ **se ta pedhiá* 'to the children' (Riemsdijk 1998:664). However, languages that demand P-D contractions still allow 'contractible' P and D to occur separately in some circumstances. For instance, in *in tutto il gruppo* 'in all the group' and *se ola ta pedhiá* 'to all the children' P and D may occur independently, due to the intervention of a quantifier. A theory must be able to account for various intricacies concerning when P-D contractions are obligatory, optional, or disallowed.

1.4 Toward a solution

I address the above issues using *optimality-theoretic Lexical-Functional Grammar* (OT-LFG, Bresnan 2000) with *lexical sharing* (Wescoat 2002). In §2, I show that the adjacency facts follow from lexical sharing. I observe in §3 that LFG offers a means of regulating the syntactic relationships between P and D with functional constraints. In §4, I discuss how OT constraints provide insights into the issue of when to use P-D contractions. I conclude in §5 with observations about other phenomena where a similar combination of mechanisms prove useful.

2 Lexical sharing and adjacency

2.1 An empirical starting point

To grasp lexical sharing, it is useful to begin with the P-D contractions of French, shown in (8). Two definite articles lack corresponding P-D contractions, giving rise to the paradigms in (9a) and (9b), where the contractions *au*, *aux*, *du*, and *des* are juxtaposed with the P-D sequences *à la*, *à l'*, *de la*, and *de l'*. Other prepositional paradigms have only separate P and D, as in (9c). To achieve uniformity within and across the paradigms in (9), one might assume that all of the highlighted expressions comprise sequences of P and D. Along these lines, pedagogical grammars traditionally relate *au*, *aux*, *du*, and *des* to pairings of free P and D; Lancelot and Arnauld describe such forms as "a contraction of the particles *de* and *à*. . . with the plural *les* and the singular *le*" (1969 [1660]:53), and Condillac asserts that "*de le* changes into *du*. . . As for *de les*, it is always transformed into *des*, *à le* into *au*, *à les* into *aux*" (1986 [1775]:219). I take these observations as motivation for the working hypothesis that *au*, *aux*, *du*, and *des* involve sequences of P and D.

(8)

'the'	le [lə]	la [la]	l' [l]	les [le(z)][1]
à [a] 'to'	au [o]	—	—	aux [o(z)]
de [də] 'of'	du [dy]	—	—	des [de(z)]

(9)

a. **au** garçon	b. **du** garçon	c. **pour le** garçon	'to / of / for the boy'
à la fille	**de la** fille	**pour la** fille	'to / of / for the girl'
à l'enfant	**de l'**enfant	**pour l'**enfant	'to / of / for the child'
aux enfants	**des** enfants	**pour les** enfants	'to / of / for the children'

[1] *Le*, *la*, and *l'* are singular; *les* is plural. *Le* is masculine, and *la* feminine; *l'* and *les* are gender-neutral. *Le* and *la* occur before consonant-initial words, and *l'* before vowel-initial ones.

I next focus on the smallest P-D contraction, *au* 'to the,' comprising the single segment /o/. Hockett remarks: "since /o/ is a single phoneme, it is hardly possible to make a cut and produce two morphs" (1947:333); i.e. discrete parts corresponding to P and D are lacking. In a generativist vein, one might attempt to derive *au* /o/ from *à le* /a lə/; however, there is no independent motivation for positing synchronic processes capable of effecting this mapping. Indeed, any such derivation would need to be constrained so as not to apply to complementizer-pronoun sequences in infinitivals, e.g. ***à le*** *faire* ∼ ****au*** *faire* 'to do it.' The only plausible relationship between *au* /o/ and *à le* /a lə/ is a historical one, mediated by the following processes (Pope 1934:154, 190, 323–325), which are no longer productive:

Enclisis: Toward the beginning of the Old French period (mid ninth to early fourteenth century) unstressed masculine or neuter pronouns and articles preceded by a vowel-final word and followed by a consonant-initial word encliticize to the preceding word. E.g. *a lə mur*>*al myr*, 'to the wall.' By the end of the Old French period, most such enclitics are lost.

Vocalization: By the early part of the twelfth century, preconsonantal *l* either is lost or vocalizes to *w*. E.g. *al myr*>*aw myr*.

Leveling: By the latter part of seventeenth century, diphthongs are leveled to an intermediate vowel. E.g. *aw myr*>*o myr*.

From a lexicalist perspective, since *au* is properly regarded as a synchronically irreducible unit, it must be recorded in the lexicon. Having hypothesized earlier that *au* is associated with both P and D, I may now appear to be on the lip of a paradox, since a lexical item is traditionally linked to one syntactic category (Chomsky 1965:84). However, I have argued (Wescoat 2002) that some phenomena are best analyzed by assuming that lexical items may indeed be associated with multiple categories; these include noun incorporation in Hindi along with auxiliary contractions and pronominal determiners (see §5) in English. This leads to an unconventional notion about the relationship between words and constituency.

Two native relations among constituents are *containment* (e.g. a PP contains a P and a DP) and *precedence* (e.g. within a PP, the P precedes the DP). Consider now the relation between P and a word like *à* 'to.' It seems unnatural to say that P 'contains' *à*; rather this is a different type of relation, which Chametzky (1996:5) calls *instantiation* (e.g. *à* instantiates P). On this view, P contains nothing; a constituent that contains no other constituent may be described as *atomic*. I assume that each atomic constituent is instantiated by a word, which may be described as that constituent's *lexical exponent*; moreover, I assume that all words instantiate atomic constituents. Though explicit discussion of the matter is somewhat rare (see Chametzky 1996:5 and references therein, as well as Bresnan 2001:92), the majority of linguists seem to work under the tacit assumption that the instantiation relation is one-to-one. However, Gruber (1976) challenges this notion, portraying instantiation as one-to-many; thus, a word may instantiate more than one atomic constituent, or, equivalently, multiple atomic constituents may 'share' the same word as their lexical exponent. Thus, I call this state of affairs 'lexical sharing.' This view readily accommodates my two 'paradoxical' hypotheses; *au* is a lexical item specified as instantiating both a P and a D.

I hypothesize that all the P-D contractions considered here involve lexical sharing; i.e. they originate in the lexicon as forms instantiating both a P and a D. Even though most P-D contractions are longer than *au* and arguably morphologically complex, they typically exhibit indications of a lexical source. For instance, the various morphophonological idiosyncrasies mentioned in connection with the Italian and German paradigms in (1) and (2) constitute such an indication. As for the instantiation of both a P and a D, one may adduce distributional arguments, since all P-D contractions of which I am aware exhibit some form of alternation with independent P and D. Facts about P-D contractions and coordination in §2.3 provide further evidence of the instantiation of both a P and a D.

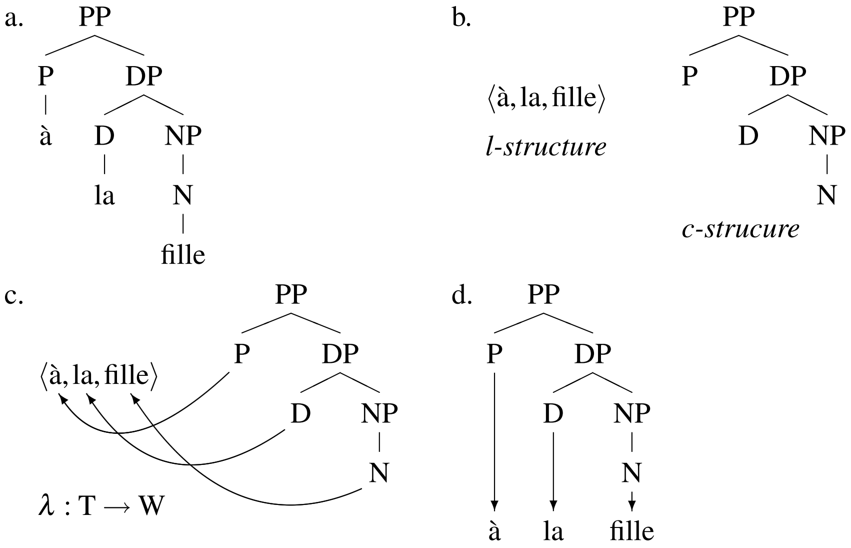
2.2 A formal model of lexical sharing

The fundamental architecture of LFG, which posits parallel *structures* related by *structural correspondences* (Kaplan 1995), provides the basis for a formal model of lexical sharing.

It is simplest to set the stage with an example involving no lexical sharing, like *à la fille* 'to the girl,' and approach the new proposal in steps, taking the conventional c(onstituent)-structure in (10a) as a starting point. In Kaplan's formulation, a c-structure comprises a set of *nodes* N, labeled with syntactic categories or words, and related by a *mother function* $\mathrm{M} : \mathrm{N} \rightarrow \mathrm{N}$ and a *precedence relation* $< \subseteq \mathrm{N} \times \mathrm{N}$. I propose to sever the nodes labeled with words from the c-structure, and to put those words into a separate representation called *l(exical)-structure*, as in (10b). Note that the nodes in c-structure are now labeled exclusively with syntactic categories. An l-structure, like $\langle \text{à}, \text{la}, \text{fille} \rangle$, consists of a linearly ordered set of words W.[2] Next one may identify the c-structure nodes that model atomic constituents; these are the members of the set of *terminals* T, comprising all non–mother nodes ($\mathrm{T} = \mathrm{N} - \mathrm{ran}\,\mathrm{M}$, where $\mathrm{ran}\,\mathrm{M}$ is the range of M). For instance, in (10b), T consists of the daughterless nodes labeled P, D and N. Now, to model the relation between atomic constituents and their lexical exponents, one may introduce a structural correspondence in the form of the *lexical exponent mapping* $\lambda : \mathrm{T} \rightarrow \mathrm{W}$, illustrated in (10c). To model the fact that all atomic constituents are instantiated by words and that all words instantiate atomic constituents, the function λ is total and onto; i.e. its domain is all of T, and its range is all of W. In the interest of transparency, (10c) follows a familiar mode of graphic representation for structural correspondences in LFG; parallel structures are side-by-side, and the correspondence mapping is rendered with curving lateral arrows. However, (10c) having served its purpose, I propose to adopt instead the more vertical format in (10d); elements of l-structure are spread out, without punctuation, below c-structure, and λ is rendered with descending arrows. The advantage of this scheme will become apparent.

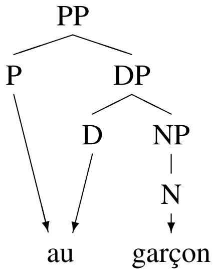
[2] In fact, I assume that the 'words' in W are abstract elements which are 'labeled' with word-forms. Thus, for $\langle \text{à}, \text{la}, \text{fille} \rangle$, W might contain w_1, w_2, w_3, labeled *à*, *la*, and *fille*, respectively, and ordered $w_1 < w_2 < w_3$. The distinction between such 'abstract' words and their labels allows l-structures in which the same word-form occurs multiple times, as in $\langle \textbf{the}, \text{dog}, \text{chased}, \textbf{the}, \text{cat} \rangle$.

(10)



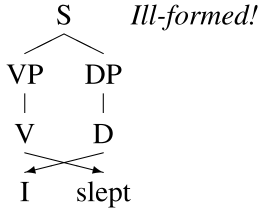
The task of representing lexical sharing is now straightforward. The mapping λ may be one-to-one, as in (10d), or it may just as easily map two or more terminals into a single word, as in (11), where the P-D contraction *au* ‘to the’ is the shared lexical exponent of P and D, and *garçon* ‘boy’ instantiates N.

(11)



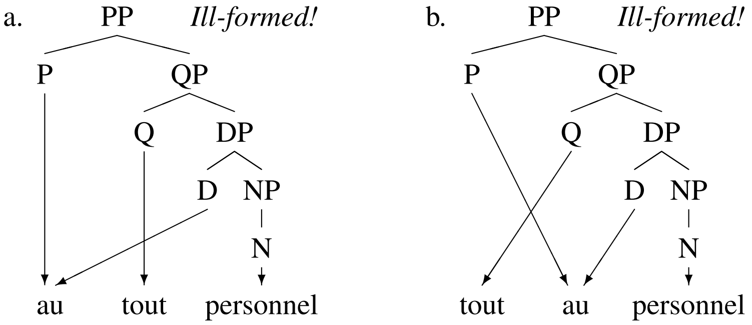
More must be said about ordering. For instance, nothing stated so far would prevent a monstrosity like (12), which appears to suggest that *I slept* may be analyzed as having verb-subject order in c-structure. Clearly there should be some consistency in ordering between c- and l-structure. This may be achieved with the introduction of the *order preservation axiom*: For all n_1 and n_2 in T, if $\lambda(n_1)$ precedes $\lambda(n_2)$, then n_1 precedes n_2. Let n_1 and n_2 be the nodes in (12) labeled D and V, respectively; $\lambda(n_1)$ precedes $\lambda(n_2)$, yet n_1 does not precede n_2, in violation of the order preservation axiom. Thus, (12) lies outside of the space of possibilities countenanced by the theory advanced here, so I have labeled the figure ‘*Ill-formed!*’ Since both T and W are linearly ordered, and since an order-preserving mapping such as λ between linearly ordered sets is technically a *homomorphism*, I call the sort of ill-formedness found in (12) a *homomorphism violation*. One may now see the advantage of the vertical format introduced in (10d); it ensures that homomorphism violations are always rendered visually conspicuous by *crossing arrows*.

(12)



The homomorphic character of λ leads to the *homomorphic lexical integrity theorem*: Only sequences of adjacent terminals may share a lexical exponent.[3] The utility of this theorem may be seen in a class of empirical predictions that I call *intermediate constituent suppression effects*: If two atomic constituents, X and Z, share a lexical exponent, then any constituent Y which the grammar would normally constrain to occur between X and Z will be blocked. A case in point is the adjacency condition on P-D contraction discussed in section 1.1. French allows [$_{\text{PP}}$ *à* [$_{\text{QP}}$ *tout* [$_{\text{DP}}$ *le personnel*]]] 'to all the personnel,' in which the Q *tout* 'all' is constrained by the grammar to fall between P and D. If, however, P and D share the P-D contraction *au* 'to the' as their lexical exponent, the presence of Q between P and D leads to a homomorphism violation, as indicated by the crossing arrows in (13a) and (13b). In other words, a Q intermediate between P and D would violate the homomorphic lexical integrity of the P-D contraction, so it is suppressed. Thus, P-D contractions are predicted to be possible only for adjacent P and D; moreover, this is an automatic consequence of the lexical-sharing analysis proposed here.

(13)



For a grammatical formalism, I employ a context-free grammar, as in (14a), to describe c-structure and a lexicon comprising *lexical-exponence rules*, as in (14b), to describe λ. A lexical-exponence rule $w \leftarrow \text{X}_1 \cdots \text{X}_n$ (note the leftward arrow) permits λ to map n adjacent terminals labeled from left to right $\text{X}_1, \ldots, \text{X}_n$ into w.

(14) a. PP $\rightarrow$ P {QP|DP}
QP $\rightarrow$ Q DP
DP $\rightarrow$ D NP
NP $\rightarrow$ N

b. à $\leftarrow$ P   la $\leftarrow$ D
au $\leftarrow$ P D   le $\leftarrow$ D
fille $\leftarrow$ N   personnel $\leftarrow$ N
garçon $\leftarrow$ N   tout $\leftarrow$ Q

[3] Proof: Let n_1, n_2, n_3 be terminals ordered $n_1 < n_2 < n_3$. Suppose the nonadjacent nodes n_1 and n_3 share w_1 as lexical exponent, while the intermediate node n_2 has a distinct lexical exponent w_2. Either w_1 precedes w_2, or vice versa; given the order preservation axiom, this contradictorily implies that n_2 should follow both n_1 and n_3 in the former case, and that it should precede both in the latter.

2.3 P-D contractions and coordination

Given a lexical-sharing analysis of P-D contractions, it is not surprising to find data like (15). A contraction is shared by a P and a D, with D heading the left-hand conjunct of a DP coordination, and with P taking scope over the entire coordinate structure, as depicted in (16). Lexical sharing permits the contractions to be treated as words, without necessitating any contortions of c-structure.

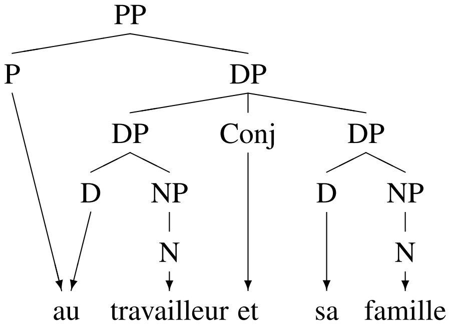
(15) a. [L]e fruit du travail revient **au travailleur et sa famille**. [Ve]
‘The fruits of labor go **to the worker and his family**.’

b. [C]eux qui sont les plus informés quant **aux médicaments et leurs effets**. . . sont parmi les moins observants d’une thérapeutique prescrite. [Br]
‘Those who are best informed with regard **to** [the] **medicines and their effects** are among the least heedful of a prescribed therapy.’[4]

c. [N]ous avons. . . employé tous les moyens qui pouvaient nous procurer le plus grand ordre. . . pour le retour **du roi et sa famille**. [Ba]
‘We did everything possible to secure the utmost order for the return **of the king and his family**.’

d. La possession et la manipulation **des germes microbiens et leurs dérivés**, quelqu’en soit le but, sont strictement réglementées. . . [Fa]
‘The possession and manipulation **of** [the] **pathogenic germs and their derivatives**, for whatever purpose, is strictly regulated.’

(16)



The analysis in (16) also applies to the Italian data in (17), where definite articles fill the position occupied by possessive pronouns in (15).

(17) a. Il problema «linguaggio e società» include questioni più particolari come quelle relative **alla storia della lingua e la storia del popolo**, . . . , **alla lingua e la nazione**, e **al linguaggio e la cultura**, **alle lingue letterarie nazionali e i dialetti**, **alla normatività delle lingue e la cultura del discorso** ecc. [Fo]
‘The problem of “language and society” includes more particular issues like those relating **to the history of the language and the history of the people**, **to** [the] **language and the nation**, and **to** [the] **language and** [the] **culture**, **to** [the] **national literary languages and** [the] **dialects**, **to the**

[4] Articles usually omitted from English translations are included in brackets to aid comparison.

prescriptivism of standard languages and the culture of speech, etc.’

b. La giovane evoca l’unione lontana **del padre e la madre**. [Gu]
‘The girl evokes the distant union **of the father and the mother**.’

c. Il riformismo…riconosce le ingiustizie del crescente divario tra le condizioni della vita **nella campagna e la città**… [Mo]
‘Reformism recognizes the injustices of the growing gap between living conditions **in the country and the city**.’

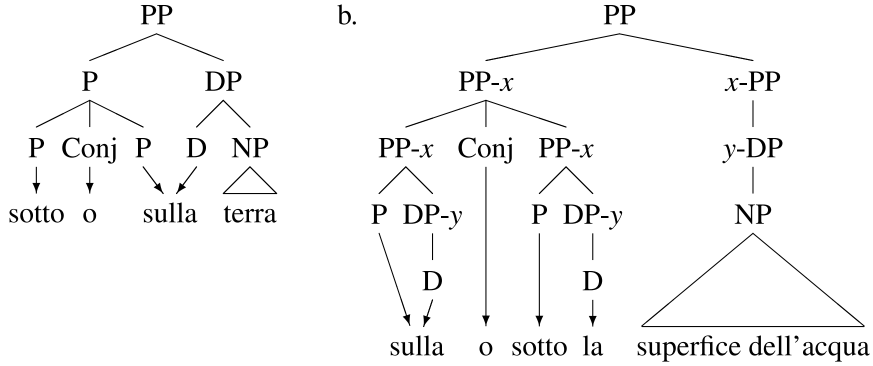
The pattern seen in (17) also arises in Catalan (A. Alsina, p.c.), Portuguese (Riemsdijk 1998:657n.), and Spanish (G.A. Broadwell, p.c.); compare (17b) with *del pare i la mare* (C), *do pai e a mãe* (P), and *del padre y la madre* (S) ‘of the father and the mother.’ Riemsdijk (1998:657) discusses similar German data.

Italian also allows coordinations like (18a) and (18b), pace Napoli and Nevis (1987:200). These too are readily analyzable in the lexical-sharing approach to P-D contractions, as shown in (19a) and (19b), respectively; (19b) exhibits right node raising, represented following the proposals of Maxwell and Manning (1996).

(18) a. [I]l poeta li immagina **sotto o sulla terra**… [Sc]
‘The poet imagines them [= hell and purgatory] **below or on the earth**.’

b. Piante galleggianti **sulla o sotto la superfice dell’acqua**… [Ma]
‘plants floating **on** [the] **or below the surface of the water**’

(19) a.



Citing (20a), Abeillé et al. (2003) report that the pattern ‘contraction NP “and” article NP’ in (17) is not generally acceptable in French. However, *au roi et la reine* ‘to the king and the queen,’ in (20b), is well attested as a frozen form, presumably left over from a time when constructions like (20c) were in use (ca. 1283).

(20) a. *J’ai parlé **au père et la mère**. (Abeillé et al. 2003:142)
‘I spoke **to the father and the mother**.’

b. [L]a nouvelle en vint jusqu’**au roi et la reine**. [Vo]
‘News of it traveled all the way **to the king and the queen**.’

c. …en la maniere que les ventes **des bois et les prevostés et les fermes** ont esté acoustumees a baillier autrefois… [Be]
‘in the manner in which it was customary to administer sales **of** [the] **woods and** [the] **provostships and** [the] **farms** in the past’

The analysis of Abeillé et al., anticipated in its broad outlines by Meigret (1888 [1550]:161–165), treats French P-D contractions as simple prepositions governing anarthrous objects. Associating determiners with NP, Abeillé et al. consider the determinerless objects to be instances of N$'$. They attribute the unacceptability of (20a) to coordination of unlike categories, [$_{\mathrm{N}'}$ *père*] *et* [$_{\mathrm{NP}}$ *la mère*] 'father and the mother.' However, the analysis relying on anarthrous objects runs into difficulties, since it predicts that the acceptable PP in (15a), *au travailleur et sa famille* 'to the worker and his family' should be no better than (20a), given that the object in (15a) would be the conjunction [$_{\mathrm{N}'}$ *travailleur*] *et* [$_{\mathrm{NP}}$ *sa famille*] 'worker and his family.'

The simplest explanation of the range of data seen here is that all the languages considered employ lexical sharing for their P-D contractions. This accounts for the pattern 'contraction NP "and" article NP' in Old French and other languages, as well as the pattern 'contraction NP "and" possessive NP' in Modern French, illustrated in (15), (17), and (20c). What remains to explain is the unacceptability of (20a); I return to this matter in §4.3, where I suggest that French has *reranked* a constraint, in the OT sense, giving rise to this pattern.

3 LFG and the syntactic relationship between P and D

3.1 LFG with lexical sharing and the statement of a functional constraint

Lexical sharing may be incorporated into LFG in three steps, which establish the relationship between l-structure and *f*(*unctional*)-*structure*, LFG's representation of grammatical functions. (*a*) The structural correspondence φ, originally envisioned as relating c-structure to f-structure (Kaplan 1995), is extended to include elements of l-structure in its domain; thus, $\varphi : \mathrm{N} \cup \mathrm{W} \to \mathrm{F}$ is a mapping from nodes and words to members of the set F of f-structures. (*b*) For convenience, one may define the metavariable $\Downarrow$ as an abbreviation for $\varphi(\lambda(*))$ 'the f-structure of the lexical exponent of the current node [$= *$].' (*c*) Finally, the right-hand sides of lexical-exponence rules are furnished with functional annotations, as in (21).

(21) a. vom ←

$$\begin{array}{cc}
\mathrm{P} & \mathrm{D} \\
(\downarrow \mathrm{PRED}) = \text{`}\mathrm{OF}\langle(\downarrow \mathrm{OBJ})\rangle\text{'} & (\downarrow \mathrm{SPEC}) = \text{`}\mathrm{THE}\text{'} \\
(\downarrow \mathrm{OBJ\ CASE}) = \mathrm{DAT} & (\downarrow \mathrm{GEND}) \neq \mathrm{F} \\
\Downarrow = \downarrow & (\downarrow \mathrm{NUM}) =_c \mathrm{SG} \\
 & (\downarrow \mathrm{CASE}) =_c \mathrm{DAT} \\
 & (\Downarrow \mathrm{OBJ\ AF}^*) =_c \downarrow
\end{array}$$

b. dem ←

$$\begin{array}{c}
\mathrm{D} \\
(\downarrow \mathrm{SPEC}) = \text{`}\mathrm{THE}\text{'} \\
(\downarrow \mathrm{GEND}) \neq \mathrm{F} \\
(\downarrow \mathrm{NUM}) =_c \mathrm{SG} \\
(\downarrow \mathrm{CASE}) =_c \mathrm{DAT} \\
\Downarrow = \downarrow
\end{array}$$

c. von ←

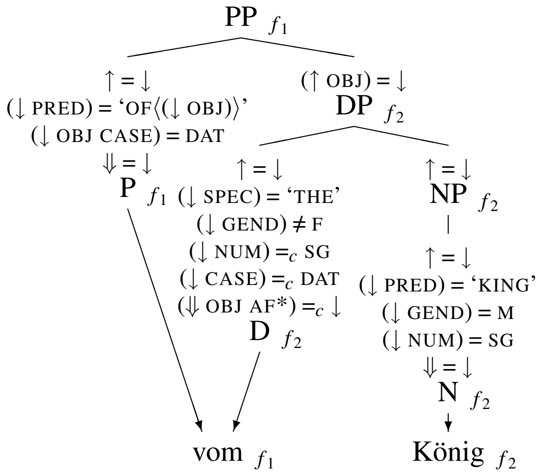
$$\begin{array}{c}
\mathrm{P} \\
(\downarrow \mathrm{PRED}) = \text{`}\mathrm{OF}\langle(\downarrow \mathrm{OBJ})\rangle\text{'} \\
(\downarrow \mathrm{OBJ\ CASE}) = \mathrm{DAT} \\
\Downarrow = \downarrow
\end{array}$$

d. König ←

$$\begin{array}{c}
\mathrm{N} \\
(\downarrow \mathrm{PRED}) = \text{`}\mathrm{KING}\text{'} \\
(\downarrow \mathrm{GEND}) = \mathrm{M} \\
(\downarrow \mathrm{NUM}) = \mathrm{SG} \\
\Downarrow = \downarrow
\end{array}$$

Seen in (22) are c-, l-, and f-structures for the German *vom König* 'of the king.' (*Von dem König*, with independent P and D, would have the same c- and f-structures, except that the annotation $\Downarrow = \downarrow$ would replace $(\Downarrow \mathrm{OBJ\ AF}^*) =_c \downarrow$ in c-structure.) The annotations $\uparrow = \downarrow$ and $(\uparrow \mathrm{OBJ}) = \downarrow$ are due to universal principles (Bresnan 2001:103); the rest are provided by lexical-exponence rules in (21).

(22)



$$f_1 \begin{bmatrix} \text{PRED} & \text{‘OF}\langle f_2 \rangle\text{’} \\ \text{OBJ} & f_2 \begin{bmatrix} \text{SPEC} & \text{‘THE’} \\ \text{PRED} & \text{‘KING’} \\ \text{GEND} & \text{M} \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

Recall from §1.2 that the felicity of German P-D contractions depends on the syntactic relationship between P and D. In (23a), *vom* ‘of the’ is blocked because D lies in an adjunct of P’s object. *Vom* is acceptable, though, when D begins a dative possessor of P’s object, as in (23b). Since recursion of possessors is allowed, as in (23c), P and D stand in a *long-distance dependency*, which may be regulated using *functional uncertainty* (Kaplan and Zaenen 1995). In (21a), the annotation ⇓ = ↓ on P says that the f-structures of P and *vom* are the same; it follows that an annotation on D can refer to P’s f-structure by referring to the f-structure of *vom* with ⇓. Thus, the constraint (⇓ OBJ AF*) $=_c$ ↓ on D in (21a) requires that D’s f-structure be reachable from P’s f-structure via a path of attributes conforming to the pattern OBJ AF*, i.e. an OBJ followed by zero or more *argument functions*, which exclude ADJUNCT (Bresnan 2001:97). The usual analysis of modifiers would assign to (23a) an f-structure along the lines of (24), where P’s f-structure would be f_1, and D’s would be f_3. Note that f_3 is reachable from f_1 only via the path OBJ ADJUNCT OBL, which does not conform to the pattern OBJ AF*. Thus, (24) violates the above constraint, so (23a) is ruled out. Since possessors may express arguments, comparable violations would not arise in the cases of (23b) and (23c).

(23) a. ***vom** König treu ergebenen Dienern [=(6b)]
of [[[*the king*] *faithfully devoted*] *servants*][5]
[of servants faithfully devoted to the King]
b. **vom** Bürgermeister seinem Gehalt [=(7b)]
of [[*the mayor*] *his* *salary*]
‘of the mayor’s salary’
c. **vom** Hans seiner Mutter ihrem Freund seinem Geld
of [[[[*the Hans*] *his* *mother*] *her* *friend*] *his* *money*]
‘of Hans’s mother’s friend’s money’ (Riemsdijk 1998:659)

[5] Since *vom* straddles phrase boundaries, I have bracketed the gloss as a proxy for the German.

(24)
$$f_1\left[\begin{array}{ll}\text{PRED} & \text{`OF}\langle f_2\rangle\text{'}\\ \text{OBJ} & f_2\left[\begin{array}{ll} \text{PRED} & \text{`SERVANT'}\\ \text{ADJUNCT} & \left\{\left[\begin{array}{ll}\text{PRED} & \text{`DEVOTED-TO}\langle f_3\rangle\text{'}\\ \text{ADJUNCT} & \left\{\left[\begin{array}{ll}\text{PRED} & \text{`FAITHFULLY'}\end{array}\right]\right\}\\ \text{OBL} & f_3\left[\begin{array}{ll}\text{SPEC} & \text{`THE'}\\ \text{PRED} & \text{`KING'}\end{array}\right]\end{array}\right]\right\}\end{array}\right]\end{array}\right]$$

3.2 Dative possessors and problems for some alternative views

Some analyses of P-D contractions founder on examples with dative possessors. In one such approach, Hinrichs (1986) treats contractions as simple prepositions inflected for definiteness, gender, number, and case; he assumes such prepositions select an N′ object onto which those features are copied. However, inclusion of a possessor is generally assumed to be incompatible with N′ status. Moreover, when the possessor's gender differs from that of the object which contains it, as in (25), the inflected P agrees with the possessor, even though one would expect the P's features to be copied onto the object, not onto one of its subconstituents.[6] In contrast, the agreement pattern in (25) is unsurprising if P-D contractions instantiate both a P and a D. As shown in (21a), P governs only the CASE of its OBJ(ect), leaving D to regulate its own GEND(er), NUM(ber), and CASE; if D comes to head a possessor, then the overarching object may have different features.[7]

(25) a. zur$_{fem}$ Prinzessin$_{fem}$ ihrem$_{neut}$ Palais$_{neut}$
to the princess her palace
'to the princess's palace'
b. am$_{neut}$ Auto$_{neut}$ seiner$_{fem}$ Stoßstange$_{fem}$
on the car its fender
'on the car's fender' (Riemsdijk 1998:658)

I next consider movement-based theories, which assume the *Y-model*, where derivation begins with *overt syntax* and branches into computations of *Logical Form* (LF) and *Phonetic Form* (PF). Computation of LF after the LF/PF branching is *covert*, so audible contraction must be in one of the other components.

Riemsdijk (1998:651–667) ascribes P-D contraction to D-to-P *Raising* in overt syntax. However, this seems to clash with the *Minimalist* notion that movement "takes place only when forced (Last Resort)," being "driven by morphological considerations: the requirement that some feature F must be checked" (Chomsky 1995:235, 262). It is counterintuitive to treat optional P-D contraction in German as an operation of 'Last Resort.' More problematic is the fact that, in languages like French, Greek, or Italian, P-D contraction is generally obligatory but can be blocked by an intervening quantifier, in which case P and D occur separately; for

[6] One informant rejects differing genders for possessor and possessum; others do not.

[7] The overarching object must, however, share the possessor's dative case (Riemsdijk 1998:659n.).

instance, recall the Italian ***in*** *tutto* ***il*** *gruppo* 'in all the group' from (3b). If the purpose of D-to-P Raising is to facilitate feature checking, then failure to move should prevent the relevant feature from getting checked and thereby cause the derivation to 'crash.' The fact that P and D survive on their own when not adjacent challenges the idea that their feature-checking needs cannot be satisfied without movement, leaving one to wonder how Raising could be 'forced' when P and D are side-by-side. Thus, theory-internal issues cast doubt on an analysis in overt syntax.

The remaining component is the computation of PF after the LF/PF branching, which Embick and Noyer (2001) call *Morphology*. This is, they claim, where trees acquire 'left-to-right' ordering via *linearization*. Operations in Morphology consequently fall into two types: *Lowering* precedes linearization, so it is sensitive only to hierarchical relations; *Local Dislocation* follows linearization and operates on the basis of linear adjacency. Focusing on French, Embick (2006) notes that either operation could be responsible for P-D contraction. However, neither shows much promise of handling the corresponding German data. Embick's preferred analysis uses Lowering, "the process which adjoins a head to the head of its complement" (2006:16), to merge P with D. However, Lowering fails to account for all the data; it cannot adjoin P to the D of a dative possessor, as in (23b) and (23c), since that D is not the head of P's complement. The remaining operation, Local Dislocation, merges adjacent X^0 elements. Hence, it would adjoin P to an adjacent D, and in the case of (23b) and (23c), the D in question would be that of the possessor. However, Local Dislocation would also incorrectly adjoin P to the adjacent D in (23a), even though the D in question lies inside of an adjunct. "Local Dislocation . . . is sensitive to relations of adjacency and precedence between constituents" (Embick and Noyer 2001:564), and as far as adjacency and precedence are concerned, there is nothing to distinguish the ungrammatical (23a) from the grammatical (23b) and (23c). To capture the data in (23), one must simultaneously enforce adjacency while retaining the ability to regulate the syntactic relationship between P and D, as in the analysis employing LFG with lexical sharing. By associating sensitivity to adjacency and sensitivity to hierarchical syntactic relationships with different stages of derivation, Embick's (2006) approach seems to deny itself the necessary combination of tools for these data.

4 OT and when to use P-D contractions

4.1 OT and obligatory P-D contractions

OT assumes two components, GEN(eration) and EVAL(uation); GEN enumerates a set of potential outputs called *candidates*, and EVAL selects the *optimal* candidate as final output. EVAL compares candidates with respect to a hierarchy of violable constraints. One candidate is more *harmonic* than another, if for some constraint the former incurs fewer violations than the latter, while for all higher-ranking constraints the two candidates incur equal numbers of violations. The optimal candidate is more harmonic than any other. In OT-LFG, Bresnan (2000) offers a version of OT where GEN is an LFG; I assume this LFG uses lexical sharing, as in §3.1.

Obligatory P-D contraction, as in the French ***au*** *garçon* ∼ *****à le*** *garçon* ‘to the boy,’ appears to be a form of *Poser blocking*, where a single lexical item is chosen over an equivalent multi-word construction (Poser 1992). Here I speculate about how Poser blocking might be treated as a constraint in the EVAL component of an OT-LFG. As an OT constraint, Poser blocking penalizes failures to exploit opportunities to effect economies of expression at the word level. Instances of Poser blocking often seem to reflect language-particular idiosyncrasies incompatible with the notion of OT constraints as universals. However, I suggest that this is because the universal constraint is fed by interword relationships that are particular to the lexicon of each language.

Clearly, a lexical treatment of P-D contractions must acknowledge the relation between these forms and independent Ps and Ds. Compare the lexical-exponence rule in (26a) for the French P-D contraction *au* ‘to the’ with the rules for *à* ‘to’ in (26b) and *le* ‘the’ in (26c). The commonalities in syntactic categories and functional specifications must be captured by some form of redundancy rules; hence, the lexicon must contain a representation of the relevant interword relations. Here I am concerned not so much with the redundancy rules as with the existence of relations between lexically shared words like *au* and ‘unshared’ words like *à* and *le*; as a shorthand, I write $\textit{à}, \textit{le} \sim \textit{au}$ to indicate that such a relation is present.

(26) a.
$$\begin{array}{lcc} \text{au} \;\leftarrow & \text{P} & \text{D} \\ & (\downarrow \text{PRED}) = \text{‘TO}\langle(\downarrow \text{OBJ})\rangle\text{’} & (\downarrow \text{SPEC}) = \text{‘THE’} \\ & \Downarrow = \downarrow & (\downarrow \text{GEND}) =_c \text{M} \\ & & (\downarrow \text{NUM}) =_c \text{SG} \end{array}$$

b.
$$\begin{array}{lc} \text{à} \;\leftarrow & \text{P} \\ & (\downarrow \text{PRED}) = \text{‘TO}\langle(\downarrow \text{OBJ})\rangle\text{’} \\ & \Downarrow = \downarrow \end{array}$$

c.
$$\begin{array}{lc} \text{le} \;\leftarrow & \text{D} \\ & (\downarrow \text{SPEC}) = \text{‘THE’} \\ & (\downarrow \text{GEND}) =_c \text{M} \\ & (\downarrow \text{NUM}) =_c \text{SG} \\ & \Downarrow = \downarrow \end{array}$$

One application of Poser blocking (I assume there are others, some not involving lexical sharing) is fed by the lexical relation $w_1, w_2 \sim w_3$. In this instance, Poser blocking applies to sequences of terminals, n_1 and n_2, with syntactic categories such that they could be instantiated either separately by w_1 and w_2 or jointly by w_3. Empirically, it seems that the categories of n_1 and n_2 must be the most restrictive ones capable of accommodating the relevant lexical exponents; for the cases examined here, I assume the categories of preposition and definite article. Poser blocking penalizes failure to exploit the relation $w_1, w_2 \sim w_3$ to achieve economy of expression. Such a failure may take two forms, either ignoring w_3 and using w_1 and w_2 to instantiate n_1 and n_2, or using w_3 in such a way that no economy is achieved, i.e. by making w_3 the lexical exponent of one but not both of n_1 and n_2. Thus, given $\textit{à}, \textit{le} \sim \textit{au}$, one subcase of Poser blocking for French is (27).

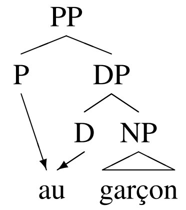
(27) For each sequence of terminals n_1, a preposition, and n_2, a definite article, count a violation if

a. n_1 and n_2 are separately instantiated by *à* and *le*, or

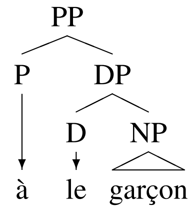
b. either n_1 or n_2 is instantiated by *au*, but not both.

Poser blocking readily predicts simple cases of obligatory P-D contractions. The LFG that enumerates candidates in GEN produces not only (28a), with lexical sharing, but also (28b) with independent P and D. In EVAL, however, (28b) violates Poser blocking—specifically, the subcase in (27a). Since (28a) incurs no violation of Poser blocking, or PBLK, it is the more harmonic candidate, as (28c) shows.

(28) a.



b.



c.

	PBLK
☞ **au** garçon	
à le garçon	*!

There are two ways to explain why usually barred combinations such as *à* 'to' and *le* 'the' arise in expressions like *à tout le personnel* 'to all the personnel.' First, Poser blocking applies to sequences of terminals, so the Q between P and D renders the constraint inapplicable. Second, the LFG in GEN cannot produce a candidate with a P-D contraction in this instance, as discussed in §2.2, so the structure with separate P and D is the only candidate, and by default the most harmonic.

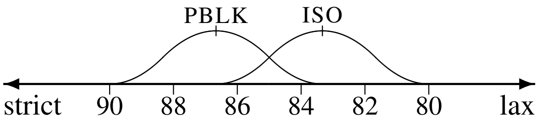
4.2 Optionality

Recall that in German, P-D contractions are optional: ***ins*** *Kino* ~ ***in das*** *Kino* 'to the cinema.' This state of affairs can be modeled in *stochastic OT*, which would allow PBLK to stand in a reversible ranking with respect to a conflicting constraint, leading to variable outputs (Bresnan et al. 2001).

I speculate that the constraint that conflicts with PBLK is one which penalizes lexical sharing. At least in the languages considered here, most words instantiate a single atomic constituent; thus, λ tends toward being one-to-one. A one-to-one mapping between the linearly ordered sets T and W would be not just a homomorphism but an *isomorphism*. The tendency toward isomorphism can be captured with a constraint, say ISO, which is violated whenever lexical sharing arises.

In stochastic OT, a constraint's rank is a value on the continuous scale of real numbers. Upon evaluation, noise in the form a random value drawn from a normal distribution is added to each rank to produce an *effective rank*; the resulting effective ranks determine an evaluation-particular ordering of constraints. Shown in (29) is an artificial example of a scale, with bell curves representing the normal distributions of the effective ranks of PBLK and ISO. If two constraints have overlapping normal distributions, as in (29), the relative ordering of their effective ranks may vary from one evaluation to another. For instance, one evaluation may give PBLK and ISO effective ranks of 88.1 and 82.4, respectively, in which case PBLK ≫ ISO follows, as in (30a), or the relevant values may be 84.0 and 85.7, respectively, yielding ISO ≫ PBLK, as in (30b). Thus, if German has a ranking like (29), P-D contractions are predicted to be optional. For languages where P-D contractions are obligatory, PBLK outranks ISO by a greater interval, leaving no overlap between the respective ranges of their effective ranks.

(29)



(30) a.

	PBLK	ISO
in das Kino	*!	
☞ **ins** Kino		*

b.

	ISO	PBLK
☞ **in das** Kino		*
ins Kino	*!	

4.3 Coordination and across-the-board effects

I next return to the issue of Modern French P-D contractions and coordination, left unresolved in §2.3. Abeillé et al. (2003) offer the facts in (31), which collectively suggest that the Poser blocking effects discussed in this section apply *across-the-board* (ATB) to coordinate structures. Recall that Poser blocking applies to 'sequences' of terminals. ATB application of the constraint amounts to this: If the sequence overlaps the leading / trailing edge of a coordinate structure, then the portion of the sequence that extends into the coordinate structure is projected onto the beginning / end of each conjunct. All the PPs in (31) feature a P that takes scope over coordinated DPs headed by definite articles, so each case involves two 'sequences' of terminals to which Poser blocking is applicable; one sequence is the P and the D of the left-hand conjunct, while the other—the problematic sequence in (31b) and (31c)—comprises the P and the D of the right-hand conjunct. In (31b), the problematic sequence is instantiated by *au* 'to the' and *la* 'the,' violating Poser blocking—specifically subcase (27b)—because *au* fails to instantiate both of the relevant terminals. In (31c), the problematic sequence is instantiated by *à* 'to' and *le* 'the,' again violating Poser blocking—subcase (27a). In the remaining examples in (31), Poser blocking is satisfied with respect to both conjuncts. Due to the above violations, (31b) and (31c) are less harmonic than alternative candidates with coordinated PPs, as indicated in (32a) and (32b), respectively.

(31) J'ai parlé… 'I spoke…' (Abeillé et al. 2003:142)
a. à la mère et la fille. 'to the mother and the daughter.'
b. *au père et la mère. 'to the father and the mother.' [= (20a)]
c. *à la fille et le garçon. 'to the girl and the boy.'
d. à la fille et l'autre garçon. 'to the girl and the other boy.'

(32) a.

	PBLK
au père et **la** mère	*!
☞ **au** père et **à la** mère	

b.

	PBLK
à la fille et **le** garçon	*!
☞ **à la** fille et **au** garçon	

Recall from (15) that French allows analogs of (31b) with a possessive in place of the definite article of the second conjunct, as in *au travailleur et* ***sa*** *famille* 'to the worker and his family.' It appears that when Poser blocking is fed by a lexical relation like *à, le*∼*au*, the effect of the constraint is limited to sequences of prepositions and definite articles. As to why possessives in the latter position do not count, I suspect that the lexicon contains some representation of paradigms like (8), correlating prepositions with definite articles, but omitting possessives and other categories, and I assume that Poser blocking constrains items figuring within the same paradigm. A more precise proposal will have to await further investigation.

Spanish offers a useful contrast, since its inventory of P-D contractions, in (33), partially parallels that of French. Despite the change from 'to' forms to 'of' forms, (34a) and (34b) are analogs of (32b) and (32c), respectively. However, the Spanish data are acceptable. This contrast follows from a difference in constraint ranking.

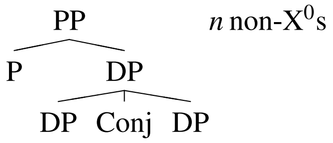
(33) [cf. (8)]

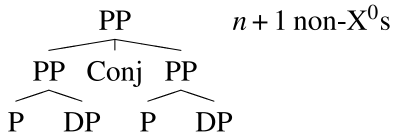
'the'	el	la	los	las[8]
a 'to'	al	—	—	—
de 'of'	del	—	—	—

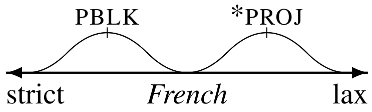
(34) a. [E]stos roles **del padre y la madre** no son exclusivos... [Bu]
'These roles **of the father and the mother** are not exclusive.'

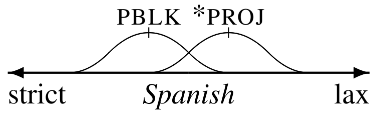
b. Yo quería actualizar estos arquetipos **de la madre y el padre**. [Vi]
'I wanted to update these archetypes **of the mother and the father**.'

Consider a rendering of Bresnan's (2001:91) principle of *Economy of Expression* as an OT constraint 'avoid projections,' symbolized *PROJ, which is violated once for every non-X^0 node. Note that a P with a conjoined object, as in (35a), contains fewer XPs than does the equivalent conjunction of PPs, as in (35b). Thus, given competing candidates exhibiting the structures in (35a) and (35b), *PROJ favors (35a) over (35b), all other things being equal, while ATB application of PBLK sometimes picks (35b) over (35a), as has been seen in connection with (32). The consistent ATB application of PBLK seen in the French data in (31) may be guaranteed by assuming that PBLK outranks *PROJ by enough of an interval to ensure that the respective ranges of their effective ranks do not overlap, as suggested in (36a), thus yielding the tableau in (37a). In contrast, if there is an overlap, as proposed for Spanish in (36b), then depending on the evaluation, either PBLK ≫ *PROJ or *PROJ ≫ PBLK may follow, giving the results in (37b) and (37c), respectively. On those evaluations where the ranking *PROJ ≫ PBLK prevails, requiring the structure in (35a), violations of PBLK involving the second conjunct may be unavoidable, but nothing will prevent PBLK from being observed with respect to the initial conjunct.

(35) a. 

b. 

(36) a. 

b. 

(37) a.

French	PBLK	*PROJ
au père et **la** mère / **à la** fille et **le** garçon	*!	6 × *
☞ **au** père et **à la** mère / **à la** fille et **au** garçon		7 × *

[8] *El* and *los* are masculine; *la* and *las* are feminine. *El* and *la* are singular; *los* and *las* are plural.

b.

Spanish	PBLK	*PROJ
del padre y **la** madre / **de la** madre y **el** padre	*!	6 × *
☞ **del** padre y **de la** madre / **de la** madre y **del** padre		7 × *

c.

Spanish	*PROJ	PBLK
☞ **del** padre y **la** madre / **de la** madre y **el** padre	6 × *	*
del padre y **de la** madre / **de la** madre y **del** padre	7 × *!	

In §2.3, I observed that Modern French differs from other languages discussed here. The available information suggests that Italian, Catalan, Portuguese, German, and Old French are like Spanish in that the ranks of PBLK and *PROJ are close enough to allow one or the other to prevail, depending on the evaluation. Modern French seems to have distinguished itself from this group by increasing the interval between the ranks of PBLK and *PROJ, so that the former is always dominant.

5 Conclusion

This study shows lexical sharing to be a useful basis for modeling P-D contractions in various languages. The contractions are treated as single words, yet they are linked to multiple elements of c-structure, yielding syntactic analyses confirmed by coordination data readily found in published sources. Lexical sharing may be incorporated into LFG, which provides effective tools for regulating the syntactic relationship between the c-structure elements associated with P-D contractions. OT extensions to LFG offer a means of predicting when P-D contractions are required, while affording insights into subtle cross-linguistic differences. The picture that emerges is one of a largely unified phenomenon with minor variations.

The proposal set out here employs a set of analytic tools that I believe may be usefully applied to other phenomena. In this regard, consider '*one*(*s*)-deletion,' discussed by Perlmutter (1970:236–237) and others, and illustrated in (38). Elsewhere (Wescoat 2002) I treat *mine* and similar *pronominal determiners* as instances of lexical sharing: *mine* ← D N. This predicts incompatibility with simple adjectives that fall between D and N as a case of intermediate constituent suppression. For many speakers, *my one* gives way to *mine* in the absence of a modifier.[9] This is suggestive of Poser blocking, requiring a single word in place of a phrase.

(38) a. **mine** b. ***my one** c. **my** blue **one** d. ***mine** blue e. *blue **mine**

Another area that may benefit from similar analytic tools concerns Danish definiteness marking. A definite suffix, e.g. *-et*, is used, unless there is a prenominal modifier, in which case an independent article is employed; see (39). If the definite suffix triggers lexical sharing, yielding *hus-et* ← D N 'the house' for example, then incompatibility with prenominal modifiers is predicted as another instance of intermediate constituent suppression. The fact that separate D and N like **det hus* 'the house' give way to suffixed forms like *hus-et* when prenominal modifiers are absent has been analyzed as Poser blocking by Hankamer and Mikkelsen (2002).

[9] C. Allen (p.c.) informs me that some speakers allow expressions like *my one*.

(39)	a. hus**et**	b. ***det** hus	c. *store hus**et**	d. **det** store hus
	house-the	*the house*	*big house-the*	*the big house*
				(Sadock 1991:115)

In sum, the combination of tools arrayed here to analyze P-D contraction shows promise of offering insights into various phenomena, including some that I lack space to mention, and is consequently worthy of further investigation.

References

Abeillé, Anne, Olivier Bonami, Danièle Godard, and Jesse Tseng (2003) 'French *à* and *de*: An HPSG analysis.' In Patrick Saint-Dizier, ed., *Proceedings of the ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and Their Use in Computational Linguistics Formalisms and Applications*. Dordrecht: Kluwer, 133–144.

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: Oxford University Press, 334–385.

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

Bresnan, Joan, Shipra Dingare, and Chris Manning (2001) 'Soft constraints mirror hard constraints: Voice and person in English and Lummi.' In Miriam Butt, and Tracy Holloway King, eds., *Proceedings of the LFG01 Conference*. Stanford: CSLI, 13–32.

Chametzky, Robert A. (1996) *A Theory of Phrase Markers and the Extended Base*. New York: State University of New York Press.

Chomsky, Noam (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT.

(1995) 'Categories and transformations.' In *The Minimalist Program*. Cambridge, MA: MIT, 219–394.

Condillac, Étienne Bonnot de (1986 [1775]) *Cours d'étude pour l'instruction du Prince de Parme: Grammaire*. Stuttgart-Bad Cannstatt: Frommann-Holzboog. [Facsimile of 1st edn, Parma: Royal Stationery Office, 1775.]

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

Embick, David (2006) *Linearization and Local Dislocation: Derivational Mechanics and Interactions*. [MS., University of Pennsylvania.]

Embick, David, and Rolf Noyer (2001) 'Movement operations after syntax.' *Linguistic Inquiry* 32, 555–595.

Gruber, Jeffrey S. (1976) 'Functions of the lexicon in formal descriptive grammars.' In *Lexical Structures in Syntax and Semantics*. Amsterdam: North-Holland, 211–367.

Hankamer, Jorge, and Line Mikkelsen (2002) 'A morphological analysis of definite nouns in Danish.' *Journal of Germanic Linguistics* 14, 137–175.

Hinrichs, Erhard W. (1986) '*Verschmelzungsformen* in German: A GPSG analysis.' *Linguistics* 24, 939–955.

Hockett, Charles F. (1947) 'Problems of morphemic analysis.' *Language* 23, 321–343.

Kaplan, Ronald M. (1995) 'Three seductions of computational psycholinguistics.' In Dalrymple et al. (1995:339–367).

Kaplan, Ronald M., and Annie Zaenen (1995) 'Long-distance dependencies, constituent structure, and functional uncertainty.' In Dalrymple et al. (1995:137–165).

Lancelot, Claude, and Antoine Arnauld (1969 [1660]) *Grammaire générale et raisonée*. Menston, Yorkshire: Scolar Press. [Facsimile of 1st edn, Paris: Pierre le Petit, 1660.]

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 First LFG Conference*. Stanford: CSLI.

Meigret, Louis (1888 [1550]) *Le tretté de la grammęre françoęse*. Heilbronn: Gebr. Henninger. [1st edn, Paris: Chrestien Wechel, 1550.]

Napoli, Donna Jo, and Joel Nevis (1987) 'Inflected prepositions in Italian.' *Phonology Yearbook* 4, 195–209.

Perlmutter, David M. (1970) 'On the article in English.' In Manfred Bierwisch, and Karl Erich Heidolph, eds., *Progress in Linguistics: A Collection of Papers*. The Hague: Mouton, 233–248.

Pope, M. K. (1934) *From Latin to Modern French with Especial Consideration of Anglo-Norman*. Manchester: Manchester University Press.

Poser, William J. (1992) 'Blocking of phrasal constructions by lexical items.' In Ivan A. Sag, and Anna Szabolcsi, eds., *Lexical Matters*. Stanford: CSLI, 111–130.

Riemsdijk, Henk van (1998) 'Head movement and adjacency.' *Natural Language and Linguistic Theory* 16, 633–678.

Sadock, Jerrold M. (1991) *Autolexical Syntax: A Theory of Parallel Grammatical Representations*. Chicago: University of Chicago Press.

Wescoat, Michael T. (2002) *On Lexical Sharing*. Ph.D. dissertation, Stanford University.

Sources

[Ba] Report by Antoine Barnave, 1791, in vol. 3 of Guillaume N. Lallement, ed., *Choix de rapports, opinions et discours prononcés à la Tribune Nationale depuis 1789 jusqu'à ce jour*. Paris: A. Eymery, 1818: 129.

[Be] Philippe de Beaumanoir, vol. 1 of *Coutumes de Beauvaisis*, ed. A. Salmon. Paris: A. Picard, 1899: 41–42. (Extracted from the BFM—Base de Français Médiéval, UMR-ICAR, ENS-LSH, http://bfm.ens-lsh.fr, 30 August 2007.)

[Br] Catherine Breton, "Croyances médicamenteuses: Aller contre ou faire avec," in André Grimaldi and Julie Cosserat, eds., *La relation médecin-malade*. Paris: Elsevier, 2004: 187.

[Bu] Juan Manuel Burgos, *Antropología: Una guía para la existencia*. Madrid: Palabra, 2003: 316.

[Fa] Jean Favelier et al., *Manuel de prévention des risques associés aux techniques biologiques: Application à l'enseignement*. Paris: Elsevier, 1995: 170.

[Fo] Lia Formigari, *Marxismo e teorie della lingua: Fonti e discussioni* Messina: La libra, 1973: 357.

[Gu] Ernesto Guidorizzi, *Echi di Goethe in Italia* Venice: Cafoscarina, 1988: 87.

[Ma] Palmer Marchi et al., *Famiglie di piante vascolari italiane: 1–30*. Rome: Università degli studi La Sapienza, 2002: 10.

[Mo] Tommaso Morlino, *Le autonomie locali nella avanzata democratica*. Rome: Cinque lune, 1962: 115.

[Sc] Nicola Scarano, *Saggi danteschi*. Livorno: R. Giusti, 1905: 44.

[Ve] Jeannette Vermeersch, *Organiser les travailleuses qui occupent une place décisives dans la nation*. Paris: PCF, 1961: 39.

[Vi] Juan Villoro et al., *Certidumbre del extravío: Entrevista con Juan Villoro*. Colima: Universidad de Colima, 2001: 80.

[Vo] Voltaire, *Zadig, ou la destinée, histoire orientale*, in vol. 21 of *Œuvres complètes de Voltaire*. Paris, Garnier, 1879: 39.