Proceedings of the LFG 06 Conference

Universität Konstanz

Editors: Miriam Butt and Tracy Holloway King

2006

CSLI Publications

ISSN 1098-6782

The Proceedings of the LFG'06 Conference

Universität Konstanz

Editors: Miriam Butt and Tracy Holloway King
2006 CSLI Publications
ISSN 1098-6782

Editors' Note

The program committee for LFG'06 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. This year one of the editors also had the role of local organizing committee, namely Miriam Butt, who is extremely thankful that the weather held, making the preconference excursion a very pleasant experience. We would like to thank all of the Konstanzer colleagues as well as a host of student assistants who helped with the conference organization, but in particular we would like to thank Ingrid Kaufmann, Carmen Kelling, Judith Meinschäfer, Bruce Mayo and most of all Zoltan Elfé, who proved to have a natural talent and by the end was running the entire conference.

Table of Contents

SPATIAL, TEMPORAL AND STRUCUTURAL USAGES OF URDU *KO*

Tafseer Ahmed
University of Konstanz

Proceedings of the LFG06 Conference
University of Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

Urdu is a New Indo-Aryan language which uses case markers to express differing semantic functions. The case marker *ko* marks accusative and dative. It is also used to express a few other spatial and temporal functions. We have studied a variety of semantic usages of *ko* and propose an unifying explantion of all the diverse usages. We assume that it originated as a spatial postposition from a Sanskrit locative. The non-spatial usages of *ko* can be explained in terms of extended meaning of its spatial origin, i.e. *ko* marks a location in a semantic field that is a spatial field by default, but can be thought as temporal or event field in a metaphorical or abstract way.

1. Introduction

Urdu-Hindi is a common term used to describe two closely related Indo Aryan languages i.e. persianzed Urdu and sanskritized Hindi spoken in Pakistan, India and many other countries. [1,2,3] We discuss different semantic usages of the case marker *ko*. We provide a history of case marking in Indo Aryan languages and try to propose a unifying explanation of all the semantic functions of Urdu-Hindi *ko*.

2. History of Case in Indo-Aryan

Old Indo-Aryan languages used morphological inflections to express case. For example, Sanskrit had eight cases whose names, in Latin grammatical terms, are: Nominative, Accusative, Dative, Ablative, Instrumental, Genitive, Locative and Vocative. These are expressed by inflections. In Middle Indo-Aryan (600 BC-1000 AD) almost all case inflections were lost.

New Indo-Aryan languages (1000 AD-present) devised a new method to mark cases. These languages mostly use clitics as case markers. The following table gives examples of different declinations of Sanskrit (an Old Indo-Aryan language) *deva*, meaning *god* (Blake 2001) and case markers of its Urdu-Hindi(a New Indo-Aryan language) equivalent *devtaa.*

Case	**Sanskrit (OIA)**	**Urdu (NIA)**
Nominative	devas	devtaa
Ergative	-	devtaa ne
Accusative	devam	devtaa ko
Dative	devaaya	devtaa ko
Instrumental	devena	devtaa se
Ablative	devaat	devtaa se
Locative	devasya	devtaa meN/par/tak
Genitive	Deve	devtaa kaa/kii/ke

It is an intersting exercise to try to establish the origin of the New Indo-Aryan case markers, especially *ko*. The present day clitics originated from Old Indo-Aryan nouns and verbs, and

[1] This research is supported by the DFG (Deutsche Forschungsgemeinschaft) via the SFB 471, Project A24.
[2] The author is thankful to Miriam Butt and Scott Grimm for their help in the analysis of data and pointing out the mistakes.

[3] Glosses used in this paper are: Acc=Accusative, Caus=Causative, Dat=Dative, Erg=Ergative, F=Feminine, Gen=Genitive, Inf=Infinitive, Inst=Instrument, Loc=Locative, M=Masculine, Obl=Oblique, Perc=Percative, Perf=Perfective, Pl=Plural, Pres=Present, Sg=Singular. For Urdu transcription, 'a', 'i' and 'u' are used for short vowels and 'aa', 'ii' and 'uu' are used for the long ones. 'ai' is used for open mid front unrounded vowel and 'ui' are for open mid back rounded vowel. Capital letters are used for retroflex consonants except capital 'S' which is used for voiceless palatal fricative. Capital 'N' used after a vowel shows nasalization. Small 'c' is used for voiceless alveolar affricate.

became postpositions and clitics during the passage of time.

According to Beames (1872), Urdu-Hindi *ko* originated from the Sanskrit noun *kaaksha* meaning 'armpit, side'. The locative of *kaaksha* is *kaakshe* which means 'in the armpit', 'at the side'. In Old Hindi, *kaaksha* became *kaakha*. Its accusative was *kaakham*. After a series of changes, it became *ko*. Beames lists early uses of *ko* to mark the recipient goal of ditransitive verbs like *give* and as an object marker of verbs like *seek*.

There seems to be a corelation between accusative/dative case marker and old Sanskrit locatives in Indo-Aryan languages. Sanskrit locative *kaakshe/kaakham* is supposed to be the origin of accusative/dative case markers of at least four other Indo-Aryan languages i.e. Sindhi (*khe)*, Siraiki (*koN*), Bengali (*ke)* and Oriya (*ku)*. Butt (2005) has pointed out that at least five other Indo-Aryan languages use words starting with l/n as accusative/dative case markers. i.e. Punjabi (*nuN*), Marathi (*laa*), Gujrati (*ne/neN*), Assamese (*ko/no*) and Napali (*laai*) These are supposed to be derived from Sanskrit locatives *laage* meaning 'stick' (Beames 1872). Butt (2005) working with Aditi Lahiri has also suggested that ergative *ne* can be related to *janniye* meaning 'for the sake of, because of'.

Few of the case markers of other (than accusative and dative) cases also have origin in locatives. For example, Urdu-Hindi and Punjabi ergative *ne* is possibly derived from the locative discussed above. The sindhi ablative *khaaN* is an oblique form of accusative/dative *khe,* derived from the Sanskrit locative *kaakshe* discussed above. The punjabi ablative is *koloN,* which can be assumed to be derived form of Punjabi word *kol,* meaning 'near'.

3. Usages of *ko*

Urdu-Hindi *ko* is widely discussed in the literature. Most of the authors have discussed accusative and dative usages of *ko*. The major issues discussed are the alternation of accusative and nominative case with objects and dative subjects. However, beyond these usages, *ko* has quite a few other functions in Urdu-Hindi. The following examples illustrate the distribution of *ko* as far as we have been able to determine.

(1) anjum=ne saddaf=ko dekhaa (Accusative Object)
Anjum.F.Sg=Erg Saddaf.F.Sg=Acc see.Perf.M.Sg
'Anjum saw Saddaf.'

(2) anjum=ne saddaf=ko haNsvaayaa (Accusative Causee)
Anjum.F.Sg=Erg Saddaf.F.Sg=Acc laugh.Caus.Perf.M.Sg
'Anjum caused Saddaf to laugh.'

(3) anjum=ne saddaf=ko ciTT^{h}ii dii (Dative Object)
anjum.F.Sg=Erg saddaf.F.Sg=Dat letter.F.Sg give.Perf.F.Sg
'Anjum gave the letter to Saddaf.'

(4) omair=ko iinaam milaa (Dative Subject)
Omair.M.Sg=Dat prize.M.Sg touch.Perf.M.Sg
'Omair got the prize.'

(5) jin=ko caSm-e-biina hai (Dative Subject)
who=Dat visionary-eye.M.Sg be.Pres.Sg
'who have vision'

(6) nadya=ko zu jaanaa paRaa (Dative Subject)
Nadya.F.Sg=Dat zoo go.Inf fall-on.Pres.Sg
'Nadya have to go to the zoo.'

(7) omair sair=ko gayaa (Adverb)
Omair.M.Sg walk.M.Sg=to go.Perf.M.Sg
'Omair went for a walk.'

(8) nadya caae banaane=ko hai (*ko* marked clause)
Nadya.F.Sg tea.F.Sg be.inf.obl=at be.pres.Sg
'Nadya is about to make tea.' (Lit:'Nadya is at tea making')

(9) saamaan g^{h}ar=ko pohoanch gayaa (Spatial Adverb)
Luggage.M.Sg home.M.Sg=at reach go.Perf.M.Sg
'The luggage reached home.'

(10) chor raat=ko aayaa (Temporal Adverb)
thief.M.Sg night.F.Sg=at come.Perf.M.Sg
'The thief came at night.'

4. Analysis of Usages of *ko*

In the previous section, we have seen differing usages of *ko*. These usages, at first glance, seem semantically unrelated, but we can try to find a unified explanation of all these usages. Most of them can be explained by assuming that the primary usage of *ko* is locative. The seemingly different usages are due to different semantic fields, e.g. temporal and eventual, in addition to the spatial field. The primary locative semantic thesis is supported by the fact that *ko* originated from a Sanskrit locative meaning 'in the armpit' or 'at the side'.

The following discussion on the different semantic usages of *ko* provides arguments to support the proposal.

4.1. Basic Spatial Usage

As shown in (11)-(13), *ko* is used to mark locative adjuncts to give semantics of 'towards the point' and 'at the point'.

(11) saamaan ghar=ko pohoanch gayaa
Luggage.M.Sg home.M.Sg=at reach go.Perf.M.Sg
'The luggage reached (at) home.'

(12) nadya laahaur=ko gayii
Nadya.F.Sg lahore.M.Sg=at go.Perf.F.Sg
'Nadya went to lahore.'

(13) kamraa andar=ko hai.
room.M.Sg inside.M.Sg=at be.Pres.M.Sg
'The room is (towards) inside.'

ko in this usage has the sense of endpoint. All the other usages of *ko* can be explained as extended usages of the spatial meaning.

4.2. Extension to Other Domains

Many of the case markers came from spatial nouns and verbs. These case markers have a core locative meaning that is extended to a different semantic domain. Before discussing *ko*, we take the example of possession and part-whole relationship that is expressed by a spatial preposition. Being present at some location is considered as possession. In English, possession can be expressed by the locative preposition "in" and the verb "be":

(14) There are two windows in the room. (English)

Similarly in Urdu-Hindi, possession can be expressed with locative postposition *paas* meaning 'near'. This is shown in (15).

(15) sadiq=ke paas aik kitaab hai.
Sadiq.M.Sg=Gen near one book.F.Sg be.Pres.Sg
'Sadiq have a book' (Lit:'Near Sadiq, is a books.')

Urdu also has a locative usage of this postposition *paas* that gives its literal meaning *near*.

(16) daryaa=ke paas aik iimaarat hai.
river.M.Sg=Gen near one building.F.Sg be.Pres.Sg
'There is a building near the river.'

Similarly, *ko* has many extended usages apart from the core locative one. Mohanan (1994) suggested that accusative, dative and locative *ko* has the same semantic configuration but different semantic fields. Croft (1991) surveyed case markers of 40 languages and observed that in many languages, ablative forms are used for antecedent oblique functions (causer, instrument etc.) and allative forms are used for subsequent oblique functions (recipient, beneficiary etc.). Differing semantic usages of Urdu *ko* is an example of locative goal used for subsequent functions.
In the next sections, we will explain the (extended) usage of *ko* to mark endpoint in temporal, mental and eventual domains.

4.3. Extension to Temporal Domain

ko is used to mark a point of time e.g. day of the week, or part of the day. This usage is shown in (17) and (18).

(17) cor mangal=ko aayaa.
thief.M.Sg Tuesday.M.Sg=at come.Perf.M.Sg
'The Thief came on tueday.'

(18) cor raat=ko aayaa
thief.M.Sg night.F.Sg=at come.Perf.M.Sg
'The thief came at night.'

In this usage, the semantic feature of *ko* is a point in temporal semantic field (in place of an endpoint in spatial field). The part of the day usage can alter with locative postposition *meN* meaning 'in'. Compare the following sentence with (18).

(19) chor rat=meN aayaa
thief.M.Sg night.F.Sg=Loc-in come.Perf.M.Sg
'The thief came during/at night.'

4.4. Extension to Causal Domain

When *ko* marks an argument of argument structure, the endpoint semantics is extended to the causal domain. *ko* marks the arguments that receives something either physical or abstract.

4.4.1. Dative Subject

The core endpoint semantics of *ko* is extended to the recipient when it marks a participant of argument structure. In (20) and (21), *ko* marks the indirect objects of ditransitive verbs.

(20) anjum=ne saddaf=ko ciTT^{h}i dii
Anjum.F.Sg=Erg Saddaf.F.Sg=Dat letter.F.Sg give.Perf.F.Sg
'Anjum gave the letter to Saddaf.'

(21) anjum=ne saddaf=ko ciTT^{h}i likhii
Anjum.F.Sg=Erg Saddaf.F.Sg=Dat letter.F.Sg write.Perf.F.Sg
'Anjum wrote a letter to Saddaf.'

In (20), the *letter* reaches the indirect object *Saddaf* marked with *ko*. In (21), she is the intended goal of the object *letter*.

In these examples, *ko* is marking a recipient. According to Grimm (p.c.), who has decomposed thematic roles into basic semantic properties (Grimm 2005), *ko* has the semantic features of a Canonical Recipient. Recipients are *sentient.* They undergo a *qualitative change* relative to the state of affairs before the onset of the event (i.e., come into possession of somebody) and they are the *endpoint* of the transfer event, i.e., a direct action. *Volitionality* (whether a recipient desires the event to occur or not) is left underspecified for Urdu-Hindi *ko*. We can say that the recipient is a location which is the goal or destination of the object.

Indirect Objects are not the only example of dative recipients. Dative Subjects involve receiving of both physical and abstract objects. In (22) and (23), Dative Subject is receiving physical and event nominal objects.

(22) omair=ko inaam milaa.
Omair.M.Sg=Dat prize.M.Sg touch.Perf.M.Sg
'Omair got the prize.'

(23) omair=ko thapaR/ghuuNsaa paRaa.
Omair.M.Sg.dat slap/punch.M.Sg fall-on.Perf.M.Sg
'Omair received a slap/punch.' (Lit: To Omair, salp/punch fell on.')

Urdu-Hindi usually has a nominative case or ergative case marker on the subject. In (24), verb *milnaa* meaning 'touch' or 'meet' is used with non-sentient nominative subject and non-sentient dative object.

(24) daryaa samandar=ko milaa.
river.M.Sg sea.M.Sg=Dat touch.Perf.M.Sg
'The river met/touched the sea.'

This traditional or canonical configuration changes, if the recipient is sentient. Sentences (13), (22) and (23) having the same verb *milnaa* show a reanalysis of the construction in which the sentient recipient becomes subject. The processing pressure in the human mind favors the subjecthood of the sentient i.e. human argument (Butt, Grimm and Ahmed 2006).

Dative Subject constructions have few other semantic usages. We will explain these usages as the (sentient) recipient receiving abstract pysch experiences in the next section (4.5).

4.4.2. Affected Agents (of Causatives)

The recipient semantics of *ko* can also be seen in Urdu-Hindi causatives. Saksena (1982) in her work on causatives introduced the concept of *affected agents*. Affected agents are subjects of intransitive and ingestive transitive verbs. Verb *parhnaa* meaning 'read/learn' can have affetced agent as shown in (25).

(25) saddaf=ne sabaq paR^{h}a
Saddaf.F.Sg=Erg lesson.M.Sg learn.Perf.M.Sg
'Saddaf learnt the lesson'.

These subjects are affected by the action. We can also say that these are the recipient of the action. The affected-agent is marked with *ko* in (26) which is the causative of the above sentence. The syntax is similar to the indirect object of a ditransitive verb. i.e. *ko* is signaling the receiving

of the lesson.

(26) anjum=ne ustaad=se saddaf=ko sabaq paR^{h}vaayaa
Anjum.F.Sg=Erg teacher.M.Sg=Inst Saddaf.F.Sg=Dat lesson.M.Sg teach.caus.Perf.M.Sg
'Anjum caused the teacher to teach the lesson to Saddaf.'

Other verbs having an unaffected agent do not allow *ko* with the causee, as that argument is not a recipient of the action. *paR^{h}naa* and few other verbs allow both affected and un-affected agents. The subject in (27) is an unaffected agent.

(27) saddaf=ne xabreN paR^{h}iiN
Saddaf.F.Sg=Erg news.F.Pl read.Perf.F.Pl
'Saddaf read the news.'

(28) anjum=ne saddaf=se/ko* (tv=par) xabreN paR^{h}vaaiiN
Anjum.F.Sg=Erg Saddaf.F.Sg=Inst/Dat* tv=Loc-on news.F.Sg read.caus.Perf.F
'Anjum caused Saddaf to read the news (on TV).'

In (28) which is the causative counterpart of (27), a causee with *ko* is not possible, because news reading is not an event of receiving. In its place, instrumental case marker *se* representing the source is used. The usage of *ko* for affected i.e. receiving agent in causatives is another example of goal and endpoint semantics of *ko*.

4.5. Extension to Mental Domain

In 4.4.1, we have seen the usage of *ko* to mark dative recipient that (usually) receives a physical object. This dative usage of *ko* is extended to the mental domain where the sentient agent receives an experience. Semantic properties like experience, (mental) state and involition are attached to these constructions. These extended usages can be explained as a metaphorical extension of the recipient semantics discussed above.

4.5.1. Experience

Dative Subject constructions are used with psych verbs and to express experience. The following examples are similar to (22) and (23), but here the received object is an experience.

(29) omair=ko xabar milii
Omair.M.Sg=Dat news.F.Sg touch.Perf.F.Sg
'Omair got the news.'

(30) omair=ko bhuuk lagii
Omair.M.Sg=Dat hunger.F.Sg stick.Perf.F.Sg
'Omair felt hungry.' (Lit: ' To Omair, Hunger came.')

Among these, Landau (2005) proposes that experiencers are (mental) locations and that an experiencer of a pysch-predicates is a locative of some sort. The reception semantics can be extended to give the notion of experience with human mind as goal, i.e. the human (mind) is the location of the experience.

The dative subject used with verb *hona* 'be' expresses experience (mental) states.

(31) sadiq=ko xushi hai
Sadiq.M.Sg=Dat happiness.M.Sg be.Pres.Sg
'Sadiq is happy.' (Lit: ' To Sadiq, is happiness.')

(32) sadiq=ko buxaar hai
Sadiq.M.Sg=Dat fever.M.Sg be.Pres.Sg
'Sadiq has fever.' (Lit: 'To Sadiq, is the fever.')

One can claim that in the above examples, the subject *Sadiq* is merely the location of the happiness or fever and it does not seem to resemble a recipient or goal. We cannot make a strong point in favor of recipient from examples of Urdu. But we can find help from another Indo-Aryan language Marathi. In Marathi, the dative case marker of subjects alternates with locative markers to give the meaning of non-integral and integral part respectively (Pandharipande 1990).

(33) tyala himmat ahe (Marathi)
3P.M.Sg.Dat courage.Sg be.Pres
'He has courage.' (Courage is non-integral-part/temporary-quality of him.)

(34) tyacyat himmat ahe (Marathi)
3P.M.Sg.Loc courage.Sg be.Pres
'He has courage.' (Courage is integral-part/permanent-quality of him.)

Pandharipande suggested that the Marathi Dative NP construction is spatial. In it, the dative marks a recipient that does not have the property for eternity, but received it at some point of time.

We can assume that Urdu counter-part of this dative construction has similar i.e. recipient or non-integeral part semantics. Even, if we disagree with this argument, then the Dative Subject with *hona* meaning 'be' verb still can be related with "point" feature i.e. the dative subject is a metaphorical point where the experience is located.

4.5.2. Volition

We have discussed in 4.4.1 that dative *ko* of Urdu is underspecified for the volitionality of the recipient. But, we find constructions with recipient *ko* and non-finite verb that exposes involition of the subject. Butt and King (1991) discussed an alternation of ergative and dative case markers in Lahori Urdu as.

(35) nadya=ne zu jaanaa hai.
Nadya.F.Sg=Erg zoo.M.Sg go.Inf.M.Sg be.Pres.3.Sg
'Nadya wants to go to the zoo.'

(36) nadya=ko zu jaanaa hai.
Nadya.F.Sg=Dat zoo.M.Sg go.Inf.M.Sg be.Pres.3.Sg
'Nadya has to go to the zoo.'

For two of the above sentences, only (36) is supposed to be grammatically correct, traditionally. But in modern Urdu-Hindi, ergative case marker is alternating or replacing the traditional use of dative case marker in this construction.[4]

Butt and King (1991) and Mohanan (1994) have argued that the ergative is associated with volitionality or the feature [+conscious choice]. Butt (2005) argued that one can receive both pleasant or unpleasant objects/events. This can be seen in (37) in which getting cold is unpleasent an involitionary event.

---

[4] Bashir(1999) studied Urdu TV dramas and found following examples of ergative marker with non-finite verb.

meN=ne Dinar=pe jaanaa t^{h}aa.
1P.Sg=erg dinner.M.Sg=loc-on go.inf.M.Sg be.past.3.M.Sg
'I was supposed to go to the dinner' (PTV drama *Tanhayian*)

aap=ne .... koii aisii baat nahiin puucnii.
2P.Sg=erg any such matter.F.Sg not ask.Perf.F.Sg
'You won't ask (me) anything like this' (PTV drama *Aanch*)

(37) Nadya got a cold/prize. (English)

Similarly in (36), one can not know whether *Nadya* likes to receive the zoo going event or not. It is underspecified for volition.

As Urdu-Hindi case marker *ne* has agentive reading, it is used to introduce volition or conscious choice, as in (35). As *ko* construction is alternating with it, it seems to contrast with *ne* to express involition and [-concious choice] as in (36) in contrast to (35).

Constructions having verb *paRnaa* with nonfinite verb also gives the meaning of involition. It is shown in (38).

(38) omair=ko zu jaanaa paRaa
Omair.M.Sg=Dat zoo go.Inf fall-on.Perf.M.Sg
'Omair had to go to the zoo.' (Lit: 'To go to the zoo, fell on to Omair.')

This construction seems to be metaphorical extension of (23). Dative *ko* is underspecified for volition in this construction. The semantics of the verb provides the involition, as an event is "falling" on the subject. The sudden reception of the event cannot be avoided and subject receives it involitionally. Hence, the construction is interpreted as being internally involitional.

4.6. Extension to Event arguments

4.6.1. *Purpose*

ko is used with clausal adjuncts to express purpose/reason of the action. It can be seen in (39).

(39) log sair/faryaad/ayaadat=ko gaae
People.M.Pl walk/complaint/visiting-sick-person=at go.Perf.M.Pl
'People went for a walk/complaint/visiting-sick-person.'

In the same construction, *ko* can also be used with an infinitival verb phrase.

(40) log Tehelne=ko gaae
People.M.Pl walk.Inf=at go.Perf.M.Sg
'People went for a walk.'

This usage is similar to the real spatial usage discussed above. The spatial domain provides a metaphor in which subject is not traveling towards a location but towards an event. This metaphorical location (event) is marked with *ko*. The semantic feature of this usage is the same as above i.e. the (metaphorical) location is an endpoint of the event.

4.6.2. Immediate Future

An interesting usage of *ko* is to express immediate future. In this construction, *ko* expresses the beginning of work in immediate future. This is shown in (41).

(41) nadya caae banaane=ko hai
Nadya.F.Sg tea.F.Sg make.Inf.Obl=at be.pres
'Nadya will make tea(in immediate future)'(Lit:'Nadya is at the act of tea making')

This usage has the semantic feature of endpoint. Metaphorically, the subject is very near to the event marked with *ko*. Here, *ko* has the semantics of very near or almost there. Hence, *ko* provides a reading of immediate future to this sentence.

4.7. Unexplained Usages

We have described a unified locative explanation of different usage of Urdu-Hindi *ko* above. There are two semantic usages that are not completely explained under the properties taken here.

4.7.1. Modal Cahiiye

We have discussed dative recipient and its extended usages in 5.1 and 5.2. Another example of extended dative usage is a construction that shows need or obligation.

(42) nadya=ko ye kitaab cahiiye.
Nadya.F.Sg=Dat this book.F.Sg want.perc
'Nadya need this book.'

(43) baccoN=ko baRoN=ka adab karnaa cahiiye.
Child.Pl=Dat Elder.Pl.Gen respect.M.Sg do.inf want.perc
'Children should respect the elders.'

Cahiiye is the percative form of verb *cahna* meaning 'want'. Percative forms are usually used in imperative sentences with nomimative subject (Platts 1909). But in (42) and (43), *cahiiye* is used as a modal. In these sentences, the combination of *ko* and *cahiiye* gives sense of need or obligation. As *ko* is underspecified for volition and Urdu-Hindi modals usually have different meanings than their main verb counterparts, we can assume that modal *cahiiye* is giving the feature of need or obligation in this construction.

4.7.2. Accusative ko

An important usage of *ko* is that it acts as an accusative case marker. Accusative *ko* is connected with a sensitivity to animacy and definite/specific interpretations. It can be seen in (44) and (45).

(44) anjum=ne saddaf=ko dekhaa
Anjum.F.Sg=Erg Saddaf.F.Sg=Acc see.Perf.M.Sg
'Anjum saw Saddaf.'

(45) anjum=ne kashtii dekhii
Anjum.F.Sg=Erg boat.F.Sg see.Perf.F.Sg
'Anjum saw a/the boat.'

In (44), the object *kashtii* meaning *boat* is neither animate nor specific, hence it is in nominative case. Allen (1951), McGregor (1972), Masica (1991), Butt (1993), Mohanan (1994) and Singh (1994) among others have discussed this issue in detail.

It is not immediately apparent that this use of ko could be connected to a spatial use. However, Mohanan (1994) has argued that the accusative is used for logical objects towards which an action or event is directed. That is, it can again be seen to mark the endpoint or goal of an action.

Boundedness is another way of analysing the accusative *ko*. The nominative object gets incoporated with the verb. It, like mass nouns, does not bound the event. While accussative *ko* marked objects bound the event or the object is end point of the event. So the specific objects put a bound on the event.

5. Case Markers/Postpositions alternating with *ko*

We have discussed locative semantics of Urdu-Hindi *ko*. We have also seen the examples in which *ko* alternates with locative case markers and postpositions. We have also seen the alternation of ergative *ne* with dative *ko* for volition and conscious choice. Two other case markers either replace or alternate with Urdu-Hindi *ko*.

5.1. Instrumental

A few verbs like *milnaa* and *kehnaa* have noun phrases marked by the instrumental/ablative case marker *se*. But in old texts, we can find examples having *ko* marking for these noun phrases. For example, the following sentence is taken from an old text (Online Urdu Dictionary, Beta version).

(46) buR^{h}aa baap **beTi=ko** milnaa caahtaa hai
Old.M.Sg father.M.Sg daughter.F.Sg=Dat meet.Inf want be.Pres.Sg
'Old father wants to meet the daughter.'

In current usage, this sentence is as in (50):

(47) buR^{h}aa baap **beTi=se** milnaa caahtaa hai
Old.M.Sg father.M.Sg daughter.F.Sg=Inst meet.Inf want be.Pres.Sg
'Old father wants to meet the daughter.'

The reason for the change of case marker is the change in semantics of the verb. *Milna* literally means 'touch' as in (24). The sentence having *ko* (46) gives the sense of a visit, when the father moves and visited the daughter. The other sentence having *se* gives sense of an interactive meeting in which both arguments are particiapating. Another example of this replacement/alternation is:

(48) ali=ne **beToN=ko** kahaa
Ali.M.Sg=Erg son.M.Pl=Dat say.Perf.M.Sg
'Ali said to the sons.'

(49) ali=ne **beToN=se** kahaa
Ali.M.Sg=Erg son.M.Pl=Inst say.Perf.M.Sg
'Ali said to the sons.'

Sentence (48) is taken from examples of old Urdu texts in Beg (1998), while (49) is more widely used today.

5.2. ke-liye(Purpose)

The postposition *(ke) liye* can be used in place of *ko*. It is shown in (50) and (51).

(50) anjum sair=ko gaaii
Anjum.F.Sg walk.F.Sg=at go.Perf.F.Sg
'Anjum went for a walk.'

(51) anjum sair=ke liye gaaii
Anjum.F.Sg walk.F.Sg=gen for go.Perf.F.Sg
'Anjum went for a walk.'

Both of the above two sentences are semantically equivalent that can be used interchangeably. Similarly, all three of the following sentences means 'Anjum asked Saddaf to come'.

(52) anjum=ne saddaf=se aane=ko kahaa
Anjum.F.Sg=Erg Saddaf.F.Sg=Inst come.Inf.Obl=at say.Perf.M.Sg

(53) anjum=ne saddaf=se aane=ke liye kahaa
Anjum.F.Sg=Erg Saddaf.F.Sg=Inst come.Inf.Obl=gen for say.Perf.M.Sg

(54) anjum=ne saddaf=se aane=ka kahaa
Anjum.F.Sg=Erg Saddaf.F.Sg=Inst come.Inf.Obl=Gen say.Perf.M.Sg

But *(ke) liye* and *ko* are not replaceable in all usages. For example, the following sentence with a beneficiary marked with *(ke) liye* can not have *ko* in its place.

(55) anjum=ne saddaf=ke liye gaaRi xariidii
Anjum.F.Sg=Erg Saddaf.F.Sg=Gen for car.F.Sg buy.Perf.F.Sg
'Anjum bought a car for Saddaf.'

What is the reason of overlapping semantic usages of these two case markers in (50)-(51) and (52)-(54)? Dative markers usually mark both goals and beneficiaries. *ko* marks the goal and

(optionally) some of the beneficiary usages. *(ke) liye* is the marker that marks all the beneficiary usages. Does *(ke) liye* replaced beneficiary usages of dative *ko*? This remains subject to further investigation.

6. Summary/Conclusion

We have analyzed different semantic usages of Urdu-Hindi *ko* that includes accusative object, dative subject and purpose of an event etc. These seemingly diverse usages can be connected to a core locative meaning. The locative usage has expanded towards other usages by involving different semantic fields. Through analysis of the differing semantic usages, we found the following three main usages of *ko*:

- Point in space as in temporal usage.
- Non-sentient endpoint in space as in spatial, purpose and immediate future usages.
- Sentient recipient as in dative and its extended usages.

It can be speculated that *ko* has entered in the language as a marker of endpoint or goal and after some time, it started marking other usages too. Further analysis of diachronic data remains to be conducted to confirm or reject this hypothesis.

7. References

Allen, W.S. 1951. A Study in the Analysis of Hindi Sentence-Structure. *Acta Linguistica Hafniensia*.

Bashir, Elena. 1999. The Urdu and Hindi Ergative Postposition *ne*: Its changing role in the Grammar. In *The Yearbook of South Asian Languages and Linguistics*, ed. Rajendra Singh. New Delhi: Sage Publications.

Beames, John. 1872–79. *A Comparative Grammar of the Modern Aryan Languages of India*. Delhi: Munshiram Manoharlal. Republished 1966.

Beg, Mirza Khalil A. 1988. *Urdu Grammar: History and Structure*. New Delhi: Bahri Publications.

Blake, Barry. 2001. *Case*. Cambridge: Cambridge University Press. Second Edition.

Butt, Miriam, and Tracy Holloway King. 1991. Semantic Case in Urdu. *In Papers from the 27th Regional Meeting of the Chicago Linguistic Society*, ed. Lisa Dobrin, Lynn Nichols, and Rosa M. Rodriguez, 31–45.

Butt, Miriam. 1993. Object Specificity and Agreement in Hindi/Urdu. In *Papers from the 29th Regional Meeting of the Chicgo Linguistic Society*, 80–103.

Butt, Miriam, and Tracy Holloway King. 2005. The Status of Case. In *Clause Structure in South Asian Languages*, ed. Veneeta Dayal and Anoop Mahajan. Berlin: Kluwer Academic Publishers.

Butt, Miraim. 2005. The Dative-Ergative Connection, In Patricia Cabredo-Hofherr (ed.) *Proceedings of the Colloque Syntax-Semantique Paris (CSSP) 2005*.

Butt, Miriam , Scott Grimm and Tafseer Ahmed. 2006. Dative Subjects. Presntation at *NWO/DFG Workshop on Optimal Sentence Processing*, Nijmegen. http://ling.uni-konstanz.de/pages/home/tafseer/usages%20of%20ko.pdf

Croft, William. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Chicago: University of Chicago Press

Grimm, Scott. 2005. *The Lattice of Case and Agency*. MSc Thesis, Universiteit van Amsterdam.

Landau, Idan. 2005. *The Locative Syntax of Experiencers*. Ms.http://www.bgu.ac.il/~idanl/files/psych.July05.pdf

McGregor, R.S. 1972. *Outline of Hindi Grammar: With Exercises*. Oxford: Clarendon Press.

Mohanan, Tara. 1994. *Argument Structure in Hindi*. Stanford, CA: CSLI Publications.

Pandharipande, R. 1990. Experiencer (Dative) NPs in Marathi, In M. K. Verma and K. P. Mohanan, eds., *Experiencer Subjects in South Asian Languages*, CSLI, Stanford, CA, 161–180.

Platts, John T. 1909. *A Grammar of the Hindustani or Urdu Language*, Crosby Lockwood and Son, London. republished 2002. Sang-meel Publications, Lahore.

Singh, Mona. 1994. *Perfectivity, Definiteness and Specifity: A Classification of Verbal Predicates in Hindi*. Doctoral dissertation, The University of Texas at Austin.

Saksena, Anuradaha. 1982. *Topics in the Analysis of Causatives with an Account of Hindi Paradigms*, University of California Press.

EXPLETIVES AND THE SYNTAX AND SEMANTICS OF COPY RAISING

Ash Asudeh and Ida Toivonen
Institute of Cognitive Science & School of Linguistics and Applied Language Studies
Carleton University

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract
We present an event semantics account of copy raising in Swedish and English. The examination of copy raising gives rise to two puzzles. We demonstrate that our event semantics analysis solves the two puzzles. We examine some challenging copy raising data from expletives, propose a solution for handling the data, and discuss consequences of the solution for the theory of expletives.

1 Introduction

Copy raising (CR) in English is demonstrated in (1) and (2):

(1) They seem like they've missed the bus.

(2) John appears as if he is tired.

This can be compared to 'standard' raising as in:

(3) They seem to have missed the bus.

(4) John appears to be tired.

Copy raising can be characterized schematically as follows:

(5) They seem like they've missed the bus.
Subject + appear/seem + like/as if/as though + finite clause containing a pronominal copy of the subject

Subject-to-subject raising from an infinitival — as in (3) and (4) — has been studied extensively in the syntactic literature. In comparison, CR is relatively unexplored, although it has been discussed somewhat, and for a variety of languages (English: Rogers (1971, 1973), Postal (1974), Potsdam and Runner (2001), Asudeh (2002, 2004), Fujii (2005); Modern Greek: Joseph (1976), Perlmutter and Soames (1979); Samoan: Chung (1978); Hebrew: Lappin (1984); Irish: McCloskey and Sells (1988); Haitian Creole: Déprez (1992); Igbo: Ura (1998); Turkish: Moore (1998); Polinsky and Potsdam (2006) discuss further languages).

A key challenge presented by copy raising is that (pre-theoretically) a single thematic role apparently corresponds to two different NPs: the CR subject and the copy pronoun.

(6) John seems like he is sleeping.

Examples such as (6) can alternate with expletive examples such as (7), just as in standard raising (8–9). These alternations are indicative of a lack of a subject thematic role.

(7) It seems like John is sleeping.

(8) John seems to be sleeping.

(9) It seems that John is sleeping.

Potsdam and Runner (2001) present evidence for the athematic status of the copy raising subject; further discussion can be found in Asudeh and Toivonen (2006b).

A second key challenge of copy raising is the obligatory presence of the copy pronoun in the complement:

(10) Jody seems like she's tired.

(11) Jody seems like her favorite show has been cancelled.

(12) *Jody seems like it's raining.

Swedish copy raising largely parallels English copy raising:

(13) Maria verkar som om hon har vunnit.
M. seems as if she has won
Subject + verka ('seem') + som om (as if) + finite clause containing a pron. copy of the subject
Maria seems like she's won.

As indicated in the gloss, the general form of a basic copy raising sentence is essentially identical to the English. The similarity continues with the obligatoriness of the copy pronoun in Swedish:

(14) * Maria verkar som om Pelle har vunnit.
M seems as if P has won

Again as in English, there is an expletive alternant of (13):

(15) Det verkar som om Maria har vunnit.
it seems as if M has won
It seems as if Maria has won.

We show elsewhere that the copy raising subject in Swedish is likewise athematic (Asudeh and Toivonen 2006b).

However, unlike English, Swedish allows specification of the source of the impression (i.e., the percept) in a copy raising sentence in an adjunct PP headed by *på* ('on') (see Asudeh and Toivonen 2006b for arguments that the PP is an adjunct):

(16) Det verkar på Elin som om Maria har vunnit.
it seems on E as if M has won
∼*Elin gives the impression that Maria has won.*

Notice that there is no equivalent of the '*på*-PP' in English, although English has the capacity to express the other part of the perceptual relation, the perceiver, in a *to*-PP:

(17) Maria seemed to me like she had won.

Swedish only marginally allows expression of the perceiver (Asudeh and Toivonen 2006b). There is thus an asymmetry between English and Swedish with respect to expression of the arguments of the perceptual relation associated with copy raising.

The adjunct *på*-PP gives rise to a puzzle that we have elsewhere discussed as 'the *på* puzzle' (Asudeh and Toivonen 2006a). The puzzle is demonstrated by the following set of sentences, the first two of which are repeated from (13) and (16) above:

(18) Maria verkar som om hon har vunnit.
M seems as if she has won
Maria seems as if she has won.

(19) Det verkar på Elin som om Maria har vunnit.
it seems on E as if M has won
∼*Elin gives the impression that Maria has won.*

(20) * Maria verkar på Elin som om hon har vunnit.
M seems on E as if she has won

The *på* puzzle is this: Why is copy raising incompatible with a *på*-PP? In particular, why can't (20) mean that Elin gives the impression that Maria gives the impression that she (Maria) has won? This is a perfectly sensible proposition, but (20) can't mean this; it is instead ungrammatical.

In Asudeh and Toivonen (2006a) we discuss another puzzle, which arises equally in English and Swedish and which we call 'the puzzle of the absent cook'. In the following scenario, where the cook is present, we see the pattern of grammaticality demonstrated in (21–23):

Scenario: You and your friend walk into John's house. You see John busy cooking in his kitchen.

(21) It seems like/that John is cooking

(22) John seems to be cooking

(23) John seems like he's cooking.

The puzzle arises when the cook is absent from the scenario:

Scenario: you and your friend walk into John's kitchen. There are pots and pans on the stove. It smells like food. It's obvious that someone is cooking. John is not in the kitchen.

(24) It seems like/that John is cooking.

(25) John seems to be cooking.

(26) */# John seems like he's cooking.

In this scenario, we see a shift in the pattern of grammaticality: the copy raising sentence (23)/(26) is now unacceptable (we leave aside for now whether this is true ungrammaticality, normally indicated by *, or semantic/pragmatic unacceptability, typically indicated with #).

The solution to both puzzles hinges on the following assumption:

(27) The copy raising subject is interpreted as the *perceptual source* (*Psource*), what gives the impression that the complement to the copy raising verb is the case.

The subject in a sentence like (26) is thus the Psource and (26) means 'It seems like John is cooking and this impression comes from John'. Sentence (26) is therefore inappropriate in an absent cook scenario where John is unavailable to give such an impression (we treat this as presupposition failure).

The postulation of a Psource similarly explains the *på* puzzle. Like the copy raising subject, the *på*-PP expresses the perceptual source. A Psource PP is incompatible with a Psource subject, due to a generalized uniqueness condition on participants in eventualities.[1] We outline this uniqueness condition and other aspects of our analysis in the next section and show how it solves the two puzzles. Then, in section 3, we look at certain theoretical challenges from expletive data in copy raising and various consequences for the theory of expletives.

2 A sketch of the analysis

The analysis we present here is based on Asudeh (2004) and Asudeh and Toivonen (2006a); many aspects of the analysis are articulated more fully in Asudeh and Toivonen (2006b). Asudeh (2004) argues that *like* and *as* in copy raising sentences are not complementizers, but are rather prepositions with clausal complements (also see Heycock 1994 and Potsdam and Runner 2001). Asudeh (2004) further argues that the subject of the *like/as*-complement to the copy raising verb is raised as the subject of the copy raising verb, using the

[1] Note that this cannot be due to the theta-criterion or the equivalent, if the copy raising subject is athematic; see Asudeh and Toivonen (2006b) for extensive discussion of this issue.

usual raising mechanism of functional control in LFG. In other words, the *like/as*-complement is treated as a predicative complement on Asudeh's analysis, which assimilates the copy raising complements to the general class of predicative raising complements:

(28) John seems/appears upset/out of his mind.

The f-structure for sentence copy raising, as in (29), is shown schematically in (30):

(29) John seems like he is upset.

(30)
$$\begin{bmatrix} \text{PRED} & \text{`SEEM/APPEAR}\langle\text{XCOMP}\rangle\text{SUBJ'} \\ \text{SUBJ} & \boxed{1}\ [\ldots] \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{`LIKE/AS}\langle\text{COMP, SUBJ}\rangle\text{'} \\ \text{SUBJ} & \boxed{1} \\ \text{COMP} & [\ldots] \end{bmatrix} \end{bmatrix}$$

Notice that the functionally controlled, raised subject is not the copy pronoun, but rather the subject of the predicative complement headed by the *like/as* preposition. The copy pronoun itself is somewhere inside the COMP of the predicative complement. The copy raising subject is related to the copy pronoun by a separate, anaphoric binding relation (Asudeh 2004).

In Asudeh's theory — which treats copy raising as a kind of resource surplus, like resumption, in Glue Semantics (Dalrymple 1999) — the copy pronoun is removed from semantic composition by a *manager resource*. The manager resource is lexically specified by the copy raising verb (*seem, appear*). The following schematic Glue proof illustrates the analysis:

(31)
$$\dfrac{A \quad \dfrac{A \multimap (A \otimes P) \qquad [A \multimap (A \otimes P)] \multimap (A \multimap A)}{A \multimap A}\ \multimap_{\mathcal{E}}}{A}\ \multimap_{\mathcal{E}}$$

Antecedent, *Pronoun*, *Manager resource* — ***Lexical contributions***

Manager resource removes pronoun

Result of pronoun removal combines with antecedent; final result is just antecedent

The manager resource's removal of the pronoun ensures that the athematic copy raising subject has a place to compose with the semantics, since it is not a semantic argument of its matrix predicate. Kokkonidis (2006) presents an alternative resource management theory for Glue Semantics, but the differences do not affect the case at hand.

Asudeh and Toivonen (2006a,b) propose an event semantics for copy raising. Copy raising verbs lexically contribute a Psource semantic role (note that we follow Bach 1981 in using 'eventuality' as a cover term for events and states):

(32) The Psource of an eventuality E is the source of perception of E (whatever gives the impression that E holds).

We below argue that other subcategorizations of raising verbs involve existential closure of the Psource (see also Asudeh and Toivonen 2006a,b).

In Asudeh and Toivonen (2006b), we argue at length that Psource is not a thematic role in the usual narrow sense, but is a *semantic role* (roughly analogous to the *thematic relations* of Parsons 1990, 1995). We will not rehearse those arguments here, but note that they essentially depend on the demonstration that 1) the copy raising subject is not a thematic argument and 2) the *på*-PP is an adjunct; thus, the realizations of Psource are athematic and Psource therefore cannot be a thematic role. We treat Psource as a function from eventualities to individuals or eventualities. This is analogous to the thematic role STIMULUS in (33) and (34):

(33) Meg saw Jack.

(34) Meg saw Jack running.

In (33), STIMULUS is a function from the seeing eventuality to the individual Jack. In (34), STIMULUS is a function from the seeing eventuality to the eventuality of Jack running.

It has been noted in the semantics literature on thematic roles that each eventuality may have only one instance of a given thematic role (Carlson 1984, Chierchia 1984, Landman 2000). Landman (2000) formulates this as the 'Unique Role Requirement':

(35) **Unique Role Requirement**
If a thematic role is specified for an event, it is uniquely specified.

Landman (2000) captures this formally by treating thematic roles as partial functions on eventualities, as anticipated above. If a thematic role is a function on its eventuality, it follows that each eventuality can have at most one instance of any thematic role. We generalize this functional definition to the Psource *semantic* role, which similarly captures this uniqueness requirement for Psource: each eventuality can only have one Psource.

We close this section with a presentation of the relevant semantic part of the lexical entry for copy raising verbs and a couple of examples of the semantics of copy raising (for more details see Asudeh and Toivonen 2006b). The Glue Semantics meaning term for a copy raising verb (leaving aside the manager resource discussed above) can be sketched in simplified form as shown in (36), where the linear logic terms are instantiated as per the f-structure (38) for sentence (37) (note that e is the event variable):

(36) $\lambda x \lambda P \lambda s.seem(s, P(x)) \wedge \textsc{Psource}(s) =_p x : f \multimap (f \multimap l) \multimap e \multimap s$

(37) Frank seemed like he was upset.

(38)
$$\begin{bmatrix} \textsc{pred} & \text{'}\textsc{seem}\langle\textsc{xcomp}\rangle\textsc{subj}\text{'} \\ \textsc{subj} & f\left[\text{"Frank"}\right] \\ s & \begin{bmatrix} \textsc{pred} & \text{'}\textsc{like}\langle\textsc{comp, subj}\rangle\text{'} \\ \textsc{subj} & \\ \textsc{xcomp} & l \\ & \begin{bmatrix} \textsc{pred} & \text{'}\textsc{like}\langle\textsc{comp, subj}\rangle\text{'} \\ \textsc{subj} & \\ \textsc{comp} & \left[\text{"he was upset"}\right] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The special equality, $=_p$, is defined as returning true or false iff the terms being compared ($\textsc{Psource}(s)$ and x) are of the same semantic type; otherwise the equality returns no truth value. The equality is therefore a kind of presuppositional equality, as discussed further below.

The following copy raising examples in English and Swedish have the semantics in (41):

(39) Tom seems like he is laughing.

(40) Tom verkar som om han skrattar.
T. seems as if he laughs
Tom seems as if he is laughing.

(41)
$$\dfrac{\dfrac{\dfrac{tom \qquad \lambda x\lambda P\lambda s.seem(s,P(x)) \wedge \textsc{Psource}(s) =_p x}{\lambda P\lambda s.seem(s,P(tom)) \wedge \textsc{Psource}(s) =_p tom} \qquad \lambda y.\exists e[laugh(e,y) \wedge \textsc{Agent}(e) = y]}{\lambda s.seem(s,\exists e[laugh(e,tom) \wedge \textsc{Agent}(e) = tom]) \wedge \textsc{Psource}(s) =_p tom}}{\exists s.seem(s,\exists e[laugh(e,tom) \wedge \textsc{Agent}(e) = tom]) \wedge \textsc{Psource}(s) =_p tom}$$

Notice that the copy raising verb's lexical entry ensures that the copy raising subject is the PSOURCE of the copy raising verb's eventuality (a state).

In contrast, infinitival raising, exemplified by the following English and Swedish examples, has the semantics in (44):

(42) Tom seems to paint.

(43) Tom verkar måla.
T. seems paint.INF
Tom seems to paint.

(44)
$$\dfrac{\dfrac{\dfrac{\lambda p\lambda s'.seem(s',p) \qquad \exists e[paint(e,tom) \wedge \textsc{Agent}(e) = tom]}{\lambda s'.seem(s',\exists e[paint(e,tom) \wedge \textsc{Agent}(e) = tom])} \qquad \lambda S\lambda s.\exists v_\delta[S(s) \wedge \textsc{Psource}(s) =_p v_\delta]}{\lambda s.\exists v_\delta[seem(s,\exists e[paint(e,tom) \wedge \textsc{Agent}(e) = tom]) \wedge \textsc{Psource}(s) =_p v_\delta]}}{\exists s\exists v_\delta[seem(s,\exists e[paint(e,tom) \wedge \textsc{Agent}(e) = tom]) \wedge \textsc{Psource}(s) =_p v_\delta]}$$

Notice that (44) contrasts with (41) in having existential closure (binding) of a variable v_δ (the type we reserve for eventualities), since infinitival-complement raising verbs do not lexically specify that their subject is a perceptual source. We motivate this existential closure in the next section.

2.1 Existential closure of Psource

Consider a standard infinitival raising sentence like the following:

(45) Maria seems to have wrecked the hotel room.

In the situation described by this sentence, something gives the impression that Maria has wrecked the hotel room, probably the state of the hotel room. This indicates that even non-copy-raising subcategorizations of *seem* have a Psource, but the Psource is not necessarily the subject. It could be Maria herself who somehow gives the impression (e.g., if she's covered in plaster and carrying a smashed-up TV), but this is not the most natural reading of (45). Notice, for example, that the corresponding copy raising sentence (46), in which Maria is lexically specified by the verb as the Psource, is decidedly odd out of context:

(46) Maria seems like she wrecked the hotel room.

The oddness of this sentence stems from the difficulty in accommodating the proposition that Maria wrecked the hotel room based on Maria being the Psource. The contrast between (46) and (45) in the null context and the intuitive meaning for (45) point to existential closure of the Psource in propositions expressed by sentences like (45), without commitment to whether the existentially closed variable is an individual or an eventuality.

Further evidence for existential closure of Psource comes from Swedish, where *på*-PPs are not only ungrammatical with copy raising — as demonstrated by the *på* puzzle data itself — but are surprisingly also ungrammatical with infinitival raising verbs:

(47) * Maria verkar på Jonas vara glad.
M. seems on J. be happy

The question is: why can't (47) mean that Jonas gives the impression that Maria seems to be happy, which is, again, a perfectly sensible proposition. If Psource is existentially bound in infinitival sentences, the ungrammaticality of (47) follows automatically from the uniqueness requirement on Psources. There are two Psources in (47) (the existentially bound Psource and the *på*-PP Psource), which violates the functional definition of Psource, as per the generalization of the Unique Role Requirement that was discussed following (35) above.

2.2 Solutions to the two puzzles

Recall that the puzzle of the absent cook concerned the ungrammaticality of copy raising sentences in scenarios like the following:

> Scenario: you and your friend walk into John's kitchen. There are pots and pans on the stove. It smells like food. It's obvious that someone is cooking. John is not in the kitchen.
>
> (48) #John seems like he's cooking.

Our analysis explains this puzzle as follows. The actual Psource in the scenario above is the state of the kitchen. However, the copy raising verb's lexically-specified Psource is John. The formal analysis of Asudeh and Toivonen (2006b) results in checking whether the Psource of (48) (the state of the kitchen) equals John, using a presuppositional equality that only returns true or false if the entities being compared have the same type. In this case, this is not true, because we are comparing a state, which has the type for eventualities, to an individual, which has the individual type. Therefore our analysis treats the unacceptability of (48) as presupposition failure (hence the use of the infelicity marker # rather than the ungrammaticality marker *). This also explains why the negation of (48) is equally odd in the given scenario:

(49) #John doesn't seem like he's cooking.

In sum, the puzzle of the absent cook is explained as presupposition failure that arises from asserting that an individual is a Psource when the Psource role is actually filled by something else.

The solution to the *på* puzzle was anticipated in the discussion of the existential closure above. *På* puzzle cases are exemplified by sentence like:

(50) * Maria verkar på Elin som om hon har vunnit.
M. seems on E. as if she has won

In such cases, there are two Psource contributors: the copy raising verb, which lexically specifies Maria as the Psource, and the *på*-PP, which specifies Elin as the Psource. Having two instances of Psource violates the uniqueness requirement.

3 The challenge of expletives

The semantics for copy raising that we have sketched thus far solves a couple of puzzles and arguably gets several aspects of the phenomenon right. Ideally, we want to maintain a consistent semantics for copy

raising verbs. However, copy raising verbs also occur with expletives, including raised expletives (illustrated below), which complicate matters considerably. Expletives present a challenge to any analysis of the syntax and semantics of copy raising. In this section, we attempt to meet this challenge and show that copy raising conversely reveals something about the syntax and semantics of expletives.

We have already noted that copy raising can occur with the standard *it*-expletive that we would generally expect with a raising verb like *seem* or *appear*:

(51) It seems like there's trouble in paradise.

(52) It seems like it's raining.

These examples illustrate that expletive choice in the lower clause is independent of the raising verb's expletive, as we would expect.

However, copy raising verbs also exhibit the expletive pattern shown here:

(53) There seems like there's trouble in paradise.

(54) *There seems like it's raining.

Many but not all speakers accept (53) as grammatical, but all speakers reject (54) as ungrammatical. This pattern of grammaticality indicates that the matrix expletive in copy raising can be dependent on the lower expletive in the copy raising verb's complement.

The pattern shown in (53) can also readily be found in attested examples:

(55) God, there seems like there's no end to the innovation we come up with, you know.
(http://www.mp3.com/features/stories/4189.html; checked 10/2006)

(56) ... there seems like there's some connection with the car jacking that took place ...
(http://transcripts.cnn.com/TRANSCRIPTS/0208/01/lol.03.html; checked 10/2006)

(57) Also, there appears as though there are less balloons in the final shot.
(www.horrorking.com/mviegoof.html; Google cached version checked 10/2006)

We follow Asudeh (2004) in analyzing copy raising with *there*-expletive subjects as an instance of double raising. The expletive is raised from the predicative *like/as*-complement's subject, as per (38) above, but it also raised, by *like* or *as*, from the sentential complement to *like/as*. This is sketched here:

(58) There seems like there's trouble in paradise.

(59)
$$\begin{bmatrix} \text{PRED} & \text{`SEEM}\langle\text{XCOMP}\rangle\text{SUBJ'} \\ \text{SUBJ} & \boxed{1} \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{`LIKE}\langle\text{XCOMP}\rangle\text{SUBJ'} \\ \text{SUBJ} & \boxed{1} \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{`BE}\langle\text{PREDLINK}\rangle\text{SUBJ'} \\ \text{SUBJ} & \boxed{1}\begin{bmatrix}\text{EXPLETIVE} & \text{THERE}\end{bmatrix} \\ \text{PREDLINK} & \begin{bmatrix}\text{“trouble in paradise”}\end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Asudeh (2004) treats the capacity for *like* and *as* to raise from their finite complements as an exceptional, lexical property.

The normal assumption is that expletives have no semantics. In our Glue Semantics treatment, this means that lexical entries for expletives contribute no resources. This presents a serious challenge for copy raising. Recall that the Glue meaning term for copy raising is as follows:

(60) $\lambda x \lambda P \lambda s.seem(s, P(x)) \ \wedge \ \textsc{psource}(s) =_p x :$
$subj \multimap (subj \multimap l) \multimap e \multimap s$

The copy raising verb contains a dependency on its subject, which it will satisfy by composing its subject with the property contributed by the predicative *like/as*-complement. However, if the expletive subject has no semantics, then this composition cannot be carried out, as shown by the following invalid Glue proof, which does not terminate in the right type for a proposition, due to the undischarged dependency on $subj$:

(61)
$$\dfrac{\dfrac{\dfrac{\dfrac{\lambda x \lambda P \lambda s.seem(s, P(x)) \ \wedge \ \textsc{psource}(s) =_p x : subj \multimap (subj \multimap l) \multimap e \multimap s}{\lambda P \lambda x \lambda s.seem(s, P(x)) \ \wedge \ \textsc{psource}(s) =_p x : (subj \multimap l) \multimap subj \multimap e \multimap s}\ \textit{curry} \qquad \begin{array}{c}\vdots\\ like : subj \multimap l\end{array}}{\lambda x \lambda s.seem(s, like(x)) \ \wedge \ \textsc{psource}(s) =_p x : subj \multimap e \multimap s} \qquad [y : subj]^1}{\dfrac{\lambda s.seem(s, like(y)) \ \wedge \ \textsc{psource}(s) =_p y : e \multimap s}{\exists s[seem(s, like(y)) \ \wedge \ \textsc{psource}(s) =_p y] : s}\ \textit{event closure}}}{\lambda y.\exists s[seem(s, like(y)) \ \wedge \ \textsc{psource}(s) =_p y] : subj \multimap s}\ {\multimap_{\mathcal{I},1}}$$

This problem is, however, more general than just a problem for Glue Semantics or our particular treatment. Any analysis that attempts to explain copy raising compositionally is potentially challenged by the ability of copy raising verbs to host both expletive subjects and apparently thematic subjects that are only licensed by virtue of being anaphorically tied to a copy pronoun.

It is initially tempting to backtrack and state that the expletive *does* actually contribute a resource, i.e. it does have a semantics. An appropriate semantics might be existential closure of the variable that corresponds to the subject in the semantics:

(62) $\lambda P.\exists x[P(x)] : (\uparrow_\sigma \ \multimap \ (\textsc{subj} \uparrow)_\sigma) \ \multimap \ (\textsc{subj} \uparrow)_\sigma$

This Glue meaning term takes a dependency on a subject — a property — and returns an existentially closed proposition. The inside-out equation in (62) states this in terms of a dependency from $\uparrow$ to the thing that $\uparrow$ is a subject of. The inside-out specification is needed due to the fact that the equation is part of the lexical specification of the expletive itself; i.e. $\uparrow$ refers to the expletive's f-structure, not the verb's.

If the expletive were to contribute this kind of meaning, then the conclusion of the proof in (61) would instead be:

(63) $\exists y[\exists s[seem(s, like(y)) \ \wedge \ \textsc{psource}(s) =_p y]] : subj \multimap s$

This is a valid Glue proof and a reasonable semantics for both the copy raising verb and the expletive. It states that something gives the impression that the proposition expressed by the complement of the copy raising verb holds.

However, the solution just sketched leads to various problems. First, if the expletive contributes the existential meaning in (62), as far as the semantics is concerned we should be able to derive the following:

(64) *There meowed.

(65)
$$\dfrac{\lambda P.\exists x[P(x)] : (s \multimap m) \multimap m \qquad \lambda y.meow(y) : s \multimap m}{\exists x[meow(x)] : m}$$

It is clear that 'There meowed' doesn't mean that something meowed: it's just ungrammatical. Independent syntactic constraints might block (64), but it is questionable whether that would be the right approach.

The question does not need to be settled, though, because the proposal suffers a much worse independent problem. The expletive raising cases illustrate that more than one *there*-expletive can be inserted from the lexicon in this construction:

(66) There seems like there is a piece missing.

If we assume a consistent semantics for both occurrences of the expletive, as would be theoretically desirable, then there would be too many subject consumers. In other words, the compositional requirements of both expletives, as per (62), could not be satisfied.

A solution suggests itself, however: instead of associating the existential closure resource with the expletive, as in (62), associate it with the head of the *like/as*-complement in its expletive raising subcategorization (notice that we have left underspecified the semantics of *like*; we return to this issue in the conclusion):

(67) *like*: $(\uparrow \text{PRED}) = \text{`like}\langle\text{XCOMP}\rangle\text{SUBJ'}$
$(\uparrow \text{PTYPE}) = \text{CLAUSAL-COMPARATIVE}$

$((\uparrow \text{SUBJ}) = (\uparrow \text{XCOMP SUBJ}))$
$(\lambda P.\exists x[P(x)] : ((\uparrow \text{SUBJ})_\sigma \multimap X) \multimap X)$

$\ldots \lambda x.like(\ldots x \ldots) : \ldots (\uparrow \text{SUBJ})_\sigma \multimap \uparrow_\sigma$

This instead associates the existential closure that (62) associated with the expletive itself with the predicator that governs the explicit raising (recall that *like/as* exceptionally raise the expletive from their complement). The two optional parts of the lexical entry for the preposition can be realized independently. This is due to the fact that the existential closure needs to be realized separately of the raising in sentences like (51), repeated here:

(68) It seems like there's trouble in paradise.

For an example like this, the existential closure is necessary to satisfy to the copy raising verb's consistent dependency on its subject, but there the *it*-expletive must be independently generated, not raised, since the lower expletive is a non-matching *there*-expletive.

The proof in Figure 1 sketches the resulting well-formed semantics. The conclusion of the proof is an atomic sentential resource with all dependencies discharged, as show above in (64). The difference is that the existential closure is contributed by the *like/as*-head of the copy raising verb's predicative complement, not by the expletive itself. This ensures successful semantic composition, because the individual expletives are not contributing multiple closures over the same variable. Furthermore, it maintains the standard semantics for expletives as contentless. Lastly, it places the exceptional semantic composition in the lexicon, where it arguably belongs, and, more particularly, in the lexical entry for *like/as*, which is exceptional for independent reasons.

Lastly, we would like to make some brief comments on LFG's Subject Condition, building on Asudeh (2004). The Subject Condition is the requirement that every predicator has a subject (Bresnan 2001). It is normally understood purely f-structurally: every predicator must have a SUBJ grammatical function at f-structure. However, expletive raising indicates that this is insufficient. Recall the sort of f-structure that is relevant:

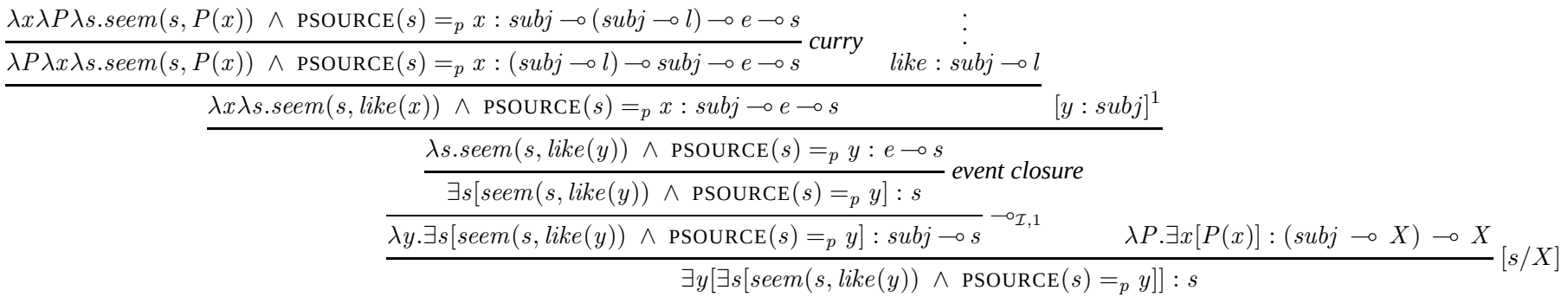
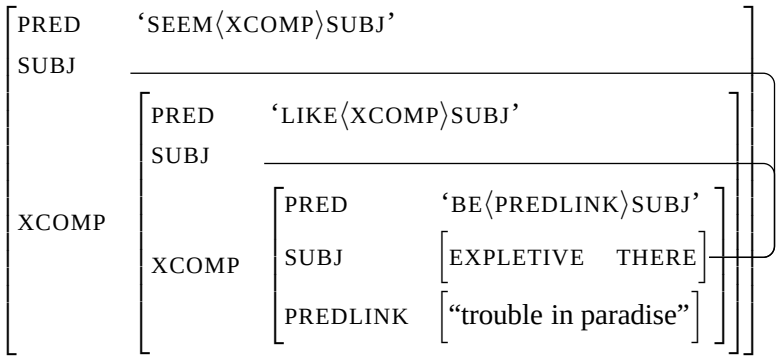


Figure 1: Semantics for copy raising with expletive subject

(69)


There are a couple of alternatives for how the functional control that handles the expletive raising could be handled.

The first option is classic LFG functional control (Kaplan and Bresnan 1982, Bresnan 1982). However, in that case the lower expletive alone should satisfy all subject requirements in the f-structure. This would overgenerate examples like the following:

(70) *Today seems like there's a problem.

Alternatively, functional control could be realized through subsumption (Zaenen and Kaplan 2002, Kaplan and Zaenen 2003), but this would mean that either the lowest expletive could satisfy all requirements, again overgenerating (70), or else the highest expletive could satisfy all requirements, which equally overgenerates:

(71) *Today there seems like is a problem.

Thus, functional control at f-structure, whether through equality or subsumption, is problematic for satisfaction of the Subject Condition at f-structure in cases of expletive raising. This suggests that there has to be a c-structural component to the Subject Condition: certain c-structural subject positions in English (specifiers of finite IP) have to be filled.

4 Conclusion and future work

Copy raising presents an intricate set of puzzles for syntax, semantics, and the syntax-semantics interface. A particular challenge, brought to the fore by expletives, is how to provide a purely compositional semantics that adequately treats all the parts while properly capturing their denotations. The solution proposed here maintains the standard analysis of expletives as semantically contentless, having shown that an attempt to attribute reasonable content to expletives is problematic.

A number of challenges remain for future work. First, what is the proper semantics for the predicative head *like/as*? These predicates of similarity can appear in various usages and with various complements, as illustrated here for *like*:

(72) John runs like Mary skips.
$\exists s[\exists s'[\exists P[P(s) \wedge P(s')] \wedge run(s,j) \wedge upset(s',j)]]$

(73) John is like Mary.
$\exists P[P(j) \wedge P(m)]$

(74) John seems like Mary.
$\exists x[\exists s[seem(s, \exists s'[\exists P[P(j) \wedge P(m)]]) \wedge \text{PSOURCE} =_p x]]$

(75) John seems like he is upset.
$\exists s[seem(s, \exists s'[\exists P[P(s,j) \ \wedge \ P(s',j)] \ \wedge \ upset(s',j)]) \ \wedge \ \text{PSOURCE} =_p j]$

We have indicated target semantics for each case. The second challenge is how to derive the variety of semantics compositionally while maintaining a core meaning for *like/as*. A promising alternative would seem to be a polymorphic analysis in which a base semantic type for *like* is specified plus a procedure for deriving the other types. The third case is particularly problematic, because it seems similar to copy raising, but with a nominal complement to *like* which has no copy pronoun. Our analysis of copy raising does not extend to such cases, since there is in fact no copy. It could arguably be a different construction, but it is surely not purely coincidental that the matrix verb is *seem*. Thus, the relationship between (74) and (75), where the latter is true copy raising, presents a third challenge. A fourth challenge is the specification of how the semantics of *like/as* interacts with the semantics of predication and comparatives (Matushansky 2002). Lastly, a fifth challenge is to explain why clausal comparatives are excluded from copular clauses:

(76) a. John seems like Mary.
b. John seems like he is upset.

(77) a. John is like Mary.
b. *John is like he is upset.

Both *seem* and *be* can occur with the *like NP* complement, but only *seem* can occur with the *like CP* complement.

We hope to have shown in this paper that copy raising is both syntactically and semantically challenging and that it opens up many avenues for further enquiry.

Acknowledgements

We would like to thank the audience at LFG 2006 for their comments and suggestions. This research was supported by SSHRC Standard Research Grant 410-2006-1650.

References

Asudeh, Ash. 2002. Richard III. In Mary Andronis, Erin Debenport, Anne Pycha, and Keiko Yoshimura, eds., *CLS 38: The Main Session*, vol. 1, 31–46. Chicago, IL: Chicago Linguistic Society.

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

Asudeh, Ash, and Ida Toivonen. 2006a. Copy Raising and Its Consequences for Perceptual Reports. In Jane Grimshaw, Joan Maling, Christopher Manning, Jane Simpson, and Annie Zaenen, eds., *Architectures, rules, and preferences: A festschrift for Joan Bresnan*. Stanford, CA: CSLI Publications.

—. 2006b. Copy Raising and Perception. Ms., Carleton University. Submitted.

Bach, Emmon. 1981. On Time, Tense and Aspect: An Essay on English Metaphysics. In Peter Cole, ed., *Radical Pragmatics*, 62–81. New York: Academic Press.

Bresnan, Joan. 1982. Control and Complementation. *Linguistic Inquiry* 13: 343–434.

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

Carlson, Gregory N. 1984. Thematic Roles and Their Role in Semantic Interpretation. *Linguistics* 22: 259–279.

Chierchia, Gennaro. 1984. Topics in the Syntax and Semantics of Infinitives and Gerunds. Ph.D. thesis, University of Massachusetts, Amherst.

Chung, Sandra. 1978. *Case Marking and Grammatical Relations in Polynesian*. Austin, TX: University of Texas Press.

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

Déprez, Viviane. 1992. Raising Constructions in Haitian Creole. *Natural Language and Linguistic Theory* 10: 191–231.

Fujii, Tomohiro. 2005. Cycle, Linearization of Chains, and Multiple Case Checking. In Sylvia Blaho, Luis Vicente, and Erik Schoorlemmer, eds., *Proceedings of Console XIII*, 39–65. Student Organization of Linguistics in Europe, University of Leiden.

Heycock, Caroline. 1994. *Layers of Predication*. New York: Garland.

Joseph, Brian D. 1976. Raising in Modern Greek: A Copying Process. In Jorge Hankamer and Judith Aissen, eds., *Harvard Studies in Syntax and Semantics*, vol. 2, 241–281. Cambridge, MA: Harvard University, Department of Linguistics.

Kaplan, Ronald M., and Joan Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*, 173–281. Cambridge, MA: MIT Press.

Kaplan, Ronald M., and Annie Zaenen. 2003. West-Germanic Verb Clusters in LFG. In Pieter Seuren and Gerard Kempen, eds., *Verb Constructions in German and Dutch*, 127–150. Amsterdam: John Benjamins.

Kokkonidis, Miltiadis. 2006. A Glue/λ-DRT Treatment of Resumptive Pronouns. In Janneke Huitink and Sophia Katrenko, eds., *Proceedings of the Eleventh ESSLLI Student Session*, 51–63. Málaga, Spain.

Landman, Fred. 2000. *Events and plurality: the Jerusalem lectures*. Dordrecht: Kluwer.

Lappin, Shalom. 1984. Predication and Raising. In Charles Jones and Peter Sells, eds., *Proceedings of NELS 14*, 236–252. Amherst, MA: GLSA.

Matushansky, Ora. 2002. Tipping the Scales: The Syntax of Scalarity in the Complement of *seem*. *Syntax* 5(3): 219–276.

McCloskey, James, and Peter Sells. 1988. Control and A-chains in Modern Irish. *Natural Language and Linguistic Theory* 6: 143–189.

Moore, John. 1998. Turkish Copy-Raising and A-Chain Locality. *Natural Language and Linguistic Theory* 16: 149–189.

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

—. 1995. Thematic Relations and Arguments. *Linguistic Inquiry* 26(4): 635–662.

Perlmutter, David M., and Scott Soames. 1979. *Syntactic Argumentation and the Structure of English*. Berkeley, CA: University of California Press.

Polinsky, Maria, and Eric Potsdam. 2006. Expanding the Scope of Control and Raising. *Syntax* 9(2): 171–192.

Postal, Paul. 1974. *On Raising*. Cambridge, MA: MIT Press.

Potsdam, Eric, and Jeffrey T. Runner. 2001. Richard Returns: Copy Raising and its Implications. In Mary Andronis, Chris Ball, Heidi Elston, and Sylvain Neuvel, eds., *CLS 37: The main session*, vol. 1, 453–468. Chicago, IL: Chicago Linguistic Society.

Rogers, Andy. 1971. Three Kinds of Physical Perception Verbs. In *Papers from the Seventh Regional Meeting of the Chicago Linguistic Society*, 206–222.

—. 1973. Physical Perception Verbs in English: A Study in Lexical Relatedness. Ph.D. thesis, UCLA.

Ura, Hiroyuki. 1998. Checking, Economy, and Copy-Raising in Igbo. *Linguistic Analysis* 28: 67–88.

Zaenen, Annie, and Ronald M. Kaplan. 2002. Subsumption and Equality: German Partial Fronting in LFG. In Miriam Butt and Tracy Holloway King, eds., *Proceedings of the LFG02 Conference*, 408–426. Stanford, CA: CSLI Publications.

EVIDENTIALITY IN SOUTH ASIAN LANGUAGES

Elena Bashir
University of Chicago

Proceedings of the LFG06 Conference,
Workshop on South Asian Languages
Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006
CSLI Publications

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

Abstract. This paper explores the encoding of the semantics of evidentiality and indirectivity in some South Asian languages. In my analysis, evidentiality is related to the complex of overlapping categories involving (i) the source of information about an event or state and (ii) its acquisition by an observer/speaker. In some languages several of these notions are morphologically encoded; in others the categories are relatively "covert" and the expression of evidentiality is distributed (Aikhenvald's "scattered") throughout the grammar. The paper summarizes previously published data on inferential systems in Tajik Persian, Kalasha, Khowar, and Nepali, and presents new data on several other languages that have morphologically encoded inferentiality--Yasin Burushaski, three Nuristani languages, and Wakhi. Additionally, other inferentiality-marking strategies are discussed for a cluster of languages including Torwali, Pashto, Shina, and Kohistani, for Hindi and Urdu, and for a cluster of South Indian languages. Evidentiality is known to be highly susceptible to language contact effects Aikhenvald (2003:21-2) and Johanson (2000:81-2). The investigations reported in this paper confirm that evidentiality marking patterns fall into recognizable areal units and sub-units in South Asia as well. Evidentiality-encoding strategies are seen to group areally with clearly identifiable northern and southern clusters and a mixed area.

1 A cognitive model for evidentiality and indirectivity (inferentiality)

A cognitive model of event structure can unify and explain various specific manifestations of the categories of evidentiality and indirectivity (including mirativity).[1] Bashir (1993) explored this idea in the context of compound verbs; I now focus on evidentiality. This analysis is based on DeLancey's cognitively based model of an event as a vector having two endpoints, interpreted at the most general level as ORIGIN and TERMINATION (DeLancey 1985:47). This generalized schema underlies varying grammatical manifestations, depending on whether one focuses on the entities involved in an event, or on its logically or temporally sequential stages. Thus the ORIGIN and TERMINATION endpoints of the vector can be associated with the concepts AGENT and PATIENT, SOURCE and GOAL, or CAUSE and RESULT (DeLancey 1982:172). This is schematized below.

ORIGIN ->	EVENT - - >	TERMINATION
Temporal onset	Event/action	Temporal conclusion
Cause (e.g. act of volition)		Resultant state
Agent		Patient
Source		Goal

An event may impinge upon an observer/speaker's awareness at any point along its causal vector. He may become aware of an event from its ORIGIN, i.e. the stage of its cause or antecedent situation, as when a situation is anticipated, feared, predicted, or actively caused by him. If the observer/speaker has access to the ORIGIN end of an event vector or to the EVENT itself, the event will be reported with a direct form. If, however, he learns about an event only by observing its resultant state–the TERMINATION of the causal chain–the event will be encoded with an indirect form or strategy. This is the central insight underlying my analysis of evidentiality and inferentiality.

Differences in the point at which an event impinges upon an observer's awareness are grammaticized in many languages, for example Tibetan (1). (1-a) would be appropriate if the speaker is involved in planning the meeting, whereas (1-b) could be uttered by someone who had learned of the meeting second hand, as by reading a notice about it. "The use of *yod* indicates an assertion made on the basis of direct knowledge of the entire vector, while *'dug* indicates direct knowledge of the result but not of the cause." (DeLancey 1986:206) That is, the choice between *yod* or *'dug* depends on the impingement point of the causal vector with the speaker.

(1-a) *gza-spen-ba la tsod-'du yod*
Saturday LOC meeting exist
'We have a meeting on Saturday.'

[1] Plungian (2001:355) says that "the recurrent polysemy of admirative and inferential and/or quotative markers needs an explanation". Discussing mirativity, DeLancey (2001) has argued, citing languages like Hare (Athapaskan) which have a mirative marker apparently independent of evidentiality, that it is a category distinct from evidentiality. In my analysis, mirativity is one of the specific semantic effects that emerges when the observer/speaker has access only to the result end of an event vector.

(1-b) *gza-spen-ba la tsog-'du 'dug*
Saturday LOC meeting exist
'There's a meeting on Saturday.' (DeLancey 1986:206)[2]

Several parameters of the impingement of an event on a sentient observer/speaker correlate with its expression: (i) source: internal (endophoric) or external (sensory); (ii) time: past (old knowledge), present (new knowledge), future (presumption [necessary overlap here with epistemic modality]); (iii) directness: direct (first-hand sensory, well-established, hence speaker-internal) or indirect (second-hand, reported; inferred). Reported (hearsay) or inferred information is necessarily new. However direct sensory experience can also be new information; this situation gives rise to uniquely mirative semantics. Indirectly acquired information (hearsay, inference) is also frequently new (mirative); hence the overlap between the categories of indirectivity and mirativity.

2 Types of evidential systems

Aikhenvald and Dixon (2003:3) present a typology of evidential systems. Type I systems "state the existence of a source for the evidence without specifying it", and a statement marked for evidentiality is characterized "by reference to its reception by a conscious subject" (Johanson 2003:274). This type of evidentiality is therefore referred to as 'indirectivity' by Johanson (2000, 2003). Type II systems "specify the kind of evidence, be it visually obtained, based on inference, or reported information." Type II systems point to the ORIGIN of an event vector, while Type I systems focus on the TERMINATION.

3. Language data

3.1 Old Indo-Aryan (OIA)

The OIA verb system grammaticized the seen/unseen distinction. Deshpande (1981:62) concludes that in Panini's language the three preterital tenses were specified as in (2). The imperfect contrasted with the perfect in that the perfect was (to be) used when an action not witnessed by the speaker is reported.[3] The +/- seen distinction appears not to have existed in the non-past tenses.

(2)

aorist	imperfect	perfect
+ past	+ past	+ past
+ recent	- recent	- recent
+/- seen	+ seen	- seen

In addition, a particle *kila* / *kira* was used in Sanskrit, Prakrit and the Pali Jatakas in senses which Emeneau (1969:244) gives as "report", "tradition", "traditional account", "general opinion, universal knowledge", "so it is heard", "as is reported", "as they say", and a secondary meaning of "irony". One example, of the traditional stories type, is given here as (3). Van Daalen (1988:11-12), analyzing Sanskrit and Prakrit texts, divides the uses of *kila/kira* into four disparate categories. However, Degener (1998:182) finds that *kila* functions in all cases as a reportative particle, seeing also in *kila* a clear mirative use and a possible historical link to Nepali *le* (see below).

(3) *vyuSitāśva iti khyāto babhūva kila pārtivah*
VyuSitāśva QUOT called exist(PERF).3s *kila* king
'There was of old, as the story goes, a king called VyuSitāśva.' [Mbh. 1.112.7] (Emeneau 1969:245)

3.2 Northern cluster

Previous research has shown that some modern IA languages–Kalasha, Khowar, and Nepali–grammaticize evidentiality in the verbal system. New work indicates that morphological inferentiality is also found in Wakhi, several Nuristani languages, and Yasin Burushaski.

3.2.1 Kalasha

In Kalasha and Khowar the old *-ta* participles took on the *parokSa* (unseen > inferential) value, while the finite preterite which developed from the aorist and imperfect retained the [+seen] specification. Thus in Kalasha and Khowar the basic [+/- seen] distinction is inherited from OIA, while a second stratum of inferential marking, accomplished with a past participle of 'become' (*birái* in Khowar, *huLa* in Kalasha) seems to be a later accretion.[4]

[2] My test sentence 'There is a meeting on Saturday' was elicited in order to make my results comparable to some degree with DeLancey's work.

[3] Cardona (2002) reaffirms Deshpande's conclusion, presenting textual evidence that the presence of the three-way distinction in the tense system described by Panini is also attested in Vedic literature.

[4] Khowar has been heavily influenced by Persian (Tajik) in many areas–lexis, syntax, and probably semantics. Turkic, with its robust indirectivity marking, may also have been important, since a ruling Chitrali dynasty came from Turkic-speaking areas. Also, until quite recently, Khowar has been in contact with Wakhi.

The basic tense-aspect forms are illustrated here with the 1st person sg. of *kárik* 'to do' (Bashir 1988a, b).

Non-past	
PRESENT/FUTURE-NON SPECIFIC (P/F-N-S) *a kár-im* 'I do, I will do'	
PRESENT/FUTURE-SPECIFIC (P/F-S) *a kár-im dai* 'I am doing, will do (at a specific time)'	
PRESENT PERFECT (P PERF) *a kai á-am* 'I have done'	
Past	
DIRECT	INFERENTIAL
PAST (PST-D) *a ár-is* 'I did.'	PST (PST-I) *a káda him* 'I did (reportedly, inadvertently.)'
PAST IMPERFECTIVE (PST IMPFV-D) *a kar-íman áy-is* 'I was doing.'	PAST IMPERFECTIVE (PST IMPFV-I) *a kar-íman ásta him* 'I was doing (reportedly.)'
PAST PERFECT (PST PERF-D) *a kai áy-is* 'I did, had done.'	PAST PERFECT (PST PERF-I) *a kai ás-ta him* 'I had done, did (reportedly.')

Past tense verb forms are obligatorily coded for the distinction between direct ("actual" in Bashir 1988a, b) and inferential (indirect) meaning. Direct subsumes such meanings as personally witnessed, or having long standing in one's conceptual repertoire, while inferential includes inference, new information, and hearsay. Present-tense forms do not have morphologically expressed inferential forms, but inferential counterparts are supplied by the addition of *huLa*, the past participle of *hik* 'to become'. When *huLa* appears in narration of directly experienced events, the meaning is mirative, i.e. that the speaker has just found out about (i.e. was not aware of before) the content of the assertion. Other specific pragmatic effects emerge, e.g. surprise, regret, or annoyance (Bashir 1988b:44). Contrastive examples follow for the past (4) and the present perfect (5). With first person agents, the inferential form gives a sense of unconscious, inadvertent, or mistaken action (5-b). This interaction effect of non-direct forms with first person has been noted for many languages. Additionally, in Kalasha specifically hearsay utterances involve a construction consisting of the infinitive of the verb expressing the semantic core of the assertion, and *ghõ'-an* 'they say' (6).

PAST - DIRECT

(4-a) *õ'je-mí par-á*
now-EMPH go(PST-A)-3s
'He just left.' (Bashir 1988b:37)

PAST - INFERENTIAL

(4-b) *a ayá a ãgár Zot káda*
I(NOM) here come(PST-A)-1s fire already do(PST-I)-3s
'I came here. (Someone) had already made the fire (unseen by me).' (Bashir 1988b:42)

PRESENT PERFECT - DIRECT

(5-a) *a põj So chaT jahás-una nisí á-am*
I(OBL) 5 6 times plane-LOC sit(PRES PERF)-1s
'I have flown (lit.'sat') in a plane five or six times.' (Bashir 1988b:41)

PRESENT PERFECT - INFERENTIAL (+ *húLa*)

(5-b) *a galatí kai á-am húLa*
I(NOM) mistake do(PRES PERF)-1s become(PST-I)-3s
'I (just realized that I) have made a mistake.' (mirative) (Bashir 1988b:44)

(6) *ne šik ghõ'-an mai putr*
not be-INAN(INF) say(P/F-NS)-3p my son
'(I hear/they say that) there isn't any, my son.' (Bashir 1988b:46)

3.2.2 Khowar

The Khowar verb system consists of marked inferential/indirect and unmarked direct forms. Its main forms are illustrated below for the third person singular of *korík-* 'to do'. (This analysis differs somewhat from Bashir (1988b).) In non-past tenses, inferential forms are constructed with an agent noun in *-ak* plus the PST-I form of *bik* 'be, become'. These forms are (partially) tense-neutral, in that they can apply to present, future, or past events. In forms built on the past participle, itself already marked as inferential, the *bik* 'become' forms add a mirative meaning.

<table>
<tr><th>DIRECT</th><th>INFERENTIAL</th></tr>
<tr><td colspan="2">Non-past</td></tr>
<tr><td>PRESENT/FUTURE, NON-SPECIFIC
korói ' S/he does, will do.'</td><td rowspan="2">PRESENT/FUTURE/PAST
korák birái 'It turns out that s/he does/will do; s/he does/will do/used to do.' (reportedly) (mirative)</td></tr>
<tr><td>PRESENT/FUTURE-SPECIFIC
koróy-an 'S/he does, is doing, will do.'</td></tr>
<tr><td>PRESENT PERFECT
korí asúr 'S/he has done.'</td><td>PRESENT PERFECT
korí asák birái S/he did, has done' (reportedly, mirative)</td></tr>
<tr><td colspan="2">Past</td></tr>
<tr><td>PAST
areér 'S/he did.'</td><td>PAST
kardú 'S/he did (unwitnessed).'</td></tr>
<tr><td>(PAST) PERFECT-1
kardú ošói 'S/he did, had done;would have done; was about to have done.'</td><td>(PAST) PERFECT-1
kardú birái 'S/he did, had done, has done' (reportedly, mirative).</td></tr>
<tr><td>PAST PERFECT-2
korí asítai 'S/he had done.'</td><td>PAST PERFECT-2
korí astaái 'S/he had done (unwitnessed, unwittingly.)'</td></tr>
<tr><td>PAST IMPERFECTIVE
koráu ošói 'S/he was doing, was about to do.'</td><td>PAST IMPERFECTIVE-1 (Chitral, Torkhow)
koráwa birái 'S/he (habitually) did, would do; was about to do (reportedly).'
PAST IMPERFECTIVE-2 (Zondrangram)
koráu astaái 'S/he was doing (reportedly, unexpectedly).'</td></tr>
</table>

Contrastive examples follow for the present/future (7) and the past tense (8).

PRESENT/FUTURE, SPECIFIC - DIRECT

(7-a) *hasé pešáur-o-te no bír-an*
he P.-OBL-DAT not go(P/F-S.3s)
'He is not going to Peshawar (known directly).'

PRESENT/FUTURE - INFERENTIAL

(7-b) *pešáur-o-te no boɣák birái*
Peshawar-OBL-DAT not go(P/F-I)3s
'He is not going to Peshawar (reportedly, new information).'

PAST - DIRECT

(8-a) *hasé lahur-o-te baɣái*
he Lahore-OBL-DAT go(PST-A)-3s
'He went to Lahore (first-hand knowledge).'

PAST - INFERENTIAL

(8-b) *awá oreéi asít-am*
I sleep(PST PERF-D-1s)
angáh hótam ki xúur kos dúr-a asteét-am
awake become(PST-D)-1s I.saw.that other someone(OBL) house-LOC be(PST-I)-1s
'I had fallen asleep. When I awoke I realized that I was in someone else's house.' (mirative)

The category of inferentiality interacts with the pragmatic dimension of politeness. For example, in (9) the telephone rings and is answered by the younger of two sisters. The caller asks whether the addressee has a certain

thing he needs. The younger sister replies in the negative with a direct form and is admonished by the older sister to use an inferential form. The inferential form signals that the speaker didn't know at first that the thing was not present, and after looking for it, found it to be absent. The direct form, however, associates the speaker with direct/prior knowledge about the status of the object and perhaps unwillingness to give it.[5]

(9) A: Question by caller ('Do you have x?')
B: Reply by younger sister: *níki* ('No, we don't have it.')
C: Admonition by older sister: *"no šak birái" ráwe* ('Say, "It turns out not to be here"'.)

3.2.3 Persian

Inferentiality in Persian is discussed most importantly in Windfuhr (1982), Lazard (1985, 1996, 2000), Utas (2000) and Jahani (2000). Afghan (Dari) Persian also displays grammaticized inferentiality, discussed in Perry (2000:230) and others. Tajik Persian is treated in Rastorgueva (1963) and Perry (2000, 2005). Since indirectivity is more highly developed in Tajik Persian, and since it has been in direct contact with Khowar and Wakhi, I summarize its evidentiality system briefly, following Perry (2005:227-234). Several tense forms are specified for indirectivity: the perfect, a non-witnessed durative, a non-witnessed past, and a non-witnessed past progressive. (i) The perfect indicative (past participle plus auxiliary 'be') functions both to indicate a resultant state and as a non-witnessed past or present. In the non-witnessed function, the perfect can indicate meanings of hearsay/quotative (10-a), mirative (10-b), and inference from observation of results. (ii) The non-witnessed durative consists of the perfect plus the prefix *me-*. This form is tense-neutral; it is frequently found in journalistic reporting, where the writer wants to establish distance from second-hand information; (11) illustrates this form with future time reference. (iii) The non-witnessed past consists of the past participle of the verb plus the perfect of *budan* 'to be' (12). (iv) The non-witnessed past progressive consists of the past participle of the verb plus the past participle of *istodan* 'to stand' grammaticized as a progressive construction, plus the perfect of *budan* 'to be' (13).

(10-a) *sayohat-ba rafta-ast*
journey-on go(PERF)
'(I heard) he went on a trip.' (Perry 2000:232) (hearsay/reportative)

(10-b) *ammo ba'd fahmid ke in ciz-i siyoh zoy buda-ast*
but then realized that this thing-EZ black crow be(PERF)
'But then he realized that this black thing (as it turned out) was a crow.' (Perry 2005:233) (mirative)

(11) *ma'lum ast ki ū pagoh me-rafta-ast*
known is that he tomorrow is.going(DUR, NON-WIT)
'It's known that he is going tomorrow.' (Perry 2005:230) (hearsay/reportative)

(12) *gonahi karde bude-ast ke sazo-yaš-raa raft*
a.sin do(PST, NON-WIT) that its.punishment-FOC he.went
'He must have done something wrong to be punished (for it).' (Perry 2000:238) (inference from result)

(13) *vai kitob xonda istoda buda-ast ki man dar-ro taq-taq kardam*
he book read(PST PPL) stand(PST PPL) be(PERF) when I door-ACC knocking did
'He was evidently reading a book when I knocked at the door.' (Perry 2005:233)

3.2.4 Nepali

Nepali has (at least) three forms marked for evidential meanings: (i) the inferential perfect, (ii) a hearsay particle *re*, and (iii) a mirative copula *rahecha*. Michailovsky (1996), citing Clark (1963), describes two forms of the perfect in Nepali: a longer form consisting of the past participle plus the genitive marker *-ko,* and a shorter form consisting of the past participle in *-e*. This short form, called "inferential" by Clark and Michailovsky, was known to Nepali grammarians as the *ajñāt bhūt* 'unknown past'. Two examples follow as (14-a) and (14-b). Note that (14-a) involves the typical context of forgetfulness or absent-mindedness associated with first-person inferentials, and that (14-b) shows inadvertent action, both contexts associated cross linguistically with first-person inferential forms.

(14-a) *tyo kāgat ta birse~ bhaneko ta khāltī-mā po hālechu*[6]
this paper TOP forget(AOR)1s QUOT TOP pocket-in on.the.contrary put(INFER)1s
'I thought I had forgotten the paper, but I find I had put it in my pocket.' (Clark 1963:248, cited in Michailovsky 1996:112)

[5] This phenomenon has also been noted for Japanese (Aoki 1986:235-6). It seems that this may be a pragmatic universal.

[6] In this paper ~ (tilde) following a vowel represents nasalization of the preceding vowel, e.g. *e~* represents nasalized e..

(14-b) *mai-le bhāvanā-lāī lukāuna košiš gare~, tara saki-nã,*
I-ERG feeling-ACC hide(INF) attempt make(AOR)1s but be.able(AOR)-not
'I tried to hide my feelings, but I could not,
musukka hā~sichu, ma kastī badmās-nī
sweetly smile(PERF)1.fs I what.a bad.girl
(and) I smiled sweetly - what a bad girl!' (Michailovsky 1996:115)

The inferential perfect of *rahanu* 'to remain, continue' supplies a specifically mirative copula *rahecha* 'why, he is', which also participates in a progressive and a marked inferential perfect (15) (Michailovsky 1996:111). *rahecha* functions as copula in sentences like (16-a) and (16-b), in which the speaker focuses on the realization of a situation of which he was previously unaware. The hearsay marker *re* appears in (17).[7]

(15) Inferential: *garecha* (do-INFERENTIAL PERFECT) 'Why, he has done!'
Progressive inferential: *gardo rahecha*
doing remain(INFERENTIAL PERFECT) 'Why, he does/is doing!'
Perfect inferential: *gareko rahecha*
done remain(INFERENTIAL PERFECT) 'Why, he has done!'

(16-a) *merā kitāb timro koThā-mā rahecha*
my book your room-in it.is(INFER)
'Oh, I see that my book is in your room.' (Matthews 1990:55)

(16-b) *āhā! kasto rāmro pokharī rahecha*
Ah! what.sort.of beautiful lake it.is(INFER)
'Ah! What a beautiful lake!' (Clark 1963:244, cited in Mikhailovsky 1996:111)

(17) *bhare pānī parcha re*
this.evening water fall(PRES INDEF) HEARSAY
'They say that it's going to rain this evening.' (Matthews 1990:87)

3.2.5 Wakhi

The most basic way of encoding inferentiality in Wakhi is the use of the perfect (perfect stem (+ pronominal clitics)). The basic indicative function of the perfect is resultative-stative; e.g. *δɨtr kɨmd vit-k* 'The sickle has become dull/is dull' (Pakhalina 1975:83), from which develop inferential and mirative senses. Compare (18-a, 18-b, and 18-c) and (19-a and 19-b). It seems that, as in Kalasha and Khowar, a second, mirative, component of meaning is achieved by adding a perfect form of 'be' or 'become' (20-b). The perfect also appears typically in the opening sentence of traditional (folk) tales about the past (21). As in other languages, volitionality distinctions often emerge from the choice between simple past (22-a) or perfect (22-b).

(18-a) *salīm pešāwar reʏd-a*
Salim Peshawar go(PST)
'Salim went to Peshawar (first-hand knowledge of speaker).'

(18-b) *salīm pešāwar reXk*
Salim Peshawar go(PERF)
'Salim went to Peshawar (unseen by speaker).'

(18-c) *salīm pešāwar reXk tiwetk*
Salim Peshawar go(PERF) be(PERF)
'Apparently Salim went/has gone to Peshawar (unseen by speaker, mirative.'

(19-a) *wuδg skpIrz mōr vit-e*
today all.day rain become-PST
'It rained all day today (first-hand observation).'

(19-b) *wuδg-i mōr dyetk*
today-ps.3s rain give(PERF)
'It has rained today.' (concluded by seeing water on ground).

[7]Michailovsky (1996) feels that the presence of *re* in the function of marking hearsay prevented the expansion of the semantic space of the inferential perfect to include hearsay. Peterson (2000) considers the category of mirativity as conceptually distinct from result-inferential. He argues that the hearsay particle *re* derives ultimately from the verb *rah-* 'stay, remain', by a development *rahécha* > *récha* > *re*, involving the loss of [h] and the erosion of the unstressed final syllable *-cha*. He finds the intermediate stage attested in written documents and gives one example. Peterson's analysis differs from that of Mikhailovsky in that he considers *re* to be a further develoment and specialization of *rahecha* rather than pre-existing the inferential development of the perfect. It also differs from Degener's analysis relating Waigali *le* and Nepali *re*.

(20-a) *yem cuán-i trešp*
this apricot-ps.3s sour
'This apricot (tree) is sour (known beforehand).'
(20-b) *yem-i trešp cuan tuétk*
this-PS.3s sour apricot be(PERF)
'This is a sour apricot (discovered after tasting it, mirative).'
(21) *yi kampir-i tiwitk yi kəS Xəy də yi xun-əv yašt haletk*
one old.woman-ps.3s be(PERF) one boy self in one house-ps.3p live(PERF)
'There was an old woman and a boy. They lived in a house.' (Mock 1998:453, 215)
(22-a) *maZ-e Xü kitôb salīm-er ðet*
I(OBL1)-OBL2 self's book Salim-DAT give(PST)
'I gave my book to Salim (intentionally).'
(22-b) *maZ-e Xü kitôb salīm-er ðet-k*
I(OBL1)-OBL2 self's book Salim-DAT give(PERF)
'I gave my book to Salim (unknowingly, mistakenly).'

3.2.6 Nuristani languages

The Nuristani languages, despite the paucity of published data on some of them and the difficulty of obtaining fresh data, show clear indications of robust inferential/indirective systems.

3.2.6.1 Waigali (KalaSa-alâ)

Waigali (self-designation *kalaSa-alâ*) has a clear "reportative" particle, *–le*, first described by Buddruss (1987:33, 37) as a particle used when a speaker reports what he has not observed himself but knows by hearsay (23). Buddruss compares its function to that of Nepali *re* .

(23) *aǰ'aa isl'am na war-'aai nüstar'a kalaS'ãã-ba kas'am prú~-Ra 'eog čar oR'oi-**le***
yet Islam NEG up-came before Waigal.people swearing give-DAT specific custom was-**le**
'Islam had not yet arrived (in the valley) when (it is said that) among the early Waigal people there was a specific custom of swearing.' (Buddruss 1987:33)

Subsequently, Degener (1998:173-182) enumerates the tense forms in which *le* has been attested and discusses its functions in several text types. She compares its semantics with Turkish *miş* and with the OIA perfect. Discussing its etymology, she compares it to Nepali *re*. Degener's own description of *-le* points to its being a mirative particle. Strand (1999), in his review of Degener (1998), says that the preterital forms of 'be' given in Degener (1998:72) are not simple preterites, but "rather a marker of what [he] has called 'Realizational Mode' for neighboring Kamviri. It indicates a past change that the speaker formerly was unaware of, but at present realizes to be true. It appears most frequently with the reportative particle *-le*. English phrases like 'I realize/see/hear that...' and 'It turns out that...' indicate a similar mode." [Strand's] data lack examples of this form as an auxiliary, but it appears to form Degener's "Imperfekt II" and "Plusquamperfekt II". Waigali appears to have an extensive set of verb forms specified as inferential/indirective, at least some of which have clear mirative semantics.

3.2.6.2 Kâmviri[8]

Strand has called a set of verb forms having mirative semantics the "Realizational Mode." His paradigm (p.c.) for the realizational mode of 'be' is given as (24). Realizational forms also appear in verbs built with *âsa-* 'be' like the progressive (25).

(24)

	Sg.		Pl.	
1.	*âs'a-o-m*	'I realize that I was.'	*âs'a-o-mi?*	'I realize that we were.'
2.	*âs'a-o-?*	'I realize that you were.'	*âs'a-o-?R*	'I realize that you [pl] were.'
3.	*âs'a-o*	'I realize that he was.'	*âs'a-â*	'I realize that they were.'

(25) *b'unâso* 'I realize that it was happening.' vs. *b'unâsi* 'It was happening.'

In Strand's words: "The basic meaning of this mode contrasts current certainty with former skepticism, disbelief, or unawareness: now I really am aware of the past action or circumstance, as opposed to my former skepticism, disbelief, or ignorance." This is a clear description of mirative semantics. Regarding other aspects of inferential/ indirective semantics, Strand says: "The Realizational mode does not appear in traditional tales, which are usually told in the retrospective imaginative mode (*e m'er bAlla* 'There was [probably] a king...'). As such tales cannot be

[8] All the information on Kâmviri is due to Richard Strand (p.c.) and http://users.sedona.net/~strand/.

verified by the speaker's experience, they would preclude the Realizational mode. And it is not used for the narration of unwitnessed events (normally in the Retrospective Perfect), unless the speaker is emphasizing his realization that the unwitnessed events were verified by his later experience. The mode does imply inference from the observation of resultant states, as does the Retrospective Perfect, but it emphasizes the speaker's change of evaluation of the event from uncertain to certain." (p.c. 6/6/06)

Kâmviri also has a reportative particle *-mma*, which may be used after past-tense verbs, except the past definite, to explicitly indicate that the speaker got knowledge of the verbal action from a source other than his own inference. This particle occurs often with the Realizational mode, to indicate that outside sources led the speaker to change his mind from skepticism to belief: *b'unaso-mma* 'I hear that it really did happen [contrary to my previous belief].' Strand sees Kâmviri *-mma* as functionally equivalent to Waigali *-le*.

3.2.6.3 Ashkun (ASkuNu)

The language of village Wama (self-designation *saNu-vīri*) is one of the dialects collectively named Ashkun. Buddruss (in press) includes three texts in this dialect, which contain a significant number of verb forms which Buddruss calls Preterite-II and Imperfect-II. These forms consist of the preterite or imperfect extended with *séi*, the present tense of *s-* 'be'. Preterite II occurs in contexts typical of inferential forms in neighboring Khowar and Kalasha. Two examples of Preterite II from Buddruss' texts appear here as (26) and (27). Morgenstierne (1934:68) gave several examples of these forms, considering their meaning uncertain. However, an example occuring in one of his texts (28) shows the form occuring in the opening sentence of a fairy tale, a typical inferential/indirective context. Also, the final sentence of the same tale shows a form with mirative meaning (29). Buddruss (in press:19) also mentions a Pluperfect II, having the paradigm shown in (30).

(26) *a sə~Rú-ə zú-es kamgə'l istrí přoti-**séi***
a man fromWama-OBL daughter-PS3s Kamgal-to wife gave(PRT II.**is**)
'A man from Wama gave his daughter in marriage to (someone in) Kamgal.' (Text 1, #1) (opening sentence of traditional tale)

(27) *zəmás batúə: "oho~, yek to son sagə-**séi***
son.in.law thought aha this so gold was(PRT II.3s)
'His son-in-law thought, "Aha, so this is gold (as is heard)!' (Text 1, #11) (mirative)

(28) *a 'bādšā 'səgə-sei sə dū ı'strīRē[s] 'səgə-sən*
one king was those two his.wives were
'There was a king; he had two wives.' (Morgenstierne 1929:232)[9]

(29) *kī mr'ākwa aRs pak'īrə tāi 'zaŋgalwa pə-k'ūcāi*
that boy's his.mother the.faqir from of.the.forest from-middle
awe'Rīara 'bādšāa i'strī a-sēi
having.brought.her the.king's wife she.is
'When the faqir brought the boy's mother from the forest (she proved to be) the king's wife.' (Morgenstierne 1929:237, 221)

(30)

	Sg.	Pl.
1.	*gestə'gə-səm* 'I had gone.', etc.	*gestə'gə-səmiš*
2.	*gestə'gə-səs*	*gestə'gə-səg*
3.	*gestə'gə-sei*	*gestə'gə-sən*

3.2.7 Yasin Burushaski

Burushaski has two main dialects–Hunza and Yasin. Yasin Burushaski has a past tense form, not found in Hunza Burushaski, in which *-ásc-* (Berger *-asċ/ast-*) is infixed between the verb stem and the personal endings. This form was first noted by Lorimer (1962:26), who described it as "producing an imperfect tense". Two of the three occurrences of this form in Lorimer's texts are the first sentences of traditional folk tales (31-a), and one indicates a mirative meaning (31-b). Later, Berger (1974:40-41) describes this form as indicating something rather "vague" or "indefinite" in that the speaker has not seen (the event) himself (32), pointing to the *-asċ/ast-* form as an 'indirect' or 'inferential' form. Following Lorimer (1962), Berger thinks that this form is an influence from Khowar, which is consistent with the phonological shape and semantics of Khowar inferential forms in *as-* 'be (animate)', e.g. *astaái* 's/he turned out to be.' Tiffou and Pesot (1989:35) also attest the *-aasc-* form, commenting that its use is highly dependent on the thought of the speaker, and that its use tends to be specific to certain speakers (p.c.). (33-a) is the beginning of a traditional tale (Tiffou and Pesot 1989:94); in (33-b), on the other hand, the fact that there was a very

[9]Morgenstierne's transcription and placement of stress marks has been maintained in his examples.

big forest is given as objective information.

(31-a) *tshoor hen wau-e hen mušʏun b-aast-imi*
long.ago a old.woman-OBL a nephew be-*aast*-PST3s
'In former times an old woman had a nephew (or grandson).' (traditional tale)

(31-b) *siia, baadšaa seni ka "uule šorba axer buT nyʌm, buT mʌza mʌne-aast-imi"*
saying, king said that well(?) soup after.all very sweet very tasty remain-*aast*-PST3s
'On his saying this, the king said, "Well, the soup was very sweet and very tasty after all."' (Lorimer 1962:294(16)) (mirative)

(32) *te zamaná-ule uTánċ buT qaimát bién-asċ-imi*
that time-LOC camels very expensive be-*asċ*-PST3p
'At that time camels were very expensive.' (Berger 1974:78[8])

(33-a) *hen zamindár hír-en b-ā'sc-imi*
a farmer man-a be-ā'sc-PST.3s.hm
'There was a peasant.'

(33-b) *han buT nyu jangák-an dulúm*
a very big forest-a be(PST)3s
'There was a very big forest.'

3.2.8 Hunza Burushaski

In Hunza Burushaski, evidential meanings do not seem to be indicated morphologically. Several evidential senses are indicated by (i) a post-verbal mirative particle *qheér* (34-b), or (in conjunction with the perfect), for inference from observation (35-b), and (ii) a form *seibáan* 'they say' for indirect information from speech-act sources (34-c) or traditional knowledge.[10]

(34-a) *guté há salíim-e y-úu-e díulai*
this house Salim-OBL his-father-ERG is.building
'Salim's father is building this house (first-hand knowledge).'

(34-b) *guté há salíim-e y-úu-e díulai qheér*
this house Salim-OBL his-father-ERG is.building *qheér*
'Salim's father is building this house (speaker just came to know about it, mirative).'

(34-c) *guté há salíim-e y-úu-e díulai seibáan*
this house Salim-OBL his-father-ERG is.building they.say
'(They say that) Salim's father is building this house.' (hearsay)

(35-a) *khuulto giílt-ulo buT-an tiS gutshárimi*
today Gilgit-in great-indef wind blow(PST)-3s.y-class
'There was a storm (here) in Gilgit today.' (direct observation) (G.M. Baig, Gilgit)

(35-b) *khuulto giílt-ulo buT-an tiS gutsharilá qheér*
today Gilgit-in great-indef wind blow(PERF).3s.y-class *qheér*
'There was a storm in Gilgit today.' (e.g. concluded after seeing broken branches)

3.2.9 Pashto

Evidentiality in Pashto appears not to be expresssed morphologically. Rather, a second-position, weak-stressed particle *xo* is used for some evidential functions. It is used, along with intonation, to report an event that represents hearsay (36-b), for new and surprising information (36-c), for inference from (visual) evidence (37-b), and to report inadvertent action (38-b) (Abid Khan, p.c.).

(36-a) *dā kor də salīm plār joR kəRay de*
this house of Salim father make(PRES PERF.ms)
'Salim's father built this house.' (if speaker saw him building it)

(36-b) *dā kor xo də salīm plār joR kəRay de*
this house *xo* of Salim father make(PRES PERF.ms)
'Salim's father built this house.' (if speaker has heard this from a third party.)

(36-c) *dā kor xo də salīm plār joR kəRay de* (with changed intonation)
this house *xo* of Salim father make(PRES PERF.ms)
'Salim's father built this house.' (speaker has just come to know this new information.)

(37-a) *nən bārān šaway de*
today rain become(PRES PERF)
'It rained today.' (If speaker saw the event of raining.)

[10] In Lorimer's (1935) texts, 22 of the 32 traditional tales included include the form *seibáan* in their introductory sentences.

(37-b) *nən xo bārān šaway de*
today *xo* rain become(PRES PERF)
'It rained today.' (If inferred by seeing water on the ground.)

(38-a) *mā xpəli Toli pesi xərc ki*
I(OBL) self's all money spend(PFV)
'I spent all my money (intentionally).'

(38-b) *mā xo xpəli Toli pesi xərc ki* (with a different intonation)
I(OBL) *xo* self's all money spend(PFV)
'I spent all my money (unwittingly, by mistake-just realized it).'

3.3 Shina and Kohistani cluster

In several Shina and Kohistani dialects, evidential distinctions are marked in the pronominal system, where the seen/unseen parameter is highly developed. Correlations of use of the different pronominal forms with tense-aspect forms have not yet been studied.

3.3.1 Palula

In Palula, an archaic variety of Shina spoken in Lower Chitral, *maní*, a non-finite form of 'say', is used sentence-finally to mark a statement as hearsay (39), or to mark the opening of traditional tales (40). It also can be used in a question about speech acts (41). Compare (41-a) and (41-b). Notice that here *maní* follows the word referring to the speech act, rather than being sentence-final.

(39) *sadár chatruuL-a the ukhaandu maní*
president Chitral-OBL to coming say
'It is said that the President is coming to Chitral.' (Bashir 1996:259)

(40) *muSTú zamanee ak bachaa he~siLu maní*
former time one king was say
'Once upon a time there was a king.' (Bashir 1996:260)

(41-a) *saliim-a gubá niveširu*
Salim-ERG what wrote
'What (thing) did Salim write?' (e.g. letter, bill, etc.) (Bashir 1996:258)

(41-b) *saliim-a gubá maní niveširu*
Salim-ERG what say wrote
'What (content/words) did Salim write?' (Bashir 1996:258)

3.3.2 Gilgit Shina

In Gilgit Shina, source of knowledge distinctions are indicated analytically; hearsay information is embedded under a synchronically transparent, finite form of 'say' as in (42). The seen/unseen distinction is, however, grammaticized in a four-valued pronominal system (Radloff and Shakil, 1998:192).

(42) *salim watun (thenan)*
Salim come(PRES PERF) (say-P/F.3p)
'(They say) Salim has come/came.' (hearsay/direct observation) (M.A. Zia)

3.3.3 Kohistani Shina dialects

Schmidt (2000) and Schmidt and Kohistani (2001) provide valuable new information on the pronominal and deictic systems of Shina dialects. In Kohistani Shina the demonstratives *aáe* 'this' and *asá* 'that' are marked both for proximity/remoteness and for *source of knowledge* [emphasis mine]: *aáe* marks visual knowledge (43-a), while *asá* marks heard knowledge (43-b) (Schmidt and Kohistani 2001:136). The deictic elements *paár* 'over there, across, away' (visible to speaker or addressee) (44-a) and *pér* 'over there, across, away' (not visible to speaker or addressee) (44-b) mark the visible/non-visible parameter. Additionally, the demonstratives *aáe* 'this' and *asá* 'that' and the deictics *paár* and *pér* combine and interact to produce various emergent meanings. In these interactions source-of-knowledge marking overrides proximity/distance marking; in such cases stress shift specifies the degree of distance (45-a, 45-b). In both (45-a) and (45-b) the knowledge-source marking (seen) in the element *-aáe* overrides distance. Relative distance is conveyed by placing the stress on *aáe* for the closer distance (45-a) and on *paár* for the farther distance (45-b) (Schmidt, 2000:210).

(43-a) *aáe jók-un*
this what-is
'What is this (thing)?

(43-b) *asá déez-i-ji patá*
that day-OBLsg-ABL after
'since that day' (Schmidt and Kohistani 2001:136)

(44-a) *phúuT th-áa-o to paár dúu tobí-in-a*
look do(PERF)3md.sg. TOP over.there [visible] two tree-pl-are(f)
'[As] he looked [he thought], "Over there are two trees" (in speaker's line of sight).'
(44-b) *pér bo waá*
away [invisible] go(IMP) EMPH
'Go away!'
(45-a) *mõ paar-aáe váari bój-m-as*
I over.there (close, seen) direction go-IMPF-1sg
'I am going over there (a short distance in the speaker's line of sight).'
(45-b) *mõ paár-aae váari bój-m-as*
I over.there (distant, seen) direction go-IMPF-1sg
'I am going way over there (a longer distance in the speaker's line of sight).'

In Tileli Shina there are four third-person pronouns (Schmidt 2000:202), specified for visible or known/invisible or unknown, and for close visible or remote visible. Schmidt (2000:212) concludes: "In both [Kohistani and Tileli Shina] three degrees of distance may be distinguished, with either visibility or line-of-sight location as an additional parameter, although these parameters are mapped on to different pronouns or deictics. Both the Tileli and Kohistani data testify to a third parameter: the source of knowledge. In Tileli, 'source' discriminates first and second-hand knowledge. First-hand knowledge is mapped onto visibility: it requires *Zo*, whereas second-hand knowledge or inference is mapped on to invisibility, and requires *so*. In Kohistani, 'source' discriminates information derived by visual means from information known by some other means. Visual source is mapped on to the proximate demonstrative, while non-visual source is mapped on to the remote demonstrative."

3.3.4 Indus Kohistani

According to Claus Peter Zoller (p.c.) the Indus Kohistani pronominal system is complicated, and the seen/unseen parameter is linked with concepts like 'inside/outside' and 'stationary/moving'.

3.3.5 Kalam Kohistani[11]

Hearsay and mirative meanings (46-b), and indirect knowledge (47-b, 48-b) are indicated by a sentence-final particle *-yer* (46-b), which appears to be from a defective verb *-ar-* 'say', which now exists only in past tense forms: *yɨ maro* 'I said', *tu aro* 'you(sg.) said', *sɨ aro* 's/he said', *ma maro* 'we said', *tha aro* 'you(pl.) said', *tām aro* 'they said'.

(46-a) *salīm-a bɔɔb-a ī~ šīT cēg*
Salim-OBL father-ERG this house(f) build(PST.f.sg.)
'Salim's father built this house (speaker witnessed event).'
(46-b) *salīm-a bɔɔb-a ī~ šīT cēg-yer*
Salim-OBL father-ERG this house(f) build(PST.f.sg.)-perhaps
'Salim's father built this house (hearsay, or new information).'
(47-a) *kyálam-a lām-mey bāra kucúr thū*
Kalam-OBL village-in many dog are (cf. *haiN*)
'There are lots of dogs in Kalam village (known to speaker from first-hand experience.).'
(47-b) *zaraafa afriqe~ za[eng]gil-mey waan-ar*
giraffes Africa jungle-in are-*yer* (cf. *hote haiN*)
'There are giraffes in Africa (presumably not direct knowledge).'
(48-a) *šə'm-aa dós-a a járga thū/wɔɔn*
Saturday-OBL day-LOC one meeting is/will be
'There's a meeting on Saturday. (If I helped to arrange it.)
(48-b) *šə'm-aa dós-a a járga wɔɔn-t-ər*
Saturday-OBL day-LOC one meeting will be-*yer*
'There's a meeting on Saturday. (If I read an announcement about it.)

3.3.6 Torwali

In Torwali, spoken in the Swat Valley, two particles indicating evidential meanings have been identified so far.[12] (i) A sentence-final particle *a* is employed in all tenses for sentences representing information acquired indirectly (49-b). A particle *ko* marks information acquired by inference from visual evidence (50-b). At this point I do not

[11] Data on Kalam Kohistani are based on field work done working with Amir Zada, an educated resident of Kalam.

[12] All the Torwali data in this paper were provided by Inam Ullah, of village Bahrain, Swat.

have enough contextualized data to say anything further about the uses of these particles.

(49-a) *miTiŋ ləu aŋa-si di chi*
meeting Saturday-of day is
'There's a meeting on Saturday.' (If speaker helped to arrange the meeting.)

(49-b) *miTiŋ ləu aŋa-si di chi-a*
meeting Saturday-of day is-*a*
'There's a meeting on Saturday.' (If speaker read about it in the newspaper.)

(50-a) *aš agha mut-tu*
today rain rain(PRES.PFV)
'It rained today.' (If event of raining seen by speaker)

(50-b) *aš agha mut-tu ko*
today rain rain(PRES.PFV) *ko*
'It rained today.' (If inferred by seeing water on the ground)

3.4 Balochi and Brahui

The status of evidentiality in Balochi and Brahui is unclear.[13]

3.5 Urdu and Hindi

In Hindi and Urdu, indication of evidentiality/inferentiality semantics is distributed throughout the grammar. It is associated with at least three morphological patterns: (i) compound verb vs. simple verb, (ii) tense marked perfective vs. simple perfective; (iii) *na* vs. *nahĪ*. Hook (1974), an analysis of the compound verb, is relevant to the meanings discussed in this paper. Hook says (1974:248), "In cases where the performance of an action is completely unforeseen by the speaker he may not use the compound verb." Again, (1976:153): "If there is no possibility of an action or event's being anticipated, it is expressed with the non-compound verb." Two of his examples appear as (51-a) and (51-b).

(51-a) *kalambas ne amrīkā kī khoj kī / *kar dī/ * kar lī*
Columbus ERG America of discovery(f.s.) do(PFV)f.s. /*do-give/*do-take
'Columbus discovered America.' (Hook 1974:240)

(51-b) *kal dūdh me~ cūhā milā*
yesterday milk in mouse(m.s.) meet(PVF)m.s.
'Yesterday we found a mouse in the milk.' (Hook 1976:153)

In (51-a) a compound verb would suggest that Columbus knew about the existence of America before discovering it; in (51-b) a compound verb (*mil gayā*) would suggest that the speaker anticipated or feared finding a mouse in the milk. In other words, mirative semantics is not compatible with the compound verb in *jānā* 'go'. In Bashir (1993), I argued that the distribution of compound verbs vis-à-vis simple verbs is related to the intersection point of an observer/speaker with an event vector. Compound verbs encode actions specified for intersection with more than one point on the vector, e.g. both origin and event, while simple verbs encode actions as an undifferentiated single-stage conception, e.g. the event itself, or the end point/resultant state. A single-stage conception including only the end point gives rise to mirative semantics.

Montaut (2001:351), comparing the semantics of the present perfect (perfective participle + present tense of 'be') with that of the "aorist" (simple perfective = perfective *-(y)ā* participle) argues that actions or events represented with the aorist are disjunct from the speaker's present (moment of speech) because of the lack of a tensed auxiliary, which would anchor the reported event to the speaker's reference time. Thus mirativity–meanings which 'are grasped through a sudden irruption in the consciousness'–emerges for the simple perfective form. Montaut's examples (52-a) and (52-b), contrast the meaning of surprise ("as when opening the door and seeing an old friend accompanied by his young son not seen for long") in the simple perfective, (52-a), with the response to it in (52-b), which is rooted in the respondent's (prior) connection to the event (Montaut 2004:106). Bashir (2003) notes that absence of the auxiliary has the same effect in the present progressive, as in (53), uttered when a speaker, telephoning someone and expecting someone to answer, is surprised when no one picks up the telephone.

[13]Rossi (1989), citing Windfuhr's (1982) discussion of inferentiality in Persian, argues on the basis of elicited Balochi sentences patterned on sentences in Windfuhr (1982) that in the Balochi of Chakansar/Kang (influenced by Dari Persian according to the informant), some verb forms are used with inferential meaning. However Sabir Badalkhan (p.c.16 April 2006) does not accept this, especially for Pakistani Balochi. Lazard (2000) discussing Barker and Mengal (1969), which is based on Pakistani Balochi, thinks that the meanings of B&M's Past II ("Past Completive") and Past Perfect II ("Past Perfect Completive") may be related to the evidential system; however, Sabir Badalkhan does not see evidential meaning in these examples. This question needs text-based research. Regarding Brahui, I have not yet been able to identify any forms or constructions which convey evidential/indirective meanings; the question remains open.

(52-a) *are ! kitnā baRā* ***ho gayā*** !
interj. how.much tall become(aor)
' Oh, he has grown so tall! / how tall he has grown! '

(52-b) *vah kāfī baRā* ***ho gayā hai***
3s fairly/rather tall become(pres.perf)
'He has grown quite tall.' (Montaut 2001:352)

(53) *koī* ***uThā nahĪ rahā***
anyone lift NEG remain (PFV-ms)
'No one is answering.' (contemporary Pakistani Urdu) (Bashir 2003)

Additionally, Bashir (2003), a study of the negative elements *na* and *nahĪ*, tentatively finds that with the loss of its unmarked/default status, in contemporary Pakistani Urdu *na* is specializing to some degree into the negative marker associated with mirativity or non-volitionality. It appears that *rahnā* 'remain' and *rakhnā* 'put' are used in some cases with mirative nuances. This awaits further investigation.

3.6 South Indian (Dravidian/Dakkhiini/Marathi) cluster

3.6.1 Malayalam

In Malayalam, evidentiality distinctions are not morphologically encoded, but are scattered throughout the grammar. Some of the means noted so far are: (i) use of a verbal noun rather than a finite past tense form (54-b); (ii) a particle *allō*, which has a range of meanings including softening a harsh statement, adding certainty, or adding surprise (54-c); (iii) the perfect. An event directly witnessed is expressed with the simple past, whereas one inferred from observation of the results is expressed with a perfect form interpretable as 'must have V-ed', an inference based on the speaker's knowledge of the world (55-b).[14]

(54-a) *Rāman-ṟe acchan ī vīTu nirmmiccu*
Raman-GEN father(NOM) this house build(PST)
'Raman's father built this house.' (Speaker saw him building it.)

(54-b) *Rāman-ṟe acchan ī vīTu nirmmiccu kēTTu*
Raman-GEN father(NOM) this house build(VERBAL NOUN)
'Raman's father built this house.' (Speaker has learned this from a third party.)

(54-c) *Rāman-ṟe acchan vīTu nirmmikkunnuNT-allō*
Raman-GEN father(NOM) house build(PRES)-*allō*
'Raman's father is building a house.' (Speaker has just come to know this.)

(55-a) *innu mazha illa*
today rain became(PST)
'It rained today.' (event of raining seen by speaker)

(55-b) *innu mazha peytāyirikkum*
today rain must.have.fallen
'It (must have) rained today.' (inferred by seeing water on the ground)

3.6.2 Tamil

Several strategies mark evidentiality. (i) The particle *-ām* marks information attributed to a third-party speech-act source, either aural or written. It functions in all tenses.[15] Compare (56-a) and (56-b). (ii) A second construction used to mark hearsay attribution involves the quotative particle *enRu*, the conjunctive participle of 'say', with the form *kēLvi* (< 'hear') marking question or hearsay (56-c). (iii) The present perfect can be used to indicate inferences, i.e. conclusions based on observation of results of an event (57). (iv) The frozen particle *-pōla* 'it seems that' can indicate mirative senses in all tenses (58). (v) A frozen form *vēNTum* 'must' functions in mirative meanings (59).

(56-a) *vīran inta vīT-ai-k kaTT-in-ān*
Viran this house-ACC build-PST-3sm
'Viran built this house.' (personally known)

(56-b) *vīran inta vīT-ai-k kaTT-in-ān-ām*
Viran this house-ACC build-PST-3sm-HEARSAY
'I gather/hear, that Viran built this house.' (hearsay)

[14] I am indebted for the Malayam examples and discussion to Nisha Kommatam, Lecturer in Malayalam, University of Chicago.

[15] The lexical source of *-ām* is not certain, but according to J. Lindholm (p.c.) it may be from the root *-āhi* 'to be, become'. I am grateful to V.J. Fedson for the Tamil examples in this section.

(56-c) *vīran inta vīT-ai-k kaTT-in-ān enRu kēLvi*
Viran this house-ACC build-PST-3sm say(CP) question/hearsay
'The *on dit* is that Viran built this house.'

(57) *vīran-kku pustakatt-ai koTuttu iru-kkir–ēn*
Viran-DAT book-ACC give(PRES PERF)-1s
'I've evidently, obviously (unknowingly/mistakenly) given Viran the book'.

(58) *vīran inta vīT-aik kaTTu-kir-ān-pōla*
Viran this house-ACC build-PRES-3s.m-it.seems
'It looks as if/seems as if Viran is building this house.' (Speaker has just learned this.)

(59) *en pustakatt-ai avan-ukku nān koTuttu irukka vēNTum*
my book-ACC he-DAT I give CP be(INF) must(frozen)
'I must have (inadvertently) given my book to him.' (For example, I've forgotten that I did.)

3.6.3 Telugu

As in Malayalam and Tamil, marking of evidential meanings is scattered, including: (i) the particle *anTa* 'saying'; (ii) a surprise particle *-ē*; (iii) the morpheme *-aTl-* 'like'.[16] *anTa* 'saying' functions to indicate hearsay (60-b), and other types of indirect knowledge. Although no meaning of reduced belief in the statement is inherent in statements with *-anTa*, it can be used as a discourse strategy to distance the speaker from responsibility for a statement, and to quote proverbs. In reporting the actions of a third person, in combination with the emphatic marker *-ē*, *anTa* can yield a mirative-like meaning (60-c); with a first-person speaker, *-ē* alone can evoke the mirative sense (61). In a case like (62-b), *-anTa* is not obligatory, and would be used only if the indirective sense is focused. *-aTl-* 'like', which follows the non-finite verbal element, can indicate indirect knowledge of events or situations acquired from sources other than (extended) speech. Thus in (63-b) it indicates inference from observation of a resulting state, while in (64-b), with a first-person agent, the nuance of inadvertent action emerges.

(60-a) *salīm vāLl-a nānna ī illu kaTT-inc-ā-Du*
Salim ones-OBL father this house build-CS-PST-3sm
'Salim's father built this house.' (Speaker saw him building it).

(60-b) *salīm vāLl-a nānna ī illu kaTT-inc-āD-anTa*
Salim ones-OBL father this house build-CS-PST(3s)-SAY
'Salim's father built this house.' (Speaker has heard this from a third party.)

(60-c) *salīm vāLla nānna ī illu kaTT-inc-āD-aTa-n-ē*
Salim ones-OBL father this house build-CS-PST-3sm-say-n-EMPH
'Salim's father built this house.' (Speaker has just come to know this information.)

(61) *ayyē Tēpī peTTu-kō-v-aDam marci-pō-y-ā-n-ē*
Oh hat put-REFL(GER) forget-go-y-1s-n-EMPH
'Oh, I forgot to put on my hat.' (Said in surprise)

(62-a) *haidarābād-u-lō gurrā-lu unn-ay*
Hyderabad-u-LOC horse-pl be(PRES)-3p.n-h
'There are horses in Hyderabad.' (presumably first-hand knowledge)

(62-b) *āfrikā-lō jirāfi-lu unn-āy-(anTa)*
Africa-LOC giraffe-p. be(PRES)-3p.n-h-(saying)
'There are giraffes in Africa.' (presumably indirect knowledge)

(63-a) *ivvāLla vāna kurisin-di*
Today rain shower(PST PPL)-3s.n-h
'It rained today.' (If the event of raining was seen by the speaker.)

(63-b) *ivvāLla vāna kurisin-aTl-un-di*
Today rain shower(PST PPL)-like-be(PRES)-3s.n-h
'It rained today.' (If the event of raining inferred by seeing water on the ground.)

(64-a) *nēnu nā pustakam salīm-ku icc-ā-nu*
I my book Salim-DAT give-PST-1s
'I gave my book to Salim (intentionally).'

[16] Telugu examples and discussion are due to Nagaraj Paturi, Fellow, Centre for Folk Culture Studies, School of Social Sciences, University of Hyderabad, India.

(64-b) *nēnu nā pustakam salīm-ku iccin-aTl-unnā-nu*
I my book Salim-DAT give(PST PPL)-like-be(PRES)-1s
'I gave my book to Salim (unknowingly, mistakenly).' (lit. 'It seems that I've given my book to Salim.')

3.6.4 Kannada

Several morphemes function to convey evidential meanings. (i) *-ante* < 'say' functions for hearsay and mirative when the new information is acquired from a speech act of someone else (65-b), and indirectly known events or states (66-b). In a case like (66-b), *-anTa* is not obligatory, and would be used only if the indirective sense is focused. (ii) *-ē* (emphatic) functions only to indicate surprise, not as a general mirative. (iii) *-ante* + *-ē* indicates new information (65-c). (iv) *-anga-* 'like' can report the traces of an unseen event if it is inferred from evidence other than (extended) hearing. In (67-b) *anga* indicates an inference from a visually observed result; while in (68-b) it indicates indavertent action. (v) A form *nōD-appa*, literally 'see-man' appears in the first-person mirative context as in (69) where it expresses surprise at an inadvertent action. All Kannada materials and judgements here are due to Nagaraj Paturi.

(65-a) *salīma-avara appa ī mane-yan-nu kaTTi-s-ida*
Salim's-OBL father this house-yan-ACC build-CS-PST3sm
'Salim's father built this house.' (Speaker saw him building it.)
(65-b) *salīma-avara appa ī mane-yan-nu kaTTi-s-ida-n-ante (< kaTTisidanu + ante)*
Salim's-OBL father this house-yan-ACC build-CS-PST3sm-*n*-say
'Salim's father built this house.' (Speaker has heard this from a third party.)
(65-c) *salīma-avara appa ī mane-y-annu kaTTi-si-da-n-anta-n-ē* (< *kaTTisidanu* + *ante*)
Salim's-OBL father this house-y-ACC build-CS-PST3s-n-say-n-SURPRISE
'Salim's father built this house.' (Speaker has just come to know this information.)
(66-a) *maisūri-n-alli kudure unTu*
Mysore-n-LOC horse are
'There are horses in Mysore.' (Presumably this is first-hand knowledge.)
(66-b) *āfrika-n-alli jīrāfi-gaLu unT-ante* (< *unTu* + *ante*)
Africa-n-LOC giraffe-pl be(PRES)3p.n-h-say
'There are giraffes in Africa.' (Presumably this is non first-hand knowledge.)
(67-a) *ivattu maLe suritu*
today rain pour(PST)3s.n-h
'It rained today.' (If the event of raining was seen by the speaker.)
(67-b) *ivattu maLe surid-ang-ide* < (*suritu* + *anga* + *ide*)
today rain pour(PST PPL)-like-be(PRES)3s.n-h
'It rained today.' (Inferred, for example, by seeing water on the ground.)
(68-a) *nānu nanna pustaka-(na) salīma-ge~ koTTe*
I my book-(ACC) Salim-DAT give(PST)1s
'I gave my book to Salim (intentionally).'
(68-b) *nānu nanna pustaka-(na) salīma-ge~ koTT-ang-iddīni* (<*koTTu* + *anga* + *iddīni*)
I my book-(ACC) Salim-DAT give(PST PPL)-like-be(PRES)1s
'I (unknowingly, mistakenly) gave my book to Salim.'
(69) *ē nānu Topi-yan-nu iTTu-koLluvaDu maratu-biTTe nōD-appā*
Oh I hat-yan-ACC put-take(GERUND) forget-leave see-man
'Oh, I forgot to put on my hat!' (Uttered in surprise.)

3.6.5 Dakkhini Urdu[17]

Several forms serve to mark evidential meanings in Dakkhini. (i) *katē* 'it is said' is obligatory in hearsay and second-hand information contexts (70-b). (ii) The particle *rē / rī* (masculine/feminine addressee) specifically indicates surprise, not merely new information (70-c). *katē* marks indirectness only. It is invariant and can occur in all tenses, including equational sentences and embedded questions (71). In (71), the speaker (A) assumes that the addressee (B) will have indirect rather than first-hand knowledge of who is to come with the bride, hence A's use of *katē*. *katē*

[17]The information on Dakkhini in this paper is due to Nagaraj Paturi. In Dakkhini, aspiration is lost, even in voiceless stops. There is no palatal sibilant, but sometimes the retroflex sibilant is heard, e.g. *pešāb* 'urine' > *peSāb*. The reflexive (*apnā*) is only rarely used. There is no agentive postposition *ne*, and the verb agrees with the subject, even in perfective transitive sentences. According to Paturi, existing Dakkhini literature includes mostly folklore, and there is no new written literature being composed in Dakknini. Dakkhini is used on the radio, but only for satire, local color, or local characters. It is, however, vital as a spoken language.

is also used to quote proverbs (72), and also in utterances involving recalled speech, e.g. (73), which represents the soliloquy of a woman recalling her husband's hurtful words. (iii) Inference from evidence to an unseen event is indicated by the invariant, sentence-final particle *sarkā* 'like', which can also indicate an impression or belief from any source: visual (74-b, 75), auditory, or the imagination. *sarkā* also occurs in the first-person mirative/non-volitional context (76-b).

(70-a) *Salīm-kā bā is gar-kō banāyā*
Salim-of father this house-ACC made
'Salim's father built this house.' (Speaker saw him building it.)

(70-b) *salīm-kā bā is gar-kō banāyā katē*
Salim-of father this house-ACC made it.is.said
'Salim's father built this house.' (Speaker has heard this from a third party.)

(70-c) *salīm-kā bā is gar-kō banāyā katē rē*
Salim-of father this house-ACC made it.is.said SURPRISE
'Salim's father built this house.' (Speaker has just learned this surprising information.)

(71) A: *baccī kē sāt kōn ārāy katē* B: *salīm katē*
bride with who is.coming it.is said Salim it.is.said
A: 'Who (do they say) is coming with the bride?' B: 'They say Salim (is coming).'

(72) *purānā marīz adā hakīm āy katē*
old patient half doctor is it.is.said
'A long-standing patient is half a doctor.'

(73) *mai sēmār-ū~ katē - vō accī ai katē - mai mar jānā katē -*
'I am lazy (he says) - She is good. - I should die.' (recalled words of husband)
mai kaiku marū~ - uskō mārkē-ī marū~gī
'Why should I die!?- I will die only after killing him/her.' (speaker's thoughts)

(74-a) *āj pānī āyā*
today water came
'It rained today. (If the event of raining was seen by the speaker.)

(74-b) *āj pānī āyā sarkāy* (< *sarkā* + *hāy*)
today water came it.is.like
'It rained today.' (lit. 'It seems like it rained today.' If the event was inferred by seeing water on the ground.)

(75) *gayā sarkāy*
went it.is.like
'It seems like he has gone.'

(76-a) *mai mērā kitāb salīm-kō diyā*
I my book Salim-DAT gave
'I gave my book to Salim (intentionally).'

(76-b) [*mai mērā kitāb salīm-kō diyā*] *sarkāy*
I my book Salim-DAT gave it.is.like
'I (unknowingly, mistakenly) gave my book to Salim.'(unknowingly, mistakenly)

3.6.6 Marathi

In Marathi, several evidential strategies are found: (i) *mhaNe*, a quotative particle from 'say' indicates hearsay information (77-b)[18] (ii) The present perfect appears with new information (77-c). (iii) The difference between direct (78-a) and indirect (78-b) knowledge is encoded by the use of the present or the imperfect, which is used for information about remote objects, not directly knowable.[19] (iv) *-ē* 'exclamation' appears in combination with *are* 'exclamation' for inference about events from visible results/mirative (79-b). An event inferred from a visible resultant state can be reported with the subjunctive (80-b) as opposed to the simple perfective/past (80-a), or with the 'surprise' particle *-e*.

[18]My information and Marathi examples are due to Philip Engblom, Lecturer in Marathi, University of Chicago. Engblom observes (p.c.): "My sense is that the *mhaNe* here is rather falling out of use in urbanized, educated Marathi. Some people do use it more consistently than others."

[19]Marathi *asNo~* 'to be' has three forms in the present tense. The first form, e.g. *mī~ āheN* 'I am' from the root *as* 'be' is used to express the existence of objects (in a location) or their properties. The second form, e.g. *mī hoy* 'I am', from the root *bhu-*, 'become' is used for affirming the qualities of objects. The third form, e.g. *mī asto* 'I (m) usually am', from the root *as* 'be', usually has the sense of present habitual or continuous action; it is called 'imperfect' by Engblom.

(77-a) *salīm-cyā vaDilā-nnī he ghar bāndh-lā*
Salim-GEN father-AG this house build(PFV)
'Salim's father built this house.' (direct knowledge)

(77-b) *salīm-cyā vaDilā-nnī he ghar bāndh-lā mhaNe*
Salim-GEN father-AG this house build(PFV).m.s. they.say
'Salim's father built this house.' (Speaker has heard this from a third party.)

(77-c) *salīm-cyā vaDilā-nnī he ghar bāndh-le-lā āhe*
Salim-GEN father-AG this house build-take-PFV be(PRES)3s
'Salim's father has built this house.' (Speaker has just come to know this.)

(78-a) *mahārāShTra-at vāgh āhet*
Maharashtra-LOC tiger be(PRES.1st.form)3p
'There are tigers in Maharashtra.' (This is presumably first-hand knowledge.)

(78-b) *āphrik-et jirāph astāt*
Africa-LOC giraffe be(IMPERF)3p
'There are giraffes in Africa.' (This is presumably indirect knowledge.)

(79-a) *majhy-ā bhāvā-nī salīm-lā patra lihi-lā*
my-OBL brother-AG Salim-DAT letter write-PST INDEF(m.s.)
'My brother wrote a letter to Salim.' (If I saw him writing it, for example.)

(79-b) *are, mājhyā bhāvā-ni salīm-lā patra lihi-lā-e!*
Oh, my(m.s.) brother-AG Salim-DAT letter write-PST INDEF(m.s.)-EXCLAM
'Oh, My brother's written a letter to Salim.' (If I learned this by seeing the letter on his desk, for example.)

(80-a) *āj pāus paD-lā*
today rain fall-PST INDEF(m.s.)
'It rained today.' (Event of raining was seen by the speaker.)

(80-b) *āj pāus paD-lā asāvā*
today rain fall-PST INDEF(m.s) be(SUBJ)m.s
'It rained today.' (If the event of raining was inferred by seeing water on the ground.)'

4. Summary

In a Northern cluster including Kalasha, Khowar, Tajik Persian, Wakhi, and perhaps Yasin Burushaski, Type I systems are found. In a Southern cluster, evidential strategies, including mixed types, include developments of 'say' into "hearsay" markers. The evidentiality/indirectivity marking systems of the southern cluster of languages are remarkably parallel (Table 1). Tamil, Kannada, Telugu and Dakkhini employ a marker renderable as 'like' in the senses of first-person mirative, and inference from evidence other than that of (extended) speech. A quotative-like form from 'say' appears in Tamil, Kannada, Telugu, Dakkhini, and Marathi. In the South Indian cluster, insofar as different markers are used for information from speech-act and non-speech sources, the system can be said to resemble a Type II system in which the source of information is specified. The pronominal systems in some dialects of Shina also appear to have some Type II-like characteristics.

Table 1. Evidentiality/indirectivity marking forms in South Indian languages

Language	form < 'say'	'like' form	surprise/ emphatic particle	'must' form	form < 'become'	perfect for inference
Malayalam	?	?	*allō*	?	?	yes
Tamil	*enRu, k ēLvi*	*pōla*	*-ē*	*v ēNTum*	*-ām*	yes
Kannada	*ante*	*anga*	*-ē*	?	?	?
Telugu	*anTa*	*aTl*	*-ē*	?	?	?
Dakkhini	*katē*	*sarkā*	*are ...rē*	?	?	?
Marathi	*mhaNe*	?	-ē	?	?	yes

Table 2 briefly compares information available to me about the expression of evidentiality and inferentiality in some South Asian languages.

Table 2. Evidential forms and meanings in some South Asian languages

	Mirative		Indirective/inferential		
Language	non-1st person	1st person	Hearsay	Inference from result	Traditional knowledge
Vedic	?	?	particle *kila*	verb sys.- perf.	?
Panini's Skt.	?	?	?	verb sys.- perf.	?
Prakrit, Pali	?	?	particle *kila*	?	?
Kalasha	verb sys.-I forms; *huLa* < 'become'	verb sys.-I forms; *huLa* < 'become'	verb sys.-I forms; *ghoan* 'they say'	verb sys.-I forms; *huLa* < 'become'	verb sys.-I forms
Khowar	verb sys.- I forms, *birái* < 'become'	verb sys.- I forms, *birái* < 'become'	verb sys.- I forms, *birái* < 'become'	verb sys.- I forms, *birái* < 'become'	verb sys.- I forms, *birái* < 'become'
Yasin Burushaski	verb sys.- infix < 'be'	?	verb sys.- infix < 'be'	?	verb sys.- infix < 'be'
Wakhi	verb sys.-perf. + perf. < 'be'	verb sys.-perf. + perf. < 'be'	verb sys.-perf.	verb sys.-perf.	verb sys.-perf.
Tajik Persian	verb sys.- perf.	verb sys.- perf.	verb sys.-perf.	verb sys.-perf.	?
Ashkun (Nuristani)	verb sys. (< 'be')	verb sys. (< 'be')	?	?	verb sys. (< 'be')
Kâmviri (Nuristani)	verb sys.- 'realizational'	verb sys.- 'realizational'	particle *mma*	verb sys.	verb sys.
Waigali (Nuristani)	verb sys.; particle *le*	verb sys.; particle *le*	particle *le* (< *kila*?)	?	particle *le*
Nepali	verb sys.- infer. perf.; infer. copula *rahecha*	verb sys.- infer. perf.; infer. copula *rahecha*	particle *re* (< *kila*?)	verb sys.- infer. perf.	?
Hunza Burushaski	particle *qheér*	particle *qheér*	*seibáan* < 'say'	particle *qheér*	*seibáan* < 'say'
Gilgit Shina	?	?	analytical < 'say'	analytical ?	?
Kohistani Shina	?	?	pronominal sys.	?	?
Tileli Shina	?	?	pronominal sys.	?	?
Palula	?	?	*maní* (< 'say')	?	*maní* (< 'say')
Torwali	?	?	particle *-a*	particle *-ko*	?
Kalam Kohistani	particle *yer* (< 'say')	intransitive construction	particle *yer* (< 'say')	?	?
Pashto	particle *xo* + intonation	particle *xo* + intonation	particle *xo* + intonation	particle *xo* + intonation	?
Tamil	*pōla* 'seems'	*pōla* 'seems'; *vēNDum* 'must'	suffix *-ām*; *enRu kēlvi*	verb sys.- perf.	?
Malayalam	surprise particle *allō*	surprise particle *allō*	verbal noun	perf.	lexical

	Mirative		Indirective/inferential		
Language	non-1st person	1st person	Hearsay	Inference from result	Traditional knowledge
Telugu	*anTa* < 'say' + *ē* (surprise)	*aTl* 'like'	*anTa* < 'say'	*aTl* 'like'	?
Kannada	*ante* < 'say'	*anga* 'like"; *noD appa* 'see man'	*ante* < 'say'	*anga* 'like'	?
Dakkhini Urdu	*sarkā* 'like, seems'; *katē* < 'say'	*sarkā* 'like, seems'	*katē* < 'say'	*sarkā* 'like, seems'	*katē* < 'say'
Marathi	pres. perf. (+ ē surprise)	adverb; intrans. constr.	*mhaNe* < 'say'	subjunctive	?
Hindi and Urdu	absence of pres. AUX; simple verb	absence of pres. AUX; simple verb	*sunā* 'heard', *kahtē haiN* 'they say'	*lagnā* 'seem, like'	*kahtē haiN* 'they say'

References

Aikhenvald, Alexandra. 2003. Evidentiality in typological perspective. In: Aikhenvald and Dixon (eds.), 1-31.

_____ and Dixon, R.M.W. (eds.) 2003. *Studies in Evidentiality.* Amsterdam: Benjamins.

Aoki, Haruo. 1986. Evidentials in Japanese. In: Chafe and Nichols (eds.), 223-238.

Barker, Abd-al-Rahman and Aqil Khan Mengal. *A Course in Baluchi.* Montreal: Mc Gill University.

Bashir, Elena. 1988a. Inferentiality in Kalasha and Khowar. *CLS 24*. Chicago: Chicago Linguistics Society, 47-59.

_____. 1988b. *Topics in Kalasha syntax: An Areal and Typological Perspective.* Ph.D. Dissertation, University of Michigan. Ann Arbor: University Microfilms. (http://wwwlib.umi.com/dissertations)

_____. 1993. Causal chains and compound verbs. In: Manindra K. Verma (ed.), *Complex Predicates in South Asian Languages*. New Delhi: Manohar.

_____. 1996. Mosaic of Tongues: Quotatives and complementizers in Northwest Indo-Aryan, Burushaski and Balti. In: Hanaway, William L. and Wilma Heston (eds.) *Studies in Pakistani Popular Culture.* Lahore: Sang-e-Meel Publications and Lok Virsa Publishing House, 187-286.

_____. 2003. *na* and *nahĪ* in Hindi and Urdu. Paper presented at the 23rd annual meeting of the South Asian Languages Analysis Roundtable (SALA 23), at the University of Texas, Austin, October 10-12, 2003.

Berger, Hermann. 1974. *Das Yasin-Burushaski*. Wiesbaden: Harrassowitz.

Buddruss, G. 1987. Ein Ordal der Waigal-Kafiren des Hindukusch. *Cahiers Ferdinand de Saussure 41*:31-43. Geneva: Librarie Droz.

_____. in press. Drei Texte in der Wama-Sprache des afghanischen Hindukusch. To appear in *Gedenkschrift Grjunberg*.

Cardona, George. 2002. The Old Indo-Aryan tense system. *Journal of the American Oriental Society* 122(2):235-241.

Chafe, W. and Nichols, J. (eds.). 1986. *Evidentiality: The Linguistic Coding of Epistemology.* Norwood, NJ: Ablex.

Clark, T.W. 1963. *Introduction to Nepali.* Cambridge: W. Heffer and Sons.

Degener, Almuth. 1998. *Die Sprache von Nisheygram im Afghanischen Hindukusch.* Wiesbaden: Harrassowitz.

Delancey, Scott. 1982. Aspect, transitivity and viewpoint. In: P. Hopper (ed.), *Tense-aspect: Between Semantics and Pragmatics*. Amsterdam: Benjamins.

_____. 1985. Lhasa Tibetan evidentials and the semantics of causation. *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society.* Berkeley, California: Berkeley Linguistics Society.

_____. 1986. Evidentiality and volitionality in Tibetan. In: Chafe and Nichols (eds.), 203-213.

_____. 2001. The mirative and evidentiality. *Journal of Pragmatics.* 33:369-382.

Deshpande. M. 1981. PāNini and the Vedic evidence: A peep into the "past". *Golden Jubilee Volume*. Poona: Vaidika Samśodhana MaNDala, 52-65.

Emeneau, Murray B. 1969. Sanskrit syntactic particles - *kila*, *khalu*, *nūnam*. *Indo Iranian Journal* 11: 241-268.
Guentchéva, Z. (ed.) *Le Énonciation Médiatisée*. Louvain and Paris: Peeters.
Johanson, Lars. 2003. Evidentiality in Turkic. In: Aikhenvald and Dixon (eds.), 273-290.
_____. 2000. Turkish indirectives. In: Johanson and Utas (eds.), 61-87.
_____ and Bo Utas (eds.). 2000 *Evidentials: Turkish, Iranian and Neighbouring Languages*. Berlin: Mouton de Gruyter.
Lazard, Gilbert. 1985. L'inferentiel ou passé distancie en persan. *Studia Iranica* 14:27-42.
_____. 1996. Le médiatif en Persan. In: Guentchéva (ed.), 21-30.
_____. 1999. Mirativity, evidentiality, mediativity, or other. *Linguistic typology* 3:91-110.
_____. 2000. Le médiatif: considérations théoriques at application à l'iranien. In: Johanson and Utas (eds.), 209-228.
Lorimer, D. L. R. 1935. *The Burushaski Language, Vol. II. Texts and Translations*. Oslo: H. Aschehoug & Co.
_____. 1962. *Werchikwar English Vocabulary*. Oslo: Universitets Forlaget.
Michailovsky, Boyd. 1996. L'inférentiel du Népali. In: Guentchéva (ed.), 109-123.
Mock, John Howard. 1998. The Discursive Construction of Reality in the Wakhi Community of Northern Pakistan. Ph.D. Dissertation, University of California, Berkeley.
Montaut, Annie. 2001. On the aoristic behaviour of the Hindi/Urdu simple past: From aorist to evidenciality. In: D. Lönne (ed.), *Tohfa-e-Dil: Festschrift Helmut Nespital*. Reinbek: Wezler, 345-364.
_____. 2004. *A Grammar of Hindi*. Munich: Lincom-Europa.
Morgenstierne, Georg. 1929. The language of the Ashkun Kafirs. *Norsk Tidsskrift for Sprogvidenskap*. 2:192-289.
_____. 1934. Additional notes on Ashkun. *Norsk Tidsskrift for Sprogvidenskap*. 7:56-115.
Pakhalina, T.N. 1975. *Vaxanskij Jazyk*. Moscow: Nauka.
Perry, John. 2000. Epistemic verb forms in Persian of Iran, Afghanistan and Tajikistan. In: Johanson and Utas (eds.), 229-257.
_____. 2005. *A Tajik Persian Reference Grammar*. Leiden, Boston: Brill.
Peterson, John. 2000. Evidentials, inferentials, and mirativity in Nepali. In: Balthasar Bickel (ed.). *Person and Evidence in Himalayan Languages. Linguistics of the Tibeto-Burman Area.* Volume I: 23.2:13-37.
Radloff, Carla F. and Shakil Ahmad Shakil. 1998. *Folktales in the Shina of Gilgit*. Islamabad: National Institute of Pakistan Studies, Quaid-i-Azam University and Summer Institute of Linguistics.
Rastorgueva, V. S. 1963. *A Short Sketch of Tajik Grammar*. trans. and ed. by Herbert H. Paper. The Hague: Mouton.
Rossi, A. 1989. L'inferenziale in Baluci. *Études Irano-Aryennes offertes À Gilbert Lazard. Studia Iranica* 7:283-291.
Schmidt, Ruth Laila. 2000. Typology of Shina pronouns. *Berliner Indologische Studien* 13/14:201-213.
_____ and Razwal Kohistani. 2001. Nominal inflections in the Shina of Indus Kohistan. *Acta Orientalia* 62:107-143.
Strand, Richard. http://users.sedona.net/~strand/
Tiffou, Étienne and Jurgen Pesot. 1989. *Contes du Yasin*. Paris: Peeters/SELAF.
Utas, Bo. 2000. Traces of evidentiality in Classical New Persian. In: Johanson and Utas (eds.), 259-274.
Van Daalen, L.A. 1988. The particle *kila/kira* in Sanskrit, Prākrit and the Pāli Jātakas. *Indo-Iranian Journal* 31:111-137.
Windfuhr, Gernot. 1982. The verbal category of inference in Persian. In: *Monumentum Georg Morgenstierne II*, 263-287. *Acta Iranica* 22. Leiden: Brill.

ALIGNMENT, PRECEDENCE, AND THE TYPOLOGY OF PIED-PIPING WITH INVERSION

George Aaron Broadwell
University at Albany, State University of New York
Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu/

ABSTRACT: Pied-piping with inversion is a phenomenon in a number of head-initial languages in which fronted interrogative phrases show an inverted, head-final word order. This paper is a typological survey of this phenomenon in nine languages. The survey supports the following conclusions: a.) Head-initial order in phrases is due to alignment constraints, b.) head-initial order in the phrases NP, PP, and QP must be due to different alignment constraints, since these phrase types often show different behavior in pied-piping with inversion contexts, and c.) alignment constraints appear to be superior to precedence constraints in describing pied-piping with inversion.

1 Introduction – pied-piping with inversion and OT

Optimality-theoretic approaches to word order have taken two different paths in formulating the relevant ordering constraints. One line of thought, following ideas that date back to GPSG (Gazdar, Pullum, Kline, and Sag 1985), uses *precedence constraints,* such as Head ≺Complement.[1] Broadwell (1999, 2001) has used constraints of this sort in OT-LFG. Another line of thought uses *alignment constraints*, such as Align (X, L, XP, L), which seek to align designated members of a phrase with the edge of that phrase. Sells (2001b), Morimoto (2001) and others have used this type of constraint in OT-LFG.

For many problems, both alignment and precedence constraints yield equivalent predictions. Consider the following data from San Dionicio Ocotepec Zapotec (SDZ). This is a VSO language, where phrases are normally head-initial:

1) Cù'á Juààny [$_{NP}$ x-pèh'cw Màríí].
com:grab Juan p-dog Mary
'Juan grabbed Mary's dog.'

*Cù'á Juààny [$_{NP}$ Màríí x-pèh'cw].
com:grab Juan Mary p-dog

In wh-questions, there is obligatory fronting of a [+wh]-phrase. If the [+wh]-phrase is contained in a NP, PP, or QP, then this phrase pied-pipes. The following example shows this for NP:

2) ¿[$_{NP}$Túú x-pèh'cw] cù'á Juààny?
who p-dog com:grab Juan
'Whose dog did Juan grab?'

[1]The idea that linear-precedence is due not to PS-rules, but to other principles of grammar is found in several syntactic theories. In Lexical Functional Grammar, such ideas are found in Falk (1983). Within early versions of Government-Binding theory, such ideas were proposed by Farmer (1980, 1984) and Stowell (1981).

*¿[$_{NP}$X-pèh'cw túú] cù'á Juààny?
p-dog who com:grab Juan

However, the pied-piped phrase now shows the order [$_{NP}$ Poss N], so the head of NP is no longer initial. This pattern is known as pied-piping with inversion (PPI), and was named and identified as an areal characteristic of Mesoamerican languages in Smith Stark (1988).

OT can provide an insightful account of PPI. This paper has two goals – a.) to discuss two alternative ways of formulating word order constraints (alignment constraints and precedence constraints), and b.) to discuss the typology of PPI with an eye towards using this phenomenon to help us decide on the better alternative.

2 Precedence and alignment compared

In Broadwell (2001), I posited the following constraints:

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

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

4) Head ≺Non-head

A head must precede its specifier.

5) Tableaux for (2) and (1)

		Wh-L	Head ≺Non-head
The interrogative order [Poss N]			
a.	☞¿[Túú x-pèh'cw] cù'á Juààny? (Whose dog grabbed Juan?)		*
b.	¿[X-pèh'cw túú] cù'á Juààny? (Dog whose grabbed Juan?)	*!	
The non-interrogative order [N Poss]			
c.	Cù'á Juààny [Màrií x-pèh'cw] (grabbed Juan Mary dog)		*
d.	☞ Cù'á Juààny [x-pèh'cw Màrií] (grabbed Juan dog Mary)		

Head ≺Non-head is a precedence constraint. We could equally propose an alignment constraint for heads of phrases, along the following lines:

6) Align (X, L, XP, L) = X-Left
Align the left edge of a head X^0 with the left edge of XP

Substituting this constraint makes no difference in prediction for these data:

7)

		Wh-L	X-Left
The interrogative order [Poss N]			
a.	☞¿[Túú x-pèh'cw] cù'á Juààny? (Whose dog grabbed Juan?)		*
b.	¿[X-pèh'cw túú] cù'á Juààny? (Dog whose grabbed Juan?)	*!	
The non-interrogative order [N Poss]			
c.	Cù'á Juààny [Màrií x-pèh'cw] (grabbed Juan Mary dog)		*
d.	☞ Cù'á Juààny [x-pèh'cw Màrií] (grabbed Juan dog Mary)		

So the questions to be examined in this paper are the following: Are there any empirical differences in the predictions that alignment and precedence constraints make about linear precedence? More generally, does the inventory of constraint types contain both alignment and precedence, or can one of these types be eliminated?

To answer these questions, this paper examines nine languages with PPI and compares alignment and precedence accounts of the phenomenon. In particular, it looks at how certain types of typological variation in these systems can be modelled in Optimality Theory.

3 Variation among phrase types in PPI systems

The phrase types that typically show PPI are NP, PP, and QP. However, it has not been sufficiently appreciated that these different phrase types may show different behavior in the PPI construction. Some phrases pied-pipe and show obligatory inversion, as is the case with NPs in San Dionicio Zapotec. However, for other phrase types in SDZ, pied-piping is found, but inversion is optional or prohibited.

Consider the following examples of pied-piped QP, where inversion is optional:

8) a) ¿[Xhíí tyóp] ù-dàw Juààny?
what two com-eat Juan

'What did Juan eat two of?'

b) ¿[Tyóp xhíí] ù-dàw Juààny?
two what com-eat Juan
'What did Juan eat two of?'

This can be modeled in OT by letting the constraint which is responsible for the head-initial position of Q overlap the Wh-Left constraint. However, since Wh-Left strictly dominates the NP-constraint we just discussed, the QP-constraint must be distinct.

Since these constraints show different positions in the constraint ranking, it is not possible to have a single X-Left or Head ≺Non-head constraint. Instead, we need different constraints which are responsible for head-initial order in NP and QP.

If we choose alignment constraints, then we appear to need two constraints like the following:

9) Align (Q, L, QP, L) = Q-Left
Align the left edge of a head Q^0 with the left edge of QP

Align (N, L, NP, L) = N-Left
Align the left edge of a head N^0 with the left edge of NP

If we pursue precedence constraints, then how do we differentiate NP and QP? In Broadwell (2001), I suggested that the relevant distinction is the X-bar theoretic status of the non-head material. So in NP, the possessor is in the Spec position. In QP, the restriction of the quantifier is in the Complement position.

That suggests the following constraints:

10) Head ≺Comp
A head must precede its complement.

Head ≺Spec
A head must precede its specifier.

Under the alignment scenario, the constraint ranking would be

11) Q-Left, Wh-Left » N-Left

Under the precedence scenario, the constraint ranking would be

12) Head ≺Comp, Wh-Left » Head ≺Spec.

Under either scenario, the relevant tableau would be as follows:

13)

		Head ≺Comp *or* Q-Left	Wh-Left
a.	☞¿Xhíí týop ù-dàù Juààny? (What two ate Juan?)	*	
b.	☞¿Tyóp xhíí ù-dàù Juààny? (Two what ate Juan?)		*

4 The problem with prepositions

Prepositions have a problematic status in many Mesoamerican languages. In Zapotec languages all or most locative prepositions are homophonous with body-part nouns. For example, SDZ *dèhjts* means both 'behind' and 'back', so a phrase like the following has two meanings:

14) dèhjts Juáàny
back/behind Juan
'behind Juan'
'Juan's back'

Other prepositions of this type are *cuèh'* 'side/beside', *lòò* 'face/to', and *nì'* 'foot/under'. I have labelled this group the *invertible prepositions*.

These body-part prepositions contrast in the PPI construction with a smaller number of prepositions such as *dèhspuèhhs* 'after', *ààxt* 'toward', *áántèhs* 'before', and *zí'cy* 'like'. This group is made up of most borrowed prepositions plus a few native non-locative prepositions, and is labelled the *non-invertible prepositions.*

Invertible prepositions show optional inversion in the PPI context:

15) a) ¿Dèhjts túú?
behind who

b) ¿Túú dèhjts?
who behind
'Behind who?'

Note that optional inversion here is found only when *dèhjts* has its prepositional interpretation. If

interpreted as a possessed body-part, it shows obligatory inversion:

16) a) ¿Túú dèhjts?
who back

b) *¿Dèhjts túú?
back who
'Whose back?'

Non-invertible prepositions show only the head-initial order:

17) a) ¿Zí'cy túú r-ù'ld Bèhjd?
like who hab-sing Pedro
'Who does Pedro sing like?'

b) *¿Túú zí'cy r-ù'ld Bèhjd?
who like hab-sing Pedro

In Broadwell (2001), I suggested that the non-invertible prepositions are purely prepositional, while the invertible prepositions show a mixture of nominal and prepositional properties. Non-invertible prepositions obey a constraint P-Left (defined as Align (Prep, L, PP, L)), and this constraint outranks Wh-L. Invertible prepositions, on the other hand, are capable of two analyses. If treated as nominal, then they show the obligatory inversion found with possessors. If treated as prepositional, then they show no inversion.

Prepositions like *dèhjts* 'behind' and *zí'cy* 'like' will have the following sorts of information in their lexical entries, using [±V] and [±N] features:

18) dèhjts - V zí'cy - V
- N

Lexical items that are [-N, -V] will obey the P-Left constraint, yielding a tableau like the following for (17):

19)

		P-Left	Wh-L
a.)	¿Túú zí'cy rù'ld Juáàny ? (Who like does Juan sing ?)	*!	
b.)	☞¿Zí'cy túú rù'ld Juáàny ? (Like who does Juan sing ?)		*

I will assume that at the time of evaluation**,** words must be fully specified for their

categorial features. There are two ways to do this for the invertible prepositions, and so they are subject to two analyses.

If they are treated as purely prepositional, then they obey the P-Left constraint, and they have a tableau like the preceding results, in which the uninverted candidate emerges as optimal. If they are treated as nominal, then they are subject to the Head ≺Spec constraint, and the inverted candidate will be optimal. The two tableaux are shown below:

20)

	(prepositional analysis)	P-Left	Wh-L	Head ≺Spec
a.)	☞¿Dèjts xhíí zúú bèh'cw? (Behind what lies dog ?)		*	
b.)	¿Xhíí dèjts zúú bèh'cw? (What behind lies dog?)	*!		inapplicable

	(nominal analysis)	P-Left	Wh-L	Head ≺Spec
a.)	¿Dèjts xhíí zúú bèh'cw? (Behind what lies dog ?)		*!	
b.)	☞¿Xhíí dèjts zúú bèh'cw? (What behind lies dog?)	inapplicable		*

This analysis gets the facts right, but note that it resorts to an alignment constraint (P-Left) to account for the head-initial property of non-invertible prepositions.

P-Left is used in this account because there does not seem to be a natural account using a precedence constraint. Head ≺Comp is already present among the constraints and overlaps Wh-Left. Since P-Left outranks Wh-Left, the ordering between non-invertible P and its complement must be due to some other constraint. But there is no good reason that the head-initial nature of PPs should be due to alignment, while head-initiality in NP and QP is due to precedence.

The solution in Broadwell (2001) uses both precedence constraints and alignment constraints to describe the tendency for heads to be initial in their phrases. But a simpler solution seems possible. If we take the alignment view, then there are three head alignment constraints (N-Left, Q-Left, P-Left), ranked as follows:

21) P-Left » Q-Left, Wh-Left » N-Left

That is a more satisfying analysis than my earlier proposal:

22) P-Left » Head ≺Comp, Wh-Left » Head ≺Spec

5 Relativized precedence?

A possible answer to this critique of precedence constraints is to change the constraints so that they do not refer generically to heads, but are relativized to particular kinds of heads. I'll call this alternative *relativized precedence*. Under this view, we do not have a single Head ≺Comp or Head ≺Spec constraint. Instead, we have separate constraints like the following:

23) N ≺Comp
P ≺Comp
Q ≺Comp
N ≺Spec
P ≺Spec
Q ≺Spec

Such an approach would clearly answer the objection to the non-uniform treatment of SDZ. We could replace P-Left with P ≺Comp. Filling in the specific heads involved for the other constraints, this would give us the following constraint ranking:

24) P ≺Comp » Q ≺Comp, Wh-Left » N ≺Spec

Now all the head-initiality is explained through precedence constraints.

However, such an account seems less satisfactory in a few particulars. First, of the six relativized precedence constraints, only three seem to be needed for the account. It is not clear what would count as a specifier of P or Q in this language.[2]

English does have a c-structure contrast between noun complements and noun specifiers in cases like the following:

25) John's picture of Mary
spec *comp*

However, we do not seem to find such a c-structure contrast in many other languages. In SDZ, the following phrase is compatible with two interpretations – one where Maria is the possessor, and another where she is the depicted:

26) x-rrètrààt Màríì

[2] Some earlier X-bar treatments of English (e.g. Jackendoff 1977) put degree adverbials like *relatively* or *quite* in specifier positions (e.g. *quite close to the edge*, *relatively few men*). It is likely that these proposals are no longer tenable. Note that they violate the Structure-Function association principles of Bresnan (2001:102ff), which say that c-structure elements in specifier position should correspond to f-structure constituents with a DF function. Since these adverbials correspond to f-structure ADJUNCTs, they ought to appear in adjoined positions, rather than Spec positions.

p-photo Maria
'The photo of Maria' *possessor*
'Maria's photo' *depicted*

There is no way to express 'Juan's picture of Maria' without using a relative clause. If we use an interrogative with a picture noun of this type, inversion is still obligatory in PPI, regardless of the interpretation:

27) a) ¿Túú x-rrètrààt?
who p-photo

b) *¿X-rrètrààt túú?
p-photo who
'Whose photo?' (*possessor or depicted interpretation*)

Relativized precedence makes the prediction that languages might show different PPI possibilities in NP due to different rankings of the N ≺Spec and N ≺Comp constraints. However, SDZ does not have such a difference, nor do any of the other languages surveyed so far. Future research may turn up such a language, but at the moment relativized precedence seems to predict a typological option which is not attested.

One might also formulate a kind of relativized precedence constraint in which heads precede everything in their phrase, disregarding the complement/specifier distinction:

28) N ≺Non-head
P ≺Non-head
Q ≺Non-head

However, this formulation seems to me to be almost indistinguishable from the alignment constraints in its predictive value. It may well be a notational variant of the alignment theory. A formulation in terms of alignment with the edge of a constituent does equally well, and is in accord with the constraint families which are already in use in Optimality Theoretic morphology and phonology.

6 Typological variation in PPI systems

A typological study of PPI in nine languages reveals a large range of variation in this construction. The data come from four Zapotec languages (San Dionicio, Tlacolula, Macuiltianguis, and Quiegolani), two Mixtecan languages (Ocotepec Mixtec and Copala Trique), two Mayan languages (Kiche and Tzotzil) and one Austronesian language (Sasak).[3] Zapotec and

[3] The survey includes my own data from San Dionicio Ocotepec Zapotec, Tlacolula Zapotec (discussed in more detail in Broadwell and Lillehaugen 2006), Macuiltianguis Zapotec, Copala Trique (discussed in more detail in Broadwell and Key 2004), and Kiche Mayan. Quiegolani Zapotec data is taken from Black (2000), Tzotzil data from Aissen (1996), Ocotepec

Mixtecan both have the time-depth comparable to Romance and they are two branches of a larger family called Oto-Manguean, which is comparable in time-depth to Indo-European. Mayan is not related to Oto-Manguean, but the two share a number of areal features.

I will describe the typological variation using alignment terminology, but return at the end to ask whether precedence constraints would serve as well.

For this study, I was interested in the following kinds of questions:[4]

a.) Which phrase types pied-pipe in questions? For these phrases, is inversion obligatory, optional, or prohibited?
b.) What is the implied constraint ranking for each of the languages?
c.) Are all orderings of the constraints possible?

All of the languages in the survey are V-initial – either VSO or VOS.

6.1 Phrase types and the possibility of inversion

A useful way to characterize the typological variation is by looking at the ranking of the Wh-Left constraint relative to the constraints which are responsible for head-initial position in the phrase. The approach followed here infers the constraint ranking from the obligatory, optional, or prohibited nature of inversion in pied-pied phrases, using the following correlation:

29)	Order in pied-piped phrase	Implied constraint ranking
	[X Wh], *[Wh X]	X-Left » Wh-Left
	[X Wh], [Wh X]	X-Left, Wh-Left (constraint overlap)
	[Wh X] , *[X Wh]	Wh-Left » X-Left

For a number of languages in the survey, we do not have information on pied-piping of QP or on possible differences among prepositions. However, all the languages in this survey show pied-piping of both NP and PP.

6.1.1 High Wh languages

Three languages in the survey (Quiegolani Zapotec, Tzotzil, Ocotepec Mixtec) report the

---

Mixtec data from Eberhart (1999), and Sasak data from Austin (2001).

[4] An additional question of interest concerns the question of whether all interrogatives behave the same in PPI. See Broadwell and Lillehaugen (2006) for more discussion of Tlacolula de Matamoros, a language in which animate interrogatives appear to be associated with a more highly-ranked Wh-Left constraint than inanimate interrogatives. Macuiltianguis Zapotec shows a similar system in which discourse-linked interrogatives (*which*) show stronger tendencies for left alignment. Space prevents fuller discussion of these issues in this paper.

same pattern – obligatory inversion in both PP and NP, resulting from an undominated Wh-Left constraint. Consider the following data from Quiegolani Zapotec (Black 2000:135):

30) **Quiegolani Zapotec**

a) [Txu lo] n-dux xnaa noo? [Wh P]
who to STAT-angry mother 1EX
'With whom was my mother angry?'

b) *[Lo txu] n-dux xnaa noo? *[P Wh]
to who STAT-angry mother 1EX

31) a) [Txu xnaa] n-dux lo de? [Wh N]
who mother STAT-angry to you
'Whose mother is angry with you?'

b) *[Xnaa txu] n-dux lo de? [*N Wh]
mother who STAT-angry to you

We find the same pattern in Ocotepec Mixtec (Eberhardt 1999):

32) **Ocotepec Mixtec**

[Ní nuu] ndée ñā? [Wh P]
where face con:sit she
'Where does she live?'

33) [Na nuu] xehē de tūtu? [Wh P]
what face COM:give he:RES paper
'To whom did he give the paper?'

34) *[Nuu na] xehē de tūtu? *[P Wh]
face who COM:give he:RES paper

35) [Na sehē] kúū xīn? [Wh N]
who child CON:be he:FAM
'Whose child is he?'[5]

The implied constraint hierarchy for Quiegolani Zapotec, Ocotepec Mixtec, and other languages with this pattern is as follows:

36) **Quiegolani Zapotec and Ocotepec Mixtec Constraint ranking**

[5] Uninverted NPs are said to be ungrammatical, but the forms are not cited.

Wh-Left » P-Left, N-Left

Note that much of the orginal work attempting to explain PPI (Aissen 1996, Trechsel 2000) assumes that this type of PPI is the only kind found cross-linguistically.

6.1.2 High P languages

Several of the Otomanguean languages in the survey (Zapotec and Trique) show systems in which the head-ordering constraint for Prepositions dominates the Wh-Left constraint. We have already seen an extended example of this for San Dionicio Ocotepec Zapotec.

Note that in these languages the majority of prepositions are homophonous with body-part nouns, so the high ranking of P-Left is seen primarily with a smaller group of non-locative, non-body-part, or borrowed prepositions.

Consider the following examples which contrast invertible and non-invertible prepositions in Copala Trique. The Copala Trique interrogative 'what' is composed of two parts – an interrogative *me*^{3} and the pronoun *ze*^{32} 'it'. For invertible prepositions there are three options – PPI, stranding of the preposition, and an unusual order I label 'disconnected' where the preposition appears between the two parts of the interrogative:

37) **Copala Trique invertible prepositions**

a) ¿Me^{3} ze^{32} xra^{1} nicun'^{3} chuvee^{4}? *PPI*
WH N behind stand dog

b) *? ¿Xra^{4} me^{3} ze^{32} nicun'^{3} chuvee^{4}? **PP w/o inversion*
behind WH N stand dog

c) ¿Me^{3} ze^{32} nicun'^{3} chuvee^{4} xra^{4} ? *P-stranding*
WH N stand dog behind

d) ¿Me^{3} xra^{1} ze^{32} nicun'^{3} chuvee^{4}? *disconnected*
WH behind N stand dog
'What is the dog standing behind?'

For the non-invertible prepositions, only pied-piping without inversion is possible:

38) **Copala Trique non-invertible prepositions**

a) ¿Naa^{13} me^{3} chuma'^{3} chee^{5} Waan^{4}? *PP w/o inversion*
toward WH town walk Juan

b) *¿Me^{3} chuma'^{3} naa^{13} chee^{5} Waan^{4}? *PPI*
WH town toward walk Juan

c) *¿Me[3] naa[13] chuma'[3] chee[5] Waan[4]? *disconnected*
WH toward town walk Juan

d) *¿Me[3] chuma'[3] chee[5] Waan[4] naa[13]? *P-stranding*
WH town walk Juan toward
'Which town did Juan walk toward?

So San Dionicio Ocotepec Zapotec, Copala Trique and other languages of this type show a partial constraint ranking of the following sort:

39) P-Left » Wh-Left

6.1.3 High N languages

Sasak, an Austronesian language of Indonesia, shows obligatory PPI in PPs, but prohibits inversion in NPs (Austin 2001):

40) **Sasak**
[Sai kance]=m bedait léq peken? [Wh P]
who with=2 meet loc market
'Who are you meeting with in the market?'

41) [Guru-n sai] yaq=m dengah léq masjit? [N Wh]
teacher-link who fut=2 listen loc mosque
'Whose teacher will you hear at the mosque?'

Sasak is very important from a typological perspective because it is the only language where N-Left dominates P-Left.

From the point of view of alignment constraints, the implied constraint hierarchy for Sasak is as follows:

42) **Sasak constraint ranking**

N-Left » Wh-Left » P-Left

Note that the relative ranking of N-Left and P-Left in Sasak is the reverse of the ranking in SDZ:

43) **San Dionicio Ocotepec Zapotec constraint ranking**

P-Left » Wh-Left, Q-Left » N-Left

That seems to show us rather clearly that no principle of Universal Grammar is forcing these

constraint rankings.

6.1.4 High Q languages

Kiche Mayan shows a pattern where Q-Left dominates Wh-Left. QPs in this language show pied-piping in questions, but the pied-piped constituent may not invert:

44) **Kiche Mayan QPs**

a) [Juntir jäs] x-u-tij l-a Xwan?
all WH:NHUM COM-3SERG-eat DET-CL Juan
'What did Juan eat all of?'

b) *[Jäs juntir] x-u-tij l-a Xwan?
WH:NHUM all COM-3SERG-EAT DET-CL Juan

PPs, by constrast, show optional inversion:

45) **Kiche Mayan PPs**

a) [Chuxe' jäs] k'o wi le tz'i'?
under:3SERG WH:NHUM exist LOC DET dog
'What is the dog under?'

b) [Jäs chuxe'] k'o wi le tz'i'?
WH:NHUM under:3SERG exist LOC DET dog
'What is the dog under?'

This implies a partial constraint ranking for Kiche Mayan like the following:[6]

46) **Kiche Mayan constraint ranking**

Q-Left » Wh-Left, P-Left

6.1.5 Summary

Space doesn't allow us to fully discuss and motivate all the constraints and their rankings

[6] It is not clear whether NPs show PPI in Kiche. Possessed NPs (e.g. *whose son*) cannot pied-pipe, but appear in a clefted construction instead. NPs of the [*Which* N] type do pied-pipe and show Wh-initial order, but it is not clear that this is an inverted order, since several types of determiners usually precede N in Kiche. See Broadwell (2005) for more discussion.

for the languages in the survey. However, the following chart shows the results of the study. (A blank cell shows that the source provides no information on this question.)

Language	**NP**	**PP [native/ body-part]**	**PP [borrowed/ non-locative]**	**QP**
San Dionicio Zapotec	obligatory	optional	prohibited	optional
Copala Trique	obligatory	obligatory	prohibited	obligatory
Tlacolula Zapotec	optional (who) prohibited (what)	optional (who) prohibited (what)	prohibited	obligatory (who) optional (what)
Macuiltianguis Zapotec	optional (whose) obligatory (which)	optional	prohibited	optional
Kiche Mayan	(no pied-piping with NP)	optional	--	prohibited
Ocotepec Mixtec	obligatory	obligatory	--	--
Quiegolani Zapotec	obligatory	obligatory	--	--
Sasak	prohibited	obligatory	--	--
Tzotzil	obligatory	obligatory	--	--

The implied constraint hierarchies are as follows:

47) **Constraint rankings for nine languages with PPI**

Type	Language	Constraint ranking
High Wh	Ocotepec Mixtec	Wh-Left » P-Left, N-Left
	Quiegolani Zapotec	
	Tzotzil	
High P	San Dionicio Zapotec:	P-Left » Wh-Left, Q-Left » N-Left
	Copala Trique	P-Left » Wh-Left » N-Left, Q-Left

	Tlacolula Zapotec	P-Left » Wh[+anim]-Left, N-Left » Wh[-anim]-Left, Q-Left
	Macuiltianguis Zapotec	P-Left, Wh[+d]-Left » Wh[-d]-Left, Q-Left, N-Left
High Q	Kiche Mayan	Q-Left » Wh-Left, P-Left
High N	Sasak	N-Left » Wh-Left » P-Left

7 Conclusions

Let us return to the question of the best way to formulate the constraints that produce head-initial order. So far I have formulated the typological results with alignments constraints. Could they have been equally formulated as precedence constraints? The answer appears to be no.

If we formulate the precedence constraints as Head ≺Spec and Head ≺Comp, then the typological results are very difficult to account for. PP and QP are the two phrase types which typically occur with interrogative complements, while NP occurs with an interrogative specifier.

So a Head ≺Comp constraint would predict that PP and QP should behave alike in PPI constructions. However, there seems to be no typological support for this. As we can see, the P-Left and Q-Left constraints do not show any tendency to be at the same position in the constraint ranking.

Looking at the question from another angle, we can ask ourselves how many head-ordering constraints are necessary to account for the orders found in PPI. We can see this in a language like San Dionicio Ocotepec Zapotec, where the constraint ranking is P-Left » Wh-Left, Q-Left » N-Left. In this language, at least three head-ordering constraints need to be ranked relative to the Wh-Left constraint. So a precedence theory with two constraints like Head ≺Spec and Head ≺Comp will not be adequate.

A relativized precedence theory (with six potential ordering constraints) can handle the data, but seems too strong in predicting typological results that are not attested. Relativized precedence constraints of the Q ≺Non-head type do not seem to make any interestingly different predictions from alignment constraints.

Alignment constraints are well-motivated in other parts of syntax, morphology, and phonology and thus the need for this constraint type is clear. What is not clear is whether there is any evidence that we need precedence constraints.

8 References

Aissen, Judith. 1996. Pied-piping, abstract agreement, and functional projections in Tzotzil. *Natural language and linguistic theory* 14:447-491.

Austin, Peter. 2001. Content questions in Sasak, eastern Indonesia: an OT syntax account.

manuscript. University of Melbourne.
Black, Cheryl. 2000. *Quiegolani Zapotec syntax: A Principles and Parameters Account.* Dallas: SIL International and UT Arlington Publications in Linguistics 136.
Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.
Broadwell, George Aaron. 1999. Focus alignment and optimal order in Zapotec. *Proceedings of the 35th Chicago Linguistics Society*. (Also available at http://www.albany.edu/anthro/fac/broadwell.htm)
Broadwell, George Aaron. 2001. Optimal order and pied-piping in San Dionicio Zapotec. In Peter Sells, ed. *Formal and empirical issues in optimality theory.* Stanford: CSLI.
Broadwell, George Aaron. 2002. Constraint symmetry in Optimality-Theoretic Syntax. In Miriam Butt and Tracy Holloway King, eds. *Proceedings of the Lexical Functional Grammar 2000 Conference*. Stanford, CA: CSLI Publications. (Online version available at http//www-csli.stanford.edu/publications/).
Broadwell, George Aaron and Brook Danielle Lillehaugen. 2006. Pied-piping with inversion in Tlacolula de Matamoros Zapotec. Presented at the Society for the Study of the Indigenous Languages of the Americas, Albuquerque, Jan 2006.
Broadwell, George Aaron and Michael Parrish Key. 2004. Pied-piping with inversion in Copala Trique. Presented at the Society for the Study of the Indigenous Languages of the Americas, San Francisco, Jan 2004.
Eberhardt, Roy. 1999. Questions and inversion in Ocotepec Mixtec. *Work Papers of the Summer Institute of Linguistics*, University of North Dakota Session 43. [http://www.und.nodak.edu/dept/linguistics/wp/1999Eberhardt.PDF]
Falk, Yehuda. 1983. Constituency, word order, and phrase structure rules. *Linguistic Analysis* 11:331-360.
Farmer, Ann. 1980. *On the Interaction of Morphology and Syntax*. Ph.D. thesis, MIT.
Farmer, Ann, 1984. *Modularity in Syntax: A Study of Japanese and English*. Cambridge, MA: MIT Press.
Gazdar, Gerald; Ewan Klein; Geoffrey Pullum; and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Cambridge, Ma: Harvard.
Hollenbach, Barbara. 1984. *The Phonology and Morphology of Tone and Laryngeals in Copala Trique.* Ph.D. thesis, University of Arizona.
Hollenbach, Barbara. 1992. A syntactic sketch of Copala Trique. In C. Henry Bradley and Barbara E. Hollenbach, eds. *Studies in the Syntax of Mixtecan languages,* vol. 4, pp. 173-431. Dallas: Summer Institute of Linguistics.
Jackendoff, Ray. 1977. *X-bar Syntax: A Study of Phrase Structure*. Cambridge, Ma: MIT Press.
Morimoto, Yukiko. 2001. Verb raising and phrase structure variation in OT. In Sells (2001a).
Sells, Peter, ed. 2001a. *Formal and Empirical Issues in Optimality Theory.* Stanford: CSLI.
Sells, Peter. 2001b. *Structure, Alignment, and Optimality in Swedish*. Stanford:CSLI.
Smith Stark, Thomas. 1988. 'Pied-piping' con inversion en preguntas parciales. Ms. Centro de estudios lingüísticos y literarios, Colegio de México y Seminario de lenguas indígenas.
Stowell, Timothy. 2001. *Origins of Phrase Structure*. Ph.D. thesis. MIT.
Trechsel, Frank. 2000. A CCG Approach to Tzotzil Pied-Piping. *Natural Language and Linguistic Theory* 18: 611-63.

Copala Trique: The orthography used is based on the practical orthography developed by Barbara and Bruce Hollenbach of SIL for their translation of the New Testament. We follow their usage in the representation of the consonants, including the following conventions: <x> = [ʃ], <xr> = [ʂ] (a retroflex alveopalatal sibilant), <ch> = [tʃ], <chr> = [tʂ], <c> = [k] (before front vowels), <qu> = [k] before back vowels, [v] = [β] and <j> = [h]. <Vn> represents a nasalized vowel. Trique has five level tones (1, 2, 3, 4, 5) and three contour tones (13, 31, 32), as discussed in Hollenbach (1984). Since the practical orthography does not distinguish all eight tones, we use the numerical superscripts of Hollenbach (1984, 1992) for our tonal representations.

Glosses use the following abbreviations: COM = completive aspect, DEC = declarative, P = possessed form.

Trique data were gathered from three Copala Trique speakers – José Fuentes, Irma Fuentes, and Roman Vidal López. I thank them, as well as Michael Parrish Key, who was my coauthor for an earlier (Broadwell and Key 2004) which dealt with these facts. I also thank Barbara Hollenbach, who graciously answered a number of questions via e-mail.

Kiche Mayan: This paper uses the conventions of the national orthography, in which <x> = a voiceless alveopalatal sibilant (IPA [ʃ]), <tz> = a voiceless dental affricate (IPA [ts]), <ch> = a voiceless alveopalatal affricate (IPA [tʃ]), <ä> = schwa (IPA [ə]), <q> is a uvular stop and apostrophe = glottal stop (following a vowel) or glottalization (following a consonant). However, Kiche dialects differ in the number of phonemic vowels and in the phonemic status of vowel length. The national orthography distinguishes long and short versions of the five cardinal vowels (thus *a, aa, e, ee, i, ii, o, oo, u, uu*). The Cantel dialect has no length distinction and instead has six phonemic vowels (*a, ä, e, i, o, u*). I write only these vowels here.

Glosses use the following abbreviations: Abs = absolutive, cl = personal classifier (markers of the age and sex of human referents), com = completive aspect, det = determiner, Erg = ergative, hum = human, inc = incompletive aspect, loc = locational focus (a particle that appears postverbally in sentences with a focussed locative phrase), nhum = nonhuman, p = plural, plain = plain status (a suffix which appears on a phrase-final verb), pass = passive, s = singular, wh= interrogative.

Data for this paper were gathered in the context of a UCLA field methods course taught by Pamela Munro in 2004-2005. I thank Pam and the participants in the class for their help and suggestions. Special thanks are due to Pedro U. Garcia Mantanic, a native speaker of the Cantel dialect of Kiche, who provided all the data cited in this paper.

Macuiltianguis Zapotec: In the orthography used here, symbols have their standard phonetic values, with the following exceptions. <c> = /k/, /x/ = /š/, <yh> = /ž/, <th> = /ϴ/, <ch> = /č/,<’> = /ʔ/, and doubled vowels are long.

The following abbreviations appear in the glosses: cl = clitic, com = completive aspect, foc = focus, hab = habitual aspect, indef = indefinite, invis = determiner for unseen things, neg = negative, pl = plural, 3 = 3rd person.

Thanks are due to John Foreman for help with understanding this data, and to Pamela Munro and Jie Zhang, who were important members of the initial working group for this language. Special thanks are due to our consultants, Ignacio Cano and Margarita Martínez, without whom none of this would be possible. All the data on PPI are due to Ignacio Cano.

San Dionicio Ocotepec Zapotec: The orthography for SDZ 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, <eh> = /ɛ/ and <ehh> = /ɛɛ/. 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.

Ordinary affixes are separated from the stem by the hyphen; clitics are separated by =. Glosses use the following abbreviations: an = animative, com = completive aspect, hab = habitual aspect, in = inanimate, neg = negative, loc = locative, p = possessed, pot = potential aspect, q = question, 1s =1st person singular, 3 = 3rd person human (ordinary respect level), 3i = 3rd person inanimate.

Special thanks to Luisa Martínez, who supplied all the data.

ON THE STATUS OF RESUMPTIVE PRONOUNS IN MODERN GREEK RESTRICTIVE RELATIVE CLAUSES

Aikaterini K. Chatsiou
University of Essex

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu

Abstract

We discuss the status of Modern Greek Resumptive Pronouns, focusing on Restrictive Relative Clauses. Several analyses have been proposed to account for the phenomenon of resumption in Modern Greek Relative Clauses arguing in favour of a similar treatment of gaps and resumptive pronouns, suggesting that Binder-Resumptive Dependencies are triggered by the same mechanism as Filler-Gap Dependencies. In this paper, it is argued that resumptive pronouns are the ordinary pronoun forms of the language and that they are not alternative manifestations of gaps, presenting evidence from Asudeh's (2004) criteria for Hebrew, Irish and Swedish. Following this, we propose an LFG analysis for resumption in Modern Greek *pu* and *o opios* Restrictive Relative Clauses, distinguishing between two types of Dependencies (Filler-Gap and Binder-Resumptive Dependencies), following Asudeh (2004)'s treatment of the syntax of resumptives in these languages.

1 Introduction

In this paper, we discuss the status of Modern Greek Resumptive Pronouns, focusing on Restrictive Relative Clauses. In particular, it is argued that resumptive pronouns are the ordinary pronoun forms of the language and that they are not alternative manifestations of gaps. Based on this, we present an LFG analysis of resumptives and gaps in Modern Greek Restrictive Relative Clauses, following Asudeh (2004), proposing a Binder-Resumptive Dependency analysis for the former as opposed to a Filler-Gap Dependency for the latter.

The paper is organised as follows: Section 2 presents an overview of the data, namely, some of the most important characteristics of Restrictive Relative Clauses and Resumptive Pronouns in Modern Greek as well as their distribution in RRCs. In Section 3 we present our observations with regard to the status of resumptive pronouns in RRCs. Finally, in Section 4 we propose an LFG analysis of resumption in *pu* and *o opios*-RRCs.

2 An overview of the data

2.1 Modern Greek Restrictive Relative Clauses (RRCs)

Modern Greek Restrictive Relative Clauses are distinct from other types of Relative Clauses (namely Non-Restrictive (Appositive) Relative Clauses and Free Relative Clauses), since they convey important information about the head element and therefore cannot be omitted without loss of information as examples (1) and (2) illustrate:[1]

(1) Oi mathites pu teliosan tin ptihiaki
the.MPL.NOM students.MPL.NOM that finished.3PL the.FSG.ACC dissertation.FSG.ACC
tus harikan.
their.MPL.GEN were.happy.3PL
'The students who finished their dissertation were happy.'

(2) Oi mathites harikan.
the.MPL.NOM students.MPL.NOM were.happy.3PL
'The students were happy.' (Which students?)

[1]The abbreviations used in the glosses are: FSG = Feminine Singular, MSG = Masculine Singular, NPL = Neuter Plural, SG = singular, 1 = first person, 3 = third person, CL = clitic pronoun, NOM = Nominative Case, GEN = Genitive Case, ACC = Accusative Case.

Other abbreviations used in the paper: RP(s) = Resumptive Pronoun(s), MG = Modern Greek, (R)RC(s) = (Restrictive) Relative Clause(s), BR-DCs = Binder-Resumptive Dependency Constructions, FG-DCs = Filler-Gap Dependency Constructions, WCO = Weak Crossover (Effects).

Further to the above, contrary to the controversy that the same issue has raised for main declarative clauses, it is generally agreed in the literature that the internal constituent order of a relative clause is relatively fixed (Tzartzanos (1963), Markantonatou (1992), Lascaratou (1998), Mackridge (1985), Theophanopoulou-Kontou (1989)): they are introduced by a relativiser (either the complementizer *pu* or the relative pronoun *o opios*), followed by a verb and zero or more phrasal elements, as illustrated in (3):

(3) Relativiser + (resumptive pronoun) + V + XP*

The RRC's position with regards to its nominal head element is also fixed: Restrictive Relative Clauses always occur postnominally, after the element they modify, as illustrated by the ungrammaticality of (4):

(4) * Pu taise ton skilo o andras.
that fed.3SG the.MSG.ACC dog.MSG.ACC the.MSG.NOM man.MSG.NOM
[intended meaning: 'The man who fed the dog.']

Another characheristic of Modern Greek Restrictive Relative Clauses is that they are introduced either by the indeclinable, unmarked for gender and number complementizer *pu* [that] or by the fully declinable for case, gender and number relative pronoun *o opios*[2] [who.MSG.NOM], which agrees in gender and number with the modifying head and gets its case depending on the grammatical function it fulfils within the relative clause:

(5) I kopela pu vrike o skilos.
the.FSG.NOM girl.FSG.NOM that found.3SG the.MSG.NOM dog.MSG.NOM
'The girl that the dog found.'

(6) I kopela tin opia vrike o
the.FSG.NOM lady.FSG.NOM the.FSG.ACC who.FSG.ACC found.3SG the.MSG.NOM
skilos.
dog.MSG.NOM
'The girl whom the dog found.'

Both *pu* and *o opios*, are normally obligatory and cannot be omitted as illustrated in examples (7) and (8) [3] :

[2]We assume that the relative pronoun *o opios* consists of the definite article *o* (the.MSG.NOM) and the pronoun *opios* (who.MSG.NOM). Alexiadou (1998), citing Hatzidakis (1907), suggests that a further decomposition of *opios* into the indefinite marker *o-* and the variation of the free relative pronoun *ópios, -pios* is possible. The particulars of this require further research involving the diachronic analysis of relative pronouns and will not be pursued here.

[3]*Pu*, however, can be omitted in certain environments, such as in Relative Clauses in subjunctive mood (1) or in the second conjunct of a coordinated relative clause construction (2). For the purposes of this paper, however, we will assume that *pu* is always obligatory:

(1) Vrike daskala (pu) na milai Yaponezika.
found.3SG teacher.FSG.ACC that SUBJUNCTIVE PART speak.3SG japanese
'S/He found a teacher that speaks Japanese [lit. to speak Japanese].'

(2) Vrikan ton skilo pu efage ti gata ke (pu) gavgize.
found.3PL the.MSG.ACC dog.MSG.ACC that ate.3SG the.FSG.ACC cat.FSG.ACC and (that) was.barking.3SG
'They found the dog which ate the cat and (which) was barking.'

(7) O pyrosvestis pu/*Ø esose to koritsi pire
the.MSG.NOM fireman.MSG.NOM that rescued.3SG the.NSG.ACC girl.NSG.ACC received.3SG
vravio.
reward.NSG.ACC
'The fireman who rescued the girl was rewarded.'

(8) To koritsi to opio / *Ø esose o
the.NSG.NOM girl.NSG.NOM the.NSG.ACC who.NSG.ACC rescued.3SG the.MSG.NOM
pyrosvestis ine kala.
fireman.MSG.NOM is.3SG well
'The girl that the fireman rescued is fine.'

2.2 Resumption in Modern Greek RRCs

Modern Greek Resumptive Pronouns have the form of the unstressed monosyllable clitic form (weak form) of the personal pronoun. Being clitics, they are declinable according to the table in (9)[4]:

(9)

Number	Case	1st person	2nd person	3rd person		
				MASC	FEM	NEUT
SINGULAR	GEN	mu	su	tu	tis	to
	ACC	me	se	ton	ti(n)	to
PLURAL	GEN	mas	sas	tus	tis	ta
	ACC	mas	sas	tus	tis	ta

As previously noted, the position of the resumptive pronoun in the Relative Clause is fixed. Resumptive pronouns are *proclitic* – that is, they immediately precede the main verb – and must follow the relativiser (and optionally any negation markers present) as illustrated in (10):

(10) O gatos pu den ton taise i kopela.
the.MSG.NOM cat.MSG.NOM that not CL.3.MSG.ACC fed.3SG the.FSG.NOM girl.FSG.NOM
'The cat that the girl did not feed'

Depending on their case-marking, resumptive pronouns can fulfil specific syntactic functions. For instance, resumptive pronouns marked for accusative case may function as direct objects, whereas those in genitive case can function as indirect objects or as complements of a preposition, as in (11):

(11) To koritsi pu tu edoses ta luludia.
the.NSG.NOM girl.NSG.NOM that CL.3.NSG.GEN gave.2SG the.NPL.ACC flowers.NPL.ACC
'The girl that you gave the flowers to.'

[4]In addition to the forms presented in table (9), there is a 3rd person Nominative Singular form of the clitic pronoun (*tos* [CL.3.MSG.NOM], *ti* [CL.3.FSG.NOM] , *to* [CL.3.NSG.NOM]), which is reserved for special uses in certain expressions following *na* and *pun* (short form of *pu ine..?* = 'where is...?') as in *pun'tos?* = 'where is he?' and *na tos* = 'there he is!'. This reserved use of the nominative case of the clitic might be an explanation as to why RRCs bearing the relativised function of a subject are ungrammatical when a RP is present, as illustrated in (1):

(1) O mathitis o opios / pu ***tos*** teliose tin ptihiaki
the.MSG.NOM student.MSG.NOM the.MSG.NOM who.MSG.NOM / that CL.3.MSG.NOM finished.3.SG the dissertation
tu.
his.MSG.GEN
'The student who/that finished his dissertation.'

Regarding their distribution, resumptive pronouns are obligatorily absent in subject position both in *pu*- and in *o opios*-RRCs, although it is not clear whether this is simply due to the fact that the form for the nominative case is reserved for specific expressions (see footnote 4). Moreover, resumption is optional in both *pu*- and *o opios*- RRCs when the relativised position is a Direct Object, whereas when it is an Indirect Object (OBJ, OBJ2) it is obligatorily present in *pu*-RRCs but obligatorily absent in *o opios*-RRCs.

The table in (12) summarises their distribution in Modern Greek RRCs (+ marks the obligatory presence of the resumptive; - marks the obligatory absence of the resumptive pronoun; +/- marks its optionality):

(12)

Relativiser	**Relativised Function**		
	SUBJ	OBJ	OBJ2
PU	-	+/-	+
O OPIOS	-	+/-	-

3 On the status of Resumptive Pronouns in Restrictive Relative Clauses

In this section we consider two issues regarding the status of Resumptive Pronouns (RPs) in Modern Greek (henceforth MG) Restrictive Relative Clauses (RRCs), namely that first of all, they are the ordinary pronouns of the language and should therefore be analysed similarly to pronouns and that secondly they are not alternative manifestations of gaps and for this purpose dependencies involving resumptives and dependencies involving gaps should receive a distinct treatment.

3.1 Resumptive pronouns are the ordinary pronouns of the language

An important observation related to RPs is McCloskey (2002)'s claim "that resumptive pronoun languages do not have resumptive-specific morphological paradigms" (Asudeh, 2004, p. 11). Although this observation does not apply to all languages[5], resumptive pronouns in Modern Greek Restrictive Relative Clauses are the normal pronouns of the language: they have the same form and syntactic distribution as the 'ordinary' pronominal clitic forms. In particular, RPs have the form of the unstressed monosyllable clitic forms of personal pronouns and are declined according to the table in (9), reproduced here for convenience as (13):

(13)

Number	**Case**	**1st person**	**2nd person**	**3rd person**		
				MASC	FEM	NEUT
SINGULAR	GEN	mu	su	tu	tis	to
	ACC	me	se	ton	ti(n)	to
PLURAL	GEN	mas	sas	tus	tis	ta
	ACC	mas	sas	tus	tis	ta

In addition to that, they have the same syntactic distribution in non-imperative clauses as the ordinary pronouns of the language[6]– they immediately precede the verb as illustated in (14a) and (14b):

[5]Not all languages behave according to McCloskey (2002)'s claim. Vata, for instance, (Koopman, 1982) has special pronouns to denote resumption and Kaqchikel (Falk, 2002), a Mayan language, appears to have a resumptive that is not a pronoun.

[6]As Philippaki-Warburton (1985, p. 82) suggests, clitics "precede the inflected non-imperative verb, but follow the imperative and gerund [forms]". Since the verb in a RRC cannot be in the imperative or the gerund form, it therefore follows that RPs may only precede the verb of the relative clause.

(14) a. **Resumptive pronoun**

I ghata pu tis edosa to gala.
the.FSG.NOM cat.FSG.NOM that CL.3.FSG.GEN gave.1SG the milk

'The cat that I gave (her) the milk.'

b. **Ordinary Clitic form of the personal pronoun**

Tis edosa to gala.
CL.3.FSG.GEN gave.1SG the milk

'I gave the milk to (her).'

3.2 Resumptive pronouns are not alternative manifestations of gaps

Another issue regarding the status of RPs in relative clauses discussed in Asudeh (2004), concerns their relationship to gaps, and in particular whether the dependency between the resumptive pronoun and its binder (Binder-Resumptive Dependency) can be analysed similarly to a Filler-Gap Dependency. Several analyses have been proposed in the literature which argue that Greek RPs are (more or less) similar to gaps. Among others, Alexiadou and Anagnostopoulou (2000) propose an analysis of RPs in MG RRCs following Kayne (1994)'s antisymmetric analysis, suggesting that RPs behave similarly to gaps and that BR-DCs are triggered by the same mechanism as FG-DCs. In addition to that, Alexopoulou (2006), following Shlonsky (1992), argues in favour of treating RPs as a variable at LF claiming that unlike Hebrew, Greek "resumptive relative clauses have the same meanings as gap relatives" (Alexopoulou, 2006, 81).

In this section we put to the test the behaviour of RPs and gaps in Modern Greek using Asudeh (2004)'s criteria for Hebrew, Irish and Swedish. Asudeh (2004) claims that resumptive relative clauses are not the same as gap relative clauses, and supports his argument by providing the reader with a number of constructions where RPs behave differently from gaps, such as island sensitivity, weak-crossover effects, across-the-board extraction from coordinated conjuncts, licensing of paracitic gaps and form-identity effects.

3.2.1 Island Sensitivity

One of the arguments that Asudeh (2004, p. 124–128) puts forward arguing against a gap-like account of resumptives involves the issue of *island sensitivity*. In particular, he suggests that resumptive pronouns occur freely in islands, or rather that "the dependency between a resumptive and its binder is island sensitive" (Asudeh, 2004, 127), whereas gaps are disallowed in the same environment. Here, we consider the two kinds of island constructions, also discussed in McCloskey (1979) for Irish: *the wh-island* (15a) and *the complex-NP island* (15b):

(15) a. Gnorisa mia gineka pu den ksero pjos tin /
met.1SG a.FSG.ACC woman.FSG.ACC that not know.1SG who.MSG.NOM CL.3.FSG.ACC
*Ø pantreftike.
married.3SG

'I met a woman that I do not know who married her.'

b. Afti ine mia glossa pu tha sevomoun ekinon pu
this.FSG.NOM is.3SG a.FSG.NOM language.FSG.NOM that would respect.1SG the one that
tha ti / *Ø miluse.
would CL.3.FSG.NOM speak.3SG

'This is a language that I would respect the one who would speak it.'

The ungrammaticality of the examples involving a gap where a RP is expected suggests that RPs, contrary to gaps, occur freely in islands, evidence supportive of the argument that MG RPs are not alternative manifestations of gapped elements.

3.2.2 Weak Crossover Effects

Further evidence supporting the claim that gaps and RPs are distinct, according to McCloskey (1990, p.236-237), comes from weak crossover (WCO) effects. In particular, sentences manifesting WCO effects are ungrammatical if a gapped element is present (16a). If the gap is replaced with a RP, however, the sentence becomes grammatical, as shown in (16b) (both examples from Alexopoulou (2006, p.26, ex.43)):

(16) a. O fititis$_i$ pu tu$_i$ estile ta vivlia i daskala tu$_{i/j}$.
the.MSG.NOM student.MSG.NOM that CL.3.MSG.GEN sent.3SG the books the teacher his.MSG.GEN
'The student that his teacher sent him the books.'

b. *? O fititis$_i$ pu Ø$_i$ estile ta vivlia i daskala tu$_{i/j}$.
the.MSG.NOM student.MSG.NOM that sent.3SG the books the teacher his.MSG.GEN
'The student that his teacher sent him the books.'

3.2.3 Across-the-board Extraction

Zaenen et al. (1981), Sells (1984) and Engdahl (1985) among others have argued in favour of a common treatment of gaps and resumptives based on evidence from across-the-board extraction, i.e. extraction from all conjuncts of a coordinate structure. In other words, if we can extract the RPs from all the conjuncts of a coordinate structure, and the output is still grammatical, then this would provide evidence in favour of a common treatment of gaps and resumptive pronouns. (17a) shows a coordinated structure where none of the resumptives is removed. If gaps and resumptives are the same, it should be possible to replace both resumptives with a gap, simultaneously maintaining the grammaticality of the sentence. This however is not the case in Modern Greek, as exemplified in (17b):

(17) a. Efige i gata pu o Jiannis tin
left.3SG the.FSG.NOM cat.FSG.NOM that the.MSG.NOM John.MSG.NOM CL.3.FSG.ACC
agapai poli ke pu tin prosehi san na ine pedi tu.
love.3SG very much and that CL.3.FSG.ACC looks after as to be child his.
'The cat that John loves very much and looks after as if it was his own child left.'

b. * Efige i gata pu o Jiannis Ø agapai
left.3SG the.FSG.NOM cat.FSG.NOM that the.MSG.NOM John.MSG.NOM love.3SG
poli ke pu Ø prosehi san na ine pedi tu.
very much and that look.3SG after like to be child his.
'The cat that John loves very much and looks after as if it was his own child left.'

The sentence's grammaticality is ameliorated if we extract the resumptive pronoun from the conjunct closer to the modifying element. This could also be related to the fact that resumptives become more obligatory the more deeply embedded in a sentence they are, as shown in (18):

(18) ? Efige i gata pu o Jiannis Ø aghapai
left.3SG the.FSG.NOM cat.FSG.NOM that the.MSG.NOM John.MSG.NOM love.3SG
poli ke pu tin prosehi san na ine pedi tu.
very much and that CL.3.FSG.ACC look.3SG after like to be child his.
'The cat that John loves very much and looks after as if it was his own child left.'

3.2.4 Parasitic Gaps

Engdahl (1985) suggests that if the RP licenses a parasitic gap, this fact can be considered as evidence in favour of the view that RPs are spelled-out gaps. Evidence from Modern Greek RRCs in (19) shows that parasitic gaps are not licensed:

(19) O mathitis pu den borusan i kathigites na tu_i
the.MSG.NOM student.MSG.NOM that not could.3PL the professors to CL.3.MSG.GEN
eksigisun oti ihe apovlithi horis na $Øp_i$ kalesun sto grafio efige.
explain.3PL that had.3SG been expelled without to invite.3PL to the office left.3SG
'The student that the professors could not explain (to him) that he had been expelled without inviting him to the office left.'

The same applies to parasitic gaps on adjuncts as in (20a), although if the parasitic gap is licensed by a gap, the grammaticality of the sentence is improved as in (20b):

(20) a. * Na ta vivlia pu ta_i edhose horis na $Øp_i$
there are the.NPL.NOM books.NPL.NOM that CL.3.NPL.ACC gave.3SG without to
dhiavasi.
read

b. ? Na ta vivlia pu $Ø_i$ edhose horis na $Øp_i$ dhiavasi.
there are the.NPL.NOM books.NPL.NOM that gave.3SG without to read
'There are the books which she gave without reading them.'

3.2.5 Form - Identity Effects

Another argument put forward by Merchant (2001) in favour of a different treatment of gaps and resumptives is that contrary to Filler-Gap Dependency constructions, Binder-Resumptive Dependency "constructions exhibit certain *form-identity effects*" (Asudeh, 2004, p. 128) such as case-marking. In other words, in a Binder-Resumptive Dependency the binder cannot receive the case of the argument position of the resumptive, since this case is assigned to the resumptive pronoun itself. On the contrary, in Filler-Gap Dependencies the filler is understood as sharing its position with the gap, and consequently receives (among other things) the case of the gap. Modern Greek exhibits this behaviour as illustrated in (21):

(21) a. Pjos itan o fititis pu tu edoses
who.MSG.NOM was.3SG the.MSG.NOM student.MSG.NOM that CL.3.MSG.GEN gave.2SG
hastuki?
slap
'Who was the student you slapped?'

b. * Pjon itan o fititis pu tu edoses
who.MSG.ACC was.3SG the.MSG.NOM student.MSG.NOM that CL.3.MSG.GEN gave.2SG
hastuki?
slap
'Who was the student you slapped?'

This argument is further reinforced by Mackridge (1985, p. 252)'s observation of cases of *anakoluthon*, where *pu* is used without a resumptive pronoun in which case ambiguity arises, as is (22):

(22) a. Tus monus pu Ø akuse i dikastis itan
the.MPL.ACC only.MPL.ACC that heard.3SG the.FSG.NOM judge.FSG.NOM were
i astinomiki.
the.MPL.NOM policemen.MPL.NOM
'The policemen were the only (people) the judge listened to.'

Mackridge (1985) suggests that in such constructions, the "antecedent, instead of a relative pronoun, indicates government by the verb of the relative clause or by a preposition which equally belongs to the relative clause" (Mackridge, 1985, p. 252). If the resumptive pronoun was in the position of the gap, the example would be ungrammatical, as illustrated in (23):

(23) *Tus monus pu tus akuse i dikastis itan
the.MPL.ACC only.MPL.ACC that CL.MPL.ACC heard.3SG the.FSG.NOM judge were
i astinomiki.
the.MPL.NOM policemen.MPL.NOM
'The policemen were the only (people) the judge listened to.'

4 LFG Analysis

As we have observed, the overwhelming majority of the test results in Section 3.2 indicate that gap and resumptive relative clauses in Modern Greek are dissimilar. Based on this evidence, we adopt an alternative approach to that of Alexiadou and Anagnostopoulou (2000) and Alexopoulou (2006): we argue in favour of a distinct treatment of resumptive pronouns and gaps. Thus, we distinguish between two types of dependencies, Binder-Resumptive Dependencies and Filler-Gap Dependencies, and outline an LFG analysis along the lines of Asudeh (2004)'s account for Irish, Swedish and Hebrew.

To begin with, based on the claim (section 3.1) that RPs in MG RRCs are the normal pronouns of the language, we define RPs in the lexicon similarly to pronouns – having, that is, 'PRO' as the value of their PRED value and bearing marking for case, number, gender and person. However, its type is contributing additional information by the $(\uparrow \text{PRONTYPE}) = \text{RP}$ equation, which indicates that it is resumptive pronoun. The lexical entry for the third person feminine RP in Genitive case, for example, is as in (24):

(24) *tis* NP
$(\uparrow \text{PRED}) = \text{'PRO'}$
$(\uparrow \text{GEND}) = \text{F}$
$(\uparrow \text{NUM}) = \text{SG}$
$(\uparrow \text{CASE}) = \text{GEN}$
$(\uparrow \text{PERS}) = 3$
$(\uparrow \text{PRONTYPE}) = \text{RP}$

In addition to that, we define the lexical entries for the relativisers *pu* and *opios* as in (39) and (40) (the lexical entry for the MSG.NOM form of the relative pronoun is shown):

(25) *pu* C
$(\uparrow \text{PRED}) = \text{'PRO'}$
$(\uparrow \text{RELFORM}) = \textit{pu}$

(26) *opios* NP
($\uparrow$ PRED) = 'PRO'
($\uparrow$ RELFORM) = *opios*
($\uparrow$ PERS) = 3
($\uparrow$ GEND) = M
($\uparrow$ NUM) = SG
($\uparrow$ CASE) = NOM
($\uparrow$ DEF) $=_c$ +

Both *pu* and *opios* have a RELFORM (RELativiser FORM) feature with different values (pu and $opios$ respectively). Contrary to *opios*, however, *pu* does not have any agreement marking for gender, case or number. Furthermore, the constraining equation ($\uparrow$ DEF)$=_c$ + on the *opios* lexical entry, ensures that it will be preceeded by a definite article.

The different grammatical category and the different value for the RELFORM feature is what differentiates *pu* from *o opios*-RRCs, which together with the case and the grammatical function specification on the resumptive pronoun node is essential to our account of the distribution of resumption in *pu* and *opios*-RRCs.

In addition to the lexical entries for the resumptive pronoun and the relativisers, we propose the following phrase structure rules for *pu* and *o opios*-RRCs. The DP rule in (27) accounts for the relationship between the modified nominal phrase (D') and the modifying RRC (CP). The modified element is the head and the set membership function $\downarrow\in$ ($\uparrow$ ADJUNCT) on the optional CP node, suggests that the relative clause will be treated as an adjunct on the head D'.

(27) DP $\rightarrow$ D' (CP).
$\uparrow=\downarrow$ $\downarrow\in$ ($\uparrow$ ADJUNCT)

The rule in (28) assumes the simplest phrase structure possible inside the nominal head-element.

(28) D' $\rightarrow$ D NP.
$\uparrow=\downarrow$ $\uparrow=\downarrow$

Appropriate agreement relations between the NP and the D are established through the appropriate agreement feature marking on the lexical entries, as shown in (29) and (30).

(29) *o* D
($\uparrow$ GEND) = M
($\uparrow$ NUM) = SG
($\uparrow$ CASE) = NOM
($\uparrow$ DEF) = +

(30) *skilos* NP
($\uparrow$ PRED) = 'DOG'
($\uparrow$ PERS) = 3
($\uparrow$ GEND) = M
($\uparrow$ NUM) = SG
($\uparrow$ CASE) = NOM

In addition to the above, the CP rule in (31) accounts for the relationships inside *pu-* and *o opios*-RRCs. In particular, it successfully accounts for the internal constituent order of the RRCs: they are introduced either by an element of grammatical category C (for complementizers like *pu* (that)) or by a DP (such as the relative pronoun *o opios* (who.MSG.NOM)) followed by an S_{rel}. The disjunction on the two grammatical categories ensures that the complementizer and the relative pronoun will be mutually exclusive.

(31) CP → { C
(↑ TOPIC) =↓
(↑ CLAUSE-TYPE) = REL

| DP
(↑ TOPIC) =↓
(↑ CLAUSE-TYPE) = REL
(↑ RELPRO) = (↑ TOPIC)
(↓ RELFORM) =$_c$ *opios*
((ADJUNCT ∈ ↑)NUM) = ((ADJUNCT ∈ ↑) ∈ ADJUNCT RELPRO NUM)
((ADJUNCT ∈ ↑)GEND) = ((ADJUNCT ∈ ↑) ∈ ADJUNCT RELPRO GEND)
{ (↑ SUBJ) =↓ (↓ CASE) = NOM
| (↑ OBJ) =↓ (↓ CASE) = ACC
| (↑ OBJ2) =↓ (↓ CASE) = GEN} }

S_{rel}.
↑=↓

In particular, the (↑ CLAUSE-TYPE) = REL specification on the C node states that the modifying element is a relative clause and the (↑ TOPIC) =↓ equation indicates that the information from the lexical entry of the relativizer will be part of the mother's TOPIC f-structure. Furthermore, as observed before, since *pu* is unmarked for number, case and gender, no agreement related information is necessary.

On the DP node, the first two equations work similarly to those appearing on the C node. Moreover, the (↑ RELPRO) = (↑ TOPIC) annotation coindexes the RELPRO f-structure with the TOPIC f-structure and the (↓ RELFORM) =$_c$ *opios* equation ensures that the DP introducing a Relative Clause is a relative pronoun and not any DP. Furthermore, we account for the fact that the relative pronoun gets its case depending on the grammatical function it fulfils in the RRC by defining a set of disjoint equations. (↑ OBJ) =↓ (↓ CASE) = ACC, for instance, ensures that if the relative pronoun is in ACC case, it will be an OBJ. On the other hand, number and gender agreement between the relative pronoun and its antecedent is accounted by inside-out functional uncertainties, reproduced in (32):

(32) ((ADJUNCT ∈ ↑)NUM) = ((ADJUNCT ∈ ↑) ∈ ADJUNCT RELPRO NUM)
((ADJUNCT ∈ ↑)GEND) = ((ADJUNCT ∈ ↑) ∈ ADJUNCT RELPRO GEND)

Finally, the {C|DP} disjunction ensures that the two relativisers will appear in mutually exclusive environments.

Last, but not least, the S_{rel} rule in (33) contains information on the elements of the RRC following the relativizers.

(33) $S_{rel} \rightarrow \{ \epsilon$
$\{ (\uparrow \text{TOPIC}) = (\uparrow \text{GF}) \ (\uparrow \text{TOPIC RELFORM}) =_c opios$
$| \ (\uparrow \text{TOPIC}) = (\uparrow \{\text{SUBJ}|\text{OBJ}\}) \ (\uparrow \text{TOPIC RELFORM}) =_c pu \ \}$
$((\text{ADJUNCT} \in \uparrow)\text{NUM}) = ((\text{ADJUNCT} \in \uparrow) \in \text{ADJUNCT RELPRO NUM})$
$((\text{ADJUNCT} \in \uparrow)\text{GEND}) = ((\text{ADJUNCT} \in \uparrow) \in \text{ADJUNCT RELPRO GEND})$

$|$ NP
$(\downarrow \text{PRON-TYPE})=_c \text{RP}$
$\{ (\uparrow\text{OBJ}) = \downarrow \ (\downarrow \text{CASE}) = \text{ACC} \ \{ (\uparrow \text{RELFORM})=_c pu \ | \ (\uparrow\text{RELFORM})=_c o\ opios \ \}$
$| \ (\uparrow\text{OBJ2}) = \downarrow \ (\downarrow \text{CASE}) = \text{GEN} \ (\uparrow \text{RELFORM})=_c pu \ \}$
$((\text{ADJUNCT} \in \uparrow)\text{NUM}) = \downarrow \text{NUM})$
$((\text{ADJUNCT} \in \uparrow)\text{GEND}) = \downarrow \text{GEND}) \ \}$

V
$\uparrow = \downarrow$

DP*.
$\{ (\uparrow \text{SUBJ}) = \downarrow \ (\downarrow \text{CASE}) = \text{NOM}$
$| \ (\uparrow \text{OBJ}) = \downarrow \ (\downarrow \text{CASE}) = \text{ACC}$
$| \ (\uparrow \text{OBJ2}) = \downarrow \ (\downarrow \text{CASE}) = \text{GEN}\}$

The S_{rel} consists of an empty string ϵ or an NP (the resumptive pronoun) followed by a V and zero or more DPs. In our analysis the distribution of RPs in *pu-* and *o opios*-RRCs is accounted by employing a disjunction over the ϵ and the NP node. The difference in the functional information contributed accounts for the difference in the distribution of resumptive pronouns and gaps in RRCs and consequently for the different status of gaps and resumptives.

In particular, with reference to the functional information on the ϵ[7], the $(\uparrow \text{TOPIC}) = (\uparrow \text{GF})$ equation (where $\text{GF} = \{\text{SUBJ}|\text{OBJ}|\text{OBJ2}\}$) ensures that the only kind of dependency the TOPIC can be involved in when a RP is absent is a Filler-Gap Dependency, where the gap shares the same f-structure information with the relevant grammatical function. In addition to the above, the absence of the resumptive pronoun is predicted by the use of a disjunction of equations (reproduced in (34)): its first part accounts for the absence of resumptives in *o opios*-RRCs whereas its second part accounts for its absence in *pu*-RRCs when the clause is in SUBJ and OBJ relativised positions.

(34) $\{ (\uparrow \text{TOPIC}) = (\uparrow \text{GF}) \ (\uparrow \text{TOPIC RELFORM}) =_c opios$
$| \ (\uparrow \text{TOPIC}) = (\uparrow \{\text{SUBJ}|\text{OBJ}\}) \ (\uparrow \text{TOPIC RELFORM}) =_c pu \ \}$

Furthermore, appropriate number and gender agreement information between the head element and the relative clause is contributed by the equations in (35):

(35) $((\text{ADJUNCT} \in \uparrow)\text{NUM}) = ((\text{ADJUNCT} \in \uparrow) \in \text{ADJUNCT RELPRO NUM})$
$((\text{ADJUNCT} \in \uparrow)\text{GEND}) = ((\text{ADJUNCT} \in \uparrow) \in \text{ADJUNCT RELPRO GEND})$

[7]The empty string ϵ represents absence of a c-structure element, but presence of f-structure information. As Dalrymple (2001, p. 175-176) points out a rule with an ϵ in it "does not license the presence the presence of an empty category or node in the c-structure tree; it simply constitutes an instruction to introduce some functional constraints in the absence of some overt word or phrase. No empty node is introduced into the tree," something which will become apparent in the examples following our analysis.

On the other hand, the NP node requires from its daughter f-structure to have a feature PRONTYPE of value RP, using the equation $(\downarrow \text{ PRON-TYPE})=_c \text{ RP}$, thus ensuring that the NP will be a resumptive pronoun. Moreover, the environments where a resumptive pronoun is present are described using a disjunction of equations (repeated in (36)). The first part of the disjunction accounts for the cases when the RP is in OBJ position in both *pu*- and *o opios*-RRCs, whereas the second part of the disjunction accounts for the presence of the RP in more oblique positions (OBJ2) in *pu*-RRCs, also ensuring appropriate case assignment depending on the grammatical function the RP fulfils within the relative clause:

(36) $\{ (\uparrow \text{OBJ}) = \downarrow \quad (\downarrow \text{CASE}) = \text{ACC} \quad \{ \quad (\uparrow \text{RELFORM})=_c \textit{pu} \quad | \quad (\uparrow \text{RELFORM})=_c \textit{o opios} \quad \}$
$| \quad (\uparrow \text{OBJ2}) = \downarrow \quad (\downarrow \text{CASE}) = \text{GEN} \quad (\uparrow \text{RELFORM})=_c \textit{pu} \quad \}$

Finally, appropriate assignment of number and gender and agreement of the resumptive pronoun with its antecedent is ensured by the use of inside-out equation in (37):

(37) $((\text{ADJUNCT} \in \uparrow)\text{NUM} = \downarrow \text{NUM})$
$((\text{ADJUNCT} \in \uparrow)\text{GEND} = \downarrow \text{GEND}) \quad \}$

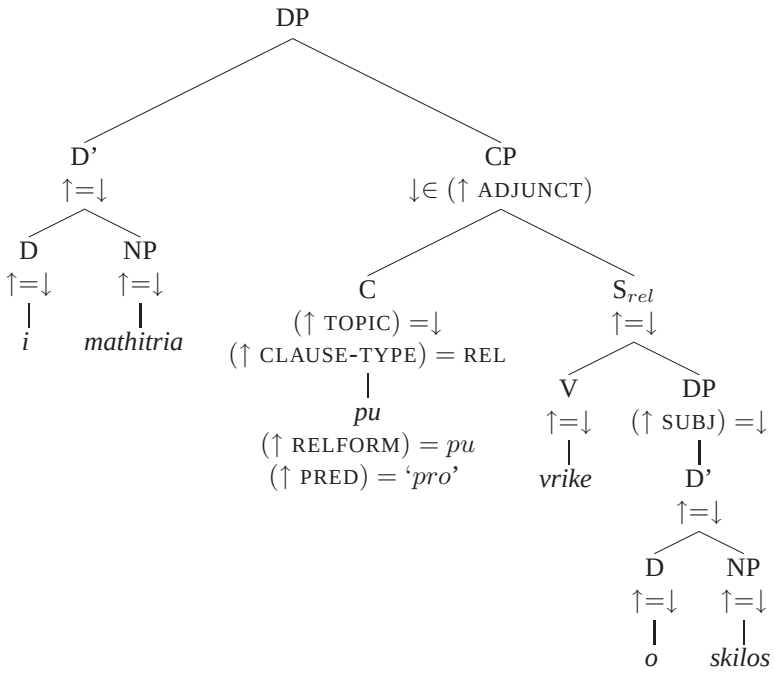
Some examples of *pu*- and *o opios*-RRCs with and without resumptives with their relevant c- and f-structures are shown in examples (38) to (41)[8]:

(38) ***pu*-RRC in Object Position with a Gap**

I mathitria pu Ø vrike o skilos.
the.FSG.NOM student.FSG.NOM that found.3SG the.MSG.NOM dog.MSG.NOM

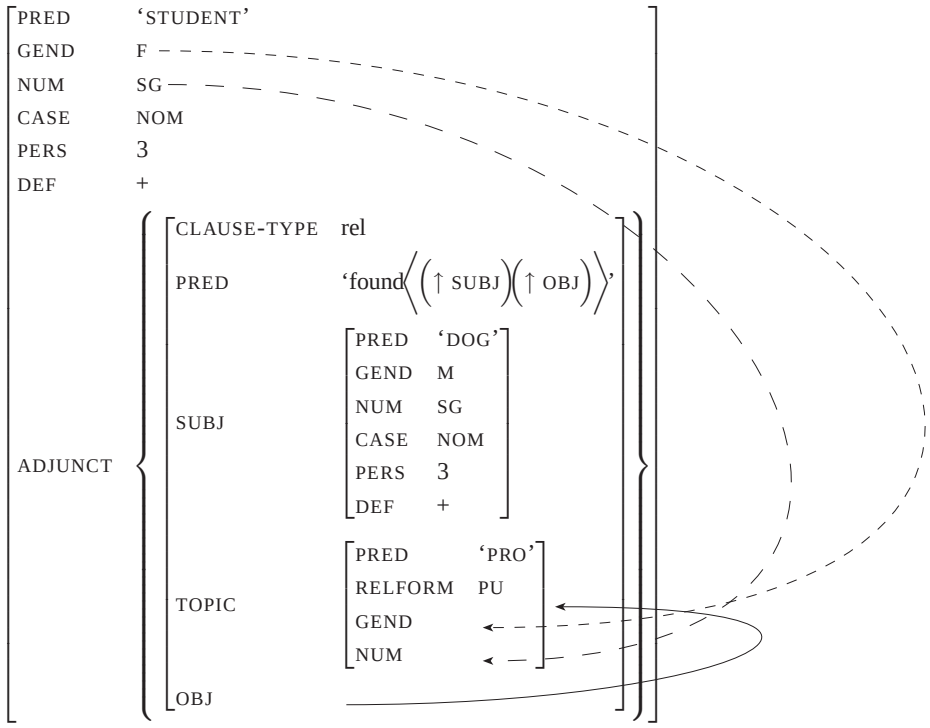
'The student that the dog found.'

a.



[8]Due to space limitations, we have only annotated in detail the nodes which play an important role in our treatment of resumption.

b.

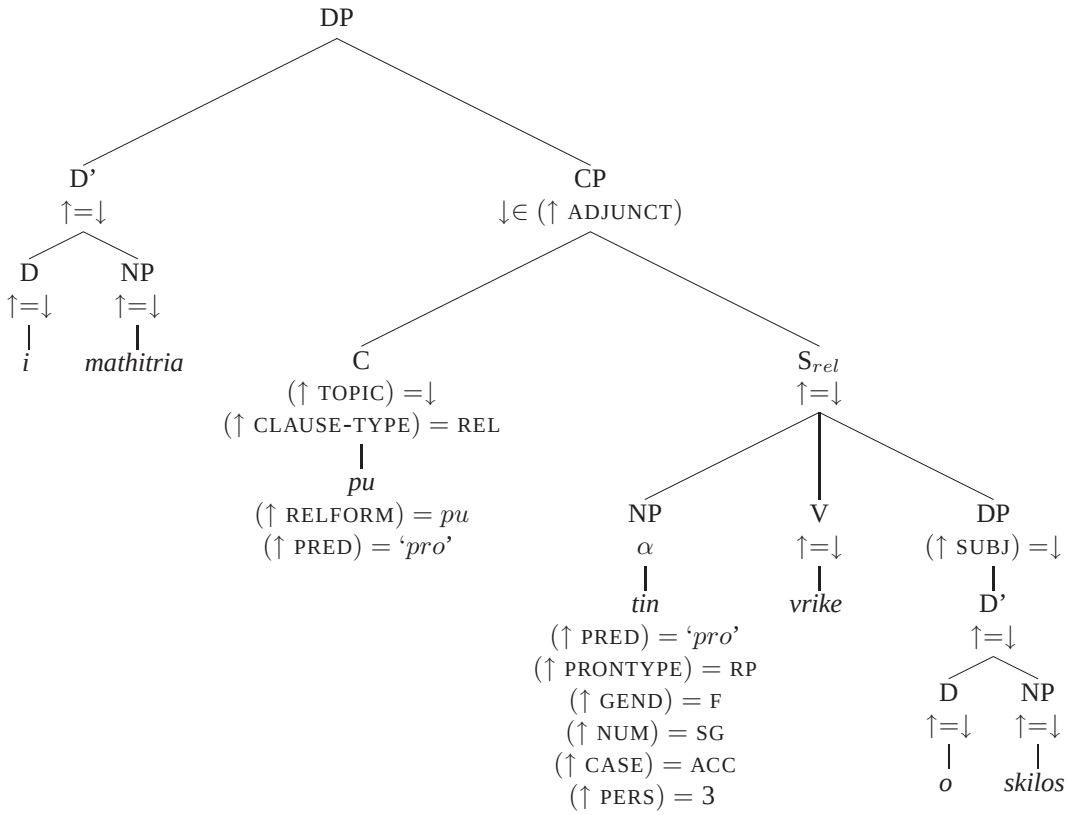


(39) ***pu*-RRC in Object Position with a RP**

I mathitria pu tin vrike o skilos.
the.FSG.NOM student.FSG.NOM that CL.3.FSG.ACC found.3SG the.MSG.NOM dog.MSG.NOM

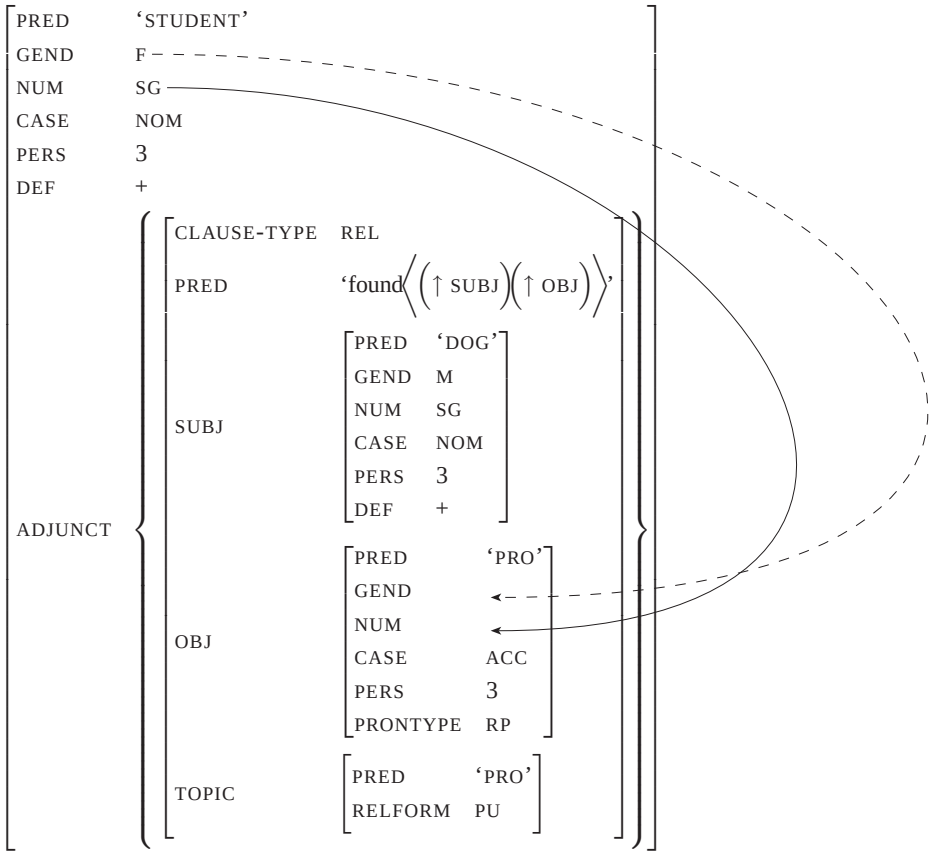
'The student that the dog found (her).'

a.



where α = $(\downarrow \text{PRON-TYPE})=_c \text{RP}$
$\{ (\uparrow \text{OBJ}) = \downarrow \ (\downarrow \text{CASE}) = \text{ACC} \ \{ (\uparrow \text{RELFORM})=_c pu \mid (\uparrow \text{RELFORM})=_c oopios \}$
$\mid (\uparrow \text{OBJ2}) = \downarrow \ (\downarrow \text{CASE}) = \text{GEN} \quad (\uparrow \text{RELFORM})=_c pu \ \}$
$((\text{ADJUNCT} \in \uparrow)\text{NUM}) = (\downarrow \text{NUM})$
$((\text{ADJUNCT} \in \uparrow)\text{GEND}) = (\downarrow \text{GEND}) \ \}$

b.

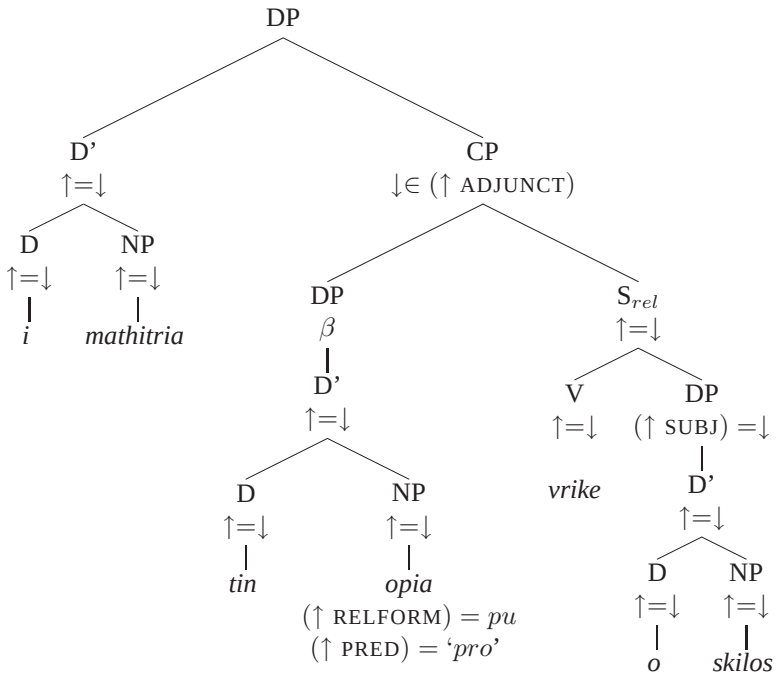


(40) ***o opios*-RRC in Object Position with a Gap**

I mathitria tin opia Ø vrike o skilos.
the.FSG.NOM student.FSG.NOM the.FSG.ACC who.FSG.ACC found.3SG the.MSG.NOM dog.MSG.NOM

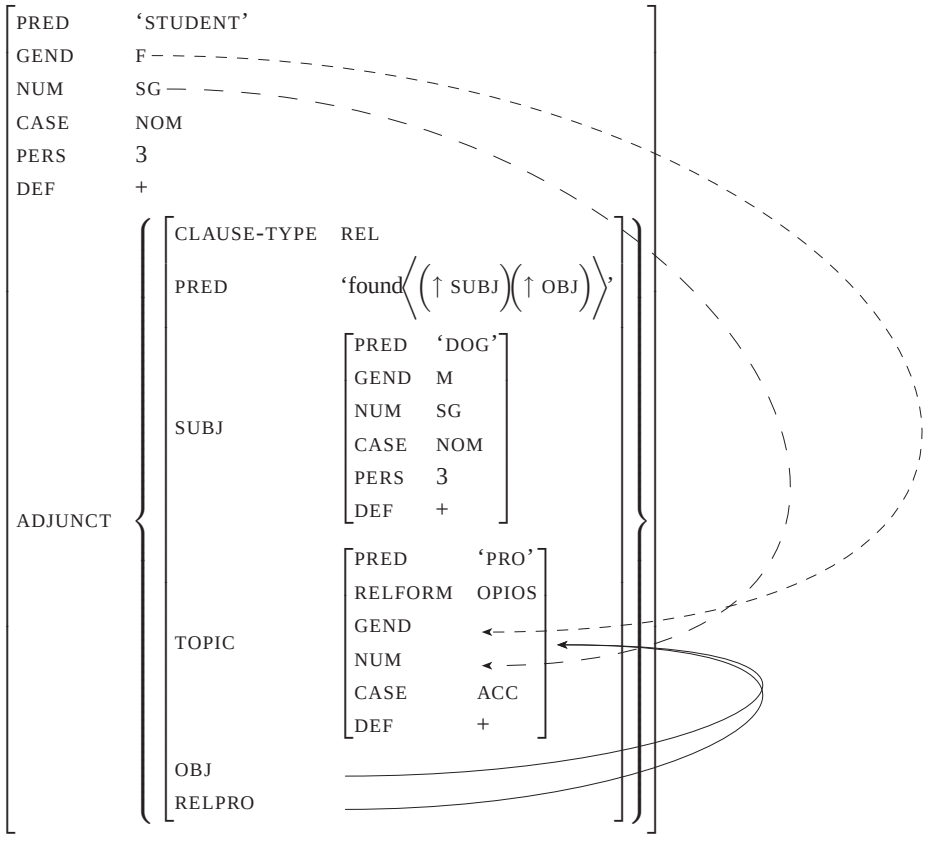
'The student that the dog found.'

a.



where β = $(\uparrow \text{TOPIC}) = \downarrow$
$(\uparrow \text{CLAUSE-TYPE}) = \text{REL}$
$(\uparrow \text{RELPRO}) = (\uparrow \text{TOPIC})$
$(\downarrow \text{RELFORM}) =_c \textit{opios}$
$((\text{ADJUNCT} \in \uparrow)\text{NUM}) = ((\text{ADJUNCT} \in \uparrow) \in \text{ADJUNCT RELPRO NUM})$
$((\text{ADJUNCT} \in \uparrow)\text{GEND}) = ((\text{ADJUNCT} \in \uparrow) \in \text{ADJUNCT RELPRO GEND})$
$(\uparrow \text{OBJ}) = \downarrow$

b.

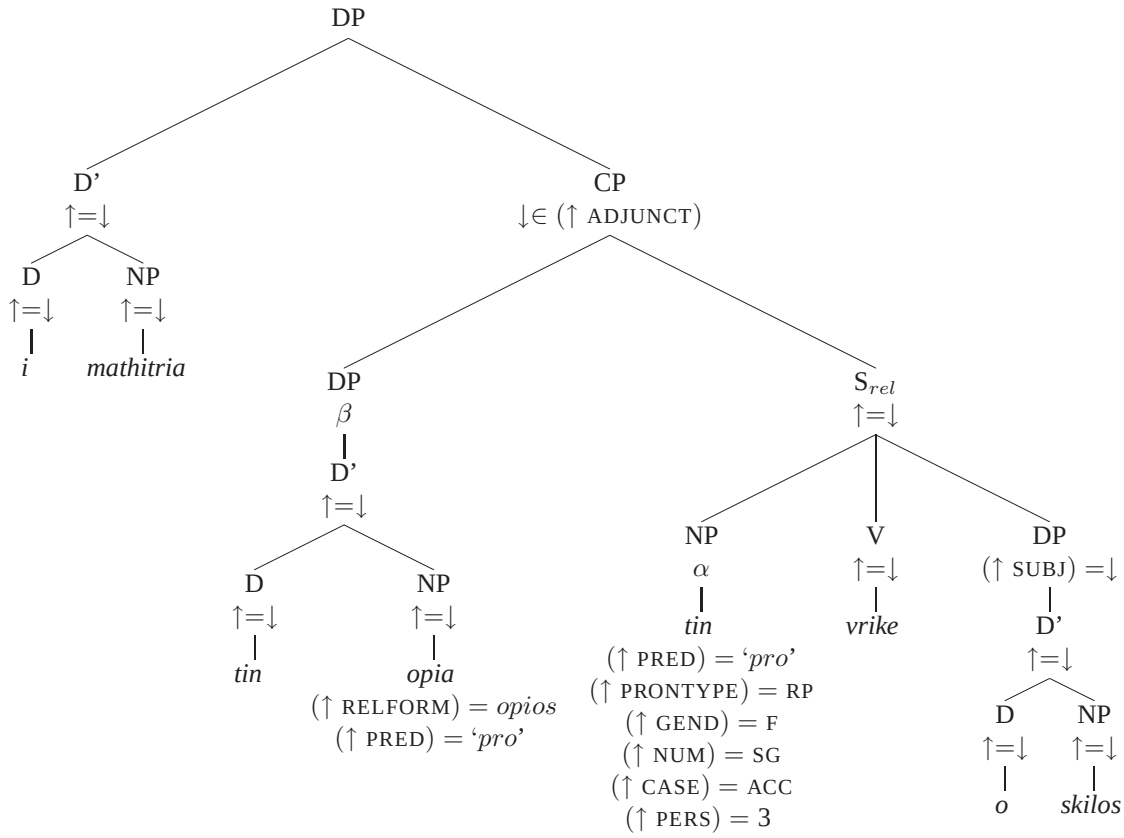


(41) ***o opios*-RRC in Object Position with a RP**

I	mathitria	tin	opia	tin	vrike
the.FSG.NOM	student.FSG.NOM	the.FSG.ACC	who.FSG.ACC	CL.3.FSG.ACC	found.3SG
o	skilos.				
the.MSG.NOM	dog.MSG.NOM				

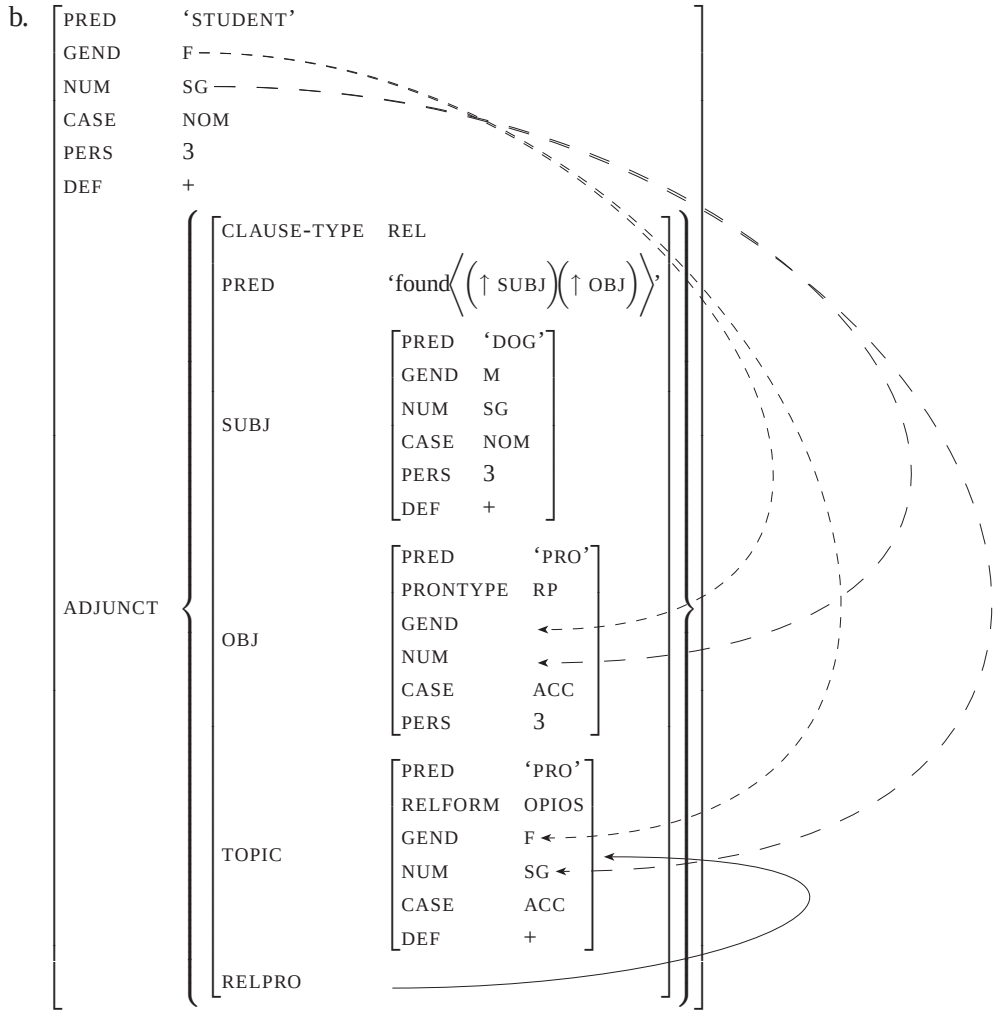
'The student whom the dog found (her).'

a.



where α = (↓ PRON-TYPE)=$_c$ RP
{ (↑ OBJ) =↓ (↓ CASE) = ACC { (↑ RELFORM)=$_c$ *pu* | (↑ RELFORM)=$_c$ *o opios* }
| (↑ OBJ2) =↓ (↓ CASE) = GEN (↑ RELFORM)=$_c$ *pu* }
((ADJUNCT ∈ ↑)NUM) = (↓ NUM)
((ADJUNCT ∈ ↑)GEND) = (↓ GEND) }

and β = (↑ TOPIC) =↓
(↑ CLAUSE-TYPE) = REL
(↑ RELPRO) = (↑ TOPIC)
(↓ RELFORM) =$_c$ *opios*
((ADJUNCT ∈ ↑)NUM) = ((ADJUNCT ∈ ↑) ∈ ADJUNCT RELPRO NUM)
((ADJUNCT ∈ ↑)GEND) = ((ADJUNCT ∈ ↑) ∈ ADJUNCT RELPRO GEND)
(↑ OBJ) =↓



5 Conclusion

In this paper we discussed the status of Modern Greek Resumptive pronouns in restrictive relative clauses. We argued that resumptive pronouns are the ordinary pronouns of the language and that they are not alternative manifestations of gaps, basing our argumentation on a series of tests put forward by Asudeh (2004). For this purpose dependencies involving resumptives and dependencies involving gaps were accounted for separately. Finally, based on these arguments, we presented an LFG analysis in which resumptive restrictive relatives and gap restrictive relatives get a distinct treatment similarly to Asudeh (2004)'s account of the syntax of resumption for Hebrew, Irish and Swedish.

Acknowledgements

I am grateful to the participants of the *LFG06 Conference*, Universität Konstanz, Germany and the *Language and Computation Day*, University of Essex, UK and in particular Doug Arnold, Ron Artstein, Ash Asudeh, Yehuda Falk, Martin Forst, Ron Kaplan, Helge Lødrup, Elizabeth Mayer, Ingo Mittendorf, Massimo Poesio, Louisa Sadler and Alexandros Tantos for insightful suggestions and comments. I would like to extend my gratitude to the local organisers, in particular Miriam Butt, Jürgen Schuhmacher and Hannah Flohr for their warm welcome and hospitality. I would also like to thank Miriam Butt, Tracy Holloway King, Miltiadis Kokkonidis, Louisa Sadler and 4 anonymous reviewers for carefully reviewing

and commenting on earlier drafts of this paper. Needless to say, I am solely responsible for any remaining errors.

This paper is part of work funded by an 1+3 Quota ESRC Award no. PTA-031-2004-00112, support which is gratefully acknowledged.

References

Alexiadou, A. (1998). On the Structure of Greek Relative Clauses. In *Studies on Greek Linguistics. Proceedings of the 18th Annual Meeting of the Dept. of Linguistics, Aristotle University of Thessaloniki*, volume 18, pages 15–29. Thessaloniki: Kiriakidis.

Alexiadou, A. and Anagnostopoulou, E. (2000). Asymmetries in the Distribution of Clitics: the Case of Greek Restrictive Relatives. In Beukema, F. and den Dikken, M., editors, *Clitic Phenomena in European Languages*, pages 47–70. Amsterdam: John Benjamins Publishing Company.

Alexopoulou, T. (2006). Resumption in Relative Clauses. *Natural Language and Linguistic Theory*, 24(1):57–111.

Asudeh, A. (2004). *Resumption as Resource Management*. PhD thesis, Stanford University.

Dalrymple, M. (2001). *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. London *et al* : Academic Press.

Engdahl, E. (1985). Parasitic Gaps, Resumptive Pronouns and Subject Extractions. *Linguistics*, 23:3–44.

Falk, Y. (2002). Resumptive Pronouns in LFG. In Butt, M. and King, T. H., editors, *The Proceedings of the LFG '02 Conference, The Proceedings of the LFG '02 Conference*. CSLI Publications (http://cslipublications.stanford.edu/).

Hatzidakis, G. (1907). *Peri tis antonimias o opios [About the pronoun o opios]*, volume B; of *Mesaionika kai Nea Ellinika*, pages 593–597.

Kayne, R. S. (1994). *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.

Koopman, H. (1982). Control from COMP and Comparative Syntax. *Linguistic Review*, 2(4):365–391.

Lascaratou, C. (1998). Basic Characteristics of Modern Greek. In Siewierska, A., editor, *Constituent Order in the Languages of Europe*, volume EUROTYP 20-1 of *Empirical Approaches to Language Typology*, pages 151–171. Berlin and New York: Mouton de Gruyter.

Mackridge, P. (1985). *The Modern Greek Language. A Descriptive Analysis of Standard Modern Greek*. Oxford: Clarendon Press.

Markantonatou, S. (1992). *The Syntax of Modern Greek NPs with a Deverbal Nominal Head*. PhD thesis, Department of Language and Linguistics, University of Essex.

McCloskey, J. (1979). *Transformational Syntax and Model Teoretic Semantics: a Case-study in Modern Irish*. Dordrecht: Reidel.

McCloskey, J. (1990). Resumptive Pronouns, A' Binding and Levels of Representation in Irish. In Randall, H., editor, *Syntax of the Modern Celtic Languages*, volume 23 of *Syntax and Semantics*, pages 199–248. San Diego, CA: Academic Press.

McCloskey, J. (2002). Resumption, Successive Cyclicity, and the Locality of Operations. In Epstein, S. D. and Seeley, D. T., editors, *Derivation and Explanation in the Minimalist Program*, pages 184–226. Oxford: Blackwell.

Merchant, J. (2001). *The Syntax of Silence*. Oxford: Oxford University Press.

Philippaki-Warburton, I. (1985). I Theoria ton Kenon Katigorion: to Ellipon Ypokimeno kai i Klitikes Antonimies sti Nea Elliniki [On the Theory of Empty Categories: the Missing Subject and the Clitic Pronouns in Modern Greek]. In *Studies in Greek Linguistics. Proceedings of the 6th Annual Meeting of the Department of Linguistics, University of Thessaloniki*, volume 6, pages 131–153. Thessaloniki: Kiriakidis.

Sells, P. (1984). *Syntax and Semantics of Resumptive Pronouns*. PhD thesis, University of Massachusetts, Amherst.

Shlonsky, U. (1992). Resumptive Pronouns as a Last Resort. *Linguistic Inquiry*, 23:443–468.

Theophanopoulou-Kontou, D. (1989). Domes tis Sinthetis OF kai Metakinisi stin Nea Elliniki [The structure of the complex NP and Movement in Modern Greek]. In *Studies in Greek Linguistics. Proceedings of the 9th Annual Meeting of the Department of Linguistics, University of Thessaloniki*, volume 9, pages 337–354. Thessaloniki: Kiriakidis.

Tzartzanos, A. (1963). *Neoelliniki Syntaxis [Modern Greek Syntax]*, volume II. Athens: Organismos Ekdoseos Didaktikon Vivlion, 2nd edition.

Zaenen, A., Engdahl, E., and Maling, J. (1981). Resumptive Pronouns can be Syntactically Bound. *Linguistic Inquiry*, 12:679–682.

Aikaterini K. Chatsiou
Department of Language and Linguistics, University of Essex
Wivenhoe Park, Colchester, CO4 3SQ, UK
e : achats@essex.ac.uk
w : http://privatewww.essex.ac.uk/∼achats/

IMPROVING TREEBANK-BASED AUTOMATIC LFG INDUCTION FOR SPANISH

Grzegorz Chrupała and Josef van Genabith
National Centre for Language Technology and School of Computing
Dublin City University

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

We describe several improvements to the method of treebank-based LFG induction for Spanish from the Cast3LB treebank (O'Donovan et al., 2005). We discuss the different categories of problems encountered and present the solutions adopted. Some of the problems involve a simple adoption of existing linguistic analyses, as in our treatment of clitic doubling and null subjects. In other cases there is no standard LFG account for the phenomenon we wish to model and we adopt a compromise, conservative solution. This is exemplified by our treatment of Spanish periphrastic constructions. In yet another case, the less configurational nature of Spanish means that the LFG annotation algorithm has to rely mostly on Cast3LB function tags, and consequently a reliable method of adding those tags to parse trees had to be developed. This method achieves over 6% improvement over the baseline for the Cast3LB-function-tag assignment task, and over 3% improvement over the baseline for LFG f-structure construction from function-tag-enriched trees.

1 Introduction

The research reported in this paper has been carried out as part of the GramLab project whose goal is to acquire multilingual wide coverage LFG resources from treebanks for several languages. We report on the ongoing work in LFG induction for Spanish.

Inducing deep syntactic analyses from treebank data avoids the cost and time involved in manually creating wide-coverage resources.

LFG f-structures provide a level of syntactic representation which is more abstract and cross-linguistically more uniform than constituency trees. F-structures include explicit encodings of phenomena such as control and raising, pro-drop and long distance dependencies: those characteristics make this level a suitable representation for many NLP applications such as transfer-based Machine Translation or Question Answering.

A methodology for automatically obtaining LFG f-structures from trees output by probabilistic parsers trained on the Penn-II treebank has been described by (Cahill et al., 2004). The f-structure annotation algorithm used for inducing LFG resources from the Penn-II treebank for English uses configurational, categorial, function tag and trace information.

Preliminary research on Spanish LFG induction was carried out by (O'Donovan et al., 2005). In the present paper we discuss several issues which became obvious while trying to expand the coverage of Spanish grammatical constructions and phenomena and while dealing with the peculiarities of the treebank that we are using. The problems arising from adapting a grammar acquisition methodology devel-

oped for one language/treebank to another language/treebank combination fall into three broad categories:

- New phenomena and constructions, successfully treated within standard LFG: clitic doubling, null subjects
- New phenomena and constructions, problematic within standard LFG: clitic climbing (i.e. complex predicates)
- Limitations of previous approach due to language/treebank specific assumptions which no longer hold: flexible constituent order and less configurational c-structures

2 Clitic doubling and null subjects

In Spanish pronominal clitics for Direct and Indirect Object can co-occur with non-clitic (full NP) objects.[1] Example 1 shows clitic doubling with Indirect Object, Example 2 with Direct Object. The non-clitic Objects are in italics; the co-occurring clitics are in bold. The clitics agree with the non-clitic arguments in person, number, gender and case.

(1) Algo parecido **les** sucede *a los hombres.*
something similar them occurs to DEF men
Something similar happens to men.

(2) Cada cual **lo** comprende *eso* a su manera.
every which it understands this to POSS manner
Everyone understands this in their own way.

Clitic doubling is quite common with Indirect Objects: in our treebank data in 23% of the cases where there is a non-pronominal Indirect Object it co-occurs with a pronominal clitic. Clitic doubling for Direct Objects is more constrained, but still relatively common at 1% of corpus occurences of non-pronominal Direct Objects.

In clitic doubling constructions, pronominal clitics should not introduce a PRED value, as that would clash with the one introduced by the non-clitic Object. However when clitics are not accompanied by non-clitic Objects, they should introduce PRED = 'pro', in order to satisfy the verb's subcategorization requirements.

[1] This phenomenon is subject to complex, dialect-dependent constraints involving animacy, specificity and information structure, especially for Direct Object. Currently we do not try to model these constraints fully.

We achieve this effect by means of optional equations, as is standard practice in LFG. Example 3 below illustrates the equations associated with the dative *le* (Indirect Object).

(3) *le* **pp3csd00**

((↑ PRED) = 'pro')
((↑ PRON-TYPE) = PERS)
((↑ PRON-FORM) = el)
(↑ CASE) = DAT
(↑ NUM) = SG
(↑ PERS) = 3

An optional equation (e) is a disjunction of e and *true*. In standard LFG the correct disjunct is chosen as follows: in a clitic-doubling context, the first disjunct is excluded because the PRED value it introduces clashes with the one introduced by the non-clitic Object, and thus the *true* disjunct applies. In non-doubling contexts, the first disjunct applies successfully, while if the second one applies, the resulting f-structure does not satisfy completeness because of the missing PRED value.

In our implementation we do not check for completeness because our PRED values lack subcategorization frames,[2] so we use a slightly different definition of optionality. An optional equation works more like a default equation: the optional equation $((f\ \ a) = v)$ holding of f-structure f is interpreted as a disjunction of the existential constraint $(f\ \ a)$ and the equation $(f\ \ a) = v$. In the clitic-doubling case the second disjunct (which introduces the PRED value) only applies if the PRED value has not been contributed by some other equation.

Another area where we use optional equations is in our treatment of null subjects (pro-drop). In Spanish explicit subjects are often absent. Subject features such as person and number are encoded in agreement morphology on the verb instead. When there is no overt subject, the PRED value that is needed to satisfy the verb's subcategorization is introduced by the inflected verb-form.

All finite verb preterminals optionally introduce a 'pro' subject. Example 4 below illustrates the annotation associated with the inflected verb form *vió* (see-3SG).

(4) *vió* **vmis3s0**

(↑ PRED)= 'ver'
((↑ PRED SUBJ) = 'pro')
(↑ SUBJ NUM) = SG

[2]The subcat frames are acquired separately in our architecture. See (O'Donovan et al., 2004).

(↑ SUBJ PERS) = 3
(↑ SUBJ TENSE) = PAST
(↑ SUBJ MOOD) = INDICATIVE
(↑ LIGHT) = −

Currently all finite verb forms receive an optional PRED equation. This is not entirely adequate as at least one Spanish verb *haber* (existential be) can never co-occur with an overt subject, so ideally it should receive an obligatory PRED equation. Similarly, weather verbs are normally ungrammatical with explicit subjects (Example 5 a and b). Exceptionally they can take modified cognate subjects (Example 5 c).

(5) (a) * Llovió lluvia.
rained rain

(b) * La lluvia llovió.
the rain rained

(c) Llovió una lluvia fina pero persistente.
rained a rain light but persistent
"A light but persistent rain rained down."

Whether it is possible to learn from treebank data which verbs do not allow overt subjects and under what conditions remains an open question for future investigation.

Our use of optionality in the treatment of Spanish clitic doubling and null subjects illustrates language-specific problems that arise for LFG induction, but for which there are standard solutions in the LFG framework. Those solutions can be adopted and adapted for our data-driven approach to grammar acquisition. They may require additional implementation effort (in this case adding appropriate optionality support to the constraint solver), but otherwise they can be easily accommodated within the existing methodology.

In the following section we discuss a phenomenon which is more problematic: it does not have a widely agreed-upon solution in standard LFG and thus is an issue in any computational implementation including our own.

3 Periphrastic constructions

In Spanish periphrastic constructions, such as in Example 6 a, verbal pronominal clitics which are understood as arguments of the "lower" verb can attach to the "higher" verb. This phenomenon, called clitic climbing, is only grammatical with

certain verbs. Others do not admit it, as illustrated in Example 6 b. The verbs that do admit clitic climbing are sometimes called *light* verbs.

(6) (a) **La** puedo *ver*. Puedo *ver***la**.
her can-1SG see can-1SG see-her

(b) * **La** insistí en *ver*. Insistí en *ver***la**.
her insisted-1SG in see insisted-1SG in see-her

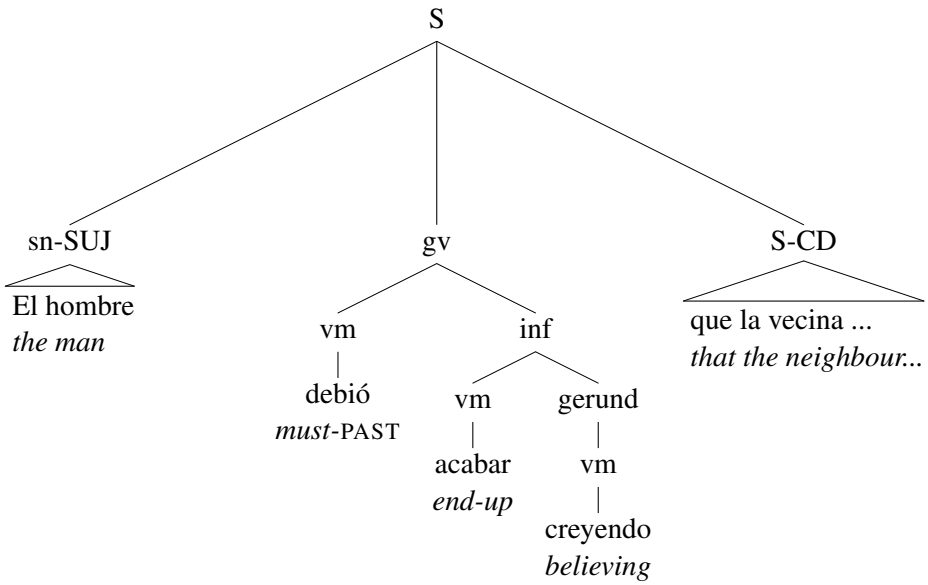
Normally only the clitic climbing versions of periphrastic constructions present difficulties for an LFG account due to the mismatch of the position of arguments in the tree and where they should end up in the f-structure. However, the configuration adopted for periphrastic constructions in Cast3LB generalizes this problematic mismatch to all contexts.

As illustrated in Figure 1, all verbs participating in the periphrastic construction are under the *gv* (Verb Group) node, with the argument of the lowest verb being attached as sister to the *gv*. This example also illustrates that periphrastic constructions can be combined with each other, so in principle the lowest non-light verb could be nested a number of levels deep.

There are several proposals of how to deal with periphrastic constructions with clitic climbing within LFG. Both (Alsina, 1997) and (Butt, 1997) propose a predicate composition analysis. As in standard LFG PRED values can never unify, this approach requires modifications to the unification operation. In (Andrews and Manning, 1999) the authors propose an even more radical departure from standard LFG and replace the projection architecture with *differential information spreading* within the f-structure.

As there seems to be no consensus as to the best treatment of Romance constructions involving light verbs, we decided in favor of a conservative approach which avoids non-standard extensions to the LFG formalism. We use functional uncertainty and a nested XCOMP configuration in our treatment of periphrastic constructions. The mechanism is illustrated in Figure 2. The *inf(initive)* and *gerund* daughters of the *gv* node constrain the f-structure corresponding to their mother nodes to be LIGHT +, and introduce their own f-structure as the value of XCOMP attribute.

Non-subject sisters of the *gv* are annotated with functional uncertainty equations which specify that their f-structure is the value of the GF attribute arbitrarily embedded in a series of XCOMPs. There is an off-path constraint that specifies that the f-structure containing each of the XCOMPs in the path has to be LIGHT +. Another off-path constraint on the f-structure containing the final GF restricts it to be LIGHT −. Together those annotations ensure that arguments are always



$$\begin{bmatrix}
\text{SUBJ} & [\text{“el hombre”}]\boxed{1} \\
\text{PRED} & \text{‘DEBER’} \\
\text{TENSE} & \text{PAST} \\
\text{LIGHT} & + \\
 & \begin{bmatrix}
\text{SUBJ} & \boxed{1} \\
\text{PRED} & \text{‘ACABAR’} \\
\text{LIGHT} & + \\
\text{XCOMP} & \begin{bmatrix}
\text{SUBJ} & \boxed{1} \\
\text{PRED} & \text{‘CREER’} \\
\text{LIGHT} & - \\
\text{COMP} & [\text{“que la vecina ...”}]
\end{bmatrix}
\end{bmatrix}
\end{bmatrix}$$

Figure 1: Periphrastic construction with two light verbs: The treebank tree, and the f-structure produced

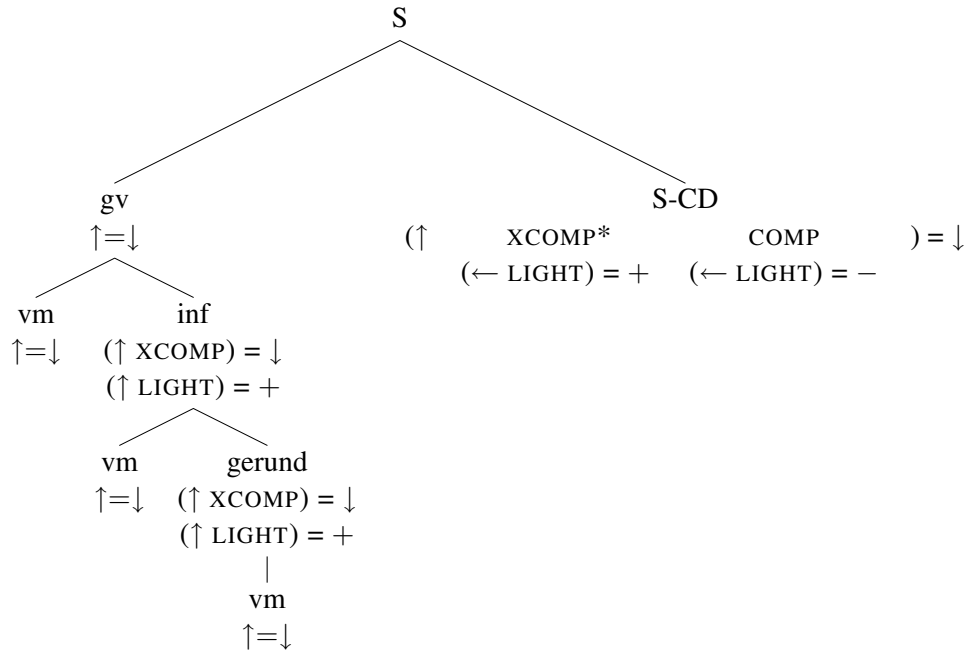


Figure 2: Treatment of periphrastic constructions by means of functional uncertainty equations with off-path constraints

attached to the lowest (non-light) verb. This is the correct analysis for the majority of periphrastic constructions.[3]

Our treatment of periphrastic constructions is not entirely satisfactory: it is a compromise solution. From a descriptive perspective it does not perfectly model the linguistic phenomena in question. Our motivation for using it is that it allows us to avoid implementing a solution which departs too far from the standard LFG formalism and for which there is no consensus among theoretical linguists.

The XCOMP-based treatment is adequate in the vast majority of cases and has the advantage that the resulting f-structure parallels the analysis that would be used in languages with no clitic climbing (such as English) for similar sentences. This could potentially be useful if our LFG resources are to be used in multilingual applications.

[3]One exception are causative constructions, where, if one insists on an XCOMP-type treatment, the causee should be the argument of the causative verb, whereas the other arguments should depend on the verb expressing the event caused (Alsina, 1997).

In the following section we discuss the particular features of our language and treebank which challenge some of the assumptions made in the design of the LFG acquisition architecture initially developed using the English Penn Treebank data.

4 Constituent order and configurationality

The method of automatic LFG induction was initially developed using the English Penn-II Treebank data. The idea behind the annotation rules is that limited configurational and categorial information should in most cases be sufficient to determine a constituent's grammatical function in the sentence: as evidenced by the good results of this approach for English, this assumption is borne out for this language. It turns out that the approach is more problematic for our Spanish Cast3LB data. Spanish allows much more variation and flexibility in major sentence constituent order than does English. Partly as a consequence of this flexibility, the treebank encoding of syntactic structure also has to be different than in the Penn Treebank.

Although the canonical word order for Spanish is SVO, in Cast3LB there are about 20% post-verbal subjects, and about 11% preverbal non-clitic direct objects. Thus the information on position relative to the verb is not a reliable predictor of grammatical function in Spanish.

Accordingly, the Spanish treebank makes extensive use of function tags to make the grammatical function of constituents more explicit. Although there are also functional tags in the Penn Treebank, their use is less necessary. In the Penn Treebank, configuration information alone is often sufficient to determine grammatical function: e.g.: left sister to *VP* is typically a Subject while right daughter to *V* is an Object.

Due to the preceding considerations the Spanish annotation algorithm has to rely on function tags much more heavily than is the case for English. It is thus important to be able to enrich parser-output trees with those tags as reliably as possible.

The initial implementation described in (O'Donovan et al., 2005) relied on the parser itself to obtain function-tagged parse trees. Bikel's parser (Bikel, 2002) was trained on trees where function tags were simply part of the category label, so instead of having one non-terminal category *sn* (Noun Phrase) there are several different NP categories e.g. *sn*-SUJ, *sn*-CD, *sn*-CI, etc. We treated this simple method as a baseline and tried to determine how much we could improve on it.

We decided to let the parser learn and output plain constituency trees and add Cast3LB function tags in a postprocessing step. The intuition behind adopting this approach is that we thus avoid the multiplication of categories (which could potentially lead to a sparse-data-related decline in performance), and also achieve

better control over the learning method and the feature set used than if we just rely on the parser.

Our method and evaluation results are described in detail in (Chrupała and van Genabith, 2006). Here we present a brief outline of this research and elaborate on some LFG-relevant aspects. Although our work is the first attempt to learn the assignment of Cast3LB function tags to parser output for Spanish, there is some existing research on enriching parse trees with Penn function tags for English (Blaheta and Charniak, 2000; Jijkoun and de Rijke, 2004). The general idea is the same in each case: function tags are added to parse tree nodes in a postprocessing step, and the assignment model is learned from treebank data.

In our research we experimented with three machine-learning methods: Memory-Based, Maximum Entropy and Support Vector Machines. The best performance was obtained with SVM and those are the results that we report below.

We treat Cast3LB function tag assignment as a classification task. Our training examples are candidate nodes in treebank trees. We treat as candidate nodes all those that are sisters to either

- *gv* (Verb Group)
- *infinitiu* (Infinitive)
- *gerundi* (Gerund)

The class label assigned to each example is its Cast3LB function tag, or the label NULL if no function tag is present.

For each example node we extract a set of features which are used by the machine-learning algorithm to build the model used to classify unseen examples. Figure 3 illustrates the features extracted from an example tree. The *focus node features* are extracted from the node labeled *sn*-SUJ. The other three nodes provide *context node features*, and the nodes included in the oval area (the head node and the mother node) are used to extract *local features*. The features encode categorial, configurational, morphological and lexical information that we considered relevant for determining functions encoded in the Cast3LB function tags:

- Node features: position relative to head, head lemma, alternative head lemma (i.e. the head of NP in PP), head POS, category, definiteness, agreement with head verb, yield (i.e. number of terminals dominated), human/nonhuman
- Local features: head verb, verb person, verb number, parent category
- Context features: node features (except position) of the two previous and two following sister nodes (if present).

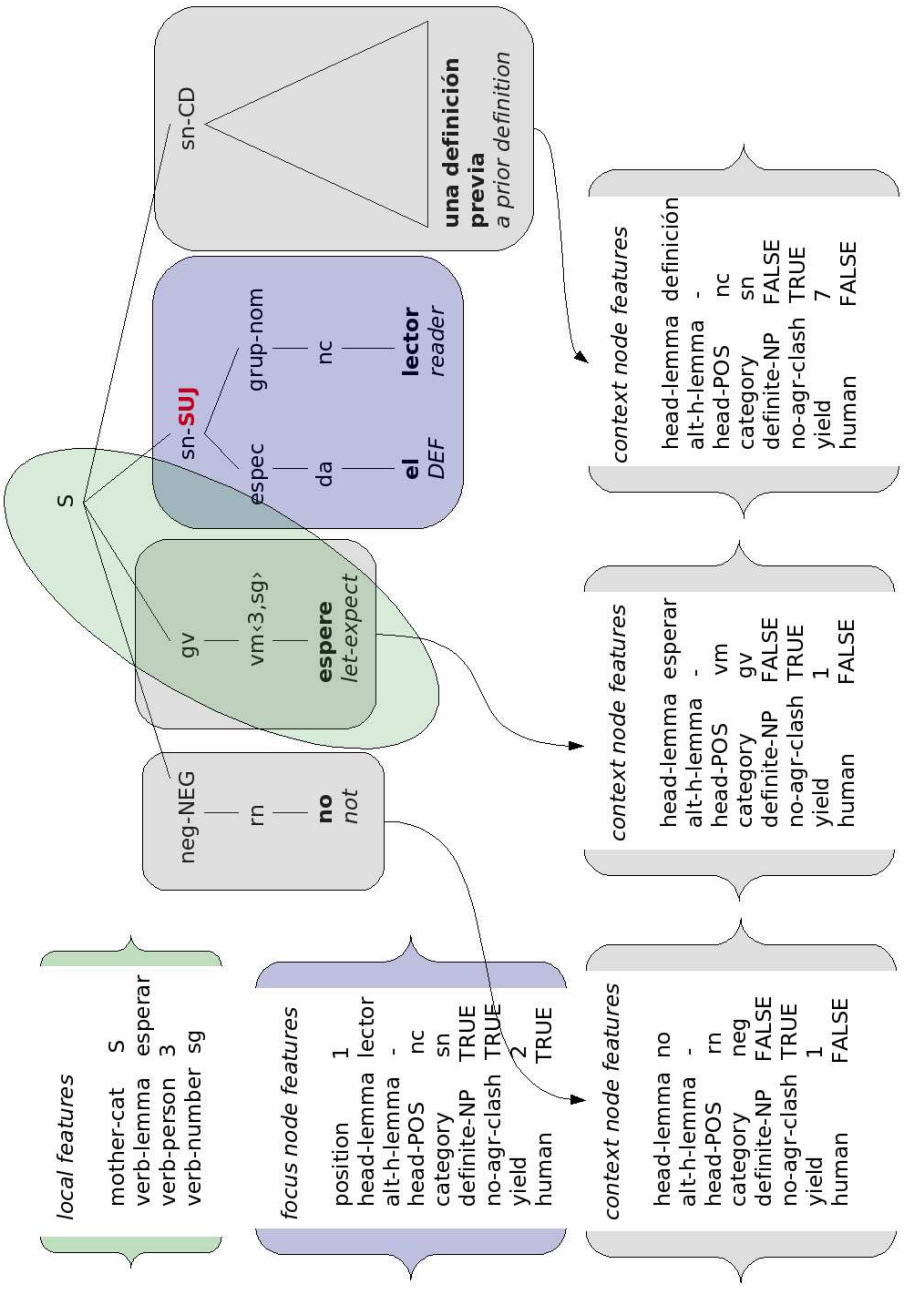


Figure 3: Examples of features extracted from an example node

	Acc.	Prec.	Recall	F-score
SVM	**89.34**	88.93	84.90	**86.87**

Table 1: Cast3LB function tagging performance for gold-standard trees

	Precision		Recall		F-score	
	all	corr.	all	corr.	all	corr.
Baseline	59.26	72.63	60.61	75.35	59.93	73.96
SVM	66.96	80.58	66.38	81.27	**66.67**	**80.92**

Table 2: Cast3LB function tagging performance for parser output

In order to evaluate the performance of the trained classifier we used the following procedure: for each function-tagged tree we first remove the punctuation tokens. Then we extract a set of tuples of the form $\langle \text{GF}, i, j \rangle$, where GF is the Cast3LB function tag and $i - j$ is the range of tokens spanned by the node annotated with this function. For example from the tree in Figure 3 the following set of tuples would be obtained: $\{\langle \text{NEG}, 1, 1 \rangle, \langle \text{SUJ}, 3, 4 \rangle, \langle \text{CD}, 5, 7 \rangle\}$. We use the standard measures of precision, recall and f-score to evaluate those sets of tuples against the ones extracted from the reference gold-standard trees.

Tables 1 and 2 contain the results of Cast3LB function tag assignment evaluation for gold trees (taken from the treebank) and for trees output by Bikel's parser. For parser trees we report the result for all nodes (all), and for the subset of nodes that were correctly bracketed (corr).

The results for parse trees, even for the correctly bracketed node subset, are still lower than for gold trees. We suspect this may be due to the fact that even for correctly bracketed nodes, the context may still contain incorrectly parsed structures. An additional consideration is the fact that we extract training data from treebank trees: perhaps an improvement can be obtained by using parsed trees for training data. We are currently experimenting with this idea.

From the perspective of LFG induction, any improvements in the Cast3LB function tag assignment task are only useful if they translate to better quality f-structures. The mapping from Cast3LB tags to LFG annotations is reasonably straightforward, but not bijective (Table 3 contains the Cast3LB function tags and specifies their correspondence to LFG features). Also LFG function tags are only available for daughters of *S* nodes. For other nodes, the annotation algorithm has

Tag	Meaning	LFG attribute
ATR	Attribute of copular verb	PREDLINK
CAG	Agent of passive verb	OBL_{ag}
CC	Compl. of circumstance	ADJUNCT
CD	Direct object	COMP for finite *S* nodes, XCOMP for non-finite *S* nodes OBJ otherwise
CD.Q	Direct object of quantity	OBJ
CI	Indirect object	OBJ2
CPRED	Predicative complement	PREDLINK
CPRED.CD	Predicative of Direct Object	PREDLINK
CPRED.SUJ	Predicative of Subject	PREDLINK
CREG	Prepositional object	OBL
ET	Textual element	ADJUNCT
IMPERS	Impersonal marker	IMPERS
MOD	Verbal modifier	ADJUNCT
NEG	Negation	NEG
PASS	Passive marker	PASSIVE
SUJ	Subject	SUBJ
VOC	Vocative	ADJUNCT

Table 3: Cast3LB function tags and corresponding LFG f-structure attributes

to rely on other evidence to come up with the correct LFG annotations.

Given those complications we compared the quality of the f-structures produced using our improved function tags against the baseline. The results of the evaluation of the f-structures produced by the two methods are given in Table 2. The difference in f-scores is smaller than in the case of Cast3LB tag assignment. This is most likely due to two facts:

- Tags are available and used for only a subset of nodes
- F-structure evaluation is less sensitive to some forms of incorrect parse trees, i.e. exact constituent boundaries are not important, only correct bracketing of heads.

We also performed a statistical significance test for these results. For each pair of methods we calculate the f-score for each sentence in the test set. For those

	Precision	Recall	F-score
Baseline	73.95	70.67	72.27
SVM	76.90	74.48	75.67

Table 4: F-structure evaluation results for parser output

sentences on which the scores differ (i.e. the number of trials) we calculate in how many cases the second method is better than the first (i.e. the number of successes). We then perform the test with the null hypothesis that the probability of success is chance ($= 0.5$) and the alternative hypothesis that the probability of success is greater than chance (> 0.5). The p-value given by the sign test was 2.118×10^{-5}: thus the improvement is statistically significant at a confidence level of 99%.

5 Conclusions and further work

We have discussed several issues which arose while adapting an automatic treebank-based LFG acquisition method developed originally for the Penn Treebank to the Spanish Cast3LB treebank. The process of porting our method to Spanish (as well as other languages we deal with within the GramLab project) has made it more obvious what are the strengths and weaknesses of our approach.

The less configurational nature of the Cast3LB data made it necessary for the LFG annotation algorithm to rely heavily on function tags, and consequently to develop better methods of obtaining function-tagged parse trees. This improved machine-learning postprocessing method is now also successfully being used for English. Thus expanding the coverage of our method to multiple languages and treebanks also benefits LFG induction for English.

Areas of current and future research include revising the LFG account of some areas of Spanish syntax:

- Replacing COMP with OBJ
- Changing the PREDLINK analysis to one which better reflects the difference between predicative complements of Direct Object vs. of Subject

We also plan to further expand grammar coverage to more kinds of constructions and linguistic phenomena.

In the area of function-tag assignment we believe there is also room for further improvement. Extracting training examples from parse trees rather than treebank

trees should lead to better performance on parser output. Trying to constrain function tag sequences to avoid impossible combinations (such as two SUJ tags) would also be desirable.

Acknowledgements

We gratefully acknowledge support from Science Foundation Ireland grant 04/IN/I527 for the research reported in this paper.

References

Alsina, A. (1997). A theory of complex predicates: evidence from causatives in Bantu and Romance. In Alsina, A., Bresnan, J., and Sells, P., editors, *Complex Predicates*, pages 203–246. Center for the Study of Language and Information, Stanford, CA, USA.

Andrews, A. D. and Manning, C. D. (1999). *Complex Predicates and Information Spreading in LFG*. Center for the Study of Language and Information, Stanford, CA, USA.

Bikel, D. (2002). Design of a multi-lingual, parallel-processing statistical parsing engine. In *Human Language Technology Conference (HLT)*, San Diego, CA, USA. Software available at `http://www.cis.upenn.edu/~dbikel/software.html#stat-parser`.

Blaheta, D. and Charniak, E. (2000). Assigning function tags to parsed text. In *Proceedings of the 1st Conference of the North American Chapter of the ACL*, pages 234–240, Rochester, NY, USA.

Butt, M. (1997). Complex predicates in Urdu. In Alsina, A., Bresnan, J., and Sells, P., editors, *Complex Predicates*. Center for the Study of Language and Information, Stanford, CA, USA.

Cahill, A., Burke, M., O'Donovan, R., van Genabith, J., and Way, A. (2004). Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Barcelona, Spain.

Chrupała, G. and van Genabith, J. (2006). Using machine-learning to assign function labels to parser output for Spanish. In *Proceedings of the COLING/ACL*

2006 Main Conference Poster Sessions, pages 136–143, Sydney, Australia. Association for Computational Linguistics.

Jijkoun, V. and de Rijke, M. (2004). Enriching the output of a parser using memory-based learning. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Barcelona, Spain.

O'Donovan, R., Burke, M., Cahill, A., van Genabith, J., and Way, A. (2004). Large-scale induction and evaluation of lexical resources from the Penn-II Treebank. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 367–374, Barcelona, Spain.

O'Donovan, R., Cahill, A., van Genabith, J., and Way, A. (2005). Automatic acquisition of Spanish LFG resources from the CAST3LB treebank. In *Proceedings of the Tenth International Conference on LFG*, Bergen, Norway. CSLI Publications.

The German Infinitival Passive: a Case for Oblique Functional Controllers ?

Philippa Cook
Zentrum für allgemeine Sprachwissenschaft (ZAS), Berlin

Proceedings of the LFG06 Conference
Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)
2006

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

Abstract

Following the standard LFG assumption that a functional controller – unlike an anaphoric controller – must bear a term GF and must be present at f-structure, one must assume that the German Infinitival Passive Construction (IPC) involves Anaphoric Control, at least for subject equi verbs. However, an Anaphoric Control analysis of the IPC with equi verbs that select a dative object fails to account for the availability of split antecedents since under Anaphoric Control split antecedents should be possible, but they are not. Instead, I pursue a Functional Control analysis of the German IPC – thus accounting for the availability of split antecedents. Equi verbs which do not license the IPC, namely accusative object equi, can be analyzed via Anaphoric Control since they prohibit split antecedents. For subject equi verbs, then, the Functional Control analysis of the IPC will require two modifications to be made to LFG's standard approach to Functional Control, both of which I will claim are independently motivated. First, one must allow a non-term argument (namely $\text{OBL}_{\text{AGENT}}$) to be a functional controller; something LFG has previously rejected. Second, the implicit $\text{OBL}_{\text{AGENT}}$ argument of passives must be represented at f-structure since this is, I claim, a functional controller. Evidence from binding facts suggests this may be required anyway. Finally, allowing non-term functional controllers actually permits an alternative account of Visser's Generalization which also captures its (partial) non-application in German.

1. The Infinitival Passive Construction (IPC) – Anaphoric or Functional Control ?

1.1. Distribution of the IPC

German subject equi verbs permit an Infinitival Passive construction (IPC), as in (1b/b') in which the [-o] (agent) argument of active *versuchen* 'try' is suppressed, as in a regular passive construction.[1] This argument can optionally occur at c-structure as an $\text{OBL}_{\text{AGENT}}$ (expressed as a *von*-PP) but, according to informants, this is pragmatically disfavoured for obvious reasons. In embedded clauses, as in (1c), which illustrates the IPC with a range of subject equi verbs, the IPC is available with both (so-called) intra- and extraposed positions of the infinitival complement, viz. (1c) and (1d). In declarative main (i.e. V2) clauses, the infinitival complement may occupy SpecCP (1b), or a placeholder *es* (cf. Berman 2003:65) may, or alternatively a locative/temporal modifier. In addition to intransitive subject equi verbs (i.e. which select just a subject and a non-finite complement and no matrix object), German also has transitive subject equi verbs which select a dative object in addition to the non-finite complement (cf. Bech 1955: 113-114). These verbs license the IPC, as in (1e):

(1) a. Hans versuchte den Turm zu erreichen
Hans tried the$_{\text{ACC}}$ tower to reach
Hans tried to reach the tower

[1] Suppression of this argument is usually taken to mean it is rendered unavailable for linking (see e.g. Dalrymple 2001:208). In standard LFG treatments of the passive, the suppressed argument maps to neither an f-structure nor a c-structure argument. In this paper I will, however, argue that the suppressed argument in a passive should map to an f-structure argument, and optionally to a c-structure argument.

b. Den Turm zu erreichen wurde versucht
the$_{ACC}$ tower to reach was tried

b.' Es wurde versucht den Turm zu erreichen
it was tried the$_{ACC}$ tower to reach

lit.: *To reach the tower was tried*

c. weil gehofft/geplant/gewagt wurde [den Turm gegen Abend zu erreichen]
since hoped/planned/dared was-SG the$_{ACC}$ tower toward evening to reach

d. weil [den Turm gegen Abend zu erreichen] gehofft/geplant/gewagt wurde
since the$_{ACC}$ tower toward evening to reach hoped/planned/dared was-S

e. weil mir von der Firma versprochen wurde [den Rohrbruch bis zum Nachmittag zu reparieren]
since me$_{DAT}$ from the firm promised was-SG the burst pipe by the afternoon to repair
intended*: since I was promised by the firm to repair the burst pipe by this afternoon*

I adopt the term IPC to distinguish this construction descriptively from the Impersonal Passive of finite intransitive verbs as in (2), in which the argument corresponding to the active subject is suppressed and there is no c-structure subject.[2]

(2) a. Gestern wurde getanzt
yesterday was danced

b. weil gestern getanzt wurde
because yesterday danced was

Berman (2003: ch.4) offers an account of the German Impersonal Passive in (2) (and other impersonal constructions) in which the lexical entry of the 3rd person singular verbal agreement affix can specify an expletive (non-thematic) SUBJ, as in (3) below, thereby satisfying the Subject Condition in the absence of a c-structurally overt subject argument. Normally the AGR information unifies with the features of the overt subject, but if no subject is present, the verbal morphology actually introduces a subject in the f-structure. The verbal morphology does not specify a PRED value, but just AGR information – and hence it is an expletive SUBJ that is projected.

(3) - t V_{infl} (↑ TENSE) = PRESENT
(↑ SUBJ) = ↓
(↓ PERS) = 3
(↓ NUM) = SG

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{TENSE} & \text{PRESENT} \end{bmatrix}$$

Given that German independently permits impersonal constructions and given Berman's analysis in (3), there are two possible analyses for the IPC:

Either

(i) the IPC is an impersonal passive construction, lacking a c-structure SUBJ but with an expletive f-structure SUBJ contributed by the 3rd person singular verb form. Under this analysis, the non-finite complement bears the GF COMP (or XCOMP), and is unaffected by passivization (i.e. it does not map to passive SUBJ).

Alternatively,

(ii) the non-finite complement is analyzed as bearing the GF OBJ in the active. This OBJ may map to SUBJ under passivization and functions as the SUBJ of the IPC. The construction thus has an overt c-structural SUBJ. The latter analysis is adopted by Lødrup (2002, 2004) for the Norwegian IPC.[3]

[1] In contrast to some other Germanic languages, German only requires an overt expletive to be inserted if the SpecCP position is not otherwise filled (Berman 2003: 60), cf. also the contrast between (1b) and (1b').

[3] Indeed, the ability to function as the subject of a passive, as in analysis (ii) above, is one of the criteria put forward by Dalrymple & Lødrop (2000) for treating a clausal complement as OBJ rather than COMP in their proposal that English, German

Which of these analyses of the IPC one adopts is, however, orthogonal to the issue of which type of control relation (Anaphoric or Functional) one must, or can, assume for the IPC, and I will therefore only comment in passing on the GF of the infinitival complement. Under either (i) the impersonal or (ii) the OBJ analysis of the IPC, it is the case that when we consider the IPC with *subject* equi, the controller is not obligatorily present at c-structure and it bears a non-term GF – namely $\text{OBL}_{\text{AGENT}}$ – and it thus appears that the control relation involved must be Anaphoric Control, cf. Lødrup (2002, 2004) for Norwegian.

I turn now to the IPC with *object* equi verbs. As shown in (4b), dative object equi verbs permit the IPC,[4] but IPC with dative object equi does not involve a structurally absent controller since the dative object (the controller) is unaffected by passivization.[5] Dative object equi verbs are thus in principle compatible with a Functional Control analysis since the controller is both overt and is a term (OBJ_θ). Accusative object equi verbs, by contrast, do not permit the IPC, hence the ungrammaticality of (5b). Any analysis of the IPC must therefore account for this distinction.

(4) a. ACTIVE (dative object equi)

weil Hans denen empfohlen/erlaubt/verboten hat [den Turm gegen Abend zu erreichen]
since Hans them$_{\text{DAT}}$ recommended/allowed/forbidden has the$_{\text{ACC}}$ tower toward evening to reach

Hans recommended/allowed/forbad them to reach the tower by evening

b. INFINITIVAL PASSIVE CONSTRUCTION (dative object equi)

weil denen empfohlen/erlaubt/verboten wurde [den Turm gegen Abend zu erreichen]
since them$_{\text{DAT}}$ recommended/allowed/forbidden was-SG the$_{\text{ACC}}$ tower toward evening to reach

(5) a. ACTIVE (accusative object equi)

weil Hans ihn gezwungen/überredet/ermuntert hat [den Turm gegen Abend zu erreichen]
since Hans him$_{\text{ACC}}$ forced/persuaded/encouraged has the$_{\text{ACC}}$ tower toward evening to reach

Hans forced/persuaded/encouraged them to reach the tower by evening

b. INFINITIVAL PASSIVE CONSTRUCTION (accusative object equi)

*weil ihn gezwungen/überredet/ermuntert wurde [den Turm gegen Abend zu erreichen]
since him$_{\text{ACC}}$ forced/persuaded/encouraged was-SG the$_{\text{ACC}}$ tower toward evening to reach

The IPC is in complementary distribution with 'regular' personal passive of object equi verbs. By 'regular' personal passive I refer to cases in which the [-r] argument which maps to (nominal) OBJ of the equi verb in the active maps to SUBJ in the passive. I use the term 'regular' since this is neither an impersonal passive construction, nor does it involve a clausal OBJ mapping to SUBJ. 'Regular' personal passive is unavailable for dative object equi verbs, viz. (6a) below, but is available for accusative object equi verbs, viz. (6b). In other words, (6a,b) contrast with (4b) and (5b) above. More generally, dative objects in German are unaffected by regular *werden*-passivization and never function as passive subject.[6] Thus it is not surprising that transitive subject equi verbs (i.e. with a matrix dative object) such as

---

and Swedish permit both OBJ and COMP clausal complements (cf. though see Alsina et al 1996, 2005, Forst 2006 for discussions of the proposal to eliminate the GF COMP from LFG entirely).

[4] It is not appropriate to analyse the dative plural *denen* in (4b) as SUBJ since German, unlike Icelandic, only permits nominative subjects. Note that *denen* does not agree with the finite verb, which is singular.

[5] Lødrup (2004: 81) discusses a similar Norwegian example with an object equi verb *anbefale* 'recommend' in which the controller is the object *dem*.

i. Det ble anbefahlt dem [å be mer]
It was recommended them to pray more

[6] The so-called Dative-Passive or *kriegen*-passive is a different construction altogether and is best not analysed as involving a passive operation. See Cook (2006) for an argument composition analysis.

versprechen 'promise' also fail to permit 'regular' personal passive, as shown in (6c) although these verbs do permit the IPC as was seen in (1e) above. The distribution of both types of passive construction across the four types of equi verb is summarized in the table in (7).

(6) a. REGULAR PERSONAL PASSIVE (dative object equi)

*weil	er	empfohlen/erlaubt/verboten	wurde	[	den	Turm	gegen	Abend	zu erreichen]
since	he_{NOM}	recommended/allowed/forbidden	was		the_{ACC}	tower	toward	evening	to reach

b. REGULAR PERSONAL PASSIVE (accusative object equi)

weil	er	gezwungen/überredet/ermuntert	wurde	[	den	Turm	gegen	Abend	zu erreichen]
since	he_{NOM}	forced/persuaded/encouraged	was		the_{ACC}	tower	toward	evening	to reach

c. REGULAR PERSONAL PASSIVE (subject equi with matrix dative)

*weil	er	versprochen	wurde	[	den	Turm	gegen	Abend	zu erreichen]
since	he_{NOM}	promised	was		the_{ACC}	tower	toward	evening	to reach

(7) Complementary Distribution of the IPC and 'regular' Personal Passive

	Infinitival Passive Construction	Regular Personal Passive
Subject equi (no matrix object)	✓	n.a.[7]
Subject equi with matrix dative	✓	*
Dative Object equi	✓	*
Accusative Object equi	*	✓

If this were the complete range of data to be accounted for, there would be no problem with adopting an Anaphoric Control analysis of these verbs, and thus of the IPC, as Lødrup (2002, 2004) did for Norwegian. In the next section, though, data concerning the availability of split antecedents reveal that an Anaphoric Control analysis cannot be upheld for the subject equi verbs with matrix dative and the dative object equi verbs.

1.2 Split Antecedents – against an Anaphoric Control analysis of IPC-licensing verbs

LFG recognizes two control relations: Functional Control and Anaphoric Control (see Bresnan 1982, 2001: ch. 13/14). It is an automatic consequence of the theory that Functional Control demands a controller that is represented at f-structure because it involves structure-sharing, i.e. identity of the f-structure of the controller and that of the control target (i.e. the complement's SUBJ in the data under consideration). This equivalence of f-structures is expressed as an identity equation in the lexical entry of an equi verb as shown in (8a) for subject equi and in (8b) for object equi respectively. This equation states that the f-structure of the SUBJ or, depending on verb type, of the OBJ of the equi verb is the same f-structure as that of the XCOMP's SUBJ. Informally, structure-sharing is sometimes represented in f-structure via a dotted line linking the f-structures of the controller and control target, as in (19) below. Since Functional Control requires identity of f-structures, the control relation must be exhaustive. This means, for instance, that split antecedents cannot function as the antecedent of the

[7] I mark this cell n.a. (not applicable) since the availability of this construction for intransitive subject equi verbs is wholly dependent on whether the non-finite complement is assumed to bear the GF OBJ or COMP (or XCOMP), cf. the two possible analyses of the IPC sketched in the main text above. If intransitive subject equi verbs do not select an OBJ, but rather a COMP, then there is no OBJ/[-r] argument that could map to SUBJ in the passive and thus there can be no 'regular' personal passive, and this cell could also be starred. Under such a COMP analysis, the IPC is a genuinely impersonal construction with an expletive f-structural SUBJ. By contrast, under an OBJ analysis of the non-finite complement, regular personal passivization would in fact yield the IPC.

control target (Bresnan 1982: 346). To clarify, note that Bresnan (1982) defines split antecedents thus: "a pronoun that refers to more than one noun phrase is said to have split antecedents; for example, in *Tom told Mary that they should leave*, *Tom* and *Mary* can be split antecedents of *they*".[8] Thus, following Bresnan, I take split antecedents to refer to antecedents which are *overtly* expressed in the matrix clause and Functional Control thus strictly prohibits a control equation of the type in (8c), which is starred to indicate that if split antecedents were to function as the antecedent in a Functional Control relation, the f-structures of both antecedents would be merged with the f-structure of the control target, leading to a clash of features and an ill-formed f-structure.[9]

(8) a. (↑ SUBJ) = (↑ XCOMP SUBJ) b. (↑ OBJ) = (↑ XCOMP SUBJ)

c. *(↑ SUBJ) ∧ (↑ OBJ) = (↑ XCOMP SUBJ)

The ban on non-term functional controllers (Bresnan 1982: 354) is motivated by the fact that Functional Control requires a controller to project its own f-structure in order for its f-structure to be identified with that of the control target. Nevertheless, this ban has a slightly stipulative quality to it. I return to this in section 3.

Anaphoric Control, by contrast, does not involve syntactic identity but, rather, requires the control target (e.g. the COMP's SUBJ), which is assumed to be a null pronominal, to find an antecedent which can provide its referent; i.e. the two are semantically related by an anaphoric binding relation. When the equi verb does not constrain the co-reference of the control target and its antecedent, one can talk of *arbitrary* Anaphoric Control, and the verb's lexical entry will include a control equation like that in (9a) which leaves it open what the antecedent of the control target is. Since in Anaphoric Control the control target finds its referent in a similar way to an ordinary pronoun (see Dalrymple 2001: 330-336), split antecedents are possible. LFG also recognizes *obligatory* Anaphoric Control in which the control target must co-refer with an argument of the matrix clause. In this case, the equi verb's lexical entry additionally includes an equation specifying which matrix argument this is, as exemplified in (9b) for an obligatory matrix SUBJ antecedent (Dalrymple 2001: 334).

(9) a. (↑COMP SUBJ PRED) = 'PRO' b. ((↑COMP SUBJ)σ ANTECEDENT) = (↑ SUBJ)

Recall that without considering split antecedents, the German IPC construction at first sight seems to require an Anaphoric Control treatment parallel to Lødrup's analysis of Norwegian IPC – at least for subject equi – since the controller is not obligatorily overt and is a non-term. However, on the basis of the availability of split antecedents, dative object equi verbs and subject equi with matrix dative verbs present evidence against an Anaphoric Control analysis.

1.2.1 IPC-licensing verbs – no split antecedents

German dative object equi verbs prohibit split antecedents, viz. (10). The same judgements were obtained for *i.a. befehlen* 'order', *untersagen* 'forbid' and *gestatten* 'allow'.[10] Similarly,

[8] Note that I am not considering Partial Control (e.g. *We thought the chair preferred to gather at noon*), in which a verb requiring a semantically plural subject occurs in the non-finite complement, as an instance of split antecedents. See Asudeh (2005: 504/5) for discussion.

[9] It is, of course, possible to have exhaustive syntactic control (i.e. no split antecedents) but to nevertheless contextually infer that some other non-overt referent is also involved in the activity expressed by the non-finite complement.

[10] Informants report that two dative object-selecting verbs permit split antecedents; *anbieten* 'offer' and *vorschlagen* 'propose'. It is interesting that the exceptions are with these two verbs because these two verbs can involve subject or object equi (as well as split antecedents) irrespective of the type of predicate in the complement (cf. Bech 1955: 114 §114, 190 §198 for this observation). Even if two separate lexical entries were assumed (i.e. one as a subject equi with dative object verb, one as a dative object equi verb), the availability of split antecedents is puzzling since both of these verbs types otherwise prohibit split antecedents. This behaviour is, however, not problematic in a lexical theory of control such as that of LFG – these verbs

split antecedents are not possible for subject equi with dative object verbs, viz. (11).[11] Informants also confirm that the use of *gemeinsam* 'together' is infelicitous in both (10) and (11). Since its use would favour a split antecedent reading, this fact is not surprising. The ban on split antecedents with both of these verb types would, of course, fall out automatically under a Functional Control analysis. By contrast, if the dative object equi and subject equi with dative object were to require Anaphoric Control, an account would still need to be sought for why split antecedents are ruled out.

(10) Ich$_i$ empfahl/riet/verbot dem Bürgermeister$_j$ den Antrag ?(gemeinsam) einzureichen
I recommended/advised/forbad the$_{DAT}$ mayor the bid together to.submit
I$_i$ recommended/advised/forbad the mayor$_j$ to submit the bid
[submitters ≠ i+j submitters = j 'the mayor' + (possibly) others but, crucially, not i+j]

(11) Ich$_i$ drohte/(zu)sicherte/schwörte dem Bürgermeister$_j$ den Antrag ?(gemeinsam) einzureichen
I threatened/assured/swore the$_{DAT}$ mayor the bid together to.submit
I$_i$ threatened/assured/swore the mayor$_j$ to submit the bid
[submitters ≠ i+j submitters = i 'Ich' + (possibly) others but, crucially, not i+j]

(12) Ich$_i$ überzeugte/drängte/überredete den Bürgermeister$_j$ den Antrag (gemeinsam) einzureichen
I convinced/urged/persuaded the$_{ACC}$ mayor the bid together to.submit
I$_i$ convinced/urged/persuaded the mayor$_j$ to submit the bid together [submitters = i+j]

It is striking that informants report unanimously that split antecedents are possible with accusative object equi verbs, viz. (12) above. The same judgements were obtained for *i.a. zwingen* 'force', *anflehen* 'beg' and *ermuntern* 'encourage'. This is interesting because a correlation emerges between the ability to license IPC and the impossibility of split antecedents, and *vice versa*, as summarized in (13)

(13) Summary:

	Infinitival Passive Construction	Regular Personal Passive	Split Antecedents
Subject equi (no matrix object)	✓	n.a.	n.a.
Subject equi with matrix dative	✓	*	*
Dative Object equi	✓	*	*
Accusative Object equi	*	✓	✓

A reviewer suggests that an *obligatory* Anaphoric Control analysis could cover these facts and thus obviate the need to modify Functional Control. Under this suggestion, then, although Anaphoric Control in principle permits split antecedents, the ungrammaticality of split antecedents with subject equi with matrix dative, and dative object equi verbs could be made to follow if the *obligatory* Anaphoric Control equation specified that *only* the matrix SUBJ and *only* the matrix dative OBJ_θ can be the antecedent of the COMP's SUBJ, for these two verb types respectively. However, the lexical entry of the passive variant of the subject equi verbs would have to include a control equation in which the matrix OBL_{AGENT} is the controller. To accommodate the lack of split antecedents with subject equi with matrix dative verbs, one would therefore also have to formulate an *obligatory* Anaphoric Control relation but – and

can be specified as involving Anaphoric Control, i.e. the lexical entry includes a control equation according to which the control target behaves parallel to an overt pronoun in resolving its reference from the context.

[11] I leave out discussion of *versprechen* 'promise' for the time being since its behaviour is more complex. It appears to 'switch' from subject equi to dative object equi when certain types of complement are embedded; namely modal, passive or beneficiary-oriented predicates. Similar facts hold for the passivized verb form of English *promised*, as is well-known (Chomsky 1965:229, Jenkins 1972:200ff, Růžička 1983). I return to the facts very briefly at the end of the paper.

this is the problem – the requirement that the controller be a term also holds for *obligatory* Anaphoric Control (see Dalrymple 2001: 344). I conclude that this is not a viable alternative.[12]

To summarize, then, an account is required of why IPC is possible for subject equi, subject equi with matrix dative, and dative object equi verbs but is ruled out for accusative object equi verbs.

I will argue that the accusative equi verbs, in contrast to the other verbs types listed, involve Anaphoric Control, and that the availability of split antecedents in (12) is therefore as expected. In turn, I will argue that the failure of these accusative equi verbs to licence IPC (viz. (5b)) is directly related to the fact that regular personal passive is available. The a-structure contains a [-r] matrix accusative OBJ which is available as a candidate to be mapped to SUBJ when the [-o] (agent) of the active is suppressed under passivization. The availability of an overt c-structure SUBJ precludes the IPC from applying because the IPC is an impersonal construction, having only an f-structure expletive SUBJ (*à la* Berman 2003). The analysis relies crucially on the fact that Berman's (2003) expletive f-structure SUBJ can only ever be projected when there is no overt subject available.

By contrast, the IPC will be shown to be available for the other verbs types precisely because such a [-r] argument, i.e. a candidate for promotion to SUBJ under passivization, is lacking. The IPC is grammatical because in the absence of any possible SUBJ-compatible argument, and importantly *only* in the absence of such an argument, German will project an expletive f-structural SUBJ.

1.3 A brief note on the lexical semantics underlying split antecedents

It has often been pointed out that the availability of split antecedents, i.e. of non-exhaustive control, is surely related to the lexical semantics of the predicates involved (cf. Sag/Pollard 1991, Culicover/Jackendoff 2005). It may thus perhaps appear that the presentation here favours a purely structural, rather than lexically-oriented, account since I am relating the possibility of split antecedents (non-exhaustive control) to case properties; namely accusative object equi. However, it should not be forgotten that the distribution of case in German is not random but has an underlying basis in lexical semantics; although this is not always synchronically transparent. I believe therefore that the distinction between predicates taking dative objects, which tend to be BENEFICIARIES, EXPERIENCERS or PATIENTS, and those taking accusative objects is underlyingly one of lexical semantics. If this is correct, it appears, then, that the availability of split antecedents in German correlates with the presence of a matrix accusative object, which tend to bear less 'affected' thematic roles than dative objects do. Thus, the distribution of exhaustive control is likely related to differences in argument-structure, i.e. lexical semantics. While a closer examination of the observed tendency towards exhaustive (object) control with more affected (i.e. BENEFICIARY, EXPERIENCER or PATIENT) objects is beyond the scope of this study, it may also have interesting connections with the availability or not of the IPC in Norwegian. Lødrup (2004) reports that some Norwegian object equi permit the IPC while others do not; a fact that he has no account for. The object equi verbs in Norwegian which do *not* allow the IPC (see Lødrup 2004: 80, his (124)) are verbs which, I sense, would correspond to accusative object equi in (many cases in) German. It seems likely, then, that an

[12] Moreover, the antecedent of a pronoun must introduce a discourse referent but it appears to be the case that the OBL_{AGENT} of a passive only introduces a discourse referent when is overt, and not when it is implicit as the following contrast shows:

i. weil vom Pförtner versucht wurde, das Schloss aufzubrechen. Er hatte Erfolg
It was attempted by the porter to break open the lock. He was successful

ii. weil versucht wurde, das Schloss aufzubrechen. #Er hatte Erfolg
intended: *it was attempted to break open the lock. #He was successful*

explanation similar to my account of the absence of IPC with German accusative object verbs, relying on the [-r] status of the object in particular, could perhaps be usefully extended to Norwegian.

2. The Functional Control Alternative

Given that IPC-licensing verbs permit split antecedents, I will pursue a Functional Control analysis of the IPC. Although the intransitive subject equi verbs offer no evidence with respect to split antecedents, I will advance a uniform analysis of all IPC-licensing verbs. Recall that a Functional Control analysis is not problematic for the dative object equi verbs since the controller is present, and the controller is a term GF, an OBJ_θ (see Cook 2006 for the motivation for assuming the dative object to bear the GF of secondary object). However, proposing a Functional Control analysis for the subject equi verbs requires one to accept (i) that non-terms can be functional controllers; something that LFG has previously rejected (Bresnan 1982). Moreover, it requires one to accept (ii) that implicit arguments of passives are represented at f-structure. I will first provide some evidence that oblique exhaustive controllers are documented elsewhere before presenting evidence from binding facts which suggest that implicit arguments of passives should indeed be represented at f-structure.

2.1 Some Evidence for oblique exhaustive controllers

LFG's claim that only term GFs may be functional (Bresnan 1982: 322) and *obligatory* anaphoric (Dalrymple 2001: 344) controllers, is not shared by other theories (and it is subject to exceptions that require further assumptions to be made, Bresnan 1982: 348). Outside LFG, it is assumed that obliques can obligatorily (or exhaustively) control, and that, for instance, implicit arguments of nouns can too, cf. (14). The examples in (15) due to Culicover/Jackendoff (2005:433) are argued to involve the object of a PP as unique controller:

(14) a. The promise by Sandy to leave the party early caused quite an uproar [Pollard/Sag 1994:289]
b. The promise to Susan by John to take care of himself/*herself [Culicover/Jackendoff 2005:435]

(15) a. John_i counted on/relied on/called upon Susan_j
to take care of herself/*himself/*oneself [controller is Susan only]

b. John's_i order/instructions/encouragement/reminder to Susan_j
to take care of herself/*himself/*oneself [controller is Susan only]

Furthermore, there have been claims in the literature that Irish involves raising to oblique (McCloskey 1984) and, since raising necessarily involves a relation of Functional Control, this too looks like potential evidence in favour of permitting non-term functional controllers. Less well-known is the argument put forward by Joseph (1979, 1990) that Modern Greek also involves raising to oblique. Full detailed investigation of these facts is beyond the scope of this paper and their mention is intended just to illustrate that there may indeed be positive evidence that non-term functional (or obligatory anaphoric) controllers are needed.

2.2. Evidence for representing implicit arguments of (German) passives at f-structure

Frey (1993: ch. 9) points out that in early LFG two lexical entries were assumed for a passivized verb: one with the $\text{OBL}_{\text{AGENT}}$ or *by*-phrase and one without. In the latter, the suppressed argument was represented by the null symbol, just like a middle variant of a verb, an inchoative, or – for an implicit *object* – a detransitivized verb. The null symbol represents the suppression

of an argument position in the lexicon and this argument is therefore not accessible for syntactic operations.

(16) a.	beaten, $V_{[part]}$: (↑PRED) = 'beat < ∅, SUBJ >'	a'.	Fred was beaten	(passive)
b.	read, V: (↑PRED) = 'read < ∅, SUBJ >'	b'.	Russian novels read easily	(middle)
c.	break, V: (↑PRED) = ' break < ∅, SUBJ >'	c'.	The vase broke	(inchoative)
d.	read, V: (↑PRED) = 'read < ∅, SUBJ >'	d'.	Fred reads infrequently	(detransitivization)

Frey, however, questions the accuracy of handling the *by*-phrase-less passive akin to the other implicit argument forms in (16) since, in contrast to the implicit arguments of types (16b-d), the implicit argument of the passive can (i) be added as an afterthought, (ii) can be a controller of an adjunct, and (iii) can be the antecedent for secondary predication. Furthermore, implicit arguments (i.e. OBL_{AGENT}) of passives, but not of middles, inchoatives and detransitivized verbs, can bind reciprocals in German. Frey (1993:132, 158) gives the following examples of binding of a reciprocal by the implicit argument (OBL_{AGENT}) of a passive in (17a-b). For completeness, I illustrate this with an example from each of the verb classes that permits subject equi in (17c-d).

(17) a. Auf Parteiversammlungen wird nur gegeneinander gekämpft
At party gatherings is only against one another fought
At party meetings all that happens is fighting against each other

b. viele Briefe wurden einander geschrieben
many letters were-PL one another written
We wrote many letters to each other

c. weil auf der Tagung versucht wurde, einander nicht zu kritisieren
since at the conference tried was one another not to criticize
since one tried not to criticize each another at the conference

d. weil ihm versprochen wurde, nicht miteinander zu streiten
since him_{DAT} promised was not with one another to argue
since one promised him not to fight with one another

Presumably, if the implicit OBL_{AGENT} can bind a reciprocal, then it has to be represented at f-structure and I thus take these data to suggest that implicit arguments of passives should project their own f-structure, even when they are c-structurally non-overt.

2.3 The representation of the implicit argument of the passive at f-structure

On the basis of the data in (17), I propose that implicit arguments of passives – in contrast to some other types of implicit argument – should be represented at f-structure. Thus, even when there is no overt *von*-phrase in c-structure, this argument nevertheless projects an f-structure. Evidence for this assumption is the availability of implicit arguments of passives as binders.

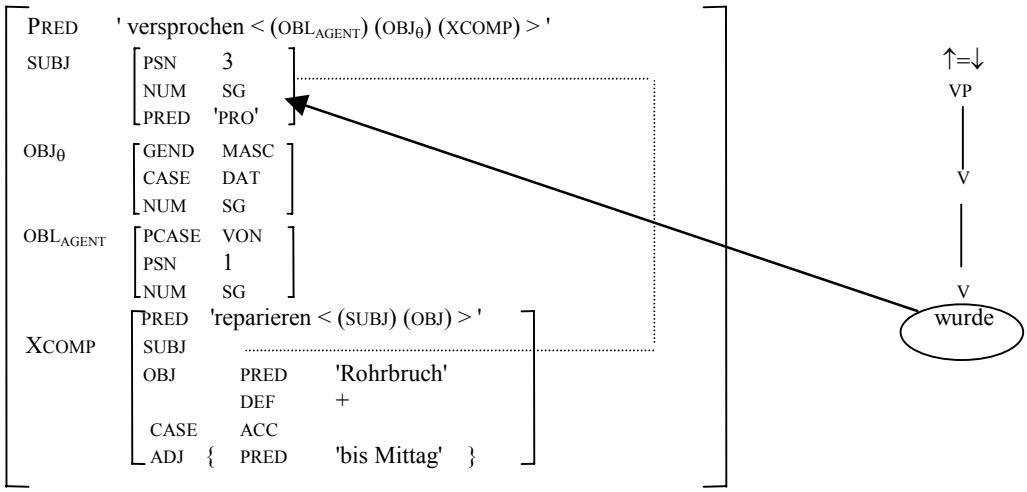
The approach I propose is parallel to LFG's treatment of pro-drop: I assume that a passivized verb always subcategorises for an OBL_{AGENT} and I propose that the lexical entry of passivized verbs includes an equation which optionally licenses the projection of an f-structure attribute OBL_{AGENT} with PRED value 'PRO'. The optional PRED value provides (minimal) semantic content for OBL_{AGENT} and satisfies Completeness when no overt *von*-phrase occurs. Sample lexical entries for passivized verbs subject equi verbs (with and without matrix dative) are given in (18a-b). Since the implicit argument will always be present in f-structure under this analysis, the Functional Control equations that I also give in the lexical entries of *versucht* 'tried' and *versprochen* 'promised' in (18a-b) are legitimate.

(18) a. 'versucht $V_{pass\ part}$ < OBL_{AGENT} XCOMP >' (tried)
(↑ OBL_{AGENT}) = XCOMP SUBJ
((↑ $OBL_{AG\ ENT}$ PRED) = 'PRO')

b. 'versprochen $V_{pass\ part}$ < OBL_{AGENT} OBJ_{θ} XCOMP >' (promised)
(↑ OBL_{AGENT}) = XCOMP SUBJ
((↑ OBL_{AGENT} PRED) = 'PRO')

In (19), I provide the f-structure of an IPC construction, assuming it to be an impersonal construction involving Functional Control (the identity of matrix SUBJ and XCOMP SUBJ is indicated here by the dotted line) in which the f-structure SUBJ is contributed via the verbal morphology (indicated by the bold arrow), following Berman's (2003) sketched in (3) above. This is in keeping with the fact that the finite verb in IPC is only ever third person singular. I am thus adopting an XCOMP analysis of the GF of the infinitival complement, although an analysis in which the IPC is not an impersonal construction and the infinitival complement maps to passive SUBJ is also compatible with the facts, as I will show below.[13] If there were no overt realization of the OBL, a PRO would be projected in the f-structure from the lexical entry above.

(19) *weil mir von der Firma versprochen wurde, den Rohrbruch bis Mittag zu reparieren*
[because me_{DAT} by the firm promised was the burst pipe by afternoon to repair]



In this way, it is possible to assume a Functional Control analysis of the verbs that license the IPC, as the facts concerning split antecedents suggest is correct. In particular, the IPC with subject equi does not force us to adopt an Anaphoric Control analysis, as it would do under LFG's standard approach to control, since the controller is represented at f-structure even when it is not overt at c-structure. This analysis also obviates the need for two distinct lexical entries (i.e.one with and one without the OBL_{AGENT}) for every passivized verb form.

2.4 Accounting for the distribution of the IPC and the regular personal passive

Given the two modifications to Functional Control introduced above, there is now no impediment to analysing those verbs forbidding split antecedents as involving Functional Control. The accusative object equi verbs, by contrast, permit split antecedents and can simply be

[13] Recall that the issue of the infinitival's GF is independent of the issue of whether Anaphoric or Functional control is assumed. Note, however, that under the analysis in which the infinitival complement maps to SUBJ, the lexical entries in (18) would not be appropriate. Instead, the a-structure of the passivized verb form would be < OBL_{AGENT}, OBJ_{θ}, SUBJ>.

analysed as involving Anaphoric Control, see (20). On the basis of the shared distribution of the IPC, I extend this analysis to the intransitive subject equi verbs too.

(20) Summary:

	IPC	Regular Personal Passive	Split Antecedents	Control Relation
Subject equi (no matrix object)	✓	n.a.	n.a.	Functional
Subject equi with matrix dative	✓	*	*	Functional
Dative Object equi	✓	*	*	Functional
Accusative Object equi	*	✓	✓	Anaphoric

Recall from section 1 that we need to account for why IPC is possible for subject equi, subject equi with matrix dative, and dative object equi verbs but is ruled out for accusative object equi verbs. I show in 2.4.1 that the ungrammaticality of the IPC with accusative object equi verbs is directly related to the fact that regular personal passive is available. In 2.4.2, by contrast, I will show that the other verb types lack a [-r] matrix argument, and lack regular personal passive, and it is this that licenses the IPC.

2.4.1 The Accusative Object Equi verbs: the [-r] matrix object prohibits the IPC

I propose that there is an argument in the a-structure of these verbs – namely a [-r] argument that maps to the matrix accusative OBJ in the active – which is available as a candidate to be mapped to SUBJ when the [-o] (agent) of the active is suppressed under passivization, viz. (21c). This is what occurs in regular personal passive where well-formedness conditions entail that the [-r] argument maps to SUBJ, thus licensing regular personal passive as in (6b). I am treating the third argument, that maps to the infinitival complement, as a state-of-affairs argument (soa). It is not clear to me that an argument that bears no thematic role should be involved in lexical mapping theory, and I thus leave this argument untreated here. Turning now to the fact that the IPC is ungrammatical with the accusative object equi verbs, as in (5b), I claim that it is the availability of an overt c-structure SUBJ that precludes the IPC from applying. This is because the IPC is an impersonal construction, having only an f-structure expletive SUBJ. Crucially, in Berman's (2003) analysis of German's expletive f-structure SUBJ, it can only be projected when there is no subject argument available. Since the [-r] argument is perfectly compatible with SUBJ status, the conditions for the IPC do not arise.

(21) a. *überreden* agent patient soa persuade (Accusative Object Equi)
b. [-o] [-r] [soa]
c. ∅ SUBJ/OBJ
✓RPP → (6b) grammatical because [-r] maps to passive SUBJ
*IPC → (5b) ungrammatical because [-r] can map to passive SUBJ, thus an impersonal passive (IPC) cannot occur

Under this analysis, the complementary distribution of the two constructions is accounted for.

Alternatively, as mentioned above, the IPC could be analysed not as an impersonal construction with an XCOMP infinitival complement, but rather as involving an infinitival complement with the GF OBJ or – not previously mentioned – OBL. For proponents of replacing COMP entirely with GFs also borne by NPs, the GF OBL (rather than OBJ) would be assumed for the

infinitival complement of accusative equi verbs because they alternate with PP (rather than NP) objects and co-occur with oblique correlatives such as *davon* 'there-from', *darauf* 'there-on' (see Berman *in press* and Forst 2006). Under this style of analysis, the complementary distribution of the two passive constructions could be accounted for if the infinitival OBL is unable to map to SUBJ in the passive. If the OBL is assigned [-o] intrinsically (cf. Berman *in press*),[14] such a mapping would be ruled out in the presence of the higher [-r] argument which can map to SUBJ. Conversely, the presence of this [-r] argument licenses regular personal passive as in (21) above.

2.4.2 The IPC-licensing verbs: the lack of the [-r] argument licenses IPC

By contrast, the IPC is available for all the other verbs types discussed, as summarized in (20). I argue that this is directly related to the fact that regular personal passivization is ungrammatical and, in this vein, I propose that these verbs *lack* an argument parallel to the [-r] argument of the accusative object equi verbs that could be promoted to SUBJ under passivization. I propose therefore that the dative object of the subject equi verbs with matrix dative, and of the dative object equi verbs is assigned [+o].

The dative object of most (standard) ditransitives in German bears the thematic role of BENEFICIARY, MALEFICIARY, EXPERIENCER or at least AFFECTED PATIENT and I have argued extensively elsewhere (see Cook 2006) that under LFG's Lexical Mapping Theory (LMT) the dative object of German ditransitives is intrinsically assigned [+o] in the presence of a [-r] theme argument, and maps to OBJ_θ. Although I suggested there that this is the German parameterization of LFG's Asymmetric Object Constraint, it is possible that such thematic roles (typically BENEFICIARY/PATIENT) should be generally intrinsically assigned [+o], i.e. in non-double object environments. I simply adopt this assumption for now although there is further evidence to support this claim for German, as discussed in Cook (2006). In (22), taking a dative object equi verb for the purposes of illustration, I outline how the [+o] analysis of the BENEFICIARY/EXPERIENCER or PATIENT role accounts for the complementary distribution of the regular personal passive and the IPC respectively.

(22)	a. *empfehlen*	agent	ben/exp	soa	recommend (Dative Object Equi)
	b.	[-o]	[+o]	[soa]	
	c.	∅	OBJ/OBJ_θ		

*RPP → (6a) ungrammatical because there is no [-r] which can map to passive SUBJ

✓IPC → (4b) grammatical because [-r] can map to passive SUBJ, thus an impersonal passive (IPC) cannot occur

Regular personal passive, then, is simply not grammatical given the absence of any [-r] argument that can map to SUBJ. IPC, by contrast, is grammatical since in the absence of any possible SUBJ-compatible argument, and importantly *only* in the absence of such an argument, German can project an expletive f-structural SUBJ according to Berman's (2003) proposal sketched in (3) above, and can thus license an impersonal passive construction. Evidently, this holds also for the intransitive subject equi verbs.

It looks, however, as if an account of these facts is also compatible with an analysis under which the non-finite complement bears the GF OBJ. Under such an analysis, if the infinitival complement bears the GF OBJ, this can map to SUBJ of the passive under the non-impersonal analysis of the IPC (although the burden of explanation rests, in my opinion, on accounting for how the soa-argument is assigned [-r] in the absence of any thematic role; but again, see

[14] Berman (*in press*) suggests that clausal complements have the same intrinsic feature assignment as their nominal counterparts, i.e. a clausal OBL would be intrinsically assigned [-o] by analogy to a nominal OBL.

Berman *in press*).[15] The dative object simply maps to OBJ_θ in the passive. Thus the IPC is licensed (in fact the IPC under this analysis corresponds to regular personal passive).
Summing up, in 2.4. I have argued that the distribution of the IPC and of regular personal passive is a consequence of the a-structure of the various equi verbs; an analysis which immediately accounts for the complementary distribution of the two constructions. In particular, lack of a [-r] (SUBJ-compatible) argument results in the ungrammaticality of regular personal passive in which case, adopting the analysis in (3), IPC is forced under passivization.

3. Visser's Generalization – an alternative analysis

The argument that *only* term arguments may function as controllers in Functional Control can be found in Bresnan (1982) and concerns what she termed Visser's Generalization, cf. Visser (1973: III.2: 2218). The reason that Bresnan proposes this restriction is that in English a transitive subject equi verb such as *promise* does not allow passivization in which the argument that maps to matrix OBJ in the active maps to SUBJ in the passive. Taking the example in (23), one might expect that the matrix OBJ *Mary* in (23a) could map to SUBJ under passivization, and thus we would expect (23b) to be grammatical, but it is not.[16] Under passivization, the controller (*John*) bears the GF OBL_{AGENT} and thus Bresnan (1982) attributes the ungrammaticality of (23b) to the fact that the controller is a non-term and functional control by a non-term is not permitted.

(23) a. John promised Mary to be on time [Bresnan 1982: 355]
b. *Mary was promised by John to be on time

The issue that needs to be resolved now is that the modifications to Functional Control proposed above can be said to 'cost' us Bresnan's account of the ungrammaticality of passive in (23b). It is for this reason, that I propose an alternative analysis of the ungrammaticality of passivization of transitive subject equi verbs in English.[17] I believe, however, that this account is perhaps superior since it concomitantly accounts for the non-application of Visser's Gener-

[15] There are, however, further complications that arise in the domain of infinitival complementation that make me hesitant to adopt Forst's (2006) proposal to replace COMP with OBJ and OBL in German. First, it is unclear that this step constitutes a major grammar writing economy since the lexical entries of verbs selecting infinitival complements require control equations, in contrast to the lexical entries of verbs selecting nominal complements – thus the lexical entries of the two types of verb cannot simply be conflated. Second, all infinitival complements, whether OBJ or OBL, permit topicalization in German, as is well-documented in the literature on coherent infinitives, e.g. Müller (2002:43) and Meurers (2000:22), and this is unexpected in Forst's (2006) account in which OBL can only topicalize when 'doubled' by a correlative. Topicalized infinitival complements of accusative object equi verbs simply do not require such doubling. Since mapping to SUBJ in the passive is inconclusive in German since German allows impersonal constructions, and because these topicalization facts are not as expected under the OBJ/OBL analysis of COMPS, the only remaining evidence for adopting the OBJ/OBL analysis is alternation with NPs vs. PPs, and is thus not very strong. Finally, the constraints on Long Distance Dependencies in German vary considerably for paths through nominal objects and through clausal complements and caution must be taken that this important distinction is not obscured by conflating OBJ and COMP. Given these problems, I prefer to adopt the analysis of the IPC as an impersonal construction, employing an XCOMP analysis, as in (19).

[16] Subject equi *promise* with an object as in (20a) is apparently marginal for many English speakers, who would prefer to use a finite *that*-clause instead (cf. Huddleston & Pullum 2002:1230, Courtenay 1998). There is clearly a deal of individual variation surrounding *promise*: (ii) and (iii) – which Bresnan (1982:355) provides to show that examples like (i) involve Anaphoric Control and are not exceptions to Visser's Generalization are marginal or even ungrammatical for many speakers:

i. Mary was never promised to be allowed to leave
ii. It was never promised to Mary to be allowed to leave
iii. To be allowed to leave was never promised to Mary

[17] I assume Visser's Generalization was only intended to cover transitive subject equi verbs since Bresnan (1982) only discusses it in relation to such verbs. At the time of Bresnan's article, the infinitival complement was assumed to have the GF COMP (not OBJ) and so the option of mapping the infinitival complement to passive SUBJ cannot have been entertained, thus the generalization could not have been intended to cover intransitive subject equi verbs. It is only more recently with the proposal that some infinitival COMPS should in fact be analyzed as OBJ that the lack of passivization of (some) intransitive subject equi verbs has become an issue at all. Since some of these verbs do, and others do not, permit passivization in English (cf. Falk 2001, although Huddleston & Pullum 2002 doubt this extraposed passive is a genuine passive construction), there is clearly more to be examined there.

alization in German with transitive subject equi verbs since transitive subject equi verbs do allow a passive construction in German; namely the IPC.

3.1 An alternative account of Visser's Generalization: The [+o] object

Let us assume for now that the object of English transitive subject equi verbs is also intrinsically assigned [+o], parallel to what was assumed for German above, see (24). Under passivization, the highest [-o] role is suppressed, but in contrast to the [-r] argument of the accusative object equi verbs, this [+o] object is not compatible with subject status, as seen above.

(24) a. promise agent beneficiary soa
b. [-o] [+o] [soa]
c. ∅ OBJ/OBJ$_\theta$
→ (23b) ungrammatical as no argument compatible with SUBJ

I propose therefore that it is this, rather than a ban on non-term functional controllers, that is the source of the ungrammaticality of (23b), i.e. passivization of transitive subject equi verbs in English is ungrammatical because the type of object involved bears a thematic role intrinsically assigned [+o], which cannot map to SUBJ ([-o/-r]) in the passive.[18] Thus, despite relaxing the ban on non-term functional controllers, an account of Visser's Generalization in English can still be offered.

Recall that German permits the IPC with transitive subject equi verbs, viz. (25b) but a 'regular' personal passive as in (25c), in which the active matrix object *mir*$_{\text{DATIVE}}$ maps to SUBJ *ich*$_{\text{NOM}}$ is ungrammatical.

(25) a. Die Firma versprach mir [den Rohrbruch bis zum Nachmittag zu reparieren]
The firm promised me$_{\text{DAT}}$ the burst pipe until afternoon to repair
The firm promised me to repair the burst pipe by this afternoon

b. weil mir **<u>von der Firma</u>** versprochen wurde den Rohrbruch bis zum Nachmittag zu reparieren
since me$_{\text{DAT}}$ from the firm promised was the burst pipe until afternoon to repair
intended: since I was promised by the firm to repair the burst pipe by this afternoon

c. *weil ich **<u>von der Firma</u>** versprochen wurde den Rohrbruch bis zum Nachmittag zu reparieren
since I$_{\text{NOM}}$ from the firm promised was the burst pipe until afternoon to repair

Assuming the same a-structure for German *versprechen* 'promise', viz. (26),[19] there are again two possible analyses of the grammaticality of (25b). First, it could be argued that such German infinitival complements bear the GF OBJ and the IPC is grammatical because there is an OBJ available in the a-structure which can map to SUBJ in the passive. A parallel construction in English would be ruled out by assuming that the infinitival complement in English must bear the GF XCOMP. Alternatively, one could argue that the IPC is an impersonal construction which lacks a c-structural subject altogether:

[18] This analysis is supported by further data from Visser (1973) and Bresnan (1982:354) illustrating other verbs predicated of the subject that disallow passive (but which have an object) since in many cases, a BENEFICIARY/EXPERIENCER analysis of the object (underlined) is plausible:
i. he strikes <u>his friends</u> as pompous/*his friends are struck as pompous (by him),
ii. Max failed <u>her</u> as a husband/*She was failed (by Max) as a husband,
iii. the vision struck <u>him</u> as a beautiful revelation/*He was struck (by the vision) as a beautiful revelation

[19] Considering the dative objects of transitive subject equi predicates such as *versprechen* 'promise', there is independent evidence that the dative object of *versprechen* is a BENEFICIARY since this verb occurs as an embedded predicate in the *kriegen*-passive (e.g. *er kriegte eine Stelle versprochen* 'he got promised a job') and the argument composition analysis in Cook (2006) requires that the embedded predicate have an a-structure < ∅/OBL$_{\text{AGENT}}$, beneficiary, theme >.

(26) a. versprechen agent beneficiary soa
b. [-o] [+o] [soa]
c. ∅ OBJ/OBJ$_\theta$
✓IPC → (25b) grammatical either (i) because soa is OBJ, and can map to SUBJ
or (ii) because German allows impersonal passives
*RPP → (25c) ungrammatical because there is no [-r] which can map to passive SUBJ

The root of the German-English contrast here, then, is either that German permits impersonal passive unlike English, or that the soa infinitival complement can bear the GF OBJ in German, but not in English.

4. Conclusion

I argued here that LFG's approach to Functional Control should be modified in two ways. First, we should allow non-term Functional Controllers. This step provides not only a satisfactory, and straightforward, account of the distribution of split antecedents of object equi verbs in German but, I believe, it ultimately permits a more satisfactory account of Visser's Generalization as it applies in English and German. This modification also requires that implicit agents of passives project an f-structure, even when they are c-structurally absent. This modification permits a Functional Control analysis of German equi verbs that license the Infinitival Passive, which is in keeping with the split antecedents facts, but also captures the fact that implicit arguments of passives can act as binders. Finally, representing implict arguments of passives at f-structure appears to resolve the issue of the (c-structural) optionality of the implicit argument in passives rather elegantly and appears to be lexically more economical since only one lexical entry is required for passivized verb forms.

References

Alsina, A., T. Mohanan and K. P. Mohanan (1996). Untitled Submission to the LFG List. 3 September 1996.

Alsina, A., T. Mohanan and K. P. Mohanan (2005). How to get rid of the COMP. In M. Butt and T. H. King (Eds.), *Proceedings of the LFG'05 Conference. University of Bergen, Norway*. Stanford: CSLI.

Asudeh, A. (2005). Control and Resource Sensitivity. *Journal of Linguistics 41*(3), 465-511.

Bech, G. (1955). S*tudien über das deutsche verbum infinitum*. Volume 1. Dan. Historisk-filologiske Meddelelser 35, 2. Copenhagen.

Berman, J. (2003). *Clausal Syntax of German*. Stanford: CSLI.

Berman, J. (in press). Functional identification of complement clauses in German and the specification of COMP. In J. Grimshaw, J. Maling, C. Manning, J. Simpson, and A. Zaenen, (Eds.), *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*. Stanford: CSLI.

Bresnan, J. (1982). Control and Complementation. In J. Bresnan (Ed.), *The Mental Representation of Grammatical relations*, pp. 282-390. Cambridge, MA: MIT Press.

Bresnan, J. (2001). *Lexical Functional Grammar*. Oxford: Blackwell.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Cook, P. (2006). The Datives that aren't born equal. Beneficiaries and the Dative Passive. In D. Hole, A. Meinunger, and W. Abraham (Eds.), *Datives and other Cases. Between argument structure and event structure*, pp 141-184. Amsterdam: Benjamins.

Courtenay, K. (1998). Subject-Control Verb PROMISE in English. Submission to the Linguist List 9.651, May 5th 1998.

Culicover, P. and R. Jackendoff (2005). *Simpler Syntax*. Oxford: Oxford University Press.

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

Dalrymple, M. and H. Lødrop (2000). The Grammatical Function of Complement Clauses. In M. Butt and T. H. King (Eds.) *Proceedings of the LFG'00 Conference. University of California, Berkeley*. Stanford: CSLI.

Falk, Y. N. (2001). *Lexical-Functional Grammar. An introduction to parallel constraint-based Syntax*. Stanford: CSLI.

Forst, M. (2006). COMP in (parallel) Grammar Writing. In M. Butt and T. H. King (Eds.) *Proceedings of the LFG'06 Conference. University of Constance*. Stanford: CSLI.

Frey, W. (1993). *Syntaktische Bedingungen für die semantische Interpretation*. Akademie Verlag: Berlin.

Huddleston, R. and G. K. Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Jenkins, L. (1972). *Modality in English Syntax*. Ph. D. thesis, MIT.

Joseph, B. D. (1979). Raising to oblique in Greek. In *Proceedings of the Fifth Meeting of the Berkeley Linguistics Society*, pp. 114-128.

Joseph, B. D. (1990). Is Raising to Prepositional Object a Possible Grammatical Rule? In B. Joseph and P. Postal (Eds.), *Studies in Relational Grammar 3*, pp. 261-276. Chicago: University of Chicago Press.

Lødrup, H. (2002). Infinitival complements in Norwegian and the form-function relation. In M. Butt and T. H. King (Eds.), *Proceedings of LFG'02. National Technical University of Athens*. Stanford: CSLI.

Lødrup, H. (2004). Clausal complementation in Norwegian. *Nordic Journal of Linguistics 27*(1), 61-95.

McCloskey, J. (1984). Raising, Subcategorization and Selection in Modern Irish. *Natural Language and Linguistic Theory 1*, 441-487.

Meurers, W. D. (2000). *Lexical Generalizations in the Syntax of German Non-Finite Constructions*. Working Papers of the SFB 340 Sprachtheoretische Grundlagen für die Computerlinguistik, Report nr. 145. Universities of Stuttgart and Tübingen.

Müller, S. (2002). *Complex Predicates. Verbal complexes, resultative constructions and particle verbs in German.* Stanford: CSLI.

Pollard, C. and Sag, I. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Růžička, R. (1983). Remarks on Control. *Linguistic Inquiry 14*(2), 309-324.

Sag, I. and C. Pollard (1991). An integrated theory of complement control. *Language 67*(1), 63-113.

Visser, F. T. (1973). *An historical syntax of the English language*, Volume 3,2. Leiden: F.J. Brill.

INFORMATION STRUCTURE AND SCOPE IN GERMAN

Philippa Cook
Zentrum für allgemeine Sprachwissenschaft (ZAS), Berlin

John Payne
University of Manchester

Proceedings of the LFG06 Conference
Universität Konstanz

Miriam Butt and Tracy Holloway King (editors)
2006

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

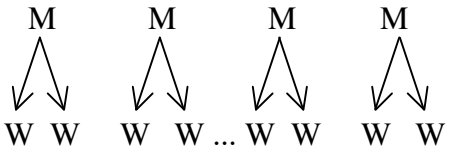
Abstract

Previous approaches to quantifier scope in German have relied on a disjunctive approach in which either higher rank on the grammatical function hierarchy or linear precedence allows a given NP to have distributive scope. In this paper, we instead tie the possibility of quantifier scope in German directly to information structure: only topics can have distributive scope. We present a new feature-based account of the information-structure concepts that are needed to predict German word order, and embed these features in f-structures, in effect amalgamating f-structure and i-structure information in a single level of representation. This amalgamated representation serves firstly as the input to a compositional logical form representation of the sentence, and secondly as a set of instructions as to how to articulate the compositional representation into, in particular, topical and non-topical components. The optional application of a distributivity operator to the topical component completes the analysis. This analysis not only obviates the need for a disjunctive approach to quantifier scope, but also neatly accounts for perceived discrepancies in the availability of particular readings with standard and non-standard predicates.

1. Initial observations

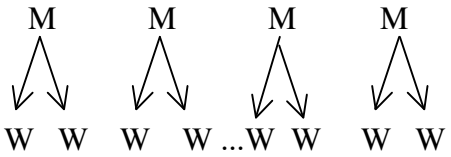
Initial observations concerning quantifier scope in German suggest that it is subject to a disjunctive condition based on (a) the grammatical function (GF) hierarchy (minimally SUBJ > OBJ) and (b) linear precedence Kiss (2001).[1] Consider the following examples, adapted from Frey (1993).

(1) a. $[\text{Viele Männer}]_{\text{SUBJ}}$ haben $[\text{zwei Frauen}]_{\text{OBJ}}$ hofiert
many men have two women courted



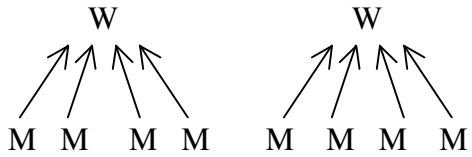
'Many men courted two women.'

b. $[\text{Zwei Frauen}]_{\text{OBJ}}$ haben $[\text{viele Männer}]_{\text{SUBJ}}$ hofiert
two women have many men courted



'Many men courted two women.'

OR



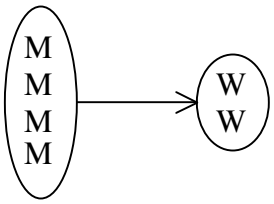
'Two women were courted by many men.

1. We acknowledge the support of the British Academy through the award of an Overseas Conference Grant (OCG44592) to Payne.

In example (1a), the NP *viele Männer* is both higher than *zwei Frauen* on the GF hierarchy and simultaneously precedes it. With neutral intonation, the only distributive reading is one in which the men distribute over the women, i.e. for each man, there is a potentially distinct set of two women that he courted. In such cases, we will say that *viele Männer* has distributive scope, and, since the scope follows linear precedence, we will call this a surface scope reading.

On the other hand, (1b), with fronting of the object, is ambiguous between two distributive readings. Firstly, *viele Männer* may have distributive scope, yielding the same interpretation as in (1a). In Kiss's (2001) approach, distributive scope in this case arises from the higher status of *viele Männer*, i.e. subject, on the GF hierarchy. Readings such as this in which scope does not follow linear order can be called inverse scope readings. However, an equally available interpretation is the surface scope reading in which *zwei Frauen* has distributive scope, i.e. for each of the two women, there is a potentially distinct set of many men who courted them. Since it is difficult to obtain this reading from the active in English, we signal this conventionally in the translation by employing the passive (even though the German construction is of course active). The distributive scope of *zwei Frauen* in this case is then attributed to its linear precedence, despite its lower rank (object) on the GF hierarchy. Thus in order for an NP to have distributive scope, it must either outrank all other elements on the GF hierarchy, or it must precede them.[2]

There is also of course a collective reading for both (1a) and (1b) in which there are at most two women and one set of many men involved in the courting event. This reading might be contextualised, for example, in a medieval setting in which a group of two women are surrounded by a sizeable group of male lute players, viz.



In collective readings there is no asymmetrical scope relationship.

Previous LFG approaches to quantifier scope within Glue Semantics (e.g., Crouch & van Genabith 1999; Dalyrmple et al. 1997) tie the possibility of scope ambiguity to the existence of multiple proofs for a single utterance. The fact that both the surface and inverse scope readings are not always equally available in English is presumed to be due to either pragmatic or plausibility restrictions. However, the rather more systematic nature of the German data – in particular the correlation between displacement and additional scope readings as in (1b) – suggests an approach in which the availability of distributive readings is linked to word order and, in turn, to information structure given

2. We note that a parallel disjunctive analysis has been adopted for binding effects in German (Choi 1995, Bresnan 1998, Berman 2003). It is conceivable that an analysis tied to information structure, analogous to the one presented here, might be successfully applied to the binding data. We leave this however as a topic for future research.

that German scrambling and other forms of displacement are clearly information-structurally driven (e.g., Lenerz 1977).

2. German Word Order and Information Structure

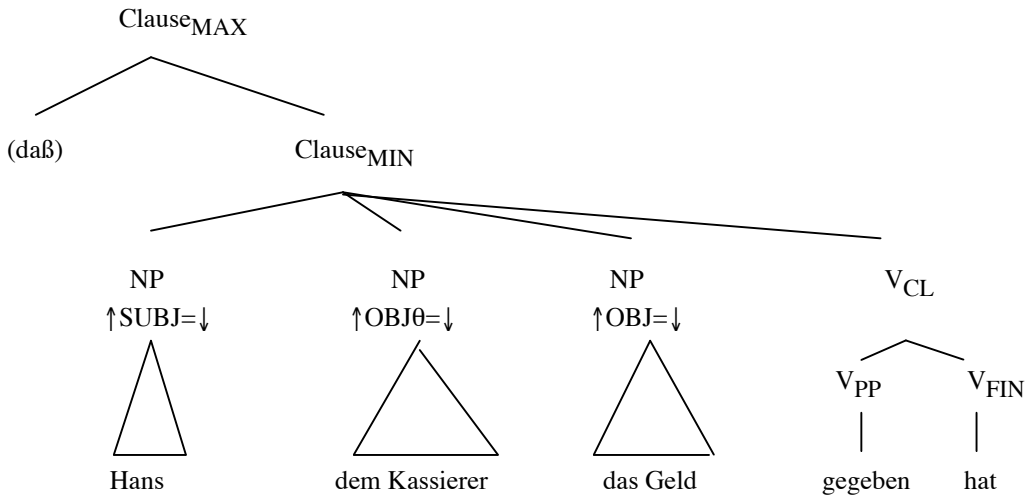
In this section, we outline our approach to German word order and information structure. In German, as is well-known, subordinate-clause word order differs, in the presence of complementizers, from main-clause word order. We consider first, in 2.1., subordinate-clause order in which the complementzer is initial and the verb or verb cluster is final. The remaining elements of the clause, i.e. arguments and adjuncts, lie between these in what is traditionally known as the "Mittelfeld" (middlefield). The order of elements in the middlefield is determined by a number of factors, but in particular the grammatical relation and information-structure status of each element. We introduce then the information-structure concepts which we believe to be word-order determinants. Secondly, in 2.2, we consider main-clause word order in which the finite verb is fronted and preceded by one element which has information-structure prominence.

2.1 Middlefield

In the spirit of Choi (1999) we assume a flat structure for the German middlefield, with word order within the middlefield determined by OT linear precedence constraints. The structure we propose for the subordinate clause in (2), with canonical word order, is (3).

(2) Ich glaube, daß [Hans] $_{SUBJ}$ [dem Kassierer] $_{OBJ\theta}$ [das Geld] $_{OBJ}$ gegeben hat]
I believe that Hans the cashier the money given has
'I believe that Hans gave the money to the cashier.'

(3)



Note that, in order to avoid controversies over category labelling, we employ here schematic labels such as Clause$_{MIN}$ for the middlefield and NP for noun phrase. The choice of category labels is essentially tangential to the issues raised in this paper.

The canonical word order in (2) follows the linear precedence constraint GF in (4), where > denotes "precedes".

(4) SUBJ > OBJθ > OBJ (GF)

Here objects are distinguished as OBJ (primary object, accusative case) and OBJθ (secondary object, dative case).[3] The domain of the constraint is the middlefield, i.e. it orders the daughters of $Clause_{MIN}$. The verb cluster V_{CL}, consisting in this example of the past participle *gegeben* 'given' and the finite auxiliary *hat* 'has', is obligatorily final.

We then use a three-term feature system to represent information structure concepts.

(5) ±T(opic), ±N(ew), ±C(ontrastive)

The feature ±T distinguishes topical from non-topical information. The concept of topic that is intended here is "aboutness topic", in the sense of Reinhart (1981). Importantly, topics do not necessarily represent old information, nor are they necessarily unique in a given utterance. The feature ±N straightforwardly distinguishes new from old information, while the feature ±C distinguishes contrastive from non-contrastive information in the sense of Frey (2006). A summary of the possible feature combinations and their English designations is given in (6).[4]

(6)

Summary of feature combinations:

{+T, –N, –C}	old-information topic
{+T, +N, –C}	new-information topic
{+T, –N, +C}	contrastive old-information topic
{+T, +N, +C}	contrastive new-information topic
{–T, +N, –C}	non-contrastive focus
{–T, +N, +C}	contrastive focus
{–T, –N, –C}	tail
{–T, –N, +C}	contrastive tail

3. For the analysis of the few ditransitives whose accusative object precedes the dative object in canonical order see Cook (2006).

4. The three-way feature system proposed here differs from the two-term system (±N, ±P) proposed by Choi (1999) in two main respects. Firstly, Choi employs a concept of topic as necessarily old information, and does not therefore have a feature ±T. Secondly, Choi employs a feature ±P (for prominent) which applies both to topics and contrastive focus: there is therefore no possibility of distinguishing between contrastive and non-contrastive topics. Seven of the eight terms permitted by the three-way system are employed in this paper. The one which is not is contrastive tail {–T, –N, +C}. However, as pointed out by Miriam Butt, a conceivable use for this term might be postverbal backgrounded phrases in Hindi/Urdu (Butt & King, to appear).

The features then play a crucial role, in addition to GF, in determining the contextually possible middlefield word orders, as shown in (7) and (8).

(7) Context: Wem hat Hans das Geld gegeben? [Who did Hans give the money to?]

Ich glaube, daß….

a. [Hans]$_{SUBJ}$ [dem Kassierer]$_{OBJ\theta}$ [das Geld]$_{OBJ}$ gegeben hat
+T, –N,–C –T, +N, +C –T, –N, –C –T, –N, –C

b. [Hans]$_{SUBJ}$ [das Geld]$_{OBJ}$ [dem Kassierer]$_{OBJ\theta}$ gegeben hat
+T, –N,–C ±T, –N, –C –T, +N, +C –T, –N, –C

c. [das Geld]$_{OBJ}$ [Hans]$_{SUBJ}$ [dem Kassierer]$_{OBJ\theta}$ gegeben hat
+T, –N,–C –T, –N, –C –T, +N, +C –T, –N, –C

(8) Context: Was hat Hans dem Kassierer gegeben? [What did Hans give to the cashier?]

Ich glaube, daß….

a. [Hans]$_{SUBJ}$ [dem Kassierer]$_{OBJ\theta}$ [das Geld]$_{OBJ}$ gegeben hat
+T, –N,–C ±T, –N, +C –T, +N, +C –T, –N, –C

b. [dem Kassierer]$_{OBJ\theta}$ [Hans]$_{SUBJ}$ [das Geld]$_{OBJ}$ gegeben hat
+T, –N,–C –T, –N, –C –T, +N, +C –T, –N, –C

c. %[Hans]$_{SUBJ}$ [das Geld]$_{OBJ}$ [dem Kassierer]$_{OBJ\theta}$ gegeben hat
+T, –N,–C –T, +N, +C –T, –N, –C –T, –N, –C

Examples (7a,b) and (8a,c) are the famous 'Lenerz data' which any account has to cover. In (7a,b), *dem Kassierer* is contrastive focus, and both object orders (OBJθ>OBJ; OBJ>OBJθ) are permitted. In (8a,c), however, *das Geld* is contrastive focus. While all speakers in this case allow the canonical order OBθ>OBJ, as in (8a), there is variable acceptance, indicated by the percentage symbol, of the OBJ>OBJθ order in (8c). In our terms, this variability depends on whether speakers allow +C information to scramble. In (7b) and (8a), note that the initial object (*das Geld* and *dem Kassierer* respectively) can be annotated either –T or +T with no effect on the ordering. If the +T annotation is chosen, there will then be two elements which have topic status in the sentence. Note also that we have added (7c) and (8b), fronting of a non-subject topic. We are making the point here that these elements can get a +T interpretation if a speaker decides to structure the answer that way.

The constraint ranking which gives these orders, disallowing (8c), is (9).

(9) $X>V_{CL}$ >> +T > –T >> {–N > +N, GF}

The constraint $X>V_{CL}$, which ensures the final position of the verb cluster, is highest ranked. The next-highest ranked is the constraint that topical information precedes non-topical information, followed by the equally ranked constraints –N > +N and GF. The equal ranking of these last two constraints allows the two alternative object orders in (7a,b), but (8c) violates both –N > +N and OBJθ > OBJ and is therefore non-optimal. For speakers who allow the scrambling in (8c), further constraints involving +C will need to be invoked. We ignore this complication here.

2.2 Front Field

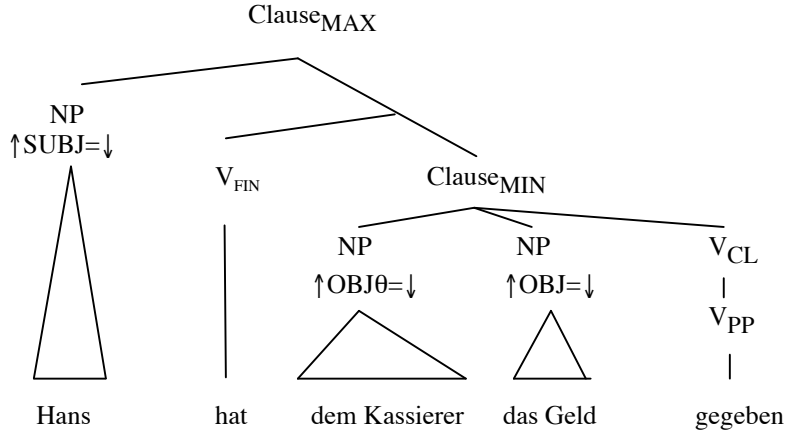
In German main clause order the front field contains a single item, either a single syntactic constituent or an information unit consisting of verb and other constituents (for the information units involved see Cook 2001, and Kaplan & Zaenen 2002). The front field is followed by the finite verb and the remaining elements of the middlefield. Consider then the following examples, given the context in (10).

(10) Context: Wem hat Hans das Geld gegeben? [Who did Hans give the money to?]
(with Hans as topic)

a.	$[\text{Hans}]_{\text{SUBJ}}$	hat	$[\text{dem Kassierer}]_{\text{OBJ}\theta}$	$[\text{das Geld}]_{\text{OBJ}}$ gegeben
	+T, –N,–C		–T, +N, +C	–T, –N, –C –T, –N, –C
b.	$[\text{Hans}]_{\text{SUBJ}}$	hat	$[\text{das Geld}]_{\text{OBJ}}$	$[\text{dem Kassierer}]_{\text{OBJ}\theta}$ gegeben
	+T, –N,–C		–T, –N, –C	–T, +N, +C –T, –N, –C
a.	$[\text{dem Kassierer}]_{\text{OBJ}\theta}$	hat	$[\text{Hans}]_{\text{SUBJ}}$	$[\text{das Geld}]_{\text{OBJ}}$ gegeben
	–T, +N, +C		+T, –N,–C	–T, –N, –C –T, –N, –C

An illustrative structure is given for (10a) in (11).

(11)



Note that either the topic *Hans* (10a,b) or the contrastive focus *dem Kassierer* (10c) can be selected for placement in the front field. Following ideas of Payne (2000), we handle the competition for placement in the front field as alignment with the left edge

of $\text{Clause}_{\text{MAX}}$. In the system proposed here, the ranking of alignment constraints then takes the form in (12).

(12) Align WH >> {Align +T, Align +C}

Highest ranked is the constraint Align WH, i.e. an interrogative WH-phrase will obligatorily occupy the front field if one is present. In the absence of an interrogative WH-phrase, the constraints Align +T and Align +C are equally ranked, allowing either a topic or a contrastive element to be fronted. The alternative object orders in (10a,b) of course follow from the middlefield linear precedence constraint rankings in (9).

The crucial role of the ±C feature in determining the eligibility of non-topic elements for front field placement can be seen examples such as (13), from Frey (2006).

(13). Context: Wo liegt Heidelberg? [Where is Heidelberg?]

a. Heidelberg liegt [am Neckar]
Heidelberg lies on.the Neckar
+T, –N,–C –T, +N,–C
'Heidelberg is on the Neckar.'

b. #[Am Neckar] liegt Heidelberg.
–T, +N,–C +T, –N,–C

The symbol # is intended here to indicate that (13b) is unacceptable in the given context. We can compare (13) with (14).

(14) Context: An welchem Fluss liegt Heidelberg? [On which river is Heidelberg?]

a. Heidelberg liegt [am Neckar]
Heidelberg lies on.the Neckar
+T, –N,–C –T, +N,+C
'Heidelberg is on the Neckar.'

b. [Am Neckar] liegt Heidelberg.
–T, +N,+C +T, –N,–C

The topic *Heidelberg* can always be placed in the front field, as in (13a) and (14a). On the other hand, it is inappropriate to place the focus *am Neckar* in the front field unless it is contrastive as in (14b), where there is a contrast with other possible rivers.

3. Information Structure and Scope

The basic claim of this paper is then that the disjunctive approach to quantifier scope in German can and should be replaced by one in which distributive quantifier scope

interpretations depend simply on information structure. The basic constraint on interpretation will be (15).[5]

(15) +T plural NPs allow distributive interpretations

In other words, only topics can have distributive scope. They can of course also be interpreted collectively.

The distributive interpretations of the examples in (1) follow straightforwardly from this constraint:

(16) a. Context: Was die Männer betrifft, wie viele von ihnen haben zwei Frauen hofiert?

[Talking about the men, how many (each) courted two women?]

[Viele Männer]$_{SUBJ}$	haben	[zwei Frauen]$_{OBJ}$	hofiert
+T, +N,+C		–T, –N, –C	–T, –N, –C
DIST			

b. Context: Was die Frauen betrifft, wie viele von ihnen haben viele Männer hofiert?

[Talking about the women, how many were (each) courted by many men?]

[Zwei Frauen]$_{OBJ}$	haben	[viele Männer]$_{SUBJ}$	hofiert
+T, +N,+C		–T, –N, –C	–T, –N, –C
DIST			

c. Context: Was die Männer betrifft, wie viele Frauen haben viele von ihnen hofiert?

[Talking about the men, how many women did many of them (each) court?]

[Zwei Frauen]$_{OBJ}$	haben	[viele Männer]$_{SUBJ}$	hofiert
–T, +N,+C		+T, –N, –C	–T, –N, –C
		DIST	

In (16a), the men are topic, but the question asks how many of them each courted two women. In the answer, the component *viele* 'many' in *viele Männer* is new and also contrastive information, since *viele* contrasts with other possible quantifiers. This is a basic information structure for (1a), in which the men have distributive scope. In (16b), we have an analogous information structure, but this time the women are the topic and the numeral *zwei* is contrastive new information. This is a basic information structure for the interpretation of (1b) in which the women have distributive scope. However, for the fronted object order there is an alternative information structure, shown in (16c), in which

5. The connection between quantifier scope interpretations and information structure, in particular topicality, has been noted in other languages. See for example van Valin (2005: 81-88) and references therein.

the subject *viele Männer* is a topic and *zwei Frauen* is a contrastive focus, the numeral *zwei* providing the answer to the question.

4. Inverse scope and prosody

For many (but not all) speakers, scope inversion, i.e. a plural NP is allowed to have distributive scope over an NP which precedes it, is possible even in the absence of displacement. The contexts are however typically quite complex. Examples are in (17).

(17) a. Context: Die Qualität der Patientenbetreuung ist normalerweise in diesem Krankenhaus sehr gut. Jeder Patient wird täglich von 3 Oberärzten besucht. Heute war es wegen des Streiks jedoch nicht so gut. [The quality of healthcare in this hospital is generally excellent. Each patient normally gets a visit by three consultants. But today, because of the strike…]

[Drei Oberärzte] $_{SUBJ}$	besuchten	[nur vier Patienten] $_{OBJ}$
three consultants	visited	only four patients
–T, –N, –C	–T, –N, –C	+T, +N, +C
		DIST

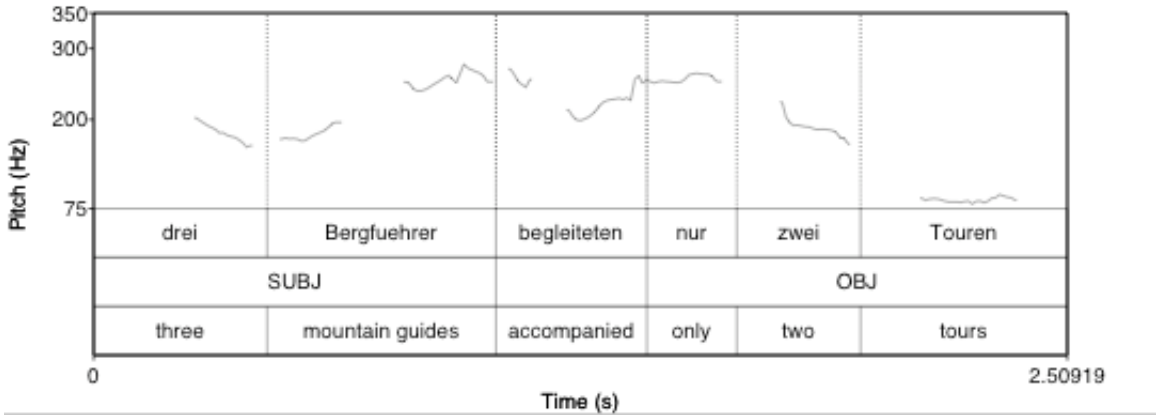
'Only four patients were visited by three consultants.'

b. Context: Die Touren sind im Allgemeinen sehr gut betreut. In der Regel hat jeder Tour mindestens drei Bergführer. Gestern war das allerdings nicht so. [The tours are generally well-staffed. As a rule, every tour has at least three mountain guides. But yesterday this didn't happen...]

[Drei Bergführer] $_{SUBJ}$	begleiteten	[nur zwei Touren] $_{OBJ}$
three mountain guides	accompanied	only two tours
–T, –N, –C	–T, –N, –C	+T, +N, +C
		DIST

'Only two tours were accompanied by three mountain guides.'

It is clear that in the given contexts these are utterances about patients and about tours, respectively. The contexts here make the patients/tours into contrastive new topics, which should normally occupy the front field. The motivation for the word order observed in (17a,b) appears to be to place new information late, but this is at the cost of placing a "tail" in the front field. This breach of the normal constraints is however prosodically marked. Such examples are associated with a special contour, as demonstrated in the following trace for (17b).



This contour may have affinities with the so-called 'hat contour', known to exist in other kinds of scope inversion examples (cf. Jacobs 1984, 1996; Büring 1997; Krifka 1998; Molnár & Rosengren 1996). However native speaker intuitions suggest that it may not be identical. We leave a fuller investigation of this issue to further research.

Scope inversion with the indicated prosodic contour is not restricted to subject-object orders. It can also occur in object-subject orders in similar kinds of context.

(18) Context: In den USA haben alle Doktoranden zwei Betreuer. Wir haben es in unserem Institut leider nicht so gut. [In the USA all PhD students have two supervisors. In our institute, we're not so fortunate…]

[Zwei Betreuer] $_{OBJ}$	haben	[nur vier Studenten] $_{SUBJ}$
two supervisors	have	only four students
–T, –N, –C	–T, –N, –C	+T, +N, +C
		DIST

'Only four students have two supervisors.'

Here again, the front field is occupied by a tail. It should be noted that from an information-structure point of view, as well as prosodically, example (18) is quite different to (16c). Both however involve a subject having distributive scope over a preceding object.

5. Semantic Representation

In this section, we consider the semantic representation of distributive scope. First of all, we follow in particular Steedman (2006) in allowing predicates to take set entities as arguments, and in taking indefinite noun phrases to denote generalized quantifiers which contain underspecified skolem terms *skolem'p*, where *p* is any property. Skolem functions map properties to entities which have that property, such that these entities are

dependent on any universal quantifier in whose scope they occur.[6] The underspecified representation of an indefinite noun phrase like *viele Männer* 'many men' will then be *λp.p(skolem'man'; many')*, denoting the set of properties which the set(s) of many men picked by the skolem term have. If a skolem term is specified outside the scope of a universal quantifier, it simply picks a constant set. Once specified, the underspecified term *skolem'man'* is converted in this case simply to $sk_{man'}$, representing the constant set picked by the skolem term. However, if a skolem term is specified within the scope of a universal quantifier which binds the variable *w*, its representation becomes $sk^{(w)}_{man'}$. That is, the skolem term in this case picks a different set of men for each value of the variable *w*. Skolem terms are a natural way to characterise the underspecified nature of the interpretation of indefinites, which depending on context either denote constant sets, corresponding to traditional "wide scope" readings, or have dependent denotations when outscoped.

Within this system, a collective reading of (19a) will have, ignoring tense and assuming saturation of the object argument first, the underspecified semantic representation (19b). There are no universal quantifiers in (19b), so when the skolem terms are specified, they will denote constant sets as in (19c). If desired, (19c) can be simplified by lambda conversion to (19d).

(19) a. $[\text{Viele Männer}]_{\text{SUBJ}}$ haben $[\text{zwei Frauen}]_{\text{OBJ}}$ hofiert

b. λp.p(skolem'man'; many')(λx.court' (skolem'woman'; two')x)

c. $\lambda p.p(sk_{man';\ many'})(\lambda x.court'(sk_{woman';\ two'})x)$

d. court'(skwoman'; two') (skman'; many')

Since *court'*, like all predicates, takes set entities as its arguments, this naturally represents the collective reading in which many men as group court two women as a group.

In order to derive the distributive readings, we then assume the optional application of a distributivity operator *D* to the semantic representation of the NP which has wide scope.[7] The underspecified representation of the distributive subject/topic interpretation of (20a) will then be (20b), exactly the same as (19b).

---

6. Winter (1997, 2001), following Reinhart (1997) has a similar analysis of indefinites in terms of choice functions, which he states as equivalent to skolem functions of arity zero. We simplify the representation of the cardinality of the sets picked out by skolem terms: a term *skolem'man'; many'* will be considered to pick sets whose cardinality is *many'*, however *many'* is defined.

7. Note that we doubt whether it is best to follow Steedman (2006) in taking quantifier distributivity in these kinds of examples to be based on multiple lexical representations of the predicate. This seems inappropriate when all arguments and indeed adjuncts can in principle scope over each other. See also Winter (1997, 2001) for arguments that both NP and predicate distributivity are in principle necessary.

(20) a. Viele Männer]$_{SUBJ}$ haben [zwei Frauen]$_{OBJ}$ hofiert

b. λp.p(skolem'man'; many')(λx.court'(skolem'woman'; two')x)

Specification of the subject/topic skolem term and application of the distributivity operator will however in this case yield (21a), which, when the function of the distributivity operator is spelled out, will be equivalent to (21b). What the distributivity operator does is to state that all properties to which the denotation of the NP applies are properties which hold of every individual member of the sets which have those properties. The distributivity operator therefore introduces a universal quantifier which will have an effect on the interpretation of any skolem term in its scope. Subsequent specification of the object skolem term in (21c) in the scope of the universal quantifier yields the interpretation in which there are separate sets of two women depending on each individual man. If desired, (21c) can again be simplified to (21d).

(21) a. D(λp.p(sk$_{man', many'}$)(λx.court'(skolem'woman';)x)

b. λp.p(sk$_{man', many'}$)(λx.∀w[w∈x→ court'(skolem'woman' two')w]$^{\{w\}}$)

c. λp.p(sk$_{man', many'}$)(λx.∀w[w∈x→ court'(sk$^{(w)}$$_{woman', two'}$)w]$^{\{w\}}$)

d. ∀w[w∈ sk$_{man', many'}$→ court'(sk$^{(w)}$$_{woman', two'}$) w]$^{\{w\}}$

It will be noted that (20b) is already in the right format to conform to a structured meaning approach (Krifka 1991) in which sentence meanings are partitioned into two discourse components, one of which applies to the other. Here the partition is, in our terms, +T(–T), i.e. the semantic representation of the topic is applied to the semantic representation of the non-topical material.[8] In order to derive the reading in which an object/topic has distributive scope, we need to manipulate the logical form so that the semantic representation of the object as topic applies to the semantic representation of the remainder of the sentence. To do this, we follow the higher order unification idea of Pulman (1997). In order to get the underspecified representation in (19b, 20b) into the right format, the equations in (22a, b) have to be solved.

(22) a. +T(–T) = λp.p(skolem'man'; many')(λx.court'(skolem'woman'; two')x)

b. +T = λq.q(skolem'woman')

In (22a), the left-hand side of the equation specifies that we need a +T(–T) partition, and the right hand side of the equation is the underspecified representation which has already been computed from the semantic components of the sentence. Equation (22b) specifies that the topical information can be identified with the semantic representation of the object, *two women*. The solution to these equations is (23), which is now in the right +T(–T) format for the object to be interpreted as topic.

8. Technically it is the semantic representation of the element marked +T which applies to the semantic representation of the element marked –T. We simplify the notation here by writing the partition as +T(–T).

(23) $+T(-T) = \lambda q.q(\text{skolem'woman'; two'})(\lambda y.\text{court' } y(\text{skolem'man'; many'}))$

We can now apply skolem specification and the distributivity operator as before, but this time to the representation of the object. This gives (24a), which is equivalent to (24b) after the function of the distributivity operator is spelled out.

(24) a. $D(\lambda p.p(sk_{\text{woman'; two'}})(\lambda y.\text{court' } y(\text{skolem'man'; many'}))$

b. $(\lambda p.p(sk_{\text{woman'; two'}})(\lambda y.\forall w[w \in y \rightarrow \text{court' } w(\text{skolem'man'; many'})]^{\{w\}}))$

Further specification of the subject skolem term now yields the representation (25a) in which we must pick a distinct set of many men for each woman. This simplifies to (25b) if desired.

(25) a. $(\lambda p.p(sk_{\text{woman'; two'}})(\lambda y.\forall w[w \in y \rightarrow \text{court' } w(sk^{(w)}_{\text{man';many'}})]^{\{w\}})$

b. $\forall w[w \in sk_{\text{woman'; two'}} \rightarrow \text{court' } w(sk^{(w)}_{\text{man'; many'}})]^{\{w\}}$

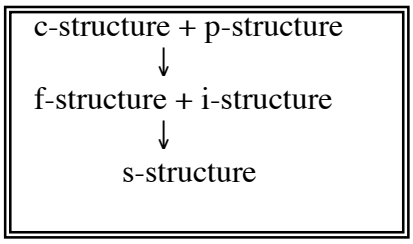
The view we have adopted here of the information-structure partitioning of semantic representations fits in well with a standard Glue approach which derives underspecified s-structure representations from f-structure predicate-argument structures. Note that in principle, the Glue approach allows the arguments of the predicate to be saturated in either order. Regardless of which order is chosen, higher order unification will be able to derive a correct +T(–T) partition from the underspecified source, and distributive scope will follow (optionally) from this partition. This approach does not tie informational partitions directly to surface structure, as in Steedman (1996). We expect that the flexibility which arises will be required in principle since distributivity and i-structure features are not generally subject to syntactic island constraints.[9]

6. Mapping

In order to simplify the number of mappings between different levels, we propose essentially three levels.

9. In particular, arguments that focus partitions are not in general subject to island constraints are given in Pulman (1997). We also note that, according to the native speaker intuitions of both authors of this paper, quantifier distributivity too is not subject to island constraints (contra Ruys 1992, Winter 2001). Our intuitions thus correspond to those reported in Abusch (1994), Geurts (2002) and Kempson & Meyer-Viol (2004).

(26)

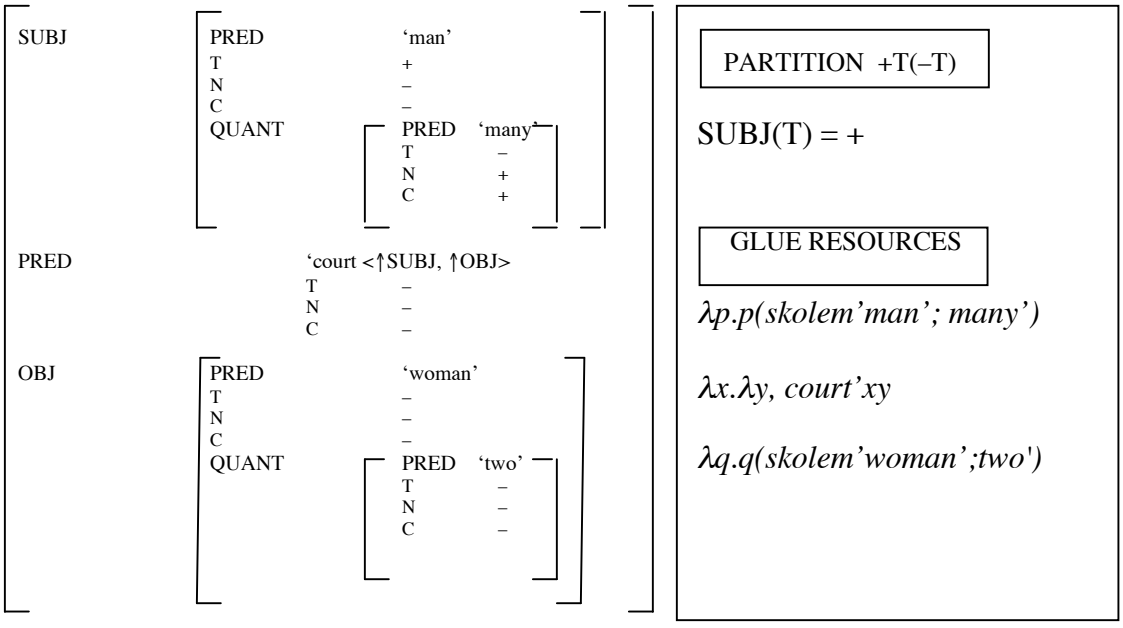


The most important notion which this diagram represents is the notion that i-structure information, in the form of i-structure features associated with individual predicates, can be amalgamated with f-structure. The position of i-structure information in the LFG architecture has been and remains subject to much debate, see especially King & Zaenen 2004). However, the amalgamation of i-structure and f-structure information ties together on the one hand the resources needed to construct a complete underspecified semantic representation, and on the other hand the i-structure information which will partition it.

Within the LFG-OT approach adopted here, f-structure and i-structure together will form the input to OT constraints which determine optimal c-structures and prosodies. Note that the prosody which is associated with scope inversion in (17) and (18) must then outrank the highly ranked constraints which disallow an initial tail. We leave however the details of the prosodic interactions for future research.

As an illustration of an amalgamated f-structure and i-structure, consider (27), which corresponds to (16a), using English names for the predicates involved and omitting tense.

(27)



The f-structure on the left contains the standard information that *many men* is subject, *two women* is object, and *court* is the main predicate. This information provides the resources needed for a Glue-based approach to semantic interpretation, as schematically shown in the box to the right, in which the representation of the predicate can combine in either

order with the representations of the subject and object. We are then left with an underspecified semantic representation such as that in (19b) and (20b). The i-structure annotations associate values of ±T(opic), ±N(ew), ±C(ontrastive) with each f-structure containing a predicate value.[10] Most importantly, the association of the feature +T with the f-structure of the subject will yield the equation SUBJ(T)= +. We assume that this implies that a +T(–T) partition must be created in which the +T information is equated with the semantic representation of the subject. Higher order unification will then partition the original semantic representation into the (still underspecified) +T(–T) format.

With the representation in this format, there are two possibilities. Either the distributivity operator is applied to the +T term, in which case we will derive an interpretation in which the subject has quantifier distributivity. Or the distributivity operator is not applied, in which case we achieve a collective interpretation. We assume that the application (or not) of the distributivity operator in German will depend on wider contextual factors.

7. Non-canonical predicates and topical objects

With verbs with regular THEME objects and AGENT subjects (canonical argument structure), the scope inversion examples of section 4 are slightly more accessible when the following argument is subject. Thus, (28) – in which the context is set up so as to force a distributive reading of the subject – is slightly more accessible than (29), in which the context is designed to force a distributive reading of the object. We use # to signal this here.

10. This notation obviates an objection made by King (1993) to locating information structure values within f-structure representations: the information structure attributed to the main predicate of the sentence will not in our system spread to its arguments. An alternative might be to invoke the subsumption approach of Kaplan & Zaenen (2002) in which information can be shared between f-structures on a partial rather than equal basis. Such an approach would involve, rather than a set of features, a set of paths linking f-structures representing topical, new and contrastive information to the basic f-structure information for the sentence. We do not exclude this approach, but observe that the feature notation adopted fits naturally with the higher order unification approach in which information structure feeds into articulations at the level of logical form. We speculate that the employment of independent f-structures to represent i-structure information might be most appropriate in cases where long-distance extractions are involved, and where island-constraints apply to the f-structure paths created. However, this is a large issue which is beyond the scope of this paper.

(28) Context: Alle Professoren wurden aufgefordert, ihre besten fünf Studenten für einen Preis zu empfehlen. Viele hatten Schwierigkeiten, überhaupt 5 Studenten zu empfehlen. Die meisten Professoren schlugen nur einen Studenten vor [All professors were requested to put forward their 5 best students for a prize. A lot had difficulty finding 5 students to recommend. Most professors just suggested one student.]

[5 Studenten]$_{OBJ}$ schlugen [nur 4 Professoren]$_{SUBJ}$ vor
Only 4 professors suggested 5 students

(29) Context: Die Qualität der Schwangerenbetreuung in Stuttgart ist sehr gut. Jede Frau hat Anspruch auf zwei Hebammen. In Frankfurt haben die Frauen es leider nicht so gut. [The quality of care for pregnant women is very good in Stuttgart. Every woman has access to two midwives. In Frankfurt the women don't have it so good]

#[2 Hebammen]$_{SUBJ}$ betreuen [nur 4 Schwangere]$_{OBJ}$
2 midwives look after only 4 pregnant women

In our account, this tendency can be seen to reflect the fact that (agentive) subjects are the default candidate for topic status (cf. Reinhart 1981).

Thematically non-canonical verbs can be seen as providing support for our topic-analysis of distributivity since we claim that EXPERIENCER objects can acquire topic status more easily than theme objects. With respect to this claim, we consider here psych verbs with STIMULUS subject and EXPERIENCER object. It is striking that with EXPERIENCER objects, it is far easier to have object distributivity in situ (inverse scope) than was the case with THEME object verbs above. Examples (30) and (31) have dative and accusative EXPERIENCERS respectively.

(30) Context: Man muss dem Jugendamt melden, wenn im Kindergarten ein Kind 5 oder mehr Unfälle in einem Monat hat. In den letzten Monaten mussten immer mehr Meldungen an das Jugendamt erfolgen. In diesem Monat war es besonders schlecht [You have to inform the Youth Services if a child has five or more accidents a month in the Kindergarten. In the last months we had to make more and more announcements to the Youth services. It was especially bad this month]

[5 Unfälle]$_{SUBJ}$ sind sogar [10 Kindern]$_{DAT\ OBJ}$ zugestoßen
5 accidents happened to 10 children this month

(31) Context: Jedes Jahr werden von den Designern neue Farben entwickelt. Die Farben werden einem Team von Gutachtern präsentiert. Normalerweise ist jedes Mitglied des Gutachterteams von ca. zwei Farben angewidert. Dieses Jahr haben die Entwürfe den Gutachtern besser gefallen. [Every year new colours are developed. The colours are presented to a panel of judges. Normally, every judge is repulsed by around 2 colours. This year, the designs appealed to the judges more.]

[2 Farben]$_{SUBJ}$ haben nur [4 Gutachter]$_{ACC\ OBJ}$ angewidert
2 colours only repulsed 4 judges

There is a further contrast between the THEME object and the EXPERIENCER object verbs: the reading in which a subject distributes over a preceding object is much 'more difficult'

to obtain with these verbs than with the agentive subject in (28) above. We indicate this again using #.

(32) Context: Die Ingenieure der verschiedenen Abteilungen der Firma kommen manchmal auf die gleichen Ideen für neue Lösungen. Dann gibt's immer Ärger. Erfreulicherweise hatten wir in letzter Zeit nicht so viel Ärger. [Engineers from different departments sometimes come up with the same idea for new solutions. Then there's always trouble. Fortunately there hasn't been so much trouble recently.]

#[Mehreren Ingenieuren]$_{\text{DAT OBJ}}$ sind nur [2 Ideen]$_{\text{SUBJ}}$ eingefallen
Only two ideas occurred to several engineers.

(33) Context: Wenn mindestens fünf Eltern sich beschweren, machen wir uns Sorgen über die Qualität unserer Produkte. Laut Firmenrichtlinien müssen die Produkte dann vorübergehend aus dem Verkauf genommen werden. Die Qualität unserer Produkte ist sehr gut. [If at least five parents complain then we worry about the quality of a product. According to firm guidelines we have to temporarily withdraw it from sale. The quality of our products is very good]

#[Fünf Eltern]$_{\text{ACC OBJ}}$ beunruhigten bislang nur [4 Produkte]$_{\text{SUBJ}}$
So far only 4 products have disturbed five parents.

The observation that thematic properties of a predicate affect scope has been made before (cf. Pafel 2006:70-74). It has, however, not previously been attributed to information structuring but has merely been stated as an extra 'factor' influencing scope. Under our analysis the facts fall out in the following way. The most typical topics are AGENTS (hence also typical animate/human) and thus, topic very often corresponds to subject. With a predicate with an AGENT argument, some contextual motivation is required for treating a non-agentive role as topic. Thus, when a THEME is topic, as in (29), some prosodic, contextual or syntactic support (or a combination thereof) is required. In the absence of an AGENT, as in the case of the STIMULUS-EXPERIENCER verbs in (30)-(31), the next highest role, namely the EXPERIENCER, is the most typical topic. Note again that this will often be an animate argument.[11] Under the disjunctive approach to scope discussed in Section 1, such facts are mysterious since that account predicts that a subject can always scope over a lower GF irrespective of linear order yet this is a dispreferred option for EXPERIENCER object psych verbs.

The availabilities of readings available with the three different types of predicates discussed here are summarised in table (33), in which *D* denotes 'distributes over'.

11. The higher a thematic role is in the thematic hierarchy, the more suitable a candidate for topic status it is. This is, of course, indirectly linked to animacy since high thematic roles such as agent, experiencer, beneficiary are typically animate. We do not, however, wish to augment the constraint set in (9) with a separate constraint concerning the linearization of animate arguments before inanimate ones since we believe any effects seemingly associated with alignment of animate arguments to be an epiphenomenon of the +T > –T constraint given in (9).

(34)	SUBJ > OBJ order	OBJ > SUBJ order
agentive subject theme object verb	**OBJ** *D* **SUBJ** available but needs some contextual and/or prosodic support, viz. (29), because THEME object is not the most typical topic	**SUBJ** *D* **OBJ** readily available, viz. (28) because AGENT subject is a typical topic
Stimulus subject Dat Experiencer Object	**OBJ** *D* **SUBJ** very easily available, (viz. 30), because EXPERIENCER is a fairly typical topic.	**SUBJ** *D* **OBJ** not readily available, viz. (32), because STIMULUS subject is not a typical topic. EXPERIENCER would be a more typical topic.
Stimulus subject Acc Experiencer Object	**OBJ** *D* **SUBJ** very easily available, (viz. 31), because EXPERIENCER is a fairly typical topic.	**SUBJ** *D* **OBJ** not readily available viz. (33), because stimulus subject is not a typical topic. EXPERIENCER would be a more typical topic

Conclusion

In this paper, we have provided a detailed analysis of quantifier scope phenomena in German in which distributive scope is directly linked to topicality. The analysis is framed in a streamlined view of the mapping between f-structure and s-structure in which information structure is amalgamated featurally with basic f-structure representations, and in which the s-structure derived compositionally from the basic f-structure representation is then partitioned into information structure components by higher order unification. The optional application of a distributivity operator to these partitioned meanings then derives the association between scope and topicality.

One of the major advantages of this approach is that it obviates the need for a disjunctive analysis based on grammatical relations and linear precedence. However, it also accounts for the varying availability of different scope readings when standard and non-standard predicates are taken into consideration. All the factors which have been implicated in the availability of distributive scope readings in addition to grammatical relations and linear precedence, e.g., higher animacy and thematic role status, fall naturally into place under the heading of topicality.

References

Abusch, Dorit (1994) The Scope of Indefinites. *Natural Language Semantics* 2. 83-135.

Berman, Judith (2003) *Clausal Syntax of German*. Stanford: CSLI.

Bresnan, Joan (1998) Morphology competes with Syntax: Explaining typological variation in weak crossover effects. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis & D. Pesetsky (eds) *Is the Best good enough? Optimality and Competition in Syntax*, 59-92. Cambridge, MA:MIT Press,

Büring, Daniel (1997) The great scope inversion conspiracy. *Linguistics & Philosophy* 20:511-545.

Butt, Miriam & Tracy Holloway King (to appear) Null elements in discourse structure. In K. V. Subbarao (ed.) *Papers from the NULLS Seminar*. Delhi: Motilal Banarasidas.

Choi, Hye-Won (1995) Weak Crossover in Scrambling languages: Precedence, rank, and discourse. Paper presented at the 1995 meeting of the LSA, New Orleans.

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

Cook, Philippa (2001) *Non-Finite Complementation and Information-Structuring in German*. PhD dissertation, University of Manchester.

Cook, Philippa (2006) The datives that aren't born equal: Beneficiaries and the dative passive. In Daniel Hole, André Meinunger and Werner Abraham (eds) *Datives and Similar Cases*: *Between argument structure and event structure*. 141-184. Benjamins. Amsterdam/Philadelphia.

Crouch, Richard and Josef van Genabith (1999) Context change, underspecification, and the structure of glue language derivations. In Dalrymple, M. (ed) (1999) *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*, 117-189. Cambridge MA: MIT Press.

Dalrymple, Mary, John Lamping, Fernando C.N. Pereira & Vijay Saraswat (1997) Quantifiers, anaphora, and intensionality. *Journal of Logic, Language and Information* 6 (3), 219-273. Reprinted in Mary Dalrymple (ed) (1999) *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*, 39-89. Cambridge MA: MIT Press.

Frey, Werner (1993) *Syntaktische Bedingungen für die semantische Interpretation*. Berlin: Akademie Verlag.

Frey, Werner (2006) Contrast and movement to the German prefield. In Valéria Molnár & Susanne Winkler (eds): *The Architecture of Focus. Studies in Generative Grammar 82*, 235-264. Berlin, New York: Mouton de Gruyter.

Geurts, Bert (2002) Specific indefinites, Presupposition, and Scope. To appear in: Rainer Bäuerle, Uwe Reyle, & T. Ede Zimmermann (eds.) *Presuppositions and Discourse*. Elsevier, Oxford.

Jacobs, Joachim (1984) Funktionale Satzperspektive und Illokutionssemantik. *Linguistische Berichte* 91:25-58.

Jacobs, Joachim (1996): Bemerkungen zur I-Topikalisierung. *Sprache und Pragmatik* 41, Lund, 1-48.

Kempson, Ruth & Wilfried Meyer-Viol (2004) Indefinites and Scope Choice. In Marga Reimer & Anne Bezuidenhout (eds) *Descriptions and Beyond*. Clarendon Press: Oxford. 558-583

King, Tracy Holloway (1993) *Configuring Topic and Focus in Russian*. Ph.D. thesis, Stanford University, Department of Linguistics.

King, Tracy Holloway & Annie Zaenen (2004) F-structures, information structure and discourse information. In Miriam Butt & Tracy Holloway King (eds) *Proceedings of LFG-04*. Sianford CA, CSLI Publications. Extended abstract: http://www-csli.stanford.edu/publications.

Kiss, Tibor (2001) Configurational and Relational Scope Determination in German. In W.D. Meurers & T. Kiss (eds) *Constraint-based Approaches to Germanic Syntax*, 141-175. Stanford: CSLI.

Krifka, Manfred (1998) Scope Inversion under the rise-fall pattern in German. *Linguistic Inquiry* 29(1):75-112.

Krifka, Manfred (1991) A Compositional Semantics for Multiple Focus Constructions. In *Informationsstruktur und Grammatik*, Sonderheft der *Linguistischen Berichte*, ed. Joachim Jacobs.

Lenerz, Jürgen (1977) *Zur Abfolge nominaler Satzglieder im Deutschen*. Studien zur deutschen Grammatik 5. Tübingen: Narr.

Molnár, Valéria & Inger Rosengren, (1996) Zu Jacobs' Explikation der I-Topikalisierung. *Sprache und Pragmatik* 41, Lund, 49-88

Pafel, Jürgen (2006) *Quantifier Scope in German*. Amsterdam/Philadelphia: John Benjamins.

Payne, John R. (2000) Verb-second in Germanic. Paper read at Australian Linguistic Society Conference, University of Melbourne.

Pulman, Stephen G. (1997) Higher Order Unification and the Interpretation of Focus. *Linguistics and Philosophy* 20:73-115.

Reinhart, Tanya (1981) Pragmatics and Linguistics: An Analysis of Sentence Topics. *Philosophica* 27: 53- 94.

Reinhart, Tanya (1997) Qunatifier scope: how labor is divided between QR and choice functions. *Linguistics and Philosophy* 20(4): 335-397.

Ruys, Eddie (1992) *The Scope of Indefinites*. Ph.D. dissertation, Utrecht University.

Steedman, Mark (1996) *Surface Structure and Interpretation*. Cambridge, Mass: MIT Press.

Steedman, Mark (2006) Surface-Compositional Scope-Alternation without Existential Quantifiers, Draft 5.1, Sept 2005. ftp://ftp.cogsci.ed.ac.uk/pub/steedman/quantifiers/journal5.pdf

Kaplan, Ronald & Annie Zaenen (2002) Subsumption and equality: German partial fronting in LFG. In Miriam Butt and Tracy Holloway King (eds) *Proceedings of the LFG02 Conference, National Technical University of Athens, Athens*. CSLI Publications: http://csli-publications.stanford.edu/

van Valin, Robert D. (2005) *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press.

Winter, Yoad (1997) Choice Functions and the Scopal Semantics of Indefinites. *Linguistics and Philosophy* 20:399-467.

Winter, Yoad (2001) *Flexibility Principles in Boolean Semantics. The Interpretation of Coordination, Plurality, and Scope in Natural Language*. Cambridge, Mass: MIT Press.

SEMANTICS VIA F-STRUCTURE REWRITING

Dick Crouch and Tracy Holloway King
Palo Alto Research Center

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)

2006

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

Abstract

This paper discusses how the XLE general purpose ordered rewrite rule system is used to produce semantic representations from syntactic f-structures. The rules apply efficiently because they operate on the packed input of f-structures to produce packed semantic structures. In addition to rules which convert the syntactic structure to a semantic one, there are rules that use external resources to replace words with concepts and grammatical functions with roles. Although the system described here could by no means be described as a theory of the syntax-semantics interface, from a practical stand point it can efficiently and robustly produce semantic structures from broad-coverage syntactic ones.

1 Introduction

This paper discusses the use of the XLE's [Crouch et al.(2006), Maxwell and Kaplan(1996)] transfer system [Crouch(2005), Frank(1999)] for mapping f-structures into semantic representations. The technique has been robustly applied to f-structures obtained by parsing open text, such as the Wall Street Journal and New York Times.

The semantics gives a flat representation of the sentence's predicate argument structure and the semantic contexts in which those predications hold. Contrast the f-structure and semantics in (1).

(1) a. Jane did not hop.

b.

```
"Jane did not hop."

 ┌                                                                          ┐
 │PRED    'hop<[1:Jane]>'                                                   │
 │        ┌                                                                ┐│
 │        │PRED  'Jane'                                                    ││
 │        │CHECK [_LEX-SOURCE morphology]                                  ││
 │SUBJ    │NTYPE ┌                                                        ┐││
 │        │      │NSEM [PROPER [NAME-TYPE first_name, PROPER-TYPE name]]  │││
 │        │      │NSYN proper                                             │││
 │      1 └      └                                                        ┘┘│
 │CASE nom, GEND-SEM female, HUMAN +, NUM sg, PERS 3                       │
 │ADJUNCT { 60 [PRED         'not'] }                                      │
 │        {    [ADJUNCT-TYPE neg  ] }                                      │
 │CHECK   [_SUBCAT-FRAME V-SUBJ]                                           │
 │TNS-ASP [MOOD indicative, PERF -_, PROG -_, TENSE past]                  │
 │30CLAUSE-TYPE decl, PASSIVE -, VTYPE main                                │
 └                                                                          ┘
```

c. alias(Jane:1,[Jane])
context_head(t,not:10)
context_head(ctx(hop:17),hop:17)
in_context(t,cardinality(Jane:1,sg))
in_context(t,role(mod(degree),ctx(hop:17),not:10,normal))
in_context(ctx(hop:17),past(hop:17))
in_context(ctx(hop:17),proper_name(Jane:1,person,Jane))
in_context(ctx(hop:17),role(Theme,hop:17,Jane:1))
lex_class(hop:17,[vnclass(run-51_3_2)])
lex_class(not:10,[sadv,impl_pn_np])
sortal_restriction(Jane:1,[7127])
word(Jane:1,Jane,noun,0,1,t,[[9482706]])
word(hop:17,hop,verb,0,17,ctx(hop:17),[[1948772], [2076532], [1823521], [2076385], [2076247], [2076113]])
word(not:10,not,adv,0,10,*t*,[[24548]])

Note how each clause of the core of the representation is set within a context (*in_context*). Contexts are introduced by clausal complements (COMP, XCOMP) in f-structure, but can also be lexically introduced, as shown by the sentential adverb *not*. Nominal and event arguments are skolemized: instead of quantifiers binding variables, terms like *hop:17* are used in place of the bound variables. In addition, roles are introduced in place of grammatical relations and PREDs are replaced by concepts which here are numbers that represent WordNet synonym sets (section 6.1).

The transfer system applies an ordered set of rewrite rules, which progressively consume the input f-structure replacing it by the output semantic representation. The system permits a form of input-limited recursion, where a rule can apply to its own ouptut provided that each application consumes some more of the rules' original input (thus ensuring termination of the recursion). This is required for capturing the contextual structure induced by the recursive embedding of complements within f-structure. The rewrite-based system can, and has, been used in place of components constructing semantic representations by more standard means, such as Glue Semantics [Dalrymple(2003)].

2 Brief System Introduction

In this section we provide a brief introduction to the XLE system that is used in producing XFR semantics. The syntactic component, including the morphology and tokenizer, is described in detail in [Riezler et al.(2002), Kaplan et al.(2004)]. XLE is described in [Maxwell and Kaplan(1996)] and many details are available in the on-line XLE documentation [Crouch et al.(2006)].

(2)

Input: *string*	
tokenizer morphology LFG grammar	XLE parser
Output of parser/Input to rewriting: *c- and f-structure*	
XFR semantics rules	XLE rewrite system
Output: *semantic facts*	

2.1 Types of Rewrite Rules

A somewhat contrived example of a rewrite rule is:

(3) PRED(%V, eat), SUBJ(%V, %S), OBJ(%V, %O), –OBL(%V, %%)
==>
word(%V, eat, verb), role(Agent, %V, %S), role(Theme, %V, %O).

This rule looks at a set of clauses describing an f-structure to see if there is some node %V (the % is used to indicate a variable), with a subject %S and object %O, but no oblique. If the left hand side of the rule is matched, the matching PRED, SUBJ and OBJ clauses are removed from the description, and are replaced by the word and role clauses on the right hand side of the rule. More generally, the format for rewrite rules is shown in figure 1.

The left hand sides of rules contain Boolean combinations of patterns over clauses. Clauses are atomic predicates heading a set of argument terms, where the terms may be non-atomic: e.g., "SUBJ(var(0), var(1))", where SUBJ is the predicate and var(0) and var(1) are the non-atomic arguments (var(#) is the XLE representation for LFG f-structure nodes). In patterns over clauses, some of the argument terms can be, or can contain, variables. For example, the pattern "SUBJ(%V, var(%Y))"

Rule	::=	LHS ==> RHS.	*Obligatory rewrite*
		LHS ?=> RHS.	*Optional rewrite*
		LHS *=> RHS.	*Recursive rewrite*
		|– Clause.	*Permanent, unresourced fact*
LHS	::=	Clause	*Match & delete atomic clause*
		+Clause	*Match & preserve atomic clause*
		LHS, LHS	*Boolean conjunction*
		(LHS | LHS)	*Boolean disjunction*
		–LHS	*Boolean negation*
		{ProcedureCall}	*Procedural attachment*
RHS	::=	Clauses	*Set of replacement clauses*
		0	*Empty set of replacement clauses*
		stop	*Abandon the analysis*
Clause	::=	Atom(Term,…,Term)	*Clause with atomic predicate*
		Atom	*Atomic clause*
		qp(Variable, [Term_i,…, Term_n])	*Clause with unknown predicate and n arguments*
Term	::=	Variable	
		Clause	

Figure 1: Format of Rewrite Rules

will match "SUBJ(var(0), var(1))", setting %V to var(0) and %Y to 1. Second-order quantification over atomic predicates is also available, where *qp(%P, [%V, %Y])* matches "SUBJ(var(0), var(1))", setting the predicate variable %P to SUBJ and the list of argument variables [%V, %Y] to var(0) and var(1).

By prefixing a clause pattern on the left hand side of a rule with a "+", you can indicate that the rule should check for the presence of a matching clause in the input without deleting the clause. Likewise, a prefix of "–" checks that a pattern is not matched by the input. Boolean combinations of clause patterns are possible, as are calls to external procedures. External procedures cannot directly manipulate the full set of input clauses; instead they allow you to perform table lookup or tests on terms, such as subsumption checking in the Cyc generalization hierarchy, or looking up the synset of a word in WordNet (section 6.1).

The right hand side of a rule can be a comma separated set of clause patterns, including the empty set represented as *0*. The right hand side can also be the directive *stop*, which means that the analysis path should be deleted.

Rules can be obligatory, optional or recursive rewrites, and can also introduce permanent non-consumable facts. If the left hand side of an obligatory rule is matched, then the consumed clauses (i.e. those not marked with a "+" or a "–") have to be removed from the set of input clauses, and replaced by the clauses on the right hand side of the rule. For an optional rule, conceptually speaking, there is a fork in the set of output clauses. On one fork the rule applies, and the consumed clauses on the left hand side are replaced by those on the right hand side. On the other fork the rule does not apply, and the set of clauses remains unchanged. But instead of forking the sets of clauses, the choice space is split to record the alternatives where the rule is and is not applied (section 2.2). A recursive rule can re-apply to its own output, provided that each recursion also consumes some of the input that was present before the first recursive application; this ensures termination of the recursion.

Rules are ordered: rule 1 applies to the input, rule 2 applies to the output of rule 1, and so on. Rule ordering can be exploited, e.g., to encode sequences of defaults, but the feeding and bleeding

behavior needs to be handled with care. The rule ordering also means that the scope of any recursion is strictly limited to a single rule. Note that the order of the discussion of the rules in this paper does not necessarily reflect their order in the system.

Clauses preceded by a |– are included as permanent, non-consumable facts. These are part of neither the input nor the output, but can be called on to provide tests or data. For example, a more sensible way of achieving the effects of (3) would be:

(4) a. |– concept-map(eat, V-SUBJ-OBJ, Agent, Theme).
|– concept-map(drink, V-SUBJ-OBJ, Agent, Theme).
…

b. PRED(%V, %P), SUBJ(%V, %S), OBJ(%V, %O), –OBL(%V, %%),
concept-map(%P, V-SUBJ-OBJ, %SR, %OR)
==>
word(%V, %P, verb), role(%SR, %V, %S), role(%OR, %V, %O).

In this way, a large number of lexical mappings can be asserted permanently (similar to lexical entries), as in (4a) and a single rule takes care of the concept mapping for transitive verbs, as in (4b) instead of having one rule for each verb.

The formalism also allows macros to be used to parameterize commonly occurring patterns in rules, and templates to parameterize commonly occurring sequences of rules. This is an alternative to using a type hierarchy [Oepen et al.(2004)] for producing compact rule sets.

2.2 Packing

Although constraints can sometimes be applied at the semantics level (or subsequent mapping to knowledge representation) to resolve syntactic ambiguities, others will pass through to the semantics level, and yet more may be introduced by things such as word sense ambiguity. Here we briefly describe the ambiguity packing mechanism used by the XLE rewrite system; for more detail on packing in the rewriting system see [Crouch(2005)] and for packing in general see [Maxwell and Kaplan(1991)].

Alternative interpretations are represented in a packed form. An example will give an idea of what these packed representations are like (shown somewhat abbreviated, e.g., cardinality and sortal restriction facts, which would all be in the top choice 1, are not shown):

(5) John saw a man with a telescope.

(6) choice: (A1 xor A2) iff 1

1 alias(John:1,[John])
1 context_head(t,see:32)
1 in_context(t,past(see:32))
1 in_context(t,specifier(man:12,a))
1 in_context(t,specifier(telescope:23,a))
1 in_context(t,proper_name(John:1,person,John))
1 in_context(t,role(Experiencer,see:32,John:1))
1 in_context(t,role(Stimulus,see:32,man:12))
A1 in_context(t,role(prep(with),man:12,telescope:23))
A2 in_context(t,role(prep(with),see:32,telescope:23))

1 word(John:1,John,noun,0,1,t,[[9487097]])
1 word(man:12,man,noun,0,12,t,[[10133569], [10423788], [10135377], [2449786], [10135514], [10135101]])
1 word(see:32,see,verb,0,32,t,[[2109658], [583923], [2109242], [1620934], [682517], [591374], [2131231], [911004]])
1 word(telescope:23,telescope,noun,0,23,t,[[4351615]]))

The standard prepositional attachment ambiguity is reflected in the semantics (6) by two alternative role restrictions: the telescope either modifies the seeing event, or the man. The two alternatives are labeled by the distinct choices *A1* and *A2*. As the first line in the representation states, *A1* and *A2* are mutually exclusive (xor = exclusive or) ways of partitioning the true choice labeled *1*. Most parts of the representation are common to both possible interpretations, and are thus labeled with the choice *1*. It is only the two role assignments for *telescope:23* that are put under distinct choice labels.

A slightly more complex case of prepositional attachment ambiguity gives rise to the following semantic representation (shown somewhat abbreviated):

(7) John saw a man in a park with a telescope.

(8)
choice: (A1 xor A2) iff 1
choice: (B1 xor B2 xor B3) iff A1
choice: (C1 xor C2) iff A2
1 alias(John:1,[John]),
1 context_head(t,see:42),
1 in_context(t,past(see:42)),
1 in_context(t,specifier(man:12,a)),
1 in_context(t,specifier(park:21,a)),
1 in_context(t,specifier(telescope:33,a)),
1 in_context(t,proper_name(John:1,person,John)),
1 in_context(t,role(Experiencer,see:42,John:1)),
1 in_context(t,role(Stimulus,see:42,man:12)),
A1 in_context(t,role(prep(in),man:12,park:21)),
A2 in_context(t,role(prep(in),see:42,park:21)),
B3 in_context(t,role(prep(with),man:12,telescope:33)),
or(B2,C2) in_context(t,role(prep(with),park:21,telescope:33)),
or(B1,C1) in_context(t,role(prep(with),see:42,telescope:33)),
1 word(John:1,John,noun,0,1,t,[[9487097]])),
1 word(man:12,man,noun,0,12,t,[[10133569], [10423788], [10135377], [2449786], [10135514], [10135101]])),
1 word(park:21,park,noun,0,21,t,[[8494974], [8495199], [2756453], [11059588], [8495445], [3847283]])),
1 word(see:42,see,verb,0,42,t,[[2109658], [583923], [2109242], [1620934], [682517], [591374], [2131231], [911004]])),
1 word(telescope:33,telescope,noun,0,33,t,[[4351615]]))

Here there are interactions between the attachments: if the location of the man is the park (*A1*), then *with a telescope* can modify either the seeing (*B1*), the park (*B2*), or the man (*B3*). But if the location of the seeing event is the park (*A2*), then *with a telescope* can only modify either the seeing (*C1*) or the park (*C2*). This is reflected in the choice structure, which says that *A1* and *A2* are a disjoint partition of *1*, and that *A1* is in turn partitioned into *B1*, *B2*, and *B3*, while *A2* is partitioned into *C1* and *C2*.

Note that the five possible readings for (7) are represented in not much more space than the two readings for (5). It is possible to count the number of readings by looking only at the choice space: *A1* has three alternatives sitting under it, *A2* has two, and *A1* and *A2* are disjoint, so there are $3+2=5$ alternatives altogether.

3 Flattening of Context-relative Predications

The semantics rules use input limited recursion to capture, and flatten, the structural embeddings in f-structure as context-relative predications. Flattening replaces embedded expressions with complex internal structure, such as clausal complements, with atomic first order terms, which are called contexts. The information about the level of embedding of an expression is preserved by associating its content with the corresponding context. Negation and intensional operators also trigger the introduction of new contexts. Contexts thus serve as scope markers since their use enables globally represented information, such as the scope of operators, to be made locally accessible.[1]

3.1 Flattening of Verbal F-structures

We will illustrate the use of recursive rules to flatten out the contextual structure implicit in f-structure. F-structures are recursive, with one node being embedded inside another, yet it is straightforward to represent this as a flat set of clauses. For (9) these might be along the (abbreviated) lines of (10).

(9) Mary knew that Ed ate turnips.

(10)
PRED(var(0), know),	
SUBJ(var(0), var(1)),	PRED(var(1), Mary)
COMP(var(0), var(3)),	PRED(var(3), eat)
SUBJ(var(3), var(2)),	PRED(var(2), Ed)
OBJ(var(3), var(4)),	PRED(var(4), turnip)
TNS-ASP(var(0), var(5)),	TENSE(var(5), past)
TNS-ASP(var(3), var(6)),	TENSE(var(6), past)

The f-structure node var(3) is embedded under var(0), and var(2) is in turn embedded under var(3). But not all of the f-structure embeddings lead to context embeddings in the semantics: in fact, it is only the nodes var(0) and var(3) that introduce semantic contexts.

The f-structure to semantics rewrite rules therefore need to make a recursive traversal of the f-structure, linking each f-structure node to the nearest dominating node that introduces a semantic context. This is achieved in three stages. First, nodes introducing a context are identified and labeled (where *c(…)* is wrapped around a node to indicate its context):

(11)
```
+COMP(%N1, %N2)
==>
new_context(%N2, c(%N2)), in_context(%N2, c(%N2)).
```

Second, immediate links between f-structure nodes are created, as in (12), so that for each sub-f-structure there is a *link* fact. This fact will be used by the final flattening rule in (13) to create the *in_context* labels.

[1] If you view the traditional semantics for sentences in terms of possible words, a context intuitively delimits a sensible chunk of a possible world, which is used to show how the bigger semantic picture is composed out of its parts.

(12) +SUBJ(%N1, %N2) ==> link(%N1, %N2).
+OBJ(%N1, %N2) ==> link(%N1, %N2).
+COMP(%N1, %N2) ==> link(%N1, %N2).
+TNS-ASP(%N1, %N2) ==> link(%N1, %N2).

Finally, a recursive rule traverses the links propagating the *in_context* labels.

(13) +in_context(%N1, %C), link(%N1, %N2), –new_context(%N2,%%)
*=>
in_context(%N2, %C).

Each recursive step will consume one of the *link(…,…)* facts, ensuring that the recursion terminates. The recursion will simultaneously start at all the nodes initially labeled as being *in_context* by rule (11), and the negative test on *new_context* ensures that nodes are only connected back to their immediately dominating context. It is important to remember that this rule only applies recursively to its own output. This is unlike more general recursion in a set of unordered rules, where rules can recursively apply to the output of other rules.

Modals such as *can* and *should*, behave similarly to other context inducing verbs, despite their distinctive syntactic structure. Since modals take XCOMPs in the f-structure, the rules described above apply relatively straightforwardly to them. Later processing, such as mapping to KR, can make further distinctions among the different modals for applications.

3.2 Negation and Other Context-inducing Adverbs

In the f-structure, sentential negation is an adverb in the ADJUNCT set. However, in the semantic structure, negation introduces a context. Thus a sentence like (14a) has a simplified f-structure like (14b) but a simplified semantics like (14c).

(14) a. Jane did not hop.

b.
$$\begin{bmatrix} \text{PRED} & \text{'hop<SUBJ>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Jane'} \end{bmatrix} \\ \text{ADJUNCT} & \left\{ \begin{bmatrix} \text{PRED} & \text{'not'} \\ \text{ADJUNCT-TYPE} & \text{neg} \end{bmatrix} \right\} \end{bmatrix}$$

c. *context_head(t*,not:10)
context_head(ctx(hop:17),hop:17)
in_context(t,role(mod(degree),*ctx(hop:17)*,not:10,normal))
in_context(*ctx(hop:17)*,role(Theme,hop:17,Jane:1))
word(Jane:1,Jane,noun,0,1,*t*,[[9482706]])
word(hop:17,hop,verb,0,17,*ctx(hop:17)*,[[1948772], [2076532], [1823521], [2076385], [2076247], [2076113]])
word(not:10,not,adv,0,10,*t*,[[24548]])

There are other context inducing adverbs: these are ADJUNCTs in f-structure but introduce a context in the semantics. The rules for these are similar to those for sentential negation and are lexicalized to apply only to adverbs of this class. Examples of such adverbs include sentential uses of *necessarily, possibly, probably, maybe,* and *certainly,* as in (15). These can combine with each other and with negation, as in (16), in which case a series of embedded contexts is created by the semantic rewrite rules.

(15) a. Jane probably left.

b. Jane certainly left.

(16) Jane certainly did not leave.

The rules to introduce the contexts work as follows. The rules need to make the clause the adverbs modifiy the first argument to the modifier. If there is a sequence of sentential modifiers, the modified clause is made the first argument of the last modifier in the sequence, which is itself the first argument of the penultimate modifier, and so on. This is done using limited recursion to build up a list of sentential modifiers. First an empty list is created, by (17) for anything that has an appropriate adjunct, is negative, or is imperative or interrogative (imperatives and interrogatives also introduce contexts).

(17)
```
+PRED(%A, %%),
( +ADJUNCT(%A,%%)
| +is_negated(%A,%%)
| +CLAUSE-TYPE(%A,imp)
| +CLAUSE-TYPE(%A,int)
)
==>
sentential_mods([], %A, %A).
```

The empty sentential modifiers list is then filled with the sentential modifiers in order by the rules in (18). (18a) first puts negation on the list. Then the recursive rule (18b) puts the other sentential modifiers on the list, checking for the relative scope of the adjuncts where the scope is provided by the f-structure *scopes* fact.[2]

(18) a.
```
sentential_mods([], %A, %A), is_negated(%A, %NMod)
==>
sentential_mods([%NMod], %A, %A).
```

b.
```
+ADJUNCT(%H,%M), in_set(%N,%M), sentential_mod(%N),
–( in_set(%N1, %M), sentential_mod(%N1), scopes(%N1, %N) ),
sentential_mods(%Mods,%H, %H)
* =>
sentential_mods([%N|%Mods],%H, %H).
```

Finally, the rules do head switching of the modifiee and the last modifier: everything that was expecting the modifiee as an argument now takes the last modifier. When the adverbs modify the main (root) clause, the rules indicate that the node of the last modifier becomes the root node since it is important that all semantic structures, like f-structures, are rooted and connected.

4 Cannonicalization of F-structures

A number of syntactic constructions are cannonicalized very early in the semantic rules. These are relatively straight-forward rewrites that then feed the more complex semantic rules (section 5).

[2] In the English LFG grammar used here, the *scopes* fact reflects the linear order of the adjuncts; other strategies could be used.

Passive constructions are turned into their corresponding actives.[3] For passives with overt agent *by* phrases, as in (19a), this results in a structure similar to that of their active counterpart. For passives without overt agents, as in (19b), a special agent pronoun is put into the subject role. The ordered rules to achieve this are shown in (20).

(19) a. The cake was eaten by John.

b. The cake was eaten.

(20) a. +VTYPE(%V, %%), +PASSIVE(%V,+), SUBJ(%V, %LogicalObj)
==>
OBJ(%V, %LogicalObj).

b. +VTYPE(%V, %%), +PASSIVE(%V,+),
OBL-AG(%V, %LogicalSubj), PFORM(%LogicalSubj,%%)
==>
SUBJ(%V, %LogicalSubj).

c. +VTYPE(%V, %%), +PASSIVE(%V,+), –SUBJ(%V,%%)
==>
SUBJ(%V,%AgtPro), PRED(%AgtPro,agent_pro), PRON-TYPE(%AgtPro, null).

(20a) takes the subject of a passive verb and makes it the object. Then (20b) takes the OBL-AG of a passive verb and makes it the subject. Finally, (20c) creates a dummy subject for any passive verb that does not have one provided by (20b).

A number of constructions are assigned null pronominal subjects in the f-structure. In many cases, the semantics substitutes in the most likely subject instead. Some example constructions are shown in (21) with the rule for (21a) shown in (22).

(21) a. Before leaving, I fixed it. (=I leaving)

b. To open it, John broke the seal. (=John to open it)

c. Having arrived early, Mary sat down. (=Mary arrived early)

d. Broken by the wind, the gate fell. (=the gate broken by the wind)

(22) +SUBJ(%Main,%MainSubj), +ADJUNCT(%Main,%Adj), +OBJ(%Adj,%AdjObj),
SUBJ(%AdjObj,%AdjObjSubj), arg(%AdjObj,1,%AdjObjSubj),
PRON-TYPE(%AdjObjSubj,null), PRED(%AdjObjSubj,%%)
==>
SUBJ(%AdjObj,%MainSubj), arg(%AdjObj,1,%MainSubj).

(22) looks for an f-structure %Main which has a subject and an adjunct. That adjunct must take an object with a null subject (in examples like (21a) *leaving* is the object of *before*). This subject is replaced by the subject of the main clause. Note that the + in front of the first fact +SUBJ(%Main,%MainSubj) ensures that the main clause subject is not deleted.

Similar cannonicalization occurs for comparatives, measure phrases, and related scalars. These have a number of different overt expressions in the syntax which are all regularized to allow the semantics to operate on them more directly.

[3]This rule can only apply after rules, such as anaphora resolution, which need to make reference to syntactic subject have applied.

(23) a. John is happier.

b. John is happier than Mary.

c. John is much happier than Mary.

(24) in_context(ctx(happy:14),comparative_diff(happy:14,John:1,Mary:23,pos,much:10))

(24) shows the cannonicalized semantic form for the comparative adjective *happier* in (24c). *Mary:23* is the comparison class, the *pos* indicates that it is more happy as opposed to less happy, and *much:10* indicates the amount of difference. The comparison class and the amount of difference can be *unspecified*, e.g., for sentences like (24a).

5 Semantic Rewrites

There are a set of rules in the semantic rewrite rules which correspond to more traditional, theoretical semantic rules. These include treating coordination by semantic instead of syntactic type, creating structures for deverbal nouns, and providing appropriate scope and cannonicalizations for quantifiers. In this section we discuss how these are done in the rewrite rules.

5.1 Quantifiers

Unlike in the Glue semantic [Dalrymple(2003)] approach to manipulating f-structures to create semantic structures, using the semantic rewrite rules does not provide a theoretically motivated treatment of quantifier scope possibilities. Instead, the rules determine one scope.[4] For example, the scope of indefinite subjects is raised relative to that of negation. Relatedly, the modals *must*, *ought*, and *should* are rescoped relative to sentential negation, while other modals are not.

One interesting rule for quantifiers derives negative sentential modifiers from downward monotone nominal arguments. The basic idea is to mark anything with a downward monotone nominal argument as negated, and then introduce a new *not* context. For example, the sentence *No girl hopped.* has a semantics as in (25) in which there are two contexts, one introduced by *no* (ctx(hop:15)), and in which there is a word fact similar to that for sentential negation.

(25) context_head(t,not:n(147)
context_head(ctx(hop:15),hop:15)
in_context(t,role(mod(degree),ctx(hop:15),not:147,normal))
in_context(ctx(hop:15),past(hop:15))
in_context(ctx(hop:15),cardinality(girl:4,sg))
in_context(ctx(hop:15),proportion(girl:4,no))
in_context(ctx(hop:15),role(Theme,hop:15,girl:4))
word(girl:4,girl,noun,0,4,ctx(hop:15),[[9979060], [9934281], [9844392], [9979885], [9979646]])
word(hop:15,hop,verb,0,15,ctx(hop:15),[[1948772], [2076532], [1823521], [2076385], [2076247], [2076113]])
word(not:147,not,adv,0,147,t,[[24548]])

[4] Multiple scopes can be produced in certain situations by using optional rules. However, this has not proven a useful or efficient strategy in using the rewrite rules to produce semantic representations.

The semantic rewrite rule to handle these cases first looks for the appropriate quantified arguments of a verb, introducing a fact that the verb is negated. Then a second rule creates the negative adjunct for the verb which then triggers the same rule that introduces the context for sentential negation (section 3.2).

5.2 Coordination

In the f-structure, coordination is represented as a set with a feature indicating what level in the c-structure the coordination occured at (e.g., N, NP). Coordination of nominals indicates the resolved number and person of the set but otherwise is identical to coordination of verbs and sentences. The f-structure for *Mary and Jane hopped.* is shown in (26).

(26)
$$\begin{bmatrix} \text{PRED} & \text{'hop<SUBJ>'} \\ \text{SUBJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'Mary'} \\ \text{NUM} & \text{sg} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{'Jane'} \\ \text{NUM} & \text{sg} \end{bmatrix}, \begin{array}{ll} \text{NUM} & \text{pl} \\ \text{COORD-FORM} & \text{and} \\ \text{COORD-LEVEL} & \text{NP} \end{array} \right\} \end{bmatrix}$$

In contrast, the semantics differentiates between nominal, verbal, and adjunct coordinations. The rules first determine which type of coordination is present, typing them as nominal, sentential, verbal, predicative, number, or adjunct. A typed PRED is then created for the coordinate structure (note in (26) that the SUBJ f-structure has not PRED of its own). It is this new PRED that will then act as an argument, with its elements listed as additional semantics facts, as in (27).

(27) context_head(t,hop:18)
in_context(t,cardinality(Jane:1,sg))
in_context(t,cardinality(Mary:10,sg))
in_context(t,cardinality(group_object:2,pl))
in_context(t,*is_element(Jane:1,group_object:2)*)
in_context(t,*is_element(Mary:10,group_object:2)*)
in_context(t,role(Theme,hop:18,*group_object:2*))
word(Jane:1,Jane,noun,0,1,t,[[9482706]])
word(Mary:10,Mary,noun,0,10,t,[[9482706]])
word(*group_object:2*,group_object,implicit,0,2,t,[[1740]])
word(hop:18,hop,verb,0,18,t,[[1948772], [2076532], [1823521], [2076385], [2076247], [2076113]])

The syntax treats all coordinators similarly. However, the semantics differentiates between coordinators like *and* which do not introduce new contexts, as seen in (27), and ones like *or* which do. Compare the analysis of *Jane or Mary hopped.* in (28) to that in (27) for *Jane and Mary hopped.*

(28) context_head(t,hop:20)
context_head(ctx(Jane:1),Jane:1)
context_head(ctx(Mary:9),Mary:9)

in_context(t,*coord_or([ctx(Jane:1),ctx(Mary:9)])*)
in_context(t,cardinality(group_object:2,sg))
in_context(t,role(Theme,hop:20,group_object:2))
in_context(ctx(Jane:1),cardinality(Jane:1,sg))
in_context(ctx(Jane:1),*is_element(Jane:1,group_object:2)*)
in_context(ctx(Mary:9),cardinality(Mary:9,sg))
in_context(ctx(Mary:9),*is_element(Mary:9,group_object:2)*)
word(Jane:1,Jane,noun,0,1,ctx(Jane:1),[[9482706]])
word(Mary:9,Mary,noun,0,9,ctx(Mary:9),[[9482706]])
word(group_object:2,group_object,implicit,0,2,t,[[1740]])
word(hop:20,hop,verb,0,20,t,[[1948772], [2076532], [1823521], [2076385], [2076247], [2076113]])

There are two contexts related to the top context *t* by the COORD_OR fact. The *group_object* is still the Theme of *hop* but which element (*Jane* or *Mary*) is in that set depends on the context.

5.3 Deverbal Nouns

Deverbal nouns, or nominalizations, can pose serious challenges for knowledge-based systems. Sentences (29) and (30) describe the same event of destruction, which has the same two participants in both cases. However, the event is expressed by a verb in the first case and a noun in the second case.

(29) Alexander destroyed the city in 332 BC.

(30) Alexander's destruction of the city happened in 332 BC.

The f-structure for these differ significantly. However, these are cannonicalized in the semantics so that both resemble events with roles from VerbNet (section 6.2) and verbal concepts from WordNet (section 6.1). That is, the goal of the rules is to take the nominal and map it to its verbal counterpart. This is in many ways a simpler task than taking a semantic representation of an event and determining how it can be syntactically realized as a nominal.

An external database of deverbal nouns is created indicating the noun, the corresponding verb, the type of deverbal (in the current set, *-ee* and *-er* deverbals are differentiated from all others; more types would be possible), and how lexicalized it is.[5] Note that gerunds are treated productively by the syntax and so do not need entries in the database.

(31) deverbal(destruction, destroy, null, only).
deverbal(parolee, parole, ee, only).
deverbal(teacher, teach, er, both).

The semantic rewrite rules first identify deverbal nouns and indicate whether they have a COMP or XCOMP argument. Then a set of rules determine for each type of deverbal (e.g., *null*, *er*) how the different specifiers and adjuncts map to the argument of the verb. In (30) the POSS maps to the subject and the *of* adjunct to the object. There are several of these rules to account for the different types of deverbals with different combinations of arguments. In certain cases, this can result in ambiguities (e.g., *the Romans' destruction*); the semantics rules will split the choice space, allowing for both analyses. Consider the rule for deverbals like *a teacher of poetry* shown in (32).

[5] Extremely lexicalized deverbals like *building* are not mapped on to events.

(32) may_be_deverbal(%N, %V, er, %HasComp),
@isPrepAdjunct(%N,of,%Obj),
@hasTransMapping(%V, %HasComp, %SubCat)
==>
is_deverbal(%N, %V, %SubCat, needs_arg, %Obj), is_er_deverbal(%N).

(33) states that a noun %N, which has been identified as an *er* deverbal, has a prepositional *of* adjunct and that the corresponding verb form %V has a transitive mapping. This creates a new fact *is_deverbal* with the transitive frame, no subject (*needs_arg*), and the *of* phrase as the object. A later rule then creates a subject for the verb from the *er* deverbal itself. That is, the subject of the *teach* event is the *teacher* while the object is the *poetry*, as in (33).

(33) in_context(ctx(teach:3),role(Agent,teach:3,teacher:3))
in_context(ctx(teach:3),role(Recipient,teach:3,implicit_arg:2))
in_context(ctx(teach:3),role(Topic,teach:3,poetry:14))
lex_class(teach:3,[vnclass(transfer_mesg-37_1-1-1)])
word(implicit_arg:2,implicit,implicit,0,0,ctx(teach:3),[[1740]])
word(poetry:14,poetry,noun,0,14,ctx(teach:3),[[6995243], [6995943]])
word(teach:3,teach,verb,0,3,ctx(teach:3),[[820277], [270355]])
word(teacher:3,teacher,noun,0,3,t,[[10533902], [5781275]])

For more details of this approach to deverbal nouns, see [Gurevich et al.(2006)].

5.4 Other Rules

There are rules which are not strictly speaking semantic from a theoretical perspective but which are useful for applications and for deeper processing such as mapping to knowledge representation ([Crouch(2005)]). An example of these is the alias fact. Proper nouns receive an alias fact which ties the skolem of that noun to the surface form. For example, the proper noun *John* would receive a fact as in (34).

(34) alias(John:67,[John])

This records the string that corresponds to the proper noun, thereby providing more identification information. This is necessary because the concept for all proper nouns referring to male people is the same (and similarly for companies, locations, and other proper noun classes).

In addition, the alias fact for multiword proper nouns contains variants of the noun that are useful for applications. For example, a name like *Mr. John Smith* would have an alias fact like that in (35).

(35) alias(John:67,[John, Smith, John Smith, Mr. John Smith])

This allows the proper noun to be more easily matched with occurrences in other sentences and texts.

6 Incorporation of Lexical-semantic Resources

This section discusses how further lexical-semantic resources (in particular, WordNet and VerbNet) can be used in the semantic rules. Incorporation of external lexical resources saves time in lexical development, but it comes at a cost both for finding ways in which to integrate the resources and in dealing with errors in those resources.

6.1 WordNet Concepts

WordNet is used to assign concepts to words. WordNet [Fellbaum(1998)] contains words with their part of speech organized into synonym sets (synsets) which represent underlying lexical concepts; these synonym sets are linked by relations such as hypernyms (e.g., *auction* is a type of *sell* which is a type of *exchange, change, interchange* which is a type of *transfer*). The semantic rewrite rules assign concepts to words by looking up the word with its associated part of speech. Some words will belong to just one synset while others belong to many. To accommodate words with more than one synset, the synset concepts are stored as a list. The concept becomes part of the word fact associated with a skolem. An example is shown in (36).

(36) word(hop:17,hop,verb,0,17,t,[[1948772], [2076532], [1823521], [2076385], [2076247], [2076113]])

The word facts contain the string, part of speech, and context as well as the skolem and synsets. This information is all stored for use in further processing, such as in the mapping to KR.

In theory, the contents of WordNet could be dumped into a lexicon of non-resourced facts or an external database that the semantic rewrite rules could refer to. However, instead the rules call the WordNet interface directly via a procedural attachment (indicated in the rewrite rules by { }). This calling of an external resource directly contrasts with the way in which VerbNet is incorporated into the rules, as described below.

In addition to the general lookup of word concepts in WordNet, certain classes of words are assigned WordNet synsets in a more constrained fashion. These lookups are done before the more general lookups since they are the more specific case and take precedence over the default case. For example, proper nouns are not looked up based on their string form[6] but instead by their proper noun type as assigned by the morphological analyzer that is used by the syntax. Proper nouns can be classed as locations, organizations, companies, male persons, female persons, etc. These are assigned a synset appropriate for this class by rules such as (37b) in conjunction with the non-resourced fact in (37a).

(37) a. |– name_synset(company, –, –, company, 7948427).

b. +in_context(%%,proper_name(%NameSk,%Type,%%)),
node_label(%V, %NameSk), NTYPE(%V,%%),
name_synset(%Type, %G, %H, %WNWord, %S),
{wn_all_hypers(%WNWord,noun,%S,%HL,%SS)},
@get_word_context(%V,%Ctx),
{%NameSk = %Pro:n(%SN,%N)}
==>
synsets(%NameSk, %SS), word(%NameSk, %Pro, noun, %N, %SN, %Ctx,%HL).

Pronouns work similarly in that they are assigned synsets based on an appropriately predefined set of features. Some of the nonresourced facts used in the pronoun mapping are shown in (38).

(38) |– pronoun_synset(he, male‘ person, 9487097).
|– pronoun_synset(she, female‘ person, 9482706).
|– pronoun_synset(we, person, 7626).
|– pronoun_synset(you, person, 7626).
|– pronoun_synset(it, entity, 1740).

[6]WordNet has entries for many proper nouns. However, these entries are spotty and can result in rather unexpected behavior in applications.

6.2 VerbNet Roles

VerbNet is used to assign roles to the arguments of verbs. VerbNet [Kipper et al.(2000)] classifies verbs according to Levin verb classes [Levin(1993)]. It includes syntactic subcategorization information, information about thematic roles (e.g., agent, patient), and basic lexical semantics. In order to use the VerbNet material, we extracted it into a Unified Lexicon (UL) ([Crouch and King(2005)]) which contained information about the XLE syntactic subcat frames, WordNet synsets, and VerbNet, as well as some lexical class information used in the later semantics to KR mapping rules.

As discussed in [Crouch and King(2005)], there were some problems in converting VerbNet into a format that could be used by the semantics rules. The first was converting VerbNet subcategorization frames into ones that were compatible with the XLE syntactic lexicon. This was difficult because the VerbNet subcategorization information is listed not as grammatical function information but as abstractions over a cannonical phrase structure tree. This extraction becomes extremely involved for verbs which take NP small clauses, particles, expletives, or verbal complements. The second issue in the VerbNet extraction was ensuring that a verb belonging to a particular VerbNet class inherited all the correct role restrictions from the classes above it. The final issue with VerbNet was that many verb frames have implicit roles. These roles are determined by looking at the semantics provided for the verb. If there is a thematic role mentioned that is preceded by a ? (question mark), e.g., ?Topic, this indicates that it is implicitly present in the verb frame and may have role restrictions on it. For example, the transcribe-25.4 class for *The secretary transcribed the speech* has an implicit Destination role which is restricted to being concrete. Note that this role is overt in other frames for this verb, as in *The secretary transcribed the speech into the record*.

In the semantics rules, the VerbNet role mapping works by looking up the head word and subcat frame in the UL to see whether there are VerbNet roles associated with the arguments. In the building of the UL, words with subcat frames which did not have a direct listing in VerbNet are sometimes assigned a VerbNet mapping from a similar frame (e.g., a version with an oblique prepositional phrase may be able to use the mapping for the base transitive, with a guessed role for the oblique). If there are roles in the UL, then the grammatical functions are converted to the relevant VerbNet roles. Thus, the f-structure SUBJ and OBJ for a sentence like *The girl ate the cake.* are mapped into the roles in (39) by a simplified rule like (40) using the UL entry in (41). The UL itself is stored as a database that the rules access ([Crouch et al.(2006)]).

(39)
```
in_context(t,role(Agent,eat:22,girl:5))
in_context(t,role(Patient,eat:22,cake:18))
```

(40)
```
+word(%VSk,%Verb,verb,%%,%%,%%,%%),
verb_mapping(%Origin, %Verb, %SubCat, %Source, %VerbClass, %WN, %VSk,%Ctx,
      %GF1, %Restr1, %Sk1,
      %GF2, %Restr2, %Sk2,
      %Mapping),
@get_subcat(%VSk, %SubCat),
@get_gf(%VSk, %Sk1, %GF1),
@check_selectional_restriction(%Sk1, %Restr1, %%C1),
@get_gf(%VSk, %Sk2, %GF2),
@check_selectional_restriction(%Sk2, %Restr2, %%C2)
==>
%Mapping,
lookup_subcat(%VSk, %SubCat, [%Origin, %Source]),
wordnet_classes(%VSk, %WN),
```

source_of_concept_mapping(%VSk, %Source),
lex_class(%VSk, %VerbClass),
sortal_restriction(%Sk1, %Restr1),
sortal_restriction(%Sk2, %Restr2).

(41) verb_map(eat, V-SUBJ-OBJ, verbnet, [vnclass(eat-39_1-1)],
[wn(1155228,verb(consumption)), wn(1157345,verb(consumption)),
wn(1168626,verb(consumption))], %Ev, [],
subj, [15024,4359], %subj,
obj, [1740], %obj,
implicit_args(%Ev, []),
concept_for(%Ev, eat),
source_of_concept(%Ev, guessed_verb),
verbnet_role(Agent, %Ev, %subj),
verbnet_role(Patient, %Ev, %obj),
vn_sem(take_in(during(%Ev), %subj, %obj))).

First consider the UL entry in (41). It corresponds to the word *eat* in its transitive V-SUBJ-OBJ use and has a VerbNet class with WordNet correspondences. The subj and obj arguments have selectional restrictions of WordNet synsets [15024, 4359] and [1740]. Then the VerbNet mapping indicates that there are no implicit arguments and that the concept of the event is the string *eat* (a semantics rewrite rule will assign the appropriate WordNet synset (section 6.1)). The *verbnet_role* facts assign the Agent role to the subject and the Patient role to the object.[7] Finally, the VerbNet semantics is recorded; this is not currently used by the semantic rules.

Even with the UL tuned for the XLE semantic rules, there are some discrepencies which must be taken into account. For example, many VerbNet mappings refer to OBL(ique) arguments. Due to the f-structure assignments, these may be either a syntactic OBL or in the ADJUNCT set as prepositional phrases. As such, the rewrite rules look first for an OBL and if none if found, then for a PP in the ADJUNCT set.

The UL entry and the semantic rewrite rule that does the mapping mention sortal restrictions(*eat* subj = [15024, 4359]; obj = [1740]). VerbNet, and other resources, provide sortal restrictions on the arguments of verbs. However, applying sortal restrictions obligatorily may be very dangerous because if the argument does not match the sortal restriction, then the only mapping of the verb can be eliminated. Mismatches on sortal restrictions can come from a variety of sources. Sometimes the concept for the restriction or for the argument is incorrect. More often, the mismatch reflects a type of coercion; for example, organizations are often treated as volitional and/or animate. As such, sortal restrictions should be implemented with an optimality mechanism similar to that used in XLE parsing ([Frank et al.(2001)]). Currently, such a mechanism is not in place in the rewrite system used by the semantic rules and so although the sortal restrictions are recorded, they are not enforced in the rules.

[7]The thematic roles in VerbNet and the semantics discussed here have the same problems that have been noticed in the theoretical linguistics literature. As such, we are working on determining which combinations of roles are most effective for applications based on the semantics.

7 Discussion, and Conclusions

7.1 Comparison to Glue

The rewrite-based system is used in place of components constructing semantic representations by more standard means, such as Glue Semantics [Dalrymple(2003)]. A comparison of Glue and rewrite semantic construction reveals a number of theoretical and practical pros and cons.

In practical terms, the transfer/rewriting system provides a relatively straightforward tool for efficiently manipulating f-structures, which should be accessible to grammarians with little or no background in formal theories of semantic construction like Glue- or Montague-semantics. This tool comes with the XLE.

Theoretically, the rewrite system imposes few interesting constraints on what kinds of representation can be constructed. Unlike Glue, for example, there is no elegant account of scope ambiguity derived from the theory of semantic construction. The rewrite approach allows semantic construction to be sensitive to features of the meaning language in a way that Glue expressly forbids. This can be convenient when quickly developing a broad-coverage system, but it provides few constraints and guides to the rule writer, allowing for theoretically unsound analyses and implementations.

The feeding and bleeding nature of the rewrite rules means that, unlike unification- or Glue-based semantics, the semantic construction rules cannot easily be run in reverse for generation purposes. Given that the XLE LFG grammars can be run in both the parsing and generation direction, having an accompanying semantics for both parsing and generation is desirable and opens a broader range of applications. Without inherent reversibility, the semantic rewrite rules have to be written in two sets and maintenance becomes a serious issue since a change in the f-structure to semantics rules can necessitate a corresponding change in the semantics to f-structure rules.

While it would be a stretch to call the rewrite approach a *theory* of the syntax-semantics interface or semantic construction, it does provide a powerful and efficient tool for the task. But the way in which they are produced has no bearing on the theoretical validity of the semantic representations themsleves: we would claim that the representations described are thoroughly defensible from the point of view of formal semantics.

7.2 Cross-linguistic Application

The technique of mapping from f-structures to semantics described in this paper can be applied cross-linguistically. Given the degree of abstraction and generality already present in f-structures for different languages [Butt et al.(1999)], one can port semantic rules from one language to another. As an experiment, an initial port of the rules to Japanese, using the Japanese ParGram grammar as a basis ([Masuichi and Ohkuma(2003)]), was successfully conducted ([Umemoto(2006)]).

Some of the rules described above, such as the context-inducing adverbs, are lexicalized. These need to be ported to apply to the corresponding lexical items in other languages. In addition, some constructions that are present in English and that need to be manipulated by the semantic rules may not be present in other languages. Leaving rules for these in the semantics will not hurt in that they will never trigger, but for clarity they should be removed. Similarly, other languages may have constructions which are not covered by the rules described here because English does not have them. We anticipate that it should be possible to use the rewrite system to process such constructions efficiently.

A more serious problem in the cross-linguistic application of this approach is in the mapping of concepts and roles which depended on WordNet and VerbNet respectively. There are WordNets, and ontologies, for a variety of languages and so for these languages it is possible to incorporate

these resources as the English semantics uses WordNet. Broad-coverage role mapping resources like VerbNet are much rarer for other languages and so if this type of role is needed for the semantics, then the lexical resources may need to be boot-strapped in some way. Fortunately, however, the semantic rules can run independently of concepts and roles, applying only the rules which convert f-structures into semantic structures, such as the context-introducing and flattening rules (section 3). For many purposes, these may be sufficient even without concepts and roles.

7.3 Summary

This paper discussed how the XLE general purpose ordered rewrite rule system is used to produce semantic representations from syntactic f-structures. The rules apply efficiently because they operate on the packed input of f-structures to produce packed semantic structures. In addition to rules which convert the syntactic structure to a semantic one, there are rules that use external resources to replace words with concepts and grammatical functions with roles. Although the system described here could by no means be described as a theory of semantic construction or the syntax-semantic interface, from a practical stand point it can efficiently and robustly produce theoretically defensible semantic structures from broad-coverage syntactic ones.

Appendix: Complex Example

Sentence: Mary does not believe that a small boy is five inches taller than John.

F-structure:

```
"Mary does not believe that a small boy is five inches taller than John."
[PRED    'believe<[1:Mary], [156:be]>'
 SUBJ    [PRED  'Mary'
          NTYPE [NSYN proper]]
       1 [HUMAN +, NUM sg, PERS 3]
 COMP    [PRED          'be<[446:tall]>[209:boy]'
          SUBJ          [PRED    'boy'
                         ADJUNCT {311 [PRED 'small']}
                         NTYPE   [NSEM [COMMON count]
                                  NSYN common]
                         SPEC    [DET [PRED 'a']]
                     209 [NUM sg, PERS 3]]
          XCOMP-PRED    [PRED    'tall<[209:boy]>'
                         SUBJ    [209:boy]
                         ADJUNCT {[PRED          'more<[468:than]>'
                                   OBL-COMPAR    [PRED   'than<[479:John]>'
                                                  OBJ    [PRED  'John'
                                                          NTYPE [NSYN proper]]
                                                      479 [HUMAN +, NUM sg, PERS 3]]
                                   ADJUNCT       {[PRED  'inch'
                                                   NTYPE [NSEM [COMMON measure]
                                                          NSYN common]
                                                   SPEC  [NUMBER 385 [PRED 'five']]
                                                  409 [NUM pl, PERS 3]]}
                                  -3 [ADJUNCT-TYPE degree, NUM pl]}
                     446 [TNS-ASP       [PERF -, PROG -, TENSE pres]
                          156 [CLAUSE-TYPE decl, PASSIVE -]]
 ADJUNCT {[PRED          'not'
          122 [ADJUNCT-TYPE neg]]}
 TNS-ASP [PERF -, PROG -, TENSE pres]
      30 [CLAUSE-TYPE decl, PASSIVE -]]
```

Semantics:

context_head(t,not:11)
context_head(ctx(be:40),be:40)
context_head(ctx(believe:71),believe:71)
context_head(ctx(tall:55),tall:55)

in_context(t,cardinality(Mary:1,sg))
in_context(t,role(mod(degree),ctx(believe:71),not:11,normal))
in_context(ctx(be:40),pres(be:40))
in_context(ctx(be:40),role(copula,be:40,ctx(tall:55)))
in_context(ctx(believe:71),pres(believe:71))
in_context(ctx(believe:71),proper_name(Mary:1,person,Mary))
in_context(ctx(believe:71),role(Agent,believe:71,Mary:1))
in_context(ctx(believe:71),role(Theme,believe:71,ctx(be:40)))
in_context(ctx(tall:55),cardinality(John:67,sg))
in_context(ctx(tall:55),cardinality(boy:36,sg))
in_context(ctx(tall:55),specifier(boy:36,a))
in_context(ctx(tall:55),measure(inch:48,inch:48,five))
in_context(ctx(tall:55),proper_name(John:67,person,John))
in_context(ctx(tall:55),role(mod(degree),boy:36,small:30,normal))
in_context(ctx(tall:55),comparative_diff(tall:55,boy:36,John:67,pos,inch:48))

lex_class(believe:71,[vnclass(consider-29_9-2),prop-attitude])
lex_class(not:11,[sadv,impl_pn_np])
sortal_restriction(ctx(be:40),[1740])
sortal_restriction(Mary:1,[7899136,15024])
alias(John:67,[John])
alias(Mary:1,[Mary])
word(John:67,John,noun,0,67,ctx(tall:55),[[9487097]])
word(Mary:1,Mary,noun,0,1,t,[[9482706]])
word(be:40,be,verb,0,40,ctx(be:40),[[2579744], [2591280], [2629830], [2578719], [2725216], [2639228], [2595485], [2422266]])
word(believe:71,believe,verb,0,71,ctx(believe:71),[[675183], [681247], [712804], [676176], [675971]])
word(boy:36,boy,noun,0,36,ctx(tall:55),[[10131706], [9725282], [10464570], [9500236]])
word(inch:48,inch,noun,0,48,ctx(tall:55),[[13469892], [13533066]])
word(not:11,not,adv,0,11,t,[[24548]])
word(small:30,small,adj,0,30,ctx(tall:55),[[1443454], [1467170], [2419704], [1708858], [1588010]])
word(tall:55,tall,adj,0,55,ctx(tall:55),[[2466583], [2088817], [786375], [678281]])

References

[Butt et al.(1999)] Butt, Miriam, Stefani Dipper, Anette Frank, and Tracy Holloway King. 1999. Writing large scale parallel grammars for English, French and German. In *Proceedings of LFG99*.

[Crouch(2005)] Crouch, Dick. 2005. Packed rewriting for mapping semantics to KR. In *Proceedings of the Sixth International Workshop on Computational Semantics.*

[Crouch et al.(2006)] Crouch, Dick, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2006. XLE documentation. http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html.

[Crouch and King(2005)] Crouch, Dick and Tracy Holloway King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes.*

[Dalrymple(2003)] Dalrymple, Mary. 2003. *Lexical Functional Grammar*. Academic Press. Syntax and Semantics, vol 34.

[Oepen et al.(2004)] et al., Stefan Oepen. 2004. Som a hoppe etter wirkola? unpublished manuscript.

[Fellbaum(1998)] Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database.* The MIT Press.

[Frank(1999)] Frank, Anette. 1999. From parallel grammar development towards machine translation. In *Proceedings of the MT Summit VII*.

[Frank et al.(2001)] Frank, Anette, Tracy Holloway King, Jonas Kuhn, and John T. Maxwell III. 2001. Optimality theory style constraint ranking in large-scale lfg grammars. In P. Sells, ed., *Formal and Empirical Issues in Optimality Theoretic Syntax*. CSLI Publications.

[Gurevich et al.(2006)] Gurevich, Olga, Richard Crouch, Tracy Holloway King, and Valeria de Paiva. 2006. Deverbal nouns in knowledge representation. In *Proceedings of FLAIRS*.

[Kaplan et al.(2004)] Kaplan, Ron, John T. Maxwell III, Tracy Holloway King, and Richard Crouch. 2004. Integrating finite-state technology with deep LFG grammars. In *Proceedings of the Workshop on Combining Shallow and Deep Processing for NLP (ESSLLI)*.

[Kipper et al.(2000)] Kipper, Karin, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *AAAI-2000 17th National Conference on Artificial Intelligence.*

[Levin(1993)] Levin, Beth. 1993. *English Verb Classes and Alternations.* Chicago University Press.

[Masuichi and Ohkuma(2003)] Masuichi, Hiroshi and Tomoko Ohkuma. 2003. Constructing a practical Japanese parser based on Lexical-Functional Grammar. *Journal of Natural Language Processing* 10:79–109. in Japanese.

[Maxwell and Kaplan(1991)] Maxwell, John and Ron Kaplan. 1991. A method for disjunctive constraint satisfaction. *Current Issues in Parsing Technologies* .

[Maxwell and Kaplan(1996)] Maxwell, John and Ron Kaplan. 1996. An efficient parser for LFG. In M. Butt and T. H. King, eds., *Proceedings of the First LFG Conference*. CSLI On-line Publications.

[Riezler et al.(2002)] Riezler, Stefan, Tracy Holloway King, Ron Kaplan, Dick Crouch, John Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

[Umemoto(2006)] Umemoto, Hiroshi. 2006. Implementing a Japanese semantic parser based on Glue approach. In *Proceedings of The 20th Pacific Asia Conference on Language, Information and Computation.*

APPLYING AN LFG PARSER IN COREFERENCE RESOLUTION: EXPERIMENTS AND ANALYSIS

Pascal Denis
Department of Linguistics
University of Texas at Austin

Jonas Kuhn
Computerlinguistik
Universität des Saarlandes

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu

Abstract

In this paper, we explore how LFG analyses as produced by the XLE parser with the English ParGram grammar can be used in a probabilistic coreference resolution system. So far, such systems have mainly relied only on information from surface-based NLP tools, reaching reasonable levels of performance while requiring only small amounts of training data. We compare these surface-based approaches with a first attempt at an LFG-based coreference system and another system using the treebank-trained probabilistic parser by Charniak. Based on the (limited) quantity of training data we used, the performance of all three approaches was quite comparable. However, there are some indications that an XLE-based approach may lead to better results if trained on larger training sets.

1 Introduction

The XLE parser coupled with the LFG grammars from the ParGram project and the log-linear disambiguation models developed at PARC (Riezler *et al.*, 2002) is one of the best available parsing systems – in particular if criteria such as depth of analysis and linguistic motivation are taken into account. One of the hopes with such a carefully engineered parsing system is that it can improve the performance of Natural Language Processing (NLP) systems on tasks that have so far been tackled mainly with linguistically unsophisticated, surface-based approaches. The work in this paper is the beginning of an exploration of the impact of using XLE analyses for machine learning based coreference resolution. The contribution of XLE on this task is compared with that of two shallower NLP tools for grammatical analysis, namely Charniak's parser and a simple part-of-speech tagger.

Coreference resolution (CR) provides an interesting testbed for such a comparative study. On the one hand, deep linguistic representations have been largely unexplored by researchers working in robust CR. Thus, most state-of-the-art machine learning systems (McCarthy and Lehnert, 1995; Morton, 2000; Soon *et al.*, 2001) rely on limited and rather shallow knowledge sources. (Some notable exceptions are (Ng and Cardie, 2002b), and more recently (Uryupina, 2006).[1]) Often, the only type of linguistic processing used is part-of-speech tagging and NP chunking. Even at this shallow level of processing and with limited sets of learning features, these systems have managed to achieve reasonably good performances with F-scores in the 60's%. This situation is somewhat at odds with the work of theoretical linguists who have identified numerous linguistic factors bearing on coreference resolution. It is worth noting in this respect that XLE makes a natural candidate for CR: the type of representations it outputs (basically, LFG f-structures) indeed gives us access to many of these factors. An obvious example are *grammatical functions*: at the center of the LFG architecture, they have been also been argued within Centering Theory (Grosz *et al.*, 1995) to play a decisive role in constraining coreference.

On the other hand, there is no guarantee that appealing to deep linguistic systems (and an extended feature set) for CR is likely to improve compared to a surface-based system. There are two issues here, one is theoretical, the other more practical. On the theoretical side, CR is ultimately an AI-complete problem: in the general case, the task involves solving extralinguistic problems for which even a perfect linguistic oracle would not help; that is, linguistic information gives us only *partial* insight. While it is true for all sub-tasks of interpretation that humans will fall back on world and

[1] (Preiss, 2002) compares Charniak and Collins parsers, but the scope of her study is rather limited, since it only deals with anaphora resolution and is not evaluated on available corpora.

situation knowledge to resolve linguistically underdetermined cases, the situation for CR may be particularly challenging for linguistic approaches since the space of possibilities left open after considering linguistic constraints is still quite considerable.[2] This in turn raises the follow-up question of what sorts of linguistic constraints are actually helpful in modeling the data. On the practical side, it is well-known that for deep linguistic analysis, increased robustness will typically go along with an increased level of noise in the analyses. So, the open question is what level of processing gives us the best results. A related question is whether the combination of information from representations of different depth of analysis will improve things. An interesting aspect of comparing XLE with the Charniak parser is that these two parsers differ not only in terms of the level of sophistication of their outputs (phrase structure trees vs. rich feature structures), but also in terms of their efficiency and robustness. While XLE surely provides more detailed information, this comes at a price: despite some coverage improvements (e.g., in the form of disambiguation and the "back-off" fragment mode, i.e., partial analyses provided for sentences that cannot be parsed completely), XLE is still less robust than Charniak's parser.

In anticipation of the results of the present study, we could not so far observe any significant improvements of overall system performance due to the addition of deeper linguistic information sources. For one thing, this shows that the baseline combination of various surface-oriented information sources established in machine learning-based work on CR already seems to strike a very effective balance of robustness and task-relevant quality, which is not easy to outperform – especially on small training sets. On the other hand, we performed some preliminary meta analyses indicating that larger quantities of training data and a more carefully designed set of learning features may bring out the strengths of deeper information sources.

The rest of this paper is organized as follows. We begin by presenting the task of coreference resolution and the type of machine learning architecture we use to model it. Then, in section 3, we discuss some of the advantages, as well as some of the potential problems, associated with using XLE for CR; there, we also briefly describe how we extracted features out of the XLE output representations. Section 4 presents the experimental set-up. The actual results along with some preliminary analyses of these results are given in section 5.

2 The task of coreference resolution

2.1 Task definition

Coreference Resolution is the automatic detection of text spans in a document that share the same referent in the real world, forming classes of coreferent text spans. Each individual text span is typically known as a *mention*; a class of coreferent mentions is called a *chain*, referring to or describing one *entity*. The present study is concerned with one particular case of coreference, namely *nominal* coreference.[3] As an illustration, the result of applying CR to the following discourse (from the ACE

[2] Note that the type of corpora typically used for training CR systems may make this problem even more acute: the annotations of the MUC and ACE corpora (from the Message Understanding Conferences and the Automatic Content Extraction program, respectively) are often debatable from a linguistic point of view, often stretching the notion of coreference to include phenomena that semanticists would not regard as coreference (e.g., nominal prediction and apposition). (See (van Deemter and Kibble, 2000) for a detailed discussion of the MUC scheme.) Hence in training the systems may have trouble detecting those linguistic generalizations that *do* exist for coreference in the narrower, linguistic sense.

[3] For some recent work on event and abstract entity coreference, see (Byron, 2002).

corpus) in (1a) is given in (1b):

(1) a. [Clinton]$_{m_0}$ told [National Public Radio]$_{m_1}$ that [his]$_{m_2}$ answers to questions about [Lewinsky]$_{m_3}$ were constrained by [Starr]$_{m_4}$'s investigation. [[NPR]$_{m_5}$ reporter Mara Liasson]$_{m_6}$ asked [Clinton]$_{m_7}$ "whether [you]$_{m_8}$ had any conversations with [her]$_{m_9}$ about [her]$_{m_{10}}$ testimony, had any conversations at all."

b. {Clinton$_{m_0}$, his$_{m_2}$, Clinton$_{m_7}$, you$_{m_8}$}$_{e_0}$,
{National Public Radio$_{m_1}$, NPR$_{m_5}$}$_{e_1}$,
{Lewinsky$_{m_3}$, her$_{m_9}$, her$_{m_{10}}$}$_{e_2}$,
{Starr$_{m_4}$}$_{e_3}$,
{NPR reporter Mara Liasson$_{m_6}$}$_{e_4}$

Thus illustrated, the task involves two main steps: (a) the identification of referring mentions,[4] and (b) the partitioning the set of mentions into chains for various entities, i.e. the resolution *per se*. In this paper, we concentrate on the latter task.

2.2 CR as a machine learning problem

Like in other areas of NLP, the last decade of research in coreference resolution has seen an important shift from rule-based systems to systems applying machine learning (ML) techniques (Mitkov, 2002). An important appeal of the latter systems of course lies in their robustness, an important precondition for their integration into larger NLP systems, such as Information Extraction, Question Answering, or Summarization systems.

In a ML setting, the task of coreference resolution is recast as a learning problem, typically a *classification* problem. Specifically, the standard approach for task (b) as addressed in section 2.1 proceeds in two distinct steps (McCarthy and Lehnert, 1995; Morton, 2000; Soon *et al.*, 2001; Ng and Cardie, 2002b,a). For the first step, a *binary* classifier is trained that determines whether or not a *pair* of nominal mentions is coreferential. (If the classifier is probabilistic in nature, it will provide a probability for a pair of mentions being coreferential.) In application, this classifier is applied to (in principle) all pairs of nominal mentions from a document. The task for the second step is to use the pairwise coreferentiality information from step one to construct a consistent partition over the entire set of mentions into chains. Although any clustering algorithm could in principle be used for this, the predominant approach is to make the assumption that a coreferent chain of mentions $m_{i_1}, m_{i_2}, m_{i_3}, \ldots, m_{i_k}$ can be effectively detected by relying only on (the coreferentiality classification of) pairs of textually adjacent mentions from that coreference chain, i.e., $\langle m_{i_1}, m_{i_2} \rangle$ and $\langle m_{i_2}, m_{i_3} \rangle \ldots \langle m_{i_{k-1}}, m_{i_k} \rangle$. (Note that this leaves out the coreferentiality status of $\langle m_{i_1}, m_{i_3} \rangle$, for instance.)[5] In other words, the chain is constructed from a sequence of direct "links" in the text. This means roughly that CR is implicitly reduced to *anaphora resolution* (i.e., the task by which an anaphoric expression is bound to its (unique) antecedent).[6]

[4] Depending on the corpus, these are often restricted to a set of predefined named entities, such as PERSON, LOCATION, ORGANIZATION, etc.

[5] A notable exception is (Kehler, 1997) who uses Dempster's rule of combination to induce a partition from the pairwise classifications.

[6] Note however that in a chain { John, he, his, John, he }, the second mention 'John' is constructed as linked to 'his'.

The most common technique for determining the links for building a chain is for each mention to go backwards in the text, pairing it with preceding mentions, until a pair is hit that is classified as coreferential by step one. (If a probabilistic classifier is used, a probability threshold can be used – e.g., threshold 0.5 to make it equivalent to a non-probabilistic classifier.) This technique is called "Closest-First" selection (e.g., (Soon *et al.*, 2001)). An alternative is to compare the (probabilistic) classifier scores for pairs from a larger text window, picking the highest-scoring pair (above a threshold, typically 0.5) to form the link. This is called "Best-First" selection ((Morton, 2000; Ng and Cardie, 2002b)). Other points of divergence exist between these systems, but they mainly concern the feature set that is used and the sample selection, i.e., the choice of actual training data from the vast number of possibilities arising from arbitrary combination of mentions in the text.[7] Some systems also use separate classifiers for different types of mentions (e.g., pronouns and proper names) (e.g. (Morton, 2000)), instead of using a single classifier.

An issue with these selection strategies as just described is that there is no treatment for mentions in the text that introduce a new referent (i.e., which form the beginning of a new chain). One could use an additional classifier that will say for each mention whether or not it is anaphoric and use the described linking technique only for mentions that are anaphoric (see e.g., (Ng and Cardie, 2002a)). As an alternative solution, (Morton, 2000) changes the original classification task from step one in such a way that non-anaphoric elements are included in the training data as being coreferential with an artificial dummy element. In application, a mention will start a new chain if the dummy element is the most probable antecedent.

2.3 Model used in this study

In this preliminary study, we used the set-up proposed by (Soon *et al.*, 2001), which is arguably one of the simplest architectures: it uses a simple sample selection method and a "Closest-First" clustering. The actual training and test procedures for this system are explained below. The main difference with the original Soon *et al* system is in the type of machine learners we used. While (Soon *et al.*, 2001) use Decision Trees, we use maximum entropy (aka, log-linear) models (Berger *et al.*, 1996). More specifically, our coreference model takes the following form, where the classes YES and NO stand for "corefer" and "don't corefer", respectively; m_i and m_j are two mentions, f_i are the features of the model and λ_i their associated parameters:

$$P(\text{YES}|\langle m_i, m_j\rangle) = \frac{exp(\sum_{i=1}^{n} \lambda_i f_i(\langle m_i, m_j\rangle, \text{YES}))}{\sum_{c\in\{\text{YES,NO}\}} exp(\sum_{i=1}^{n} \lambda_i f_i(\langle m_i, m_j\rangle, c))} \quad (2)$$

Parameters were estimated using the limited memory variable metric (LMVM) algorithm implemented in the Toolkit for Advanced Discriminative Modeling (Malouf, 2002).[8] We regularized our model using a Gaussian prior of variance of 1000 — no attempt was made to optimize the prior for each data set. Maxent models are well-suited for the coreference task, because they are able to handle many different, potentially overlapping learning features without making independence assumptions.

[7] Because coreference is a very "rare" relation, looking at all possible pairs of mentions yield a very skewed class distribution.

[8] Available from `tadm.sf.net`.

Previous work on coreference using maximum entropy includes (Kehler, 1997; Morton, 1999, 2000).[9]

For the LFG audience, it may be interesting to note that there is a close parallelism between the Maxent approach and Optimality Theory (compare also (Johnson, 1998; Goldwater and Johnson, 2003)): one can can think of OT as a restricted class of a binary classifiers, where the learning features are called *OT constraints* and a tableau of n candidates corresponds to n classifier decisions. For the coreference task, the OT input would be a particular mention for which we seek an appropriate "linking point", i.e., preceding mention. Each candidate is a pair of the input mention and a potential antecedent. Now harmony evaluation – based on the constraint violation profile of the candidates and the ranking in the grammar – will determine the harmony for each candidate and output the most harmonic one as the winner, i.e., the predicted link. The main difference between OT and the more general Maxent model used in our work is that OT assumes a *strict ranking* of the constraints: that is, lower-ranked constraints are not allowed to "gang up" to beat an higher-ranked constraint. The weighting of the parameters in the Maxent model (= the "strength" of the violable "constraints") is less restricted so that ganging-up effects can happen.

The training and testing procedures proposed in (Soon *et al.*, 2001) are as follows. For training, the text is scanned from left to right and for each anaphoric mention α: (i) a *positive instance* is created between α and its *closest* antecedent m_i, (ii) *negative instances* are created between α and all the (non-coreferential) mentions m_j intervening between α and m_i.

Once trained, the classifier is used to build coreference chains in the following way. For each mention m_i in the text, the preceding text is scanned from right to left, generating pairs of m_i with each of its preceding mentions m_j. Each such pair is submitted to the classifier, which returns a number between 0 and 1 representing the probability of the two mentions to be coreferential. (Soon *et al.*, 2001) use "Closest-First" clustering, which means that the process terminates as soon as the first coreferring mention (i.e., one with probability > 0.5) is found or the beginning of the text is reached.

2.4 Potential limitations of the classification approach

There are at least two potential limitations to the classification approach, both related to the very strong independence assumptions. First, the classifier considers antecedent candidates independently from each other, since only a *single* candidate pair is evaluated at a time. An alternative allowing different NP candidates to be directly compared is to use a *ranker*; this option is explored for pronoun resolution by (Denis and Baldridge, 2007). A second possible limitation has to do with the clustering used: the "Closest-First" and "Best-First" selection algorithms are extremely greedy. They assume that coreference decisions for chain building are independent from one another (McCallum and Wellner, 2003). To take a simple example, consider the following set of mentions {Mr Clinton, Clinton, he}. Under a pairwise classification scenario, the decision regarding the pair ⟨Clinton, he⟩ is done independently from the decision regarding the pair ⟨Clinton, Mr Clinton⟩, although this earlier decision is likely to provide important information for the second decision (e.g., that Clinton is a male). An attempt to solve this problem is provided by (Morton, 2000) and relies on using a discourse model. But this approach is again likely to be greedy, since mistakes made at the beginning are likely to propagate.

[9] In the context of XLE, Maxent models have been used by (Johnson *et al.*, 1999) and (Riezler *et al.*, 2002) for parse selection.

3 Incorporating XLE information

In this section, we motivate the use of XLE for CR by examining a simple example taken from the ACE corpus (from the Automatic Content Extraction program). We also come back to some potential issues that arise when using a deep parser such as XLE. Finally, we discuss the strategy used to extract features from XLE output representations.

3.1 Motivation

The main advantage given by using XLE lies in the richness of the output representations returned by this parser. These representations are rich enough to give us (at least indirect) access to many of the relevant factors identified by linguists as influencing anaphora resolution. In particular, they provide us with morpho-syntactic information (via gender, number, person, and case attributes), syntactic information (via grammatical functions and f-structure configurations[10]), as well as shallow lexical semantics (in the form of animacy, count/mass attributes). As is well-know, an interesting aspect of grammatical functions (GFs) is that they are also correlated to some degree with salience, therefore also giving some partial access into pragmatics. Thus, certain GFs (e.g., subjects) often make more likely antecedents than others. Furthermore, certain "transitions" over GFs (e.g. subject-subject, subject-object) are also potentially useful for coreference in giving us shallow access to discourse structure: parallelism (or contrast) can to a certain extent be captured at the level of grammatical functions. For these reasons, GFs (along with f-structure "paths") will provide most of the features in this pilot study.

To show the importance of GFs for CR, consider the following example from the ACE data:

(3) [He]knew [Brosius]was coming off a bad year, and **he** knew Brosius would be in line to make a decent salary.

In the context of this example, the pronoun he could be resolved to two mentions, namely either Brosius or the preceding pronoun He. Based on surface-based features alone, the pronoun is likely to be resolved to Brosius, since this expression is the closest mention corresponding to the same type of named entity (i.e., a person). Access to the XLE analysis in figure 1 provides us with information that may lead to a better prediction. Intuitively, the first pronominal mention is more "salient" than than the proper name; this is encoded grammatically by the fact that that mention is the subject of the main clause. Note that there is also a parallelism effect here, since the same subject is maintained across clauses. Features based on grammatical functions and transitions over grammatical functions appear to have a potential for correcting some of the mistakes that would be made by simply relying on surfacy features.

3.2 Potential Issues

However, there are a number of places where things can go wrong when using the outputs of deep parsing systems such as XLE. For one thing, statistical disambiguation potentially introduces a level

[10] In LFG, binding principles are stated in terms of the notion of f-command, a relation which is defined directly over f-structures.

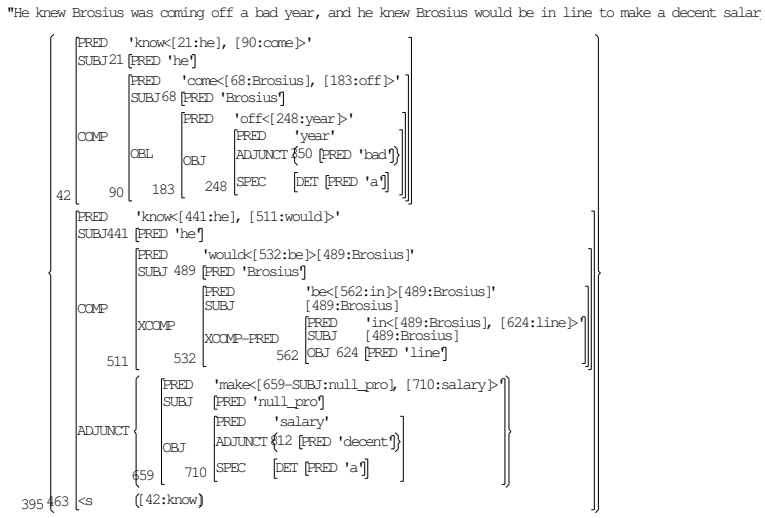


Figure 1: XLE output for sentence (3)

of noise by potentially filtering out correct analyses. This is likely to affect us, since here we only consider the unique *most probable* parse for each sentence (rather than the whole parse forest or even the n best parses). Second, XLE simply fails to produce output for some sentences. In our experiments, we found no parse for 4.3% and 5.1% in training data and in test data, respectively.[11] By comparison, note that Charniak parser only missed less than 0.1% in both the training data and the test data.[12] Also of interest is the fact that the XLE parser outputs "fragment" parses for a significant number of the parsed sentences: 24.4% in training data and 24.7% in test. This in turn raises the following questions: (i) for XLE, to what extent will the additional precision gained by using the XLE representations be able to out-weigh the noise, and (ii) more generally, are the richer outputs still more useful than shallower, but more robust, representations?

As a concrete illustration of these issues, consider the case of grammatical functions and the problem of their identification. GFs can potentially be identified (or at least approximated) using representations reflecting various levels of processing. But crucially, the shallower the processing is, the higher the recall of the identification will be, but the lower its precision will be. Thus, GFs in English at least can be first approximated in terms of part-of-speech (POS) contexts: for instance, subjects are often found before a verbal form, objects after a verbal form, while obliques tend to occur after prepositions. While entirely robust, this strategy of solely relying on linear order is prone to make many errors. For instance, some embedded NPs will be wrongly identified as subjects (say, in relative clauses), while others will be wrongly treated as obliques (say, in PP modifying a head noun). Some of these errors will be handled by going one level up in terms of linguistic processing, and using actual phrase structure configurations to capture GFs: e.g., `[S [NP VP]]` vs. `[S [VP [NP]]]` for the subject/object contrast. While more reliable, these representations are harder to obtain with precision: these sorts of configurations are more reliable, but they are not error-free: e.g., the first NP in a dative-shift construction will be wrongly treated as a direct object. At the end of the spectrum, in XLE GFs

[11] We used one of the latest releases of XLE (June 18, 2006) and of the English grammar (December 5, 2005). The parser was used with its default parameters.

[12] We used the August 16, 2005 version (`ftp.cs.brown.edu/pub/nlparser/`); we also used the default parameters.

can be simply read off the output as attributes, but the problem here is that they might not always be available.[13] There is no general answer to the question how the trade-off between robustness and quality will affect a particular practical performance task.

A possible issue we are facing for the more linguistically sophisticated approaches lies in the set of learning features: too small a feature set might not give us enough to properly model the data. We are likely to suffer from this problem, since we only focus on GFs and GF paths here. One of the main goals of the present study was to set up the machinery for incorporating rich linguistic learning features in the CR task. There is a large space of sophisticated features and feature combination that should be carefully explored.

In our experiments we are also only doing manual feature selection (i.e., filtering of the vast number of feature combinations that are possible), which is typically inferior to automatic feature selection techniques.

Finally, there are some potential issues of a more fundamental kind. We mentioned the AI-completeness point in the introduction already. For a linguistics-rich approach this means that even perfect syntactic information may have a limited effect on performance. Since the various linguistic factors involved in coreference resolution are not sufficient for specifying a deterministic procedure, it is not necessarily the case that the richer linguistic information sources (with the unavoidable noise in the output of any parser) can add task-relevant information that is not already accessible in a more surface-oriented approach. Surface-oriented approaches may actually have an advantage picking up patterns correlating to extra-linguistic factors, without an intermediate representation that may add noise.

A further potential issue has to do with the size of the training data: for the surface-oriented learning features used in most machine-learning based work on CR, learning curves show that already a relatively small quantity of training data already provides sufficient information to acquire the relevant generalizations. Performance figures tend to plateau when adding more training data. Now, if more sophisticated features (and in particular combinations of features) are used, it is quite possible that considerably larger training sets would be required to pick up certain patterns. In our experiments, a number of features that appear interesting from a theoretical point of view were only instantiated in very few training examples; so, data sparsity issues are likely to influence the results.

Somewhat related to the previous two issues, we may note that the CR task is of a somewhat peculiar nature: a considerable proportion of the coreference linking decisions are almost trivial, some of the remaining decisions follow clear linguistic patterns, but a fairly large proportion is controlled by a highly complex interaction of constraints. Thus, a surface-oriented approach has a fair chance of getting up to a certain level of performance and will even get some of the hard cases right ("by chance", so to speak). Ideally, a more sophisticated approach should keep up the quality of the simple technique for the easier cases, but avoid some of the errors for the harder ones. However due to the complex interactions of factors, picking up certain valid deeper patterns may have the effect of breaking a favorable behavior in certain other cases, which may overall balance out the gain from deeper insights.[14]

[13] Note a final advantage of XLE: since GFs are not tied in LFG to particular structural configurations, the strategy used for their identification will work for other languages.

[14] For instance, a surface-based system will typically exclude person-shifts as shown for e_0 (Clinton) in (1). But a more sophisticated system may pick up circumstances under which they *are* possible. It is quite likely however that this pattern will overgenerate to some extent, thus leading to misclassifications in cases considered almost trivial with a surface-based system.

Data-set	train	test
BNEWS	216	51
NPAPER	76	17
NWIRE	130	29

Table 1: # of documents

Dataset	train	test
BNEWS	3740	950
NPAPER	2453	615
NWIRE	2724	608

Table 2: # of sentences

Dataset	train	test
BNEWS	10086	2608
NPAPER	11410	2504
NWIRE	10868	2630

Table 3: # of mentions

3.3 From the XLE output to learning features

How do we extract information from XLE output for creating our features? Among the various formats available, XLE outputs its analyses in Prolog, where c-structure subtrees and f-structure constraints are represented as lists of Prolog facts. (The mapping function ϕ from c- to f-structure is also captured this way.) This is illustrated in figure 2, which is the output for sentence (3).

More specifically, these representation encode: (i) the character offsets of each token, (ii) the c-structure projections for each token as well as the mapping from each subtree to its f-structure node, and finally (iii) the constraints associated with each f-structure node (i.e., a full description of the f-structure). The way we were able to map each mention to its corresponding f-structure was first unpacking the different Prolog facts into various data structures, then mapping the different tokens making up the mention to their corresponding surface forms in the XLE representation. Once identified, the different surface forms could be mapped to an actual f-structure node (and to the associated set of AVMs). In the case of multi-word mentions, the highest node in the graph, i.e., the node corresponding to the maximal projection, was used. Each mention is furthermore associated with a f-structure path from the main (i.e., ROOT) f-structure to its f-structure node.

4 Experimental setup

In order to evaluate the contribution of XLE for the coference task, we ran comparative experiments with various feature sets extracted from analyses provided by the three different "syntactic" analyzers with different depths of processing and degrees of robustness: (i) a part-of-speech tagger (we used OpenNLP Maxent POS tagger), (ii) a Penn Treebank trained phrase structure parser (namely, the Charniak parser), and (iii) XLE parser, which is a full-blown implementation of LFG.

4.1 Corpus and evaluation

For training and evaluation, we used the datasets from the ACE corpus (Phase 2). This corpus is composed of three parts, corresponding to different genres: broadcast news transcripts (BNEWS), newspaper texts (NPAPER), and newswire texts (NWIRE).[15] Each of these is split into a `train` part and a `devtest` part. We used the `devtest` material only once, namely for final testing. Progress during the development phase was estimated only by using cross-validation on the training set for the NPAPER section. Statistics for the different datasets are given in tables 1-3.

In our experiments, we restricted ourselves to the *true* ACE mentions, i.e., rather than trying to identify candidate phrases for coreference resolution automatically (task (a) addressed in section 2.1), we

[15]The mentions in ACE2 are restricted to 7 types of entities:FACility, GPE (geo-political entity), LOCation, ORGanization, PERson, VEHicle, WEApons.

```
fstructure('He knew Brosius was coming off a bad year, and
           he knew Brosius would be in line to make a decent salary.',
% Properties:
[
'xle_version'('XLE release of Aug 15, 2006 15:04.'),
...
'statistics'('15+30 solutions, 0.83 CPU seconds, 1119 subtrees unified'),
'rootcategory'('ROOT')
],
...
],
% Constraints:
[
cf(1,in_set(var(1),var(0))),
cf(1,in_set(var(22),var(0))),
cf(1,eq(attr(var(0),'COORD'),'+_')),
...
cf(1,eq(attr(var(1),'PRED'),semform('know',2,[var(19),var(3)],[]))),
cf(1,eq(attr(var(1),'SUBJ'),var(19))),
cf(1,eq(attr(var(1),'COMP'),var(3))),
...
cf(1,eq(attr(var(19),'PRED'),semform('he',0,□,[]))),
cf(1,eq(attr(var(19),'CASE'),'nom')),
cf(1,eq(attr(var(19),'GEND-SEM'),'male')),
cf(1,eq(attr(var(19),'HUMAN'),'+')),
cf(1,eq(attr(var(19),'NUM'),'sg')),
cf(1,eq(attr(var(19),'PERS'),'3')),
cf(1,eq(attr(var(19),'PRON-TYPE'),'pers')),
...
cf(1,eq(attr(var(3),'PRED'),semform('come',11,[var(15)],[]))),
cf(1,eq(attr(var(3),'SUBJ'),var(15))),
...
cf(1,eq(attr(var(15),'PRED'),semform('Brosius',3,[],□))),
cf(1,eq(attr(var(15),'CASE'),'nom')),
cf(1,eq(attr(var(15),'NUM'),'sg')),
cf(1,eq(attr(var(15),'PERS'),'3')),
...
],
% C-Structure:
[
cf(1,subtree(13969,'ROOT',15617,899)),
cf(1,phi(13969,var(0))),
...
cf(1,terminal(460,'he',[441])),
cf(1,phi(460,var(53))),
...
cf(1,terminal(77,'Brosius',[68])),
cf(1,phi(77,var(15))),
...
cf(1,terminal(38,'he',[21])),
cf(1,phi(38,var(19))),
...
cf(1,surfaceform(68,'Brosius',9,16)),
cf(1,surfaceform(42,'knew',4,8)),
cf(1,surfaceform(21,'^ he',1,3))
]).
```

Figure 2: XLE Prolog (abbreviated) output for sentence (3)

relied on the gold standard phrases/mentions marked manually in the corpus annotation. We made this decision because our focus is on comparing features between different knowledge sources, rather than on building a full-fledged resolution system. It is worth noting that previous work tends to be vague about mention detection: details on mention filtering or providing performance figures for identification are rarely given.

Following common practice in coreference resolution, we report our main results in terms of Recall-Precision at the level of chains partitioning the set of all mentions in the text. In particular, we use the model-theoretic metric proposed by (Vilain *et al.*, 1995). This method operates by comparing the equivalence classes defined by the resolutions produced by the system with the gold standard classes: these are the two "models". Roughly speaking, the scores are obtained by determining the minimal perturbations needed to transform one model into the other model. Recall is computed by trying to transform the predicted chains into the true chains, while precision is computed the other way around.

4.2 Feature sets

Overall, we actually used four systems, based on four different feature sets. In our baseline feature set, we used features obtainable from shallow processing; the corpus was preprocessed with the OpenNLP Toolkit[16], which includes a sentence detector, a tokenizer, and a POS tagger. These features include **NP type** features for the anaphor candidate and the antecedent candidate (i.e., whether the mention is a pronoun, a proper name, a definite description, etc.), **locality** features (encoded in the form of various distance features), **morpho-syntactic agreement** features (i.e., gender, number, and person compatibilities), **semantic compatibility** features (this is captured in terms of the named entity types), salience-based features (e.g., number of times a mention has been seen in the previous context), as well a number of **ad hoc features** for specific NP types (e.g., string matching, apposition and acronym). These features are summarized in table (4.2).

In addition to the simple features described above, we used various composite features by "crossing" some of the basic features above. For the baseline feature set, we simply combined distance features with the type of the anaphor (e.g., pronoun, definite NP, proper names).

The second feature set expands on the baseline by encoding more linguistically-motivated features (mainly features appromixating GFs), but which are based solely on the outputs of the POS tagger. The third feature set incorporates features that use the output of Charniak, while the fourth feature set includes features derived from the XLE output. With both parsers, we used the unique most probable parse for each sentence. These new features fall into four main categories: **GF**, **GF transitions**, **Binding**, and **Syntactic context**. They are presented in detail in the form of templates in table 4.2.

In addition to these base features described above, we added composite features of the following types: (i) distances and GFs, (ii) distances and syntactic context of the antecedent candidate, (iii) distances and binding, (iv) anaphor type and syntactic embedding of the antecedent candidate, and (v) distances, anaphor type, and syntactic context of the antecedent.

[16] Available from `opennlp.sf.net`.

Feature type	Feature Name	Description
NP type	ANA_PRO	T if m_i is a pronoun; else F
	ANA_SPEECH_PRO	T if m_i is a speech pronoun; else F
	ANA_REFL_PRO	T if m_i is a refl. pron.; else F
	ANA_PN	T if m_i is a PN; else F
	ANA_DEF	T if m_i starts with *the*; else F
	ANTE_PRO	T if m_j is a pronoun; else F
	ANTE_PN	T if m_j is a PN; else F
	ANTE_DEF	T if m_j starts with *the*; else F
	ANTE_INDEF	T if m_j starts with *a(n)*; else F
Locality	S_DIST	binned values for S distance between m_i and m_j
	NP_DIST	binned values for NP distance between m_i and m_j
Morphosynt. Agreement	NUM_AGR	T if m_i and m_j agree in number; else F
	GEN_AGR	T if m_i and m_j agree in gender; else F;
Salience	ANA_M_CT	# of times m_i has been seen
String match	STR_MATCH	T if the strings of m_i and m_j match; else F
Semantic Agreement	NE_AGR	T if m_i and m_j correspond the same NE; else F
	ANTE_NE_&_ANA_GEN	the NE of m_j and the gender of m_i
Quotes	ANA_IN_QUOTES	T if m_i is within quotation marks; else F
	ANTE_IN_QUOTES	T if m_j is within quotation marks; else F
Acronym	ACRONYM	T if one NP is an acronym of the other; else F
Apposition	APPOSITION	T if m_i is an apposition of m_j; else F

Figure 3: Baseline feature set

5 Results and Analysis

This section presents the results of our various experiments, as well as some initial elements of analysis. Table 5 summarizes the results of our main experiment on the three ACE datasets.

The results tell us a number of things. First, the addition of the new features appears to yield a small drop in overall f-score; the differences are however not statistically significant (at $p < .01$) for any of the feature sets. Second, the actual pattern found for the different feature sets is that the addition of the new features produces a gain in recall, but this gain is accompanied by a corresponding drop in precision. From our statistical testing, we however found that although the decreases in precision were significant for all the features (at $p < .01$), the increase in recall is significant only with the XLE features. How do we interpret these results? One can start by considering more closely the different types of errors made by the new systems. One can break down the types of mistakes made by a CR system into three categories: (i) *missing* mentions (i.e., mentions that are not treated as anaphoric when they should), *spurious* mentions (i.e., mentions that treated as anaphoric when they should not), (ii) (correctly identified anaphoric) mentions that are *wrongly resolved*. The first two categories concern the (non-)anaphoricity of a mention, while the third one concerns the resolution *per se*. Also note that the first category only affects recall (these are the false negatives), the second category only affects precision (these are the false positives), while the latter affects both. Looking first at the distributions of the different types of mistakes in the baseline, one first finds that almost 2/3 of the recall mistakes are due to missing anaphoric mentions (the other third is due to wrong resolutions). On the precision side, one finds the opposite pattern: only 1/3 of errors are due to spurious anaphora. As for the effect of the new features, one finds that 2/3 of the recall error reduction comes from a reduction of the missing anaphora; that is, only 1/3 comes from rectifying wrong resolutions.

Looking at the actual predictions, one finds that the XLE features allow the system to identify new,

Feature Type	Feature Name	Description
GFs	ANA_SUBJ	m_i has subject POS context/tree config./SUBJ attr.
	ANA_OBJ	m_i has object POS context/config./OBJ attrib.
	ANA_OBL	m_i has oblique POS context/tree config./OBL attr.
	ANTE_SUBJ	m_j has subject POS context/tree config./SUBJ attr.
	ANTE_OBJ	m_j has object POS context/tree config./OBJ attr.
	ANTE_OBL	m_j has oblique POS context/tree config./OBL attr.
GF	BOTH_SUBJ	m_i and m_j are both subjects
Transitions	SAME_GR	m_i and m_j have the same GF
Binding	C-/F-COMMAND	m_j c-/f-commands m_i
Context	ANTE_PATH_SUFFIX_N	last n nodes (n in {1,2,3}) in m_j's FS/tree path[17]
	ANA_PATH_SUFFIX_N	last n nodes (n in {1,2,3}) in m_i's FS/tree path
	ANTE_PATH_LN	binned value for number of nodes in m_j's FS/tree path

Figure 4: New feature templates

	BNEWS			NPAPER			NWIRE			Overall		
Feature Set	R	P	F	R	P	F	R	P	F	R	P	F
Baseline	53.4	**84.0**	65.3	55.8	**84.3**	**67.1**	51.6	**80.5**	62.9	51.6	**80.5**	62.9
Tagger	54.5	80.6	65.1	56.6	81.0	66.6	53.2	78.7	63.5	53.2	78.7	63.5
Charniak	55.6	79.8	65.5	56.1	80.3	66.1	54.8	80.0	**65.0**	54.8	80.0	**65.0**
XLE	56.2	76.8	64.9	**58.8**	77.1	66.8	55.2	76.0	63.9	55.2	76.0	63.9
All	**57.6**	76.4	**65.7**	57.9	76.9	66.0	**56.3**	76.1	64.8	**56.3**	76.1	64.8

Figure 5: Results for the 3 ACE datasets

more subtle coreferential configurations, but these features tend to be unreliable. To give an illustration of this tendency, note for instance that one finds more correct long distance resolutions, but at the same time one also finds errors showing number and gender mismatches (e.g., $\langle he, she \rangle$, $\langle he, Mrs.Anderson \rangle$).

Although the performances are fairly similar for all the new systems, there is however one dataset where XLE seems to be better than the baseline, namely the NPAPER dataset.[18] Interestingly, it is also on this the corpus that XLE shows the best parsing performances (especially in training), with only 3% (against an average is 4.3%) of parses missing and 18.7% of fragment parses (against an average is 24.4%) for training, and 2% (against an average is 5.1%) of parses missing and 23% of fragment parses (against an average is 24.7%) for test. This would suggest that there is a correlation between the amount of sentences given a full parse and the coreference performances.

A similar conclusion emerges from looking at learning curves for this dataset. These are given for the three feature sets on the in figure 6. These curves are encouraging for XLE in suggesting that this system would benefit the most from additional training data; indeed, it is the only curve among the three that does not appear to converge. This indicates that as speculated in section 3.2, the deeper approaches may benefit more from larger training sets than the surface-oriented approaches.

A final, interesting question is whether the systems did differently for different types of mention. Here, we consider three main types, namely mentions that are headed by a pronoun, a proper name (PN), or a common noun (CN). The results below are given in terms of a slightly different evaluation

[18]The difference is not statistically significant however.

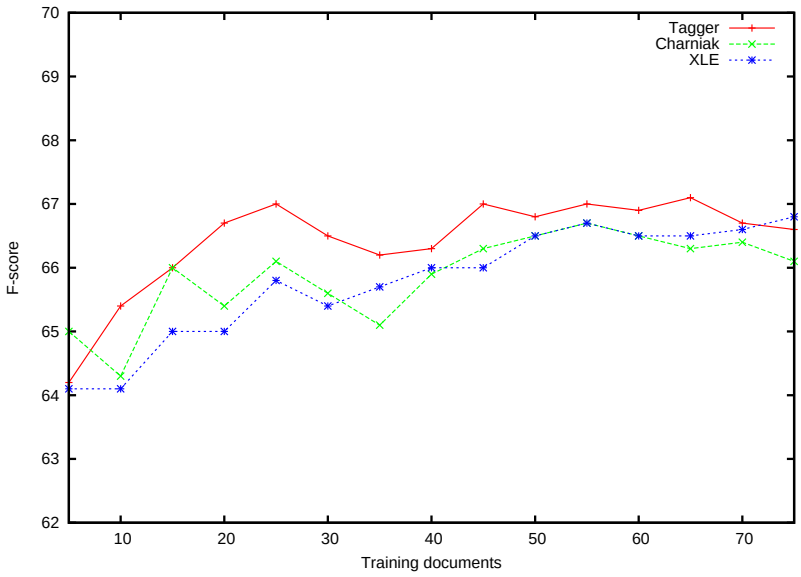


Figure 6: Learning curves for the NPAPER dataset

scheme, namely these are *anaphora resolution* scores. Roughly, one looks at individual links rather than comparing the entire chains.[19] Under this metric, recall is the number of mentions (of the given type) that are correctly resolved divided by the total number of anaphoric mentions (of that same type). And precision is the number of mentions (of the given type) that are correctly resolved divided by the number of mentions that are resolved. The results for the different mention types are given in table 7.

These results are rather inconclusive: the Charniak features seem to make a stronger contribution with pronouns, while the XLE features yield improvements with proper and common nouns. Note that the general low scores for the latter type is explained by the fact that our system does not incorporate a lot of lexical semantic information, which is so critical for these (e.g., definite descriptions).

6 Conclusions and Future Work

By way of various experiments, this study has compared the use of feature sets encoding various depths of linguistic processing for the task of robust coreference resolution. We have in particular compared three main feature sets, extracted from a simple POS tagger, Charniak parser, and the XLE parser. The main conclusions are as follows. The addition of the new features gives rise to an increase in Recall, but don't lead to an overall increase in f-score. We take this to indicate that the new features permit the detection of more coreference configurations, but the extra information is not reliable yet. XLE seems to offer a better improvement potential than Charniak or the POS tagger, but only when it

[19]This type of evaluation is coarser than Vilain's metric in that it misses potential "implicit" links (cf. coreference is an equivalence relation), but it makes it easier to compare different NP types.

	BNEWS			NPAPER			NWIRE		
NP type	R	P	F	R	P	F	R	P	F
Pronouns									
Baseline	67.8	77.0	72.1	**67.2**	**74.8**	**70.8**	60.6	70.8	65.3
Tagger	68.0	76.9	72.1	65.7	72.8	69.1	61.5	71.3	66.0
Charniak	**69.2**	**77.1**	**72.9**	65.2	72.0	68.4	**67.4**	**74.2**	**70.6**
XLE	67.2	75.6	71.1	65.2	66.9	66.1	63.0	69.1	65.9
PNs									
Baseline	47.6	84.6	60.9	56.6	**87.6**	68.8	58.2	87.8	70.0
Tagger	48.4	**84.8**	61.6	56.5	**87.6**	68.7	58.6	**87.9**	**70.3**
Charniak	49.3	84.7	62.3	56.5	87.3	68.6	58.6	87.7	70.2
XLE	**50.5**	82.7	**62.7**	**57.6**	86.6	**69.1**	**59.6**	83.5	69.6
CNs									
Baseline	27.8	**86.0**	**42.0**	25.6	**89.3**	**39.8**	27.4	**75.9**	40.2
Tagger	30.9	64.5	41.8	27.3	61.6	37.8	30.7	63.8	41.5
Charniak	30.3	57.5	39.7	27.0	62.6	37.7	30.2	65.9	41.4
XLE	**33.7**	51.7	40.8	**30.4**	54.0	38.9	**33.2**	61.0	**43.0**

Figure 7: Results per mention types

achieves good parsing performances. XLE also seems more likely to benefit from additional training data.

In section 3.2, we speculated about a lot of potential issues that may preclude a straightforward improvement of the surface-oriented CR techniques by simply adding more linguistically sophisticated knowledge sources. Presumably several of them do hold true. By setting up a flexible system for integrating linguistic resources, we established a basis for further explorations of the interactions.

There are various natural ways to extend this work. First, by using the unique most probable parses, our experiments have not used the two parsing systems to their full potentials. For instance, one would like to take advantage of the "packed" representations provided by XLE, instead of just using a single parse. Second, a lot of extensions are possible regarding feature design: we only scratched the surface in considering only GFs and GF paths. Third, there are more effective ways of combining the different feature sets, instead of just adding them together in a unique model; a better alternative would be to use ensemble models. Finally, there is also the possibility that more "global" models and less greedy search strategies will make better use of the rich features extracted from the deep parses.

References

Berger, A., Pietra, S. D., and Pietra, V. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**(1), 39–71.

Byron, D. K. (2002). Resolving pronominal reference to abstract entities. In *Proceedings of the ACL '02*, pages 80–87.

Denis, P. and Baldridge, J. (2007). A ranking approach to pronoun resolution. In *Proceedings of IJCAI-07*.

Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, , and Ö. Dahl, editors, *Proceedings of the Stockholm Workshop*

on 'Variation within Optimality Theory. April 26-27, 2003 at Stockholm Univ. Sweden, pages 111–120.

Grosz, B., Joshi, A., and Weinstein, S. (1995). Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, **2**(21).

Johnson, M. (1998). Optimality-theoretic Lexical Functional Grammar. In *Proceedings of the 11th Annual CUNY Conference on Human Sentence Processing, Rutgers University*.

Johnson, M., Geman, S., Canon, S., Chi, Z., and Riezler, S. (1999). Estimators for stochastic "unification-based" grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), College Park, MD*, pages 535–541.

Kehler, A. (1997). Probabilistic coreference in information extraction. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 163–173.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Workshop on Natural Language Learning*, pages 49–55, Taipei, Taiwan.

McCallum, A. and Wellner, B. (2003). Toward conditional models of identity uncertainty with application to proper noun coreference. In *Proceedings of IJCAI Workshop on Information Integration on the Web*.

McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. In *IJCAI*, pages 1050–1055.

Mitkov, R. (2002). *Anaphora Resolution*. Longman, Harlow, UK.

Morton, T. (1999). Using coreference for question answering. In *Proceedings of ACL Workshop on Coreference and Its Applications*.

Morton, T. (2000). Coreference for NLP applications. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong.

Ng, V. and Cardie, C. (2002a). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*.

Ng, V. and Cardie, C. (2002b). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 104–111.

Preiss, J. (2002). Choosing a parser for anaphora resolution. In *Proceedings of DAARC 2002*, pages 175–180.

Riezler, S., Crouch, D., Kaplan, R., King, T., Maxwell, J., and Johnson, M. (2002). Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Pennsylvania, Philadelphia*.

Soon, W., Ng, H., and Lim, D. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, **27**(4), 521–544.

Uryupina, O. (2006). Coreference resolution with and without linguistic knowledge. In *Proceedings of LREC 2006*, pages 893–898.

van Deemter, K. and Kibble, R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, **26**(2), 629–637.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings fo the 6th Message Understanding Conference (MUC-6)*, pages 45–52, San Mateo, CA. Morgan Kaufmann.

ON THE REPRESENTATION OF CASE AND AGREEMENT

Yehuda N. Falk
Department of English
The Hebrew University of Jerusalem

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract
Case and agreement are typically modeled in LFG as f-structure phenomena. Inadequacies of this account suggest an alternative, in which grammatical marking is modeled as a separate projection from f-structure, called g-structure.

1. Case and Agreement in LFG

1.1. Agreement

It has become clear over the past few years that the standard LFG treatment of agreement is inadequate, a fact which was the focus of the agreement workshop at the LFG 05 conference. This inadequacy is a combination of several factors, but it is primarily a conceptual failure. An adequate theory of agreement must begin with a conceptual understanding of the nature and purpose of agreement. This paper[1] will argue that agreement, along with Case marking, involves an additional level of representation in the LFG projection architecture, one which we will call grammatical marking structure, or g-structure.

At the outset, a caveat is in order. It is beyond the scope of this study to discuss every agreement problem that has been discussed in the vast literature on the subject. Thankfully, linguistics has moved beyond the point where agreement is a mere footnote, and discussion of agreement patterns and unusual agreement phenomena currently abound in the theoretical and typological literature. An anonymous reviewer of this paper comments that "the proposal that a g-structure will solve all our problems is unconvincing: we do not really know what our problems are, so why should g-structure solve them?" We demur. Even if we limit our attention to problems that have already been discussed in the LFG literature (as we do for the most part in this paper), there is much to be said about problems facing a theory of agreement. No theory in linguistics has yet solved all problems in *any* area. Our goal is much more modest: we suggest that the proposal to be made here results in a better theory of agreement than the standard LFG approach, and we expect that it will prove fruitful in dealing with other problematic cases of agreement. However, as is the case with any theoretical proposal, this expectation must face the test of further empirical investigation.

The standard LFG analysis of agreement treats it as an f-structure phenomenon. This embodies the claim that agreement is based on grammatical functions. While grammatical functions do play a major role in agreement, a fact we incorporate by treating g-structure as a projection from f-structure, there are additional considerations, primarily c-structural and i-structural, which also play a role.

One non-f-structure factor which plays a role in agreement systems is linear order. This has been discussed in the LFG literature primarily in regard to agreement with coordinate structures. As noted by Sadler (1999), many languages exhibit agreement not with the resolved features of a coordinate structure, but rather with the conjunct which is closest to the agreeing head. This is true in Welsh (1a,b), where the verb precedes the agreement trigger, and in Swahili (1c), where it follows:

(1) a. Roedd Mair a fi i briodi.
was.3SG Mair and I to marry
'Mair and I were to marry. '

b. Roeddwn i a Mair i briodi.
was.1SG I and Mair to marry
'I and Mair were to marry.'

[1]I would like to thank Aaron Broadwell, Tracy King, Rachel Nordlinger, Nigel Vincent, and especially Mary Dalrymple for comments on this paper.

c. Ki- ti na m- gurr wa meza u- mevunjika.
7- chair and 3- leg of table 3- be.broken
'The chair and the leg of the table were broken.'

More surprisingly, Sadler observes that there are also languages in which the farthest conjunct triggers agreement, such as Slovenian:

(2) Groza in strah je prevzela vso vas.
horror(FSG) and fear(MSG) has seized.FSG the.whole village
'Horror and fear have seized the whole village.'

A particularly interesting case of closest conjunct agreement, to which we will return later, is found in Portuguese attributive adjectives (Sadler & Villavicencio 2005).

Linear order shows up in other contexts as well. For example, in Standard Arabic the features in which the verb agrees with a non-pronominal subject depend on whether the subject precedes or follows the verb: if the subject precedes, the verb agrees in both gender and number; if the subject follows, the verb agrees in gender only (Aoun, Benmamoun, & Sportiche 1994).[2]

(3) a. Naama l- ʔawlaad- u.
slept.3MSG the- children- NOM
'The children slept.'

b. ʔal- ʔawlaad- u naamuu.
the- children- NOM slept.3MPL
'The children slept.'

c. *Naamuu l- ʔawlaad- u.
slept.3MPL the- children- NOM

d. *ʔal- ʔawlaad- u naama.
the- children- NOM slept.3MSG

The relevance of the c-structure relation of precedence is surprising under an f-structural account of agreement.

Another non-f-structure factor in agreement is i-structure. One example of a language in which i-structure plays a role is Maithili, discussed by (Dalrymple & Nikolaeva 2005). In Maithili, the verb agrees in person and honorific status with the subject and with one other element. This additional element is chosen not by grammatical function—it can be OBJ (4a), OBL (4b), SUBJ POSS (4c), or OBJ POSS (4d)—but rather by discourse status: it must be a secondary topic.

(4) a. Həm to- ra kitab d- əit ch- iəuk.
I you.HN- OBJ book give- PART be- 1.2NH
'I gave a book to you$_{\text{non-honorific}}$.'

b. Tō hunkā- sa kiae khisiael chahun?
you him.H- INST why angry are.2MH.3H
'Why are you angry with him$_{\text{honorific}}$?'

[2]In addition, closest conjunct agreement (which is not obligatory in Standard Arabic) is only possible when the subject follows the verb.

c. Tohar bābu Mohan- ke dekhalthun.
your.NH father.H Mohan- OBJ saw.3H.2NH
'Your$_{\text{non-honorific}}$ father$_{\text{honorific}}$ saw Mohan.'

d. O tora: bāp- ke dekhalthun.
he.H your.NH father- OBJ saw.3H.2NH
'He$_{\text{honorific}}$ saw your$_{\text{non-honorific}}$ father.'

There are further problems with the idea that agreement is an f-structure phenomenon. The agreement of attributive adjectives with their nominal heads, a phenomenon familiar from many languages, is puzzling. Generally, heads agree with their dependents (verbs with subjects and objects, prepositions with objects, nouns with possessors), leading to the expectation that the function of agreement is to mark certain information on heads. However, attributive adjective agreement appears to be the agreement of a dependent with a head. While modeling attributive adjective agreement formally is not a problem, its existence raises conceptual problems in understanding the nature of agreement. One possible solution would be to model agreement at a level of representation in which head-dependent relations are not the same as those at f-structure.

It is even possible for an agreement trigger to not be an element of f-structure. This can be seen, for example, in Hindi-Urdu, where a verb is third person masculine singular if both the subject and object are Case-marked (Butt 1993, inter alia).

(5) Naadyaa ne ciṭṭ^{h}ii ko lik^{h}- aa hai.
Nadya(F) ERG note(F) ACC write- PERF.**M**.SG be.PRES.3SG
'Nadya has written a (particular) note.'

In this sentence, the verbal forms *lik^{h}aa* and *hai* agree with a third-person masculine singular element, but there is no such element in the f-structure of the sentence. This is a case of what is generally treated as "default agreement", but in the absence of a theory of default agreement, the name is simply a dodge. Crucially, default agreement is not the absence of agreement, as one might expect; rather, it is agreement with a specific set of features which the grammar of the language specifies as the default. This suggests that agreement should be modeled as an aspect of a level of structure which can have non-f-structure elements in it; the default can be defined in terms of elements at this additional level of structure.

None of these phenomena is unexpressible using the standard LFG analysis and formalism. However, the conceptual basis of the theory of agreement is stretched by the analyses of these phenomena. In some cases, particularly those involving linear order, the conceptual inappropriateness leads to a formal expression which is rather convoluted and uninsightful.

1.2. Case

Another aspect of agreement which is not expressed by standard LFG treatments of agreement is the relationship between agreement and Case. That agreement and Case are related to each other is well known. As pointed out by Nichols (1986), Case and agreement are alternative ways of marking the same sorts of head-dependent relations. The difference between Case and agreement is that the former is marked on the dependent element in the relation, while the latter is marked on the head. Without losing sight of the differences between Case (dependent-marking) and agreement (head-marking),[3] the two should be analyzed as elements of a single dimension of language. LFG does not do this.

The Nichols-type conceptualization of Case and agreement as alternative markings of the same

[3]As is done in standard transformational accounts, which treat Case and agreement as reflexes of the same SPEC-head relationships.

relationships receives strong support from the grammars of many languages which employ both. In many languages of this kind, one or more elements of the sentence are left without Case (or with the unmarked Case, nominative or absolutive). The verb agrees with one element: a Caseless element. The facts are particularly clear in Hindi-Urdu (Butt 1993), where both the SUBJ and the OBJ are sometimes Case-marked. The agreement facts are that the verb agrees with the highest Caseless NP; as we have already seen, if both SUBJ and OBJ are Case-marked, the verb displays (default) third-person singular masculine agreement.

(6) a. Naadyaa xat lik^{h}- tii hai.
Nadya(F) letter(M) write- IMPF.F.SG be.PRES.3SG
'Nadya writes a letter.'

b. Naadyaa ne xat lik^{h}- aa hai.
Nadya(F) ERG letter(M) write- PERF.M.SG be.PRES.3SG
'Nadya has written a letter.'

c. Naadyaa ne ciṭṭ^{h}ii lik^{h}- ii hai.
Nadya(F) ERG note(F) write- PERF.F.SG be.PRES.3SG
'Nadya has written a note.'

d. Naadyaa ne ciṭṭ^{h}ii ko lik^{h}- aa hai.
Nadya(F) ERG note(F) ACC write- PERF.**M**.SG be.PRES.3SG
'Nadya has written a (particular) note.' (=(5) above)

However, this phenomenon is not limited to Hindi-Urdu. For example, in Icelandic, a verb with a dative SUBJ and nominative OBJ or OBJ$_{\theta}$ agrees with the nominative (examples from Otoguro 2005 and Andrews 1982).

(7) a. Henni leiddust strákarnir.
her.DAT bored.3PL the.boys.NOM
'She found the boys boring.'

b. Henni voru sýndir bílarnir.
her.DAT were.PL shown.MPL the.cars
'She was shown the cars.'

Another such situation is the possessive sentence in Modern Hebrew. Possessive sentences in Hebrew have the structure: 'be' – possessor (in the dative) – possessed. Historically, the possessed nominal was the subject. It thus was unmarked for Case and triggered agreement on the verb. Such usage is still considered normative. However, in actual spoken Hebrew, the possessed nominal has been reinterpreted as an object. This means that it is marked with accusative Case (only when definite, as is always the case in Hebrew). As observed by Ziv (1976), the presence or absence of accusative Case is correlated with the absence or presence of agreement in colloquial Hebrew.

(8) a. Hayta li mexonit kazot.
be.PAST.3FSG DAT.1SG car(F) such
b. ?Haya li mexonit kazot.
be.PAST.3MSG DAT.1SG car(F) such
'I had such a car.'

(9) a. ?Hayta lanu et ha- mexonit hazot od kše garnu
be.PAST.3FSG DAT.1PL ACC the- car(F) this still when live.PAST.1PL
be- tel aviv.
in- Tel Aviv

b. Haya lanu et ha- mexonit hazot od kše garnu
be.PAST.3MSG DAT.1PL ACC the- car(F) this still when live.PAST.1PL
be- tel aviv.
in- Tel Aviv
'We had this car when we were living in Tel Aviv.'

The result again clearly correlates agreement with the absence of Case. The widespread occurrence of such correlations suggests that there is a representational relationship between head- and dependent-marking, a relationship which is not expressed in the standard LFG analysis.

A Nichols-inspired approach to the relation between Case and agreement allows us to make sense of the complementarity between the two found in such languages as Hindi-Urdu, Icelandic, and Hebrew. The purpose of these morphosyntactic markings is to provide an overt indication of the relation between a head's argument slots and the elements that fill them. Agreeing with a Case-marked element is uneconomical, as it provides double marking for a single element.

Independently of agreement phenomena, the standard LFG analysis of Case as an f-structure feature of NPs does not provide adequate expression of the nature of Case marking. While Case often reflects grammatical functions, as would be expected if it is part of f-structure, it can also reflect thematic roles, topicality, and the like. This kind of information goes beyond the usual uses of agreement (although, as we have seen, agreement in Maithili can mark topicality), but it is to be expected that dependent marking is richer than head marking. The primary burden of head marking is to point to the argument-filling elements, which is done by expressing their pronominal features. In the case of dependent-marking, the full element is there, so additional information can be encoded by the morphological marking.

The phenomenon of differential Case marking also suggests a different kind of representation for Case than is standard. The basic concept behind differential marking is that more prototypical arguments do not need to be identified by morphological marking (Comrie 1989); this is best expressed theoretically in terms of a formal concept of grammatical marking of argument status.

2. The Proposal

In light of problems of this kind, we propose that the LFG projection architecture be enriched by the addition of a level of representation which, as noted above, we propose to call g-structure (grammatical marking structure). The relatively close relationship between grammatical marking and grammatical functions is expressed by modeling g-structure as a projection from f-structure, the γ projection; however, other levels of representation also have a hand in constraining g-structure.[4]

We represent g-structure as an AVM in which the attributes are names of grammatical markings: AGMT, ACC, ERG, etc. The "head" of the g-structure is a representation of the marking head, represented with the attribute name LEX(EME). Each marking lexeme heads its own g-structure; unlike f-structure, subordinate lexemes are not embedded under superordinate lexemes unless they are themselves subject to grammatical marking.

The predicate of a simple English sentence such as (10a) will have a g-structure representation (10b).

(10) a. She sees him.

[4]G-structure bears superficial resemblance to the level of morphosyntactic structure (m-structure) which has been hypothesized in some of the LFG literature. However, there is a fundamental difference between the two: m-structure, whose existence we reject, purports to represent all inflectional features. The crucial property of g-structure is that it represents a very narrowly defined set of inflections, those involved in grammatical marking of arguments.

b.
$$\begin{bmatrix} \text{LEX} & \text{SEE} & \\ \text{AGMT} & \begin{bmatrix} \text{LEX} & \text{SHE} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{ACC} & \begin{bmatrix} \text{LEX} & \text{HE} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

This representation expresses the fact that both arguments of the verb are involved in grammatical marking: *she* is the agreement trigger and *him* is in accusative Case. As in many analyses of Case, we assume that the "unmarked" Case (nominative or absolutive) is formally the absence of Case; there is no g-structure attribute NOM. However, further investigation is required to determine whether this is the correct analysis for all languages; the Case splits in the morphologically ergative languages of Australia may require a more complicated analysis. In any case, "unmarked" here is not a matter of morphological marking, but rather formal status (as shown by such properties as being the citation form).

Case is an attribute rather than a value under this approach. Since Case is not a feature value, the usual specification of the Case of a noun cannot be used. Instead, inside-out designators can be used, as in Constructive Case (Nordlinger 1998):

(11) *him*: $(\text{ACC}\ \uparrow_{\gamma})$

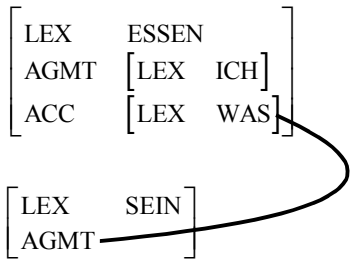
This formalization of Case as a g-structure attribute rather than as an f-structure (or g-structure) value provides a solution for the problem of feature indeterminacy, discussed by Dalrymple & Kaplan (2000). Dalrymple & Kaplan note that in the following German sentence, the form *was* is simultaneously the accusative OBJ of *gegessen* and the nominative SUBJ of *übrig war*:

(12) Ich habe gegessen was übrig war.
I have eaten what left was
'I ate what was left.'

They observe that *was* cannot simply be analyzed as using a disjunction to specify its Case, since in this sentence it is both accusative and nominative, not either nominative or accusative. They opt for representing the value of the f-structure CASE feature as a set: [CASE {NOM, ACC}]. This works, but at a cost. The cost is that every use of *was*, even in cases where it is completely unambiguous, will have the same dual value for the CASE feature. Since forms like *was* are often used unambiguously, this analysis, while technically adequate, is less appealing than a disjunction analysis, which says that each instance of *was* can be identified as either nominative or accusative (or both). Under the present proposal, a disjunctive analysis is possible. The lexical entry of *was* will include:

(13) $(\text{AGMT}\ \uparrow_{\gamma}) \vee (\text{ACC}\ \uparrow_{\gamma})$

In this particular sentence, there are two inside-out g-structure paths originating at *was*: one through ACC and one through AGMT.

(14)


This is allowed by the disjunction.[5]

G-structure represents actual Case and agreement, not "abstract Case". An OBJ which is not marked accusative will not be the value of the ACC attribute. Thus, the g-structure of (15a) is (15b).

(15) a. She sees the hippo.

b.
$$\begin{bmatrix} \text{LEX} & \text{SEE} \\ \text{AGMT} & \begin{bmatrix} \text{LEX} & \text{SHE} \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

In this sentence, *the hippo* has no grammatical marking, and thus does not appear in g-structure. This makes g-structure a suitable level on which to define the conditions for differential Case marking. In English, where (roughly) pronouns are Case-marked and lexical nouns are not, the lexical entry of a transitive verb like *see* will include the following specification:

(16) $(\uparrow \text{OBJ PRED}) = \text{'PRO'} \Rightarrow (\uparrow \text{OBJ})_{\gamma} = (\uparrow_{\gamma} \text{ACC})$

Other conditions for differential marking can be similarly encoded.

G-structure shares certain features with argument structure; not surprisingly, since both are representations of head-dependent relations. For example, just as each argument-taking element has its own a-structure, each marking head has its own g-structure. Unlike c- and f-structure, a- and g-structure are not representations of whole sentences. Thus, as noted above, a subordinate clause which is not itself grammatically marked will correspond to an independent g-structure, not a g-structure embedded in the main clause g-structure.

More interesting is the question of well-formedness constraints on g-structure, and their relation to well-formedness constraints on a-structure. The main well-formedness constraints on a-structure are Biuniqueness, which imposes a one-to-one mapping between arguments and grammatical functions, and the Subject Condition, which requires every verb to have a SUBJ at f-structure. These conditions may be options rather than universally applicable; at least the Subject Condition can be shown not to be active in every language (for example, it does not apply in Modern Hebrew: Falk 2004.) We propose that there is a g-structure analog of Biuniqueness, and perhaps also an analog of the Subject Condition. As with the a-structure Subject Condition, these appear not to be universal.

We attribute the common pattern linking agreement to Caselessness to the g-structure analog of Biuniqueness. Consider the Hindi-Urdu agreement pattern illustrated in (6) above, and repeated here.

(17) a. Naadyaa xat lik^{h}- tii hai.
Nadya(F) letter(M) write- IMPF.F.SG be.PRES.3SG
'Nadya writes a letter.'

[5]As stated, this is too permissive, as it would also allow the form *was* to be the value of DAT or GEN, as long as it is the value of AGMT or ACC. Formally, this could be handled by conjoining the specification with negative statements for the other Cases. Alternatively, the inside-out specifications given for g-structure might be interpreted as implicitly including negative statements for all other possible attributes. At this stage I am inclined to leave the options open.

b. Naadyaa ne xat lik^{h}- aa hai.
Nadya(F) ERG letter(M) write- PERF.M.SG be.PRES.3SG
'Nadya has written a letter.'

c. Naadyaa ne ciṭṭ^{h}ii lik^{h}- ii hai.
Nadya(F) ERG note(F) write- PERF.F.SG be.PRES.3SG
'Nadya has written a note.'

d. Naadyaa ne ciṭṭ^{h}ii ko lik^{h}- aa hai.
Nadya(F) ERG note(F) ACC write- PERF.**M**.SG be.PRES.3SG
'Nadya has written a (particular) note.' (=(5) above)

We propose that the basic rule in Hindi-Urdu is that the agreement trigger (the element which is the value of the AGMT attribute at g-structure) is the highest available [$-r$] argument:[6]

(18) $(\uparrow_{\gamma}$ AGMT$) = (\uparrow [-r])$, subject to the Relational Hierarchy

Now suppose that there is a Biuniqueness condition on g-structure which states:[7]

(19) In the g-structure of a head, a dependent can be the value of only one attribute and an attribute must have a unique value.

In (17a), the highest core function is SUBJ. The f-structure SUBJ corresponds to the g-structure AGMT, resulting in the verb agreeing with the SUBJ. In (17b), on the other hand, such a choice would result in the following g-structure:

(20)
$$\begin{bmatrix} \text{LEX} & \text{LIK}^{\text{H}}\text{-} \\ \left.\begin{matrix}\text{ERG}\\ \text{AGMT}\end{matrix}\right\} & \begin{bmatrix} \text{LEX} & \text{NAADYAA} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

This g-structure violates the proposed Biuniqueness condition. Choosing the OBJ as the agreement trigger is therefore the only option available:

(21)
$$\begin{bmatrix} \text{LEX} & \text{LIK}^{\text{H}}\text{-} \\ \text{ERG} & \begin{bmatrix} \text{LEX} & \text{NAADYAA} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{AGMT} & \begin{bmatrix} \text{LEX} & \text{XAT} \\ \text{GEND} & \text{M} \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

In other words, OBJ is the highest unrestricted argument which can be picked for the agreement trigger that will result in a well-formed g-structure.

The Biuniqueness condition on g-structure is essentially an economy condition on grammatical marking. It disallows multiple marking of the same element. As such, it is a natural condition on

[6]We use the notation $(\uparrow[-r])$ to refer to the class of grammatical functions which have the feature [$-r$]; i.e. {SUBJ, OBJ}., without regard for whether the arguments in question are [$-r$] at a-structure.

[7]Only one of these directions actually needs to be stipulated here: the requirement that every g-structure attribute have a unique value, like f-structure Uniqueness, is a consequence of the operation of unification.

g-structure. On the other hand, not all languages obey Biuniqueness. For example, in Warlpiri agreement is independent of Case marking. It appears, then, that g-structure Biuniqueness is an option rather than an absolute requirement.[8]

For the Hindi-Urdu example, (17d), Biuniqueness rules out both SUBJ and OBJ as agreement triggers, as they are both Case marked, i.e. both correspond to values of Case attributes in g-structure. However, Hindi-Urdu requires the verb to be marked with agreement. While this is, at its core, a morphological requirement, it is possible that it is enforced at g-structure by an analog of the Subject Condition, which we can call the Agreement Trigger Condition. Whether a purely morphological matter or the result of a g-structure condition, the consequence is that there is an AGMT which corresponds to nothing in f-structure.

(22)
$$\begin{bmatrix} \text{LEX} & \text{LIK}^{\text{H}}\text{-} \\ \text{ERG} & \begin{bmatrix} \text{LEX} & \text{NAADYAA} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{ACC} & \begin{bmatrix} \text{LEX} & \text{CIṬṬ}^{\text{H}}\text{II} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{AGMT} & \begin{bmatrix} \text{GEND} & \text{M} \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

The grammar of Hindi-Urdu will specify that a g-structure element with no corresponding f-structure must have the features [GEND M, NUM SG], thus providing a formal expression of the status of masculine singular in Hindi as the default agreement.

3. Adjectives

3.1. Attributive Adjectives

The very common phenomenon of agreement between a noun and its attributive adjectives presents a conceptual challenge to theories of agreement. Such agreement is exemplified below in Hebrew.

(23) a. bayit xadaš
house(M) new.M
'a new house'

b. dira xadaš- a
apartment(F) new- F
'a new apartment'

c. batim xadaš- im
houses(MPL) new- MPL
'new houses'

d. ha- dirot ha- xadaš- ot
DEF- apartments(FPL) DEF- new- FPL
'the new apartments'

There are several puzzling facets to this kind of agreement.

The first puzzle, as noted above, is that this agreement exists at all. Agreement is head-marking.

[8]Given the observation in the previous footnote, it is only the requirement that a dependent can be the value of only one attribute that is violable.

Here, however, the agreeing element is not the syntactic head at either c-structure or f-structure, but rather an adjunct; the head is the element triggering the agreement. As a result, given standard assumptions, one must either analyze this kind of agreement as a completely different kind of phenomenon from other cases of agreement, a kind of dependent marking, or analyze attributive adjectives as if they are predicational, with a SUBJ that is identified with the head of the larger NP. We take it as self-evident that the former approach is undesirable, as it takes what is essentially a unified phenomenon and splits it. The latter approach is less obviously wrong, but it glosses over differences between attributive and predicative adjectives. A g-structure-based analysis allows us to overcome this puzzle: since mismatches between levels can be shown to exist, and even be an integral part of a parallel-architecture conception of language, it is possible for the adjective to be a g-structure head, even if it is not a head at c-structure or f-structure.

A second puzzle concerns the agreement features. As the examples show, the Hebrew attributive adjective agrees with the head noun in number, gender, and definiteness. While agreement in number and gender is expected—agreement generally involves pronominal features, and number and gender are pronominal features—agreement in definiteness is not. In this respect, attributive adjective agreement should be contrasted with predicative adjective agreement, where number and gender agreement remain obligatory but there is no agreement for definiteness.

(24) a. Ha- bayit haya xadaš.
DEF- house was new.MSG
'The house was new.'

b. Ha- dirot hayu xadaš- ot.
DEF- apartments were new- FPL
'The apartments were new.'

c. *Ha- dirot hayu ha- xadaš- ot.
DEF- apartments were DEF- new- FPL
'The apartments were new.'

d. *Ha- dirot hayu xadaš.
DEF- apartments were new.MSG
'The apartments were new.'

This contrast in agreement features is not unique to Hebrew; it is found in other languages in which the attributive adjective agrees with the noun in definiteness (such as other Semitic languages and Scandinavian languages).

Contrasts between attributive adjective agreement and predicative adjective agreement are not limited to languages with definiteness agreement. Even in other languages, there appears to be a preference to have more agreement with attributive adjectives than with predicative adjectives. For example, in German attributive adjectives agree in gender, number, and Case, and predicative adjectives do not.

The picture that emerges from these considerations is the following. Since attributive adjectives agree with the nouns they modify, and we are modeling agreement as a property of (g-structure) heads, attributive adjectives must be represented as heads at g-structure. However, their agreement properties are different from those of predicative adjectives: they are often richer, and may involve definiteness, which is a property one expects to find on NPs, not APs. This suggests to us that the relation between the adjective and noun is closer than the relation between head and dependent. We propose that the attributive adjective is a co-head with the NP, and that the "agreement" morphology on the adjective is an extension of the nominal morphology on the head noun. The g-structures of (23d) and the adjectival predicate in (24b) are as follows:

(25) a. (=23d)

$$\begin{bmatrix} \text{LEX} & \{\text{XADAŠ, DIRA}\} \\ \text{GEND} & \text{F} \\ \text{NUM} & \text{PL} \\ \text{DEF} & + \end{bmatrix}$$

b. (=24b)

$$\begin{bmatrix} \text{LEX} & \text{XADAŠ} & \\ \text{AGMT} & \begin{bmatrix} \text{GEND} & \text{F} \\ \text{NUM} & \text{PL} \\ \text{DEF} & + \end{bmatrix} \end{bmatrix}$$

More precisely, the g-structure of the adjective-noun complex subsumes the g-structures of the noun and the adjective. The phrase structure rule adjoining the adjective copies the g-structure of the noun.

(26) $\text{NP} \rightarrow$ NP AP

$\uparrow = \downarrow$ $\quad \downarrow \in (\uparrow \text{ADJ})$

$\uparrow_\gamma = \downarrow_\gamma$ $\quad \downarrow_\gamma \subseteq \uparrow_\gamma$

The relevant elements in the lexical entries for the adjectival forms *haxadašot* and *xadašot* are:

(27) a. *haxadašot*

$(\uparrow_\gamma \text{LEX}) \subseteq \text{XADAŠ}$
$\%\text{TRIGGER} = (\uparrow_\gamma (\text{AGMT}))$
$(\%\text{TRIGGER GEND}) = \text{F}$
$(\%\text{TRIGGER NUM}) = \text{PL}$
$(\uparrow_\gamma \text{DEF}) = +$

b. *xadašot*

$(\uparrow_\gamma \text{LEX}) \subseteq \text{XADAŠ}$
$\%\text{TRIGGER} = (\uparrow_\gamma (\text{AGMT}))$
$(\%\text{TRIGGER GEND}) = \text{F}$
$(\%\text{TRIGGER NUM}) = \text{PL}$
$(\uparrow_\gamma \text{DEF}) \neq +$

3.2. Case Agreement

Another kind of agreement which is sometimes exhibited by adjectives is Case agreement, as in these examples from Modern Greek.

(28)

a.	o	arxeos	ðromos
	the.M.NOM	ancient.M.NOM	street.NOM
b.	ton	arxeo	ðromo
	the.M.ACC	ancient.M.ACC	street.ACC
c.	tou	arxeou	ðromou
	the.M.GEN	ancient.M.GEN	street.GEN

Since Case is dependent-marking, its use in what appears to be head-marking is puzzling. We will outline an analysis under which adjective agreement is a kind of extended dependent-marking, drawing heavily on the analysis of Nordlinger (1998).

Nordlinger observes that in the Australian language Wambaya, as in many dependent-marking languages, attributive adjectives agree in Case with the nouns they modify. In Wambaya, as in other Australian languages, attributive adjectives can be separated from the nouns they modify. The Case

agreement is required whether the adjective forms a constituent with the noun or not:

(29) a. [Galalarrinyi- ni bugayini- ni] gini- ng- a dawu.
dog.M- ERG big.M- ERG 3MSG.SUBJ- 1OBJ- NFUT bite
'The big dog bit me.'

b. [Galalarrinyi- ni] gini- ng- a dawu [bugayini- ni].
dog.M- ERG 3MSG.SUBJ- 1OBJ- NFUT bite big.M- ERG
'The big dog bit me.'

In the discontinuous structure, the dependent-marking nature of the Case agreement is clear. Both 'dog' and 'big' need to be marked with ergative Case so they can be identified as parts of the same dependent element. While such marking is in some sense redundant in the cross-linguistically more common continuous structure, it presumably serves the same basic function; perhaps it is an aid to parsing.

Formally, in Nordlinger's f-structure–based analysis, Case inflection carries with it an inside-out designation specifying the grammatical function that the noun bears; in our g-structure–based analysis, it carries an inside-out designation of the Case (g-structure attribute) that it marks. For example, ergative Case marking carries the following specification in the two theories:

(30) a. Nordlinger's analysis (f-structure): (SUBJ $\uparrow$)
b. Our analysis (g-structure): (ERG $\uparrow_{\gamma}$)

The extension to attributive adjectives requires an extension of the starting point of the inside-out path from $\uparrow$ to (ADJ $\uparrow$).

(31) (ERG (ADJ $\uparrow$)$_{\gamma}$)

The designation (ADJ $\uparrow$)$_{\gamma}$ points to the value of ERG in *dawu*'s g-structure. Under both Nordlinger's formulation and ours, Case marking on attributive adjectives is thus a kind of extended dependent-marking, and thus not really agreement.

Case agreement on predicative adjectives is, as Nordlinger observes, a different situation. It is not a core case of Case agreement, but rather an extension of it. The functional purpose of extending Case agreement to predicate adjectives is clear, but as a more peripheral use of the construction, we might expect to find languages in which attributive adjectives agree in Case but predicate adjectives do not. One such language is Classical Arabic (Ferguson & Barlow 1988: 11), where the Case on the predicate adjective is governed by the verb. The usual Case is nominative, but the verb *kaana* 'be' governs accusative.

(32) a. al- muʕallim- i al- jadiid- i
DEF- teacher(MSG)- GEN DEF- new.MSG- GEN
'of the new teacher'
b. Muḥammad- u kabiir- u.
Muhammad- NOM old- NOM
'Muhammad is old.'
c. Kaana Muḥammad- u kabiir- an/*un.
was Muhammad- NOM old- ACC/*NOM
'Muhammad was old.'

In Lithuanian (Timberlake 1988) agreement of predicate adjectives in Case is optional, and governed by semantic factors.

(33) a. Karas padarė jį ?neturtingą / neturtingu.
war made him.ACC ?poor.ACC / poor.INSTR
'The war made him poor.'
b. Visi įtarė jį kaltą / kaltu.
all suspect him.ACC guilty.ACC / guilty.INSTR
'Everyone suspected him of being guilty.'

We take this to confirm the correctness of an analysis which treats this kind of agreement as more peripheral than that of attributive adjectives. Formally, we analyze it as combining head- and dependent-marking (formally, outside-in and inside-out designators):

(34) (ERG ($\uparrow_{\gamma}$ AGMT))

As long as ERG is an attribute of the verb and AGMT is an attribute of the adjective, the resulting g-structure does not violate Binuiqueness.

3.3. Non-subject Relative Participles

Participles used as noun-modifiers are a relative-clause-like construction in which the participial head functions similarly to adjectives.[9] As observed by Doron & Reintges (2005), a participial form which relativizes a non-subject displays a conflict in its agreement patterns. On the one hand, it could be expected to agree with its subject, like a verb. On the other hand, due to its adjective-like nature, it might be expected to agree with the relative head, i.e. the noun to which it is an adjunct.

The theory of g-structure makes specific predictions about agreement in this situation. The participial relative is an f-structure adjunct to the noun; in g-structure it is a co-head with the noun. Thus, in languages with definiteness agreement the participial relative might exhibit definiteness agreement, but it would be with its g-structure co-head: the head of the relative construction. Definiteness agreement with the subject is impossible, just as predicative items never agree in definiteness with their subjects. Case agreement is also expected to be with the head noun; while Case agreement with subjects is also possible, it is a less central use of Case agreement. On the other hand, agreement in number and gender is predicted to be an area in which languages may differ. Languages in which there is agreement with the head noun in Case and/or definiteness might be expected to prefer agreement with the subject in number and gender, in order to include more varied information in the agreement.

Doron and Reintges' survey bears this out. In Standard Arabic, the participle agrees with the relative head in Case and definiteness, and with subject in number and gender.

(35) Qaabal- tu l- marʔat- a [l- jaalis- a
meet.PERF- 1SG DEF- woman(FSG)- ACC DEF- sitting.MSG- ACC
zawj- u- haa].
husband(MSG)- NOM- POSS.3FSG
'I met the woman whose husband is sitting.'

On the other hand, in Older Egyptian there is no agreement in Case and definiteness. The participle agrees with the relative head in number and gender.

[9]These constructions do not involve relative pronouns, and the discussion here does not extend to "agreement" patterns for relative pronouns.

(36) Mxʔt tw n(j)t rʕ [fʔ- ʔ (-w)- t mʔʕt
scale.FS this.FSG of.FSG Re carry- IMPF- PASS- PRTC.FSG justice
jm- s rʕ nb]
in- 3FSG day every
'this scale of Re in which justice is carried every day'

4. Agreement, Coordination, and Linear Order

We turn now to the role of linear order of conjuncts in agreement. As noted above, a common situation is for the agreeing head to agree with the features of the closest conjunct rather than with the resolved features of the conjoined noun phrase. In standard LFG this is problematic: if the verb agrees with the SUBJ, it should reflect the features of the SUBJ, not of part of the SUBJ. While it is possible to mechanically specify the correct agreement features by, for example, allowing the coordinate structure to resolve to the features of the closest conjunct, such analyses have an artificial feel to them. We agree with Sadler's (1999) evaluation that in such analyses "the intuition that the target really does agree with the first conjunct is captured only indirectly". Sadler's proposed solution is similar to ours: agreement is represented at a level of representation distinct from f-structure: m-structure in her case, g-structure in ours.

Under the g-structure analysis, the Welsh verb will specify its agreement target as follows:

(37) $(\uparrow_{\gamma} \text{AGMT}) = f_{\gamma}, \quad f = (\uparrow \text{SUBJ}(\in)) \ \wedge \ \exists f', f' = (\uparrow \text{SUBJ}(\in)) \wedge \uparrow \prec_f f' \prec_f f$

The agreement trigger is either equal to the SUBJ or a subset of it, the former if it is a simple SUBJ, the latter if it is a coordinate structure. However, there cannot be a closer SUBJ.

This analysis puts the burden of choosing the appropriate features for agreement on the head, rather than on the coordination construction itself. This is correct from the conceptual perspective, since agreement is about the relation between head and dependent. It also allows us to make sense of the pattern of adjective agreement in Portuguese noun phrases (Sadler & Villavicencio 2005).

In Portuguese noun phrases, both feature resolution and closest conjunct agreement are possible. The two may be mixed in a single adjective, with closest conjunct agreement for gender and feature resolution for number.

(38) a. o sofrimento e a experiência vividas
the suffering(MSG) and the experience(FSG) lived.FPL
'the lived suffering and experience'

b. … para um país com fome de capitais e tecnologia
to a country with hunger for capital(MPL) and technology(FSG)
externas
external.FPL
'…to a country in need of external capital and technology'

Under the analysis proposed here, the adjective can choose its features individually from different noun phrases, rather than picking a single noun phrase with which to agree. This can be accomplished by an adjective-introducing phrase structure rule such as the following. The local variable %CLOSEST is defined in terms of the closest nominal to which the AP is an adjunct, and the linear order of the choices represents what Sadler and Villavicencio report to be the preferred option.

(39) $\text{NP} \rightarrow \underset{\uparrow=\downarrow}{\text{NP}} \quad \underset{\downarrow \in (\uparrow \text{ADJ})}{\text{AP}}$

$$\%\,\text{CLOSEST} = f_\gamma,\ f = (\text{ADJ}\ \downarrow) \wedge \nexists f',\ f' = (\text{ADJ}\ \downarrow) \wedge f \prec_f f' \prec_f \downarrow$$
$$(\downarrow_\gamma \text{ GEND}) = (\%\,\text{CLOSEST GEND}) \vee (\uparrow_\gamma \text{ GEND})$$
$$(\downarrow_\gamma \text{ NUM}) = (\uparrow_\gamma \text{ NUM}) \vee (\%\,\text{CLOSEST NUM})$$

Prenominal adjectives are similar, although there is a greater tendency for number agreement to also be with the closest conjunct. Naturally, the closest conjunct is the last one for postnominal adjectives and the first one for prenominal adjectives.

Since each AP picks its agreement features independently, it is also possible to have different agreement on different elements within the same noun phrase. Sadler and Villavicencio report such examples involving prenominal determiners and postnominal adjectives.

(40) a. os mitos e lendas brasileiras
the.MPL myths(MPL) and legends(FPL) Brazilian.FPL
'the Brazilian myths and legends'

b. os programas e instituições brasileiras
the.MPL programs(MPL) and institutions(FPL) Brazilian.FPL
'the Brazilian programs and institutions'

Villavicencio (personal communication) reports that intuitions vary on whether this can be done with both a prenominal and postnominal adjective. There is a tendency for speakers to interpret such a sentence with each adjective scoping over only the closer noun. However, it is apparently possible for some speakers to interpret this with each adjective scoping over the entire coordination.

(41) nas gélidas salas e corredores escuros do
in.the.FPL frozen.FPL rooms(FPL) and corridors(MPL) dark.MPL of.the
Centro de Arte Rainha Sofia
Center of Arts Queen Sofia
'in the dark, frozen rooms and corridors of the Queen Sofia Arts Center'
'in the frozen rooms and dark corridors of the Queen Sofia Arts Center'

5. Syntactic and Semantic Agreement

The approach we are taking to agreement distances it somewhat from the core of the syntactic system. The well-known fact that agreement is sometimes determined by semantic properties rather than morphosyntactic features suggests that this is correct.

The semantic aspect of agreement can be seen, for example, in the preferred agreement patterns in Portuguese. Portuguese prefers feature resolution in coordination for number; this is presumably a consequence of the fact that coordinated elements are semantically plural. Since a coordination of a masculine and a feminine is not semantically masculine, there is no semantic pressure to disprefer closest conjunct agreement.

Another set of cases of this kind is discussed by Corbett (1988) in Slavic languages. The Russian noun *vrač* 'doctor' is formally masculine, but may refer to a woman. When it does, both (morphosyntactic) masculine and (semantic) feminine agreements are possible, with the morphosyntactic agreement preferred for attributive modifiers and semantic agreement for predicates.

(42) Novyj vrač skazala …
new.MSG doctor(M) said.FSG
'The new woman doctor said …'

The Serbo-Croatian nouns like *gazda* 'master' are morphosyntactically feminine but semantically masculine. In the singular masculine agreement is used, while in the plural both are possible. Stacked modifiers may even exhibit different agreement, with the one closer to the head agreeing morphosyntactically and the one farther from the head agreeing semantically.

(43) a. naši / naše gazde
our.MPL / our.FPL masters(F)
'our masters'

b. ovi privatne zanatlije
these.MPL private.FPL artisans(FPL)
'these private artisans'

We have already seen the possibility of different agreements on attributes to the same noun in Portuguese. Since each adjective-noun combination is a distinct g-structure head, this does not pose a challenge.

6. A Beginning…

As noted at the outset (and by one of the reviewers of this paper), we have not covered the full range of agreement phenomena in language. However, we believe that the phenomena we *have* discussed support the approach to agreement (and Case) that we have proposed here.

The core concept at the center of LFG's approach to the structure of language is the idea that language consists of multiple parallel dimensions of representation. The levels of structure hypothesized in LFG analyses model these dimensions. The claim of this paper is that the marking of head-dependent relations is one such dimension of linguistic structure, and that where LFG analyses of Case and agreement have gone wrong in the past in not recognizing the status of grammatical marking, seeing it instead as part of the network of functional relations.

It is our hope that this paper will serve as the beginning of new studies of Case and agreement phenomena, and thus will be a positive addition to the ongoing discussion within LFG concerning the proper analysis of agreement.

References

Andrews, Avery (1982) "The Representation of Case in Modern Icelandic." in Joan Bresnan, ed., *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press. 427–503.

Aoun, Joseph, Elabbas Benmamoun, and Dominique Sportiche (1994) "Agreement, Word Order, and Conjunction in Some Varieties of Arabic." *Linguistic Inquiry* 25: 195–220.

Butt, Miriam (1993) "Object Specificity and Agreement in Hindi/Urdu." *CLS* 29: 89–103.

Comrie, Bernard (1989) *Language Universals and Linguistic Typology* (2nd edition). Chicago: University of Chicago Press.

Corbett, Greville G. (1988) "Agreement: A Partial Specification Based on Slavonic Data." in Michael Barlow and Charles A. Ferguson, ed., *Agreement in Natural Language: Approaches, Theories, Descriptions*. Stanford, Calif.: Center for the Study of Language and Information. 23–53.

Dalrymple, Mary, and Ronald M. Kaplan (2000) "Feature Indeterminacy and Feature Resolution." *Language* 76: 759–798.

Dalrymple, Mary, and Irina Nikolaeva (2005) "Agreement and Discourse Function." LFG 05.

Doron, Edit, and Chris H. Reintges (2005) "On the Syntax of Participial Modification." ms.

Falk, Yehuda N. (2004) "The Hebrew Present-Tense Copula as a Mixed Category." in Miriam Butt and Tracy Holloway King, ed., *Proceedings of the LFG 04 Conference, University of Canterbury*. On-line: CSLI Publications. 226--246. http://cslipublications.stanford.edu/LFG/9/lfg04.html

Ferguson, Charles A., and Michael Barlow (1988) "Introduction." in Michael Barlow and Charles A. Ferguson, ed., *Agreement in Natural Language: Approaches, Theories, Description*. Stanford, Calif: Center for the Study of Language and Information. 1–22.

Nichols, Johanna (1986) “Head-Marking and Dependent-Marking Grammar.” *Language* 62: 56–119.
Nordlinger, Rachel (1998) *Constructive Case: Evidence from Australian Languages*. Stanford, Calif.: CSLI Publications.
Otoguro, Ryo (2005) “Agreement Path in Icelandic.” Presented at LFG 05 conference, University of Bergen.
Sadler, Louisa (1999) “Non-Distributive Features in Welsh Coordination.” in Miriam Butt and Tracy Holloway King, ed., *Proceedings of the LFG 99 Conference, The University of Manchester*. On-line: CSLI Publications. http://cslipublications.stanford.edu/LFG/4/lfg99.html
Sadler, Louisa, and Aline Villavicencio (2005) “Agreement, Coordination, and Portuguese NPs.” Presented at LFG 05 conference, University of Bergen. handout at: http://privatewww.essex.ac.uk/~louisa/agr/NPagreement.html
Timberlake, Alan (1988) “Case Agreement in Lithuanian.” in Michael Barlow and Charles A. Ferguson, ed., *Agreement in Natural Language: Approaches, Theories, Descriptions*. Stanford, Calif.: Center for the Study of Language and Information. 181–199.
Ziv, Yael (1976) “On the Reanalysis of Grammatical Terms in Hebrew Possessive Constructions.” in Peter Cole, ed., *Studies in Modern Hebrew Syntax and Semantics: The Transformational-Generative Approach*. Amsterdam: North Holland.

GF-DOP: Grammatical Feature Data-Oriented Parsing

Ríona Finn, Mary Hearne, Andy Way and Josef van Genabith

National Centre for Language Technology,
School of Computing,
Dublin City University

Proceedings of the LFG06 Conference

Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006

CSLI Publications

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

Abstract

This paper proposes an extension of Tree-DOP which approximates the LFG-DOP model. GF-DOP combines the robustness of the DOP model with some of the linguistic competence of LFG. LFG c-structure trees are augmented with LFG functional information, with the aim of (i) generating more informative parses than Tree-DOP; (ii) improving overall parse ranking by modelling grammatical features; and (iii) avoiding the inconsistent probability models of LFG-DOP. In a number of experiments on the HomeCentre corpus, we report on which (groups of) features most heavily influence parse quality, both positively and negatively.

1 Introduction

This paper proposes an extension of Tree-DOP [e.g. Bod, 1992, 1998] which approximates the LFG-DOP model [Bod and Kaplan, 1998, 2003]. GF-DOP combines the robustness of the DOP model with some of the linguistic competence of LFG. LFG c-structure trees are augmented with LFG functional information, with the aim of (i) generating more informative parses than Tree-DOP; (ii) improving overall parse ranking by modelling grammatical features; and (iii) avoiding the inconsistent probability models of LFG-DOP. In a number of experiments on the HomeCentre corpus, we report on which (groups of) features most heavily influence parse quality, both positively and negatively.

The remainder of the paper is organised as follows. In section 2, we describe the Tree-DOP model. An overview of LFG-DOP is provided in section 3, and the new GF-DOP model is described in section 4. We motivate the experiments carried out on the HomeCentre corpus in section 5, and provide results and evaluation in section 6. Finally, we conclude and list a number of avenues for further work.

2 Tree-DOP

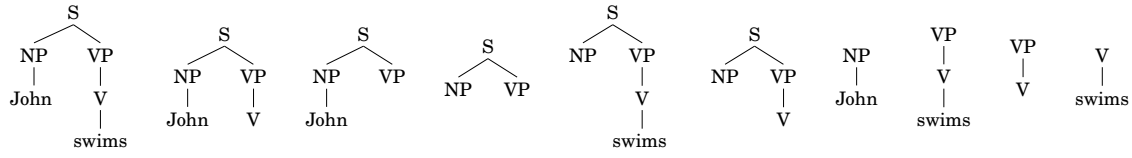
Data-oriented models of language [e.g. Bod, 1992, 1998] are based on the assumption that humans perceive and produce language by availing of previous language experiences rather than abstract grammar rules. These models exploit large treebanks comprising linguistic representations of previously occurring utterances. Analyses of new input sentences are produced by combining fragments from the treebank; the most probable analysis is determined using the relative frequencies of these fragments.

The tree fragments used in Tree-DOP are called subtrees. Two decomposition operators are used in order to produce subtrees from sentence representations:

1. the *root operator* which takes any node in a tree to be the root of a subtree and deletes all nodes except this new root and all nodes dominated by it;

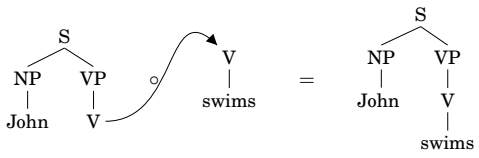
2. the *frontier operator* which selects a (possibly empty) set of nodes in the newly created subtree, excluding the root, and deletes all subtrees dominated by these nodes.

As an example, the complete set of DOP fragments which can be derived from the representation of *John swims* is shown in (1).



(1)

Representations for new input are formed by combining other fragments using the composition operator, namely leftmost substitution, which ensures that each derivation in DOP is unique. The composition of trees t_1 and t_2 ($t_1 \circ t_2$) is only possible if the leftmost frontier node of t_1 and the root node of t_2 are of the same category. The resulting tree is a copy of t_1 where t_2 has been substituted at its leftmost nonterminal frontier node, as demonstrated in (2).



(2)

The probability of a derivation is the joint probability of choosing each of the subtrees involved in that derivation. Letting $|e|$ be the number of times subtree e occurs in the corpus and r(e) be the root node category of e, the probability assigned to e is as in (3).[1]

$$P(e) = \frac{|e|}{\sum_{u:r(u)=r(e)} |u|} \tag{3}$$

The probability of a derivation is the product of the probabilities of choosing each of the subtrees involved in that derivation. Thus, the probability of a derivation $t_1 \circ ... \circ t_n$ is given by (4).

$$P(t_1 \circ ... \circ t_n) = \prod_i P(t_i) \tag{4}$$

A parse tree can potentially be generated by many different derivations, each of which has its own probability of being generated. Therefore, the probability of a parse tree T is the sum of the probabilities of its distinct derivations as in (5).

$$P(T) = \sum_{D \ derives \ T} P(D) \tag{5}$$

3 LFG-DOP

The LFG-DOP model differs from the Tree-DOP model in that the assumed corpus is annotated with LFG representations, i.e. <c,ϕ,f> triples comprising c-structure, ϕ-links and f-structure. The definitions of the fragmentation operators and composition operator, along with the probability model, must be adapted accordingly.

The fragmentation operators for LFG-DOP are extensions of those used in Tree-DOP as we wish to extract exactly the same set of generalised c-structure fragments as before. However, we also wish to extract the corresponding f-structure fragment to go with each c-structure. Consequently, the original *root* and *frontier* operators must be extended to take f-structure into account. Many different extensions can be envisaged; those defined in [Bod and Kaplan, 1998, 2003, Bod, 2000b,a] are as follows:

Root Given a copy of the example-base representation <c,ϕ,f> named <c_{copy},ϕ_{copy},f_{copy}>:

1. select a node in c_{copy} to be *root* and delete all nodes except this node and the nodes it dominates;
2. delete all links in ϕ_{copy} which link deleted c-structure nodes to f_{copy};

[1] We use this estimation method for all experiments presented in this paper, although others (e.g. [Zollmann and Sima'an, 2005]) are possible.

3. delete all f-structure units in f_{copy} which are not ϕ-*accessible* from c_{copy};

4. delete all semantic forms in f_{copy} which are local to f-structure units corresponding to erased c-structure terminals.

Frontier Given a representation of the form $<c_{copy},\phi_{copy},f_{copy}>$ created by the root operation:

1. select a (possibly empty) set of nodes in c_{copy} to be *frontier* nodes and delete all nodes dominated by these newly-created frontier nodes;

2. delete all links in ϕ_{copy} which link deleted c-structure nodes to f_{copy};

3. *(delete all f-structure units in f_{copy} which are not* ϕ-accessible *from c_{copy};)*

4. delete all semantic forms in f_{copy} which are local to f-structure units corresponding to erased c-structure terminals.

Step 3 of the frontier operation is given here for the sake of completeness; as a consequence of the definition of ϕ-accessibility given in [Bod and Kaplan, 1998, 2003, Bod, 2000b,a] and described below, the root node of c_{copy} (selected during the root operation) accesses all f-structure units in f_{copy} regardless of which nodes are selected by the frontier operation. This definition of ϕ-accessibility is as follows:

ϕ-accessibility An f-structure unit f is ϕ-accessible from c-structure node n if and only if

1. n is ϕ-linked to f, i.e. $\phi(n) = f$, or

2. n is ϕ-linked to f_x and f contains f_x i.e. there is a chain of attributes leading from f to f_x.

The extended root and frontier operations for LFG-DOP yield precisely the same c-structures as are yielded by the root and frontier operations defined for Tree-DOP. However, it is also possible to extract further fragments from each fragment yielded by the root and frontier operations via the *discard* operation. Discard is used to delete attribute-value pairs from the f-structure whose values are not ϕ-linked to remaining nodes in the c-structure and are not surface forms corresponding to c-structure terminals. Thus, when discard is applied to any fragment $<c,\phi,f>$, a new fragment $<c,\phi,f_{d_x}>$ is extracted; the c-structure and ϕ-links are unchanged but f_{d_x} differs from f in that all attribute-value pairs in f_{d_x} are also in f but the reverse does not hold, i.e. it is not the case that all attribute-value pairs in f are also in f_{d_x}.

The LFG-DOP composition operation involves two stages: leftmost substitution over c-structure and recursive unification over f-structure such that the ϕ-links are not broken. That is, a pair of c-structures are first composed exactly as for Tree-DOP and, subsequently, the f-structure parts of those fragments are unified. Any sequence of composition operations yielding a complete derivation (i.e. one which contains no open c-structure substitution sites) is only valid if that derivation's f-structure adheres to the LFG well-formedness conditions. However, the presence in the fragment base of discard-generated fragments means that many input strings which are ill-formed with respect to the corpus can also be parsed.

We can estimate LFG-DOP fragment probabilities as for Tree-DOP, i.e. compute their empirical frequencies conditioned on the root node. However, this estimator draws no distinction between fragments generated by the root and frontier operators and discard-generated fragments. Fragments generated by discard effectively relax the constraints specified in the f-structure to allow the fragment to be used in a wider variety of contexts and so are very useful in constraint-based DOP

parsing. Intuitively, however, they should only be considered when no parse can be produced which satisfies all relevant constraints, i.e. they should be used only when the input is ill-formed with respect to the corpus. Consequently, this estimator does not seem entirely appropriate. Way [1999] and Bod and Kaplan [2003] propose an alternative method – termed 'discounted RF', as opposed to 'simple RF' – of estimating fragment probabilities whereby root and frontier fragments are treated as seen events and discard fragments as unseen events. The fragment set is partitioned using this distinction and two separate probability distributions induced. The probabilities of seen events are estimated by their relative frequencies as before. However, these probabilities are then discounted and the discounted mass distributed amongst the unseen events, i.e. the discard-generated fragments.

In the Tree-DOP model, valid derivations are constructed by composing fragments such that the category-matching condition is fulfilled. Each valid derivation is assigned a probability by calculating the product of the probabilities of the fragments used in the construction of that derivation. The probabilities of all valid derivations which can be constructed from a given DOP grammar for all of the strings which it recognises sum to 1.

Each LFG-DOP derivation probability is also calculated as the product of the probabilities of the fragments used in the construction of that derivation. In the LFG-DOP model, however, we have seen that valid derivations are constructed by composing fragments such that the category-matching, uniqueness, completeness and coherence conditions are fulfilled. If we calculate LFG-DOP fragment probability distributions in the same way as we did for Tree-DOP – i.e. define distributions over root node category – then the probabilities of all derivations for all the strings recognised by the grammar which adhere to the category-matching condition will sum to 1. However, it is not the case that all of these derivations are valid according to the LFG-DOP model as they may not fulfil the uniqueness, completeness and coherence conditions. Consequently, the probabilities of all *valid* derivations which can be constructed from a given LFG-DOP grammar for all of the strings which it recognises no longer sums to 1 and, therefore, do not constitute a probability distribution. [Bod and Kaplan, 1998, 2003] handle this by normalisation: parse probability is divided by the sum of the probabilities of all *valid* parses for the input string. However, Abney [1997] observes that normalisation serves only to mask the fact that, unlike for the context-free case, establishing probabilities for grammars encoding context-sensitive dependencies using relative frequency estimation does not yield the best weights.

4 GF-DOP

The GF-DOP model can be seen as an extension of the Tree-DOP model, and an approximation towards LFG-DOP. It combines the robustness of the DOP model with some of the linguistic competence of LFG. This model exploits a treebank transformed by the addition of further linguistic information: features are extracted from f-structures and appended to the c-structure category labels to form a new, extended set of c-structure category labels. As this model extends the Tree-DOP model, category-matching is the only restriction imposed on fragments which are candidates for composition. No restrictions are placed on the category labels, so labels which incorporate features incur no extra processing and no changes to the model are required to handle the new set of extended category labels. The Tree-DOP model is applied to the transformed treebank. This model can be as accurately and efficiently implemented as the Tree-DOP model, and produces linguistically informed output based on identification and incorporation of grammatical functions and features.

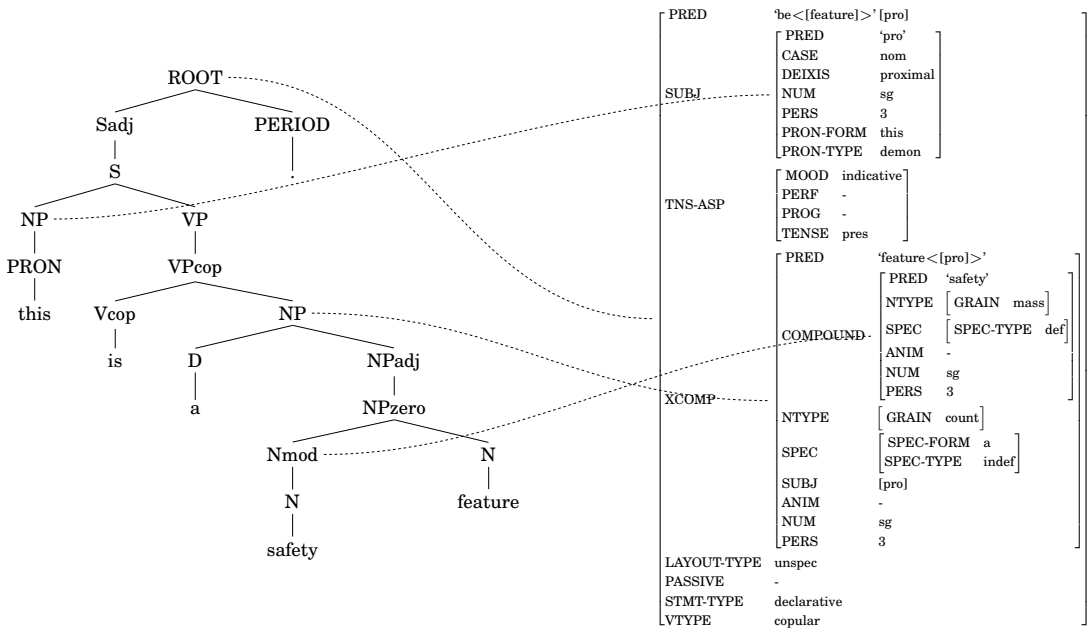


Figure 1: A c-structure with its corresponding ϕ-linked f-structure, from which we extract features.

4.1 Feature Classification

F-structures contain informative features (for example LAYOUT-TYPE may specify that this sentence is a header, a list item or is unspecified) and functional information (such as SUBJ and OBJ) which describe the grammatical functions of the constituents in question. A linked c-structure and f-structure representation can be seen in Figure 1; this representation contains examples of each of the 5 feature classes identified:

- grammatical functions, e.g. SUBJ, XCOMP,
- atomic features, e.g. NUM=sg, PERS=3,
- lexical features, e.g. PRON-FORM=this, SPEC-FORM=a,
- 'super features' (or non-grammatical function features) which have an f-structure containing a group of features as their values,
  e.g. TNS-ASP[MOOD=imperative, PERF=-, PROG=-], NTYPE[GRAIN=count],
- predicates, e.g. PRED 'be<[feature]>'[pro], PRED 'feature<[pro]>'.

4.2 Annotating with Grammatical Features

4.2.1 Grammatical functions

Using ϕ-linked f-structure units, we identify functions of constituents within the c-structure. For example, the leftmost NP in the c-structure representation in Figure 1 functions as the SUBJ of *'be'*. We transform the tree by appending this information to the syntactic category label, giving 'NP_SUBJ'. We place the annotation on the topmost node in the constituent which corresponds to the

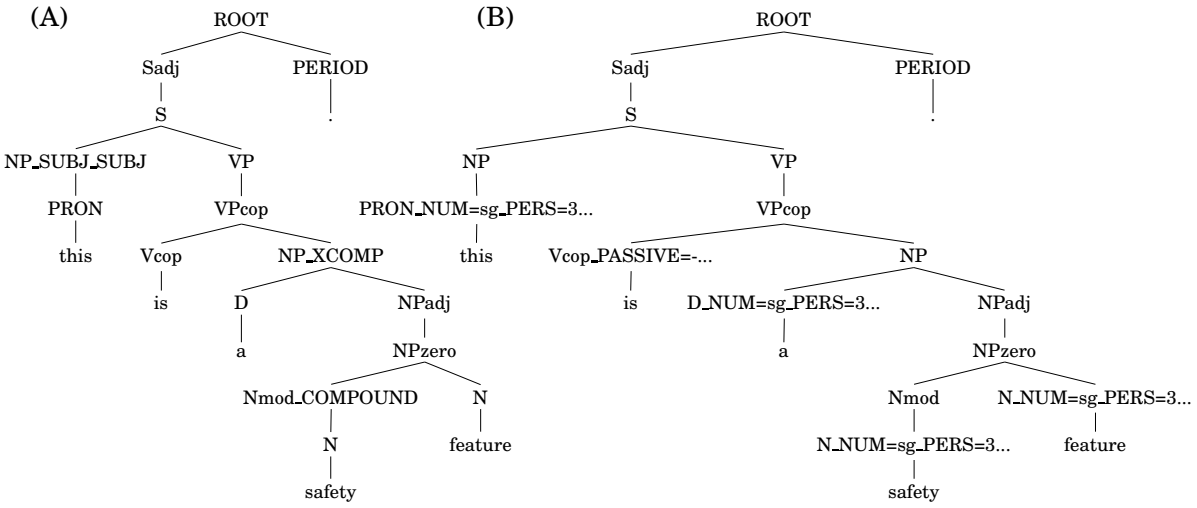


Figure 2: (A) Illustration of a c-structure with functional annotations on the top-most nodes of appropriate constituents. (B) Illustration of a c-structure with atomic annotations on the preterminal nodes.

function in question. All nodes dominated by this annotated node are part of the constituent which fulfils this function.

Where a constituent fulfils more than one function in the sentence, we append a label for each function to the top-most node in the constituent which serves that function. Upon further examination of the f-structure, we see that the NP node also functions as the SUBJ of *'feature'*; this label becomes 'NP_SUBJ_SUBJ'. A c-structure annotated with functions can be seen in Figure 2 (A).

4.2.2 Atomic features

The second class of features is atomic features. These features have a small set of closed class items as possible values; for example the feature PERS can only ever have the value 1, 2 or 3. We annotate pre-terminal nodes with atomic features, as these nodes are closest to the terminals to which the annotations specifically apply. A single atomic feature may apply to more than one node; in this case, each such node receives the atomic annotation.

Looking at the ϕ-linked f-structure for the sentence in Figure 1, we see that the outermost f-structure is linked to the ROOT node, which dominates all other nodes. If we consider the features which lie within this f-structure unit, but outside other inner units, it might appear that the features LAYOUT-TYPE, PASSIVE, STMT-TYPE and VTYPE should be annotated on all pre-terminal nodes, even to those which, logically, we know to be unrelated; for instance, we know that determiners, such as the terminal *a*, do not have a PASSIVE quality. However, this does not occur in a practical implementation of the GF-DOP model. Nodes which correspond to inner f-structure units are ϕ-linked to their respective f-structure units, rather than the outermost unit which dominates them. In the c-structure shown in Figure 1, only the Vcop node receives these annotations, as illustrated in Figure 2 (B).

Although the pre-terminal PERIOD is also dominated by this f-structure unit, and not ϕ-linked to any other unit, we do not annotate pre-terminals of punctuation.

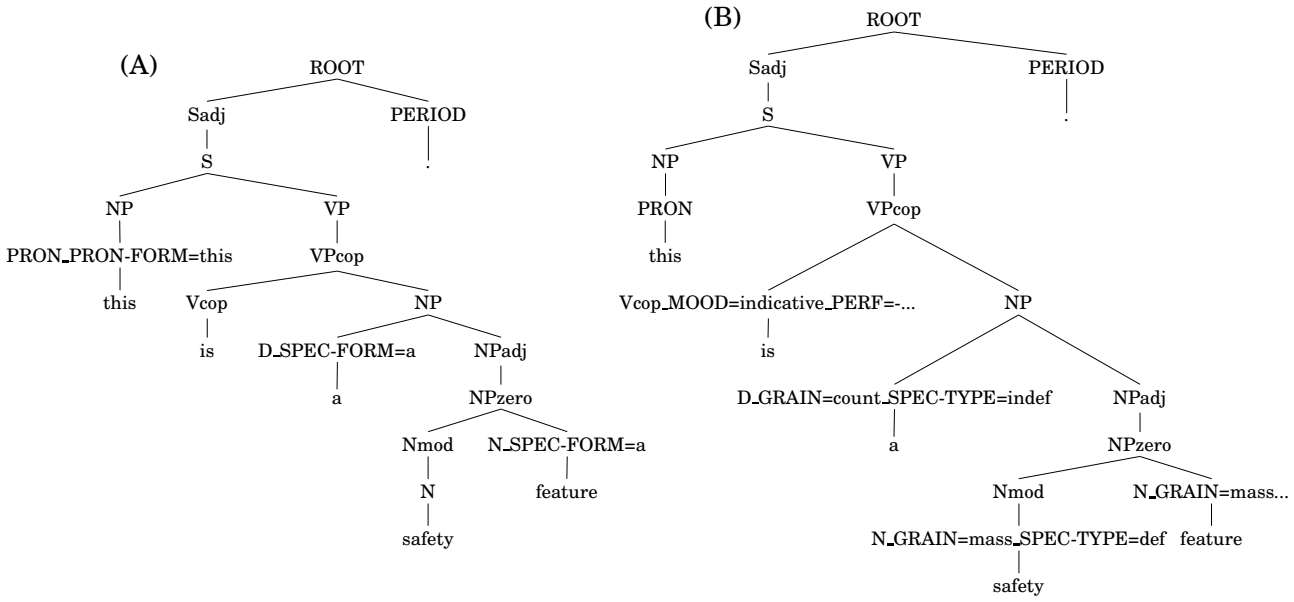


Figure 3: (A) Illustration of a c-structure with lexical annotations on pre-terminal nodes. (B) Illustration of a c-structure with super-feature value atomic annotations on pre-terminal nodes.

4.2.3 Lexical features

The third class of features is lexical features. These features have one of a small number of lemmas as their values; for example, CONJ-FORM can have one of *and*, *or*, *and-or*, *then*, *plus* or *null* as its value. Lexical annotations are placed on the pre-terminal nodes dominating the terminals they relate to. An example of a lexically annotated c-structure can be seen in Figure 3 (A); the PRON-FORM is specified as *this*. The SPEC-FORM used with *feature* is specified as *a*, also indicated on this c-structure. Where a lexical feature applies to two or more nodes, each node is annotated with this feature.

4.2.4 Super-features

We call the fourth class 'super-features'. These features have a set of atomic features as their value. Intuitively, it is more useful to annotate the node with the contents of the super-feature's f-structure value: rather than identifying that a node has, for example, tense and aspect, denoted by the feature TNS-ASP, we annotate it with the features which define the tense and aspect: Vcop_MOOD=indicative_PERF=-_PROG=-_TENSE=pres. These features are added in the same manner as the other atomic features, as described in section 4.2.2. An example of these annotations can be seen in Figure 3 (B); the features which describe the tense and aspect are annotated on the Vcop node, rather than on the VP or VPcop nodes. If these features were added to the VP or VPcop nodes, it would imply that the other constituents dominated by VP or VPcop also have tense and aspect.

4.2.5 Predicates

A final, single feature class contains the PRED feature. This feature has a lemma as its value, but it differs from the lexical features described in section 4.2.3 because lexical features can have only

a small number of lemmas, essentially a closed class set, as their values, while PRED can have any lemma as its value. The PRED feature may also have subcategorisation arguments; this is a list of arguments which are required by the predicate.

Let us consider the annotation possibilities for this feature. From the PRED we can establish the lexical word, the list of obligatory arguments, and adjuncts. There is perhaps no great advantage in extracting the lexical word from the value, as this word features in the c-structure as a terminal. However, the list of arguments might be used to specify the context in which this word can appear. For example, if we encounter a sentence with the word *eat*, we might use the subcategorisation requirements to check that the sentence also has some node which is labelled SUBJ of *eat* and a node labelled OBJ of *eat*. However, the GF-DOP model presented here does not make use of subcategorisation information.

4.3 Preserving Robustness

Data-sparseness is a prominent issue in parsing and is exacerbated by the highly specific node labels in GF-DOP. The GF-DOP model's use of additional feature information could lead to reduced coverage, i.e. there may be sentences which can be parsed by the Tree-DOP model but not the GF-DOP model. This would constitute a weakness in the GF-DOP model. To address this issue and preserve robustness, we further extend the GF-DOP model by introducing a 'backing-off'.

We achieve this via a two-step training procedure: we extract grammars from both the annotated treebank and a version of the treebank with the annotations stripped away. We merge these grammars by assigning the majority of the probability mass (W_1) to the annotated grammar and the remainder ($W_2 = 1 - W_1$) to the unannotated grammar. This ensures that the GF-DOP model maintains the same level of coverage as the Tree-DOP model.

4.4 How does GF-DOP improve on Tree-DOP?

When compared to the Tree-DOP model, GF-DOP has the following advantages:

1. it has the capacity to generate more informative parses;
2. its probability model is sensitive to grammatical feature information which can help to improve the overall parse ranking;
3. it displays the above advantages without losing any of the coverage of the Tree-DOP model.

The Tree-DOP model is limited by the representations it assumes. The GF-DOP model trains on data with a greater degree of linguistic information than the Tree-DOP model and, consequently, the parses it generates contain more information than those of Tree-DOP. This additional information also plays a part in determining the basic phrase-structure tree assigned to the input string: better parses which are supported by the additional grammatical feature information will have their probabilities boosted and, correspondingly, those which are not supported by these features will be ranked lower. Finally, we note that these advantages do not, as might have been expected, come at a cost in terms of robustness. Back-off is integral to the model and it ensures that all sentences which can be parsed by Tree-DOP can also be parsed by GF-DOP.

4.5 How does GF-DOP improve on LFG-DOP?

The GF-DOP model also improves on the LFG-DOP model:

1. Due to difficulties in establishing a valid probability model, there is currently no satisfactory realisation of an LFG-DOP system. In contrast, the GF-DOP model maintains the integrity of the original Tree-DOP probability model.

2. The implementation of discard in LFG-DOP is computationally expensive: exponentially many more fragments are generated than for Tree-DOP. Although the integration of back-off into the GF-DOP model increases the number of fragments, the resulting grammar contains at most double the number of fragments of the Tree-DOP model.

LFG-DOP's strength comes from the information contained in the assumed representations and the corresponding extensions to fragmentation and composition. The unification of features ensures well-formed grammatical parses are generated. However, not all LFG-DOP derivations unify globally, or they may fail to meet (one or more of) the three well-formedness conditions required to produce a valid parse. Because these ill-formed derivations are excluded as they are encountered, probability mass is lost; the probability distribution of derivations does not correspond to the probability model.

The GF-DOP model, in contrast, enforces only category-matching during composition. As a result, only valid derivations are constructed. In this way, we make use of available feature and functional information, while avoiding the probabilistic difficulties which arise due to the generation of invalid parses.

A correspondence can be drawn between the 'backing-off' technique employed in the GF-DOP model and the 'discounted RF' technique employed in the LFG-DOP model in that both assign a limited proportion of the available probability mass to fragments which are unseen in the treebank. There is, however a crucial difference: exponentially many fragments are generated using discard because all possible combinations of attribute/value pair deletion are applied whereas this is not the case for GF-DOP back-off because all feature annotations are deleted simultaneously. This renders the GF-DOP model less powerful but more computationally tractable.

4.6 The GF-DOP hypothesis

Having outlined the theoretical characteristics and advantages of the GF-DOP model, we hypothesise that this new model will:

1. give us better phrase-structure tree parse accuracy than the Tree-DOP model and

2. allow us to learn grammatical features with a high degree of accuracy.

5 Experiments

We have carried out a set of experiments in order to determine whether the theoretical advantages of GF-DOP outlined in Section 4 are reflected in the performance of the model. In this section, we describe the data and parser used to carry out these experiments.

5.1 The Data

The treebank used in the experiments presented here is the English section of the Xerox HomeCentre corpus. This treebank consists of 980 sentences annotated with c-structures and corresponding ϕ-linked f-structures. In this data-set, 75 features (excluding PRED) were identified. These features were divided into 4 classes. This classification can be seen in Table 1. As stated in section

Functions		Atomic Features				Lexicalised Features	Super-Features
ADJUNCT	OBL-AGT	ABBREV	DEIXIS	NUM	PSEM	COMP-FORM	ARG-EXT
APP	OBL-COMP	ACONSTR	EMPH	NUMBER-TYPE	PTYPE	CONJ-FORM	ARGS-INT
COMP	PRON-INT	ADEG-DIM	EMPHASIS	PASSIVE	SPEC-TYPE	CONJ-FORM-COMP	ASPECT
COMP-EX	PRON-REL	ADEGREE	FIN	PERF	STMT-TYPE	PCASE	DEP
COMPOUND	SUBJ	ADJUNCT-TYPE	GEND	PERS	TEMPORAL	PRECONJ-FORM	NON-DEP
OBJ	TOPIC-INT	ADV-TYPE	GERUND	POL	TENSE	PREDET-FORM	NTYPE
OBJ2	TOPIC-REL	ANIM	GRAIN	PREDET-TYPE	TIME	PRON-FORM	PREDET
OBL	XCOMP	ATYPE	INF	PROG	TYPE	PRT-FORM	SPEC
		AUX-FORM	LAYOUT-TYPE	PRON-TYPE	VFORM	SPEC-FORM	TNS-ASP
		CASE	MOOD	PROPER	VTYPE		
			NEG-FORM				

Table 1: Classification of 75 features identified in the English section of the data set.

4.2.4, we do not annotate with super-features explicitly; rather we use the features listed within their f-structure values. In addition to these, there are 5 other features we do not use:

- AUX-FORM: although this feature is a form like most of the lexical features, only one value is possible: *contracted*. This feature is used to indicate that an auxiliary form is contracted, for example *here's* rather than *here is*, or *you're* instead of *you are*. This feature occurs only 11 times in the data-set. We manually 'cleaned up' the corpus by removing all contracted forms from the c-structures, thus this f-structure feature is no longer relevant. In addition, this step helps slightly counteract the effect of data-sparseness.

- NEG-FORM: like AUX-FORM, NEG-FORM has *contracted* as its only value. This feature works in the same way as AUX-FORM: it indicates that a negative form has been contracted, for example *doesn't* rather than *does not*, or *don't* in place of *do not*. This feature occurs only 14 times in the data-set. We removed occurrences of contracted negative forms from the c-structures, making this f-structure feature redundant, and again modestly reducing the effects of data-sparseness.

- VFORM: despite this feature being called a form, it appears to behave more like an atomic feature in that it has a small set of non-lexical values: *presp*, *base*, *passp* and *perfp*. Upon examination of the corpus, we found that this feature was contained in f-structure units which were neither linked to the main f-structure unit, nor to any c-structure nodes. As this feature is not connected to c-structure nodes either directly, via ϕ-links, or indirectly, through another f-structure unit which is ϕ-linked to some c-structure node, we do not generate a corpus annotated with this feature. Any such corpus would essentially be the same as the baseline (original, unannotated) corpus.

- FIN: this atomic feature occurs in f-structure units which are not linked to the main f-structure, and are not linked to any c-structure nodes. Thus we do not generate a corpus annotated with this feature.

- INF: this atomic feature occurs in the same situations as FIN, i.e. in f-structure units which are not linked to the main f-structure, or linked to c-structure units. We do not generate a corpus annotated with this feature.

Thus, the number of features we use in generating treebanks annotated with only a single feature is 61, and these features are listed and classified in Table 2. In addition, we generate 3 multi-feature

Functions		Atomic Features				Lexicalised Features
ADJUNCT	OBL-AGT	ABBREV	DEIXIS	NUMBER-TYPE	PSEM	COMP-FORM
APP	OBL-COMP	ACONSTR	EMPH	PASSIVE	PTYPE	CONJ-FORM
COMP	PRON-INT	ADEG-DIM	EMPHASIS	PERF	SPEC-TYPE	CONJ-FORM-COMP
COMP-EX	PRON-REL	ADEGREE	GEND	PERS	STMT-TYPE	PCASE
COMPOUND	SUBJ	ADJUNCT-TYPE	GERUND	POL	TEMPORAL	PRECONJ-FORM
OBJ	TOPIC-INT	ADV-TYPE	GRAIN	PREDET-TYPE	TENSE	PREDET-FORM
OBJ2	TOPIC-REL	ANIM	LAYOUT-TYPE	PROG	TIME	PRON-FORM
OBL	XCOMP	ATYPE	MOOD	PRON-TYPE	TYPE	PRT-FORM
		CASE	NUM	PROPER	VTYPE	SPEC-FORM

Table 2: List and classification of the 61 features for which we generated singly-annotated treebanks.

treebanks: one with all the functional annotations listed in Table 2, one with the five functions most prevalent in the data (ADJUNCT, OBJ, SUBJ, COMPOUND and XCOMP) and one annotated with the SUBJ and OBJ functions.

5.2 Experimental set-up

From the training treebank we generated eight training sets of 890 sentences each and eight corresponding test sets of 90 sentences each. The splits were generated at random[2] such that every word in the test set is present in the corresponding training set, thus avoiding the issue of unknown words at this time.

For each feature presented in Table 2, an annotated corpus was created and the eight pre-established splits applied. For each split, the parser is trained on the training set, tested on the test set and evaluated on the corresponding reference set. Scores are then averaged over the eight splits for each annotation type.

5.3 Parser details

Training our DOP parser involves extracting the DOP fragment set and associating probabilities with each fragment. Testing then involves submitting one or more sentences to the parser, applying the fragment set to establish a parse forest and selecting the best parse from that forest to output. There are a number of methodologies available to us (e.g. [Bod, 1995], [Sima'an, 1995, 1999], Goodman [1996, 1998, 2003]) in implementing our DOP system. Details of the parser used to perform the GF-DOP experiments presented in this paper are given below.

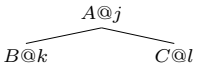
5.3.1 Training

Goodman [1996, 1998, 2003] describes a method by which the DOP grammar projected from a treebank in which all trees are binary branching is reduced to a PCFG containing at most eight rules for each node in the training data. This PCFG is equivalent to the DOP grammar in that a) it generates the same strings with the same probabilities and b) it generates the same parse trees with the same probabilities, although one must sum over several PCFG trees for each DOP tree.

Goodman PCFG-reductions are constructed as follows. Every node in every tree in the treebank is assigned a unique address: $A@k$ is the node labelled A at address k. One new non-terminal A_k

[2]The eight test sets of 90 sentences each are not disjoint; because they were extracted at random, it is entirely possible that they overlap to some extent.

is created for every node in the treebank; such non-terminals are called "interior" nodes and the original nodes "exterior" nodes. a_k is the number of subtrees with root node $A@k$ and a the number of subtrees with root node label A, i.e. $a = \sum_j a_j$. Given node $A@k$ with a set CH of two or more children $CH = \{B@l...C@m\}$, the number of fragments a_k which have root node $A@k$ is calculated by multiplying the numbers of fragments yielded by each of its children: $a_k = \prod_{X@n \in CH}(x_n + 1)$.

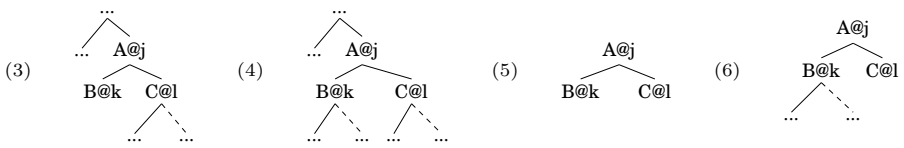


(6)

For any node grouping such as the one in (6), the eight PCFG rules and their corresponding probabilities in (7) are then extracted; Goodman provides proofs by induction that the rule probabilities are valid.

$$
\begin{array}{llllllll}
(1) & A_j & \longrightarrow & BC & (\frac{1}{a_j}) & (5) & A & \longrightarrow & BC & (\frac{1}{a}) \\
(2) & A_j & \longrightarrow & B_kC & (\frac{b_k}{a_j}) & (6) & A & \longrightarrow & B_kC & (\frac{b_k}{a}) \\
(3) & A_j & \longrightarrow & BC_l & (\frac{c_l}{a_j}) & (7) & A & \longrightarrow & BC_l & (\frac{c_l}{a}) \\
(4) & A_j & \longrightarrow & B_kC_l & (\frac{b_k c_l}{a_j}) & (8) & A & \longrightarrow & B_kC_l & (\frac{b_k c_l}{a})
\end{array}
\tag{7}
$$

These rules correspond to the eight possible contexts in which the node grouping in (6) can occur in fragments extracted from the corresponding treebank tree; each of the three nodes can be either interior or exterior (i.e. root node or substitution site) to any fragment in which the grouping occurs. The examples in (8) illustrate the contexts to which rules (3) – (6) in (7) correspond. Node *A@j* is an interior (i.e. non-root) node in rules 3 and 4 and an exterior (i.e. root) node in rules 5 and 6 – the parent node of any grouping (the node which appears on the left-hand side of the rule) corresponds to either a root or internal node but not a substitution site. Conversely, the child nodes of each grouping, which appear on the right-hand side of the corresponding rules, can be either internal nodes or substitution sites but never root nodes as shown in (8). As previously stated, Goodman's PCFG reduction requires the projection of *at most* eight rules for each node in the treebank. The maximum number of rules are projected from each node which is internal to a treebank tree and dominates two non-terminal children; four rules are projected from each node corresponding to the root node of a treebank tree, as this node can never be internal to a fragment, and two rules are projected from nodes dominating a single terminal symbol as terminal symbols are never substitution sites.



(8)

A PCFG-reduction derivation is isomorphic to a DOP derivation if for every substitution of a DOP fragment there is a corresponding sub-derivation in the PCFG. In other words, each PCFG sub-derivation yielding a subtree whose internal nodes are all of the form X_y, whose root node is of the form X and whose frontier nodes are either of the form X or are terminal symbols, corresponds exactly to a DOP fragment when the subscripts are removed. Furthermore, each such PCFG sub-derivation has exactly the same probability as the DOP fragment to which it corresponds.

Thus, for each of the eight training splits for each annotation type, we induced a GF-DOP grammar as follows:

1. binarise the training trees;

2. extract a PCFG-reduction from the annotated trees;

3. strip away the feature annotations from the treebank and extract a PCFG-reduction from the unannotated trees;

4. merge the extracted grammars with weights $W_1 = 0.99$ and $W_2 = 0.01$.

5.3.2 Parsing

During parsing, the GF-DOP grammar described above is applied to the input string and a derivation forest is built using the CKY and Viterbi algorithms in combination. Viterbi allows us to prune the derivation space as it is built such that the final derivation space contains the single best derivation for the input string. This is achieved by pruning sub-derivations with low probabilities from the PCFG-reduction derivation space in a bottom-up manner. Two different sub-derivations which have the same root node and span the same portion of the input string are used in building derivations of the entire input string in exactly the same way. This means that parses containing the more probable of these sub-derivations will always be more probable than those derivations containing the less probable sub-derivation. Consequently, the less probable sub-derivation will never be used to build the most probable derivation/parse and can be removed from the derivation space.

To this Viterbi derivation space we then apply the method of Jiménez and Marzal [2000] in order to determine the n-best derivations for the input string, where we set n to 2,000. We sum over the probabilities of the parses yielded by the n-best derivations, and return the one with the highest probability.

6 Evaluation

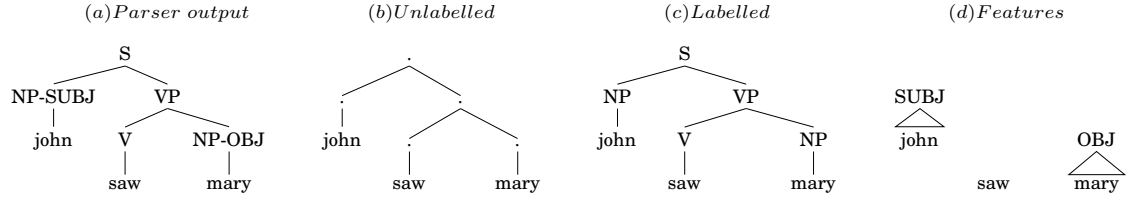
We evaluate our parser output using the standard precision, recall and f-score metrics as given in (9), (10) and (11).

$$Precision = \frac{\#\ correct\ constituents}{\#\ parse\ constituents} \tag{9}$$

$$Recall = \frac{\#\ correct\ constituents}{\#\ reference\ constituents} \tag{10}$$

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{11}$$

We transform our parser output and reference trees in three different ways and then apply the above metrics to each. The purpose of these transformations is to facilitate evaluation of unlabelled parse accuracy, labelled parse accuracy and feature detection accuracy. The transformations are illustrated in (12), where (12)(a) gives the parser output and (12)(b) – (d) the three transformations applied to the parser output. In (12)(b) all labels have been replaced with the generic '.' label. The application of the evaluation metrics to these transformations gives unlabelled parse accuracy. In (12)(c) all feature annotations have been stripped away, allowing evaluation of labelled parse accuracy. Finally, in (12)(d) all labels with feature annotations are stripped of syntactic category so that only those feature annotations remain, and all other constituents are deleted. This last transformation allows us to evaluate feature annotation accuracy.



(12)

The results of our experiments with the GF-DOP model are given in Tables 3, 4 and 5. The second and third colums in each of these tables (marked unlabelled and labelled) give the results for parse accuracy while the fourth column (features) gives the results for grammatical feature detection. The rightmost column in each (marked occ) gives the number of occurrences of feature annotations in the reference representations. Essentially, this column tells us how many feature annotations we were looking to identify[3] across the 8 test sets (720 test sentences in total) for each annotation run. In each of the tables, the first line of results corresponds to the baseline, i.e. the results for the run with no grammatical feature annotations on the treebank. The baseline scores in each of the tables are identical, and are repeated for convenience only. Coverage for all runs including the baseline was 93.89%.[4]

6.1 Functional annotation

The results of our experiments with functional annotation are given in Table 3. Focusing firstly on single-function annotations, we observe that the unlabelled f-score is higher for every annotated treebank than it is for the baseline. OBJ and ADJUNCT give the highest improvements over the baseline (of 0.2279% and 0.1749% respectively). We observe similar trends for labelled f-score in that OBJ and ADJUNCT again give the highest improvements over the baseline (of 0.4219% and 0.3347% respectively). A single annotation type, SUBJ, yields a decrease in labelled f-score over the baseline (of -0.0321%); all others yield an increase. The f-scores for grammatical feature detection range from 64.2857% (OBL) up to 100% (OBL-COMP, TOPIC-INT). If we focus on those features with reference set occurrences of at least 100 (ADJUNCT, OBJ, SUBJ, COMPOUND, XCOMP) this range of accuracy narrows to 68.7783% – 87.4267%, with highest accuracy for ADJUNCT and lowest for XCOMP. In fact, we score significantly worse for XCOMP than for the next lowest performing feature, which is SUBJ at 80.9938%.

Multi-function annotation results are given in the last three lines of Table 3. Annotation type ALL refers to the treebank annotated with all 16 of the functions listed as single annotations. Annotation type TOP5 refers to the treebank annotated with the 5 most frequently occurring functions which, as above, are ADJUNCT, OBJ, SUBJ, COMPOUND and XCOMP. Finally, annotation type SUBJ_OBJ refers to the treebank annotated with those two functions only. We observe firstly that the unlabelled f-scores for these annotated treebanks are not only higher than the baseline (by 0.2573% for ALL, by 0.2967% for TOP5 and by 0.2135% for SUBJ_OBJ) but also higher than all single annotations with the exception of OBJ, which outperforms SUBJ_OBJ by 0.0144%. The same observations hold for labelled f-score, where ALL improves over the baseline by 0.6454%, TOP5 by 0.6080% and SUBJ_OBJ by 0.4077%. Furthermore, we note that this improvement in tree structure accuracy does not cause feature detection accuracy to suffer: the f-scores for grammatical feature detection hold up well, ranging between 84.4528% and 85.1899%.

6.2 Annotation with lexical features

The results of our experiments with lexical feature annotation are given in Table 4. We observe that all but one of the annotations (SPEC-FORM, decrease of -0.0122%) give an improvement over the baseline in terms of unlabelled f-score. The greatest improvement is gained by annotating with

[3] An occurrence count of 0 for an annotation type means that the feature annotation occurred in the training data but never in the test/reference data. Grammatical feature detection scores are nevertheless given as they reflect the presence or absence of false positives.

[4] Those sentences which could not be fully parsed were assigned the most probable sequence of partial parses linked together by a 'TOP' node.

	unlabelled			labelled			features			occ
	precision	recall	fscore	precision	recall	fscore	precision	recall	fscore	#
BASELINE	96.0619	96.3603	96.2109	92.6076	92.8953	92.7512	—	—	—	—
ADJUNCT	96.2365	96.5447	96.3903	92.9373	93.2350	93.0859	87.8173	82.9736	85.3268	834
APP	96.1010	96.4088	96.2547	92.7148	93.0117	92.8630	100.000	100.000	100.000	0
COMP	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	77.2727	77.2727	77.2727	22
COMP-EX	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	100.000	100.000	100.000	0
COMPOUND	96.0913	96.3991	96.2450	92.6471	92.9438	92.7952	87.2852	81.6720	84.3854	311
OBJ	96.2848	96.5932	96.4388	93.0244	93.3223	93.1731	88.4058	86.4691	87.4267	776
OBJ2	96.0910	96.3894	96.2399	92.7141	93.0020	92.8578	100.000	100.000	100.000	0
OBL	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	75.0000	56.2500	64.2857	16
OBL-AGT	96.0913	96.3991	96.2450	92.7051	93.0020	92.8533	100.000	100.000	100.000	0
OBL-COMP	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	100.000	100.000	100.000	3
PRON-INT	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	100.000	91.6667	95.6522	24
PRON-REL	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	100.000	75.0000	85.7143	4
SUBJ	96.0542	96.3991	96.2263	92.5532	92.8856	92.7191	85.1175	77.2512	80.9938	422
TOPIC-INT	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	100.000	100.000	100.000	25
TOPIC-REL	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	100.000	66.6667	80.0000	3
XCOMP	96.1014	96.4185	96.2597	92.6478	92.9535	92.8004	77.5510	61.7886	68.7783	123
ALL	96.3049	96.6320	96.4682	93.2385	93.5553	93.3966	87.5106	81.6351	84.4708	2532
TOP5	96.3257	96.6903	96.5076	93.1831	93.5359	93.3592	87.2460	81.8329	84.4528	2466
SUBJ_OBJ	96.2658	96.5835	96.4244	93.0057	93.3126	93.1589	87.1616	83.3055	85.1899	1198

Table 3: Evaluation of the DOP model with training data annotated with grammatical functions.

PCASE, where the increase is 0.2196%. When we look at the labelled f-scores we see that annotating with SPEC-FORM again leads to a tiny decrease in accuracy, this time of -0.0099%. In contrast, the greatest improvement is gained by annotating with COMP-FORM, where the increase is 0.2817%. The f-scores for grammatical feature detection range from 51.2821% to 100%. If we focus on those features with reference set occurrences of at least 100 (SPEC-FORM, PCASE, COMP-FORM, CONJ-FORM, PRON-FORM) this range of accuracy narrows to 70.4782% – 92.1833%, with highest accuracy for PRON-FORM and lowest for PCASE.

6.3 Annotation with atomic features

The results of our experiments with atomic feature annotation are given in Table 5. Of the 36 different atomic grammatical features we generated treebanks for, 27 give an increase in unlabelled f-score over the baseline and 9 a decrease. Those which give the greatest increases are ADJUNCT-TYPE (0.2082%), PERF (0.1690%), VTYPE (0.1593%), ANIM (0.1497%) and PASSIVE (0.1400%). Those which give the greatest decreases are STMT-TYPE (-0.0667%), PSEM (-0.0520%) and NUM (-0.0517%). When we focus on labelled f-score, we note that 33 of the features give an increase in accuracy and only 3 give a decrease. Those which give decreases are GRAIN (-0.2318%), NUM (-0.1469%) and PROPER (-0.0782%). The greatest increases in labelled f-score are gained by annotating with PASSIVE (0.5369%), VTYPE (0.4206%) and PERF (0.4109%). The f-scores for grammatical feature detection range from 0% to 100%. If we narrow our focus to those features with reference set occurrences of at least 100, this range of accuracy narrows to 81.0526% – 95.1879% with highest accuracy for PERS and lowest for ADJUNCT-TYPE.

	unlabelled			labelled			features			occ
	precision	recall	fscore	precision	recall	fscore	precision	recall	fscore	#
BASELINE	96.0619	96.3603	96.2109	92.6076	92.8953	92.7512	—	—	—	—
COMP-FORM	96.1401	96.4573	96.2984	92.8799	93.1865	93.0329	82.4468	63.5246	71.7593	244
CONJ-FORM	96.0910	96.3894	96.2399	92.6560	92.9438	92.7997	94.7368	75.0000	83.7209	240
CONJ-FORM-COMP	96.1010	96.4088	96.2547	92.7148	93.0117	92.8630	0.0000	0.0000	0.0000	6
PCASE	96.2394	96.6223	96.4305	92.7978	93.1670	92.9820	72.9032	68.2093	70.4782	497
PRECONJ-FORM	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	100.000	100.000	100.000	0
PREDET-FORM	96.1107	96.4185	96.2644	92.6858	92.9826	92.8340	100.000	50.0000	66.6667	12
PRON-FORM	96.1494	96.4573	96.3031	92.7632	93.0603	92.9115	96.6102	88.1443	92.1833	194
PRT-FORM	96.1021	96.4379	96.2697	92.7749	93.0991	92.9367	83.3333	37.0370	51.2821	27
SPEC-FORM	95.9896	96.4088	96.1987	92.5396	92.9438	92.7413	95.1774	86.5178	90.6412	1209

Table 4: Evaluation of the DOP model with training data annotated with lexical features.

6.4 Discussion

Looking first to general parse accuracy, we note that the best overall unlabelled f-scores are achieved using the TOP5 (96.5076%) and ALL (96.4682%) annotation types. Furthermore, best overall labelled f-scores are also achieved using the ALL (93.3966%) and TOP5 (93.3592%) annotation types. We conclude from this that annotating with multiple grammatical functions gives the greatest improvement in phrase-structure tree parse accuracy. In addition, we conclude that the GF-DOP model generally helps phrase-structure tree parse accuracy rather than hindering it. This conclusion is based on the following: of the 64 annotated runs we carried out, 84.37% gave improvements over the baseline in terms of unlabelled f-score and 93.75% gave improvements over the baseline in terms of labelled f-score.

We do reasonably well at detecting grammatical functions, particularly the 5 most frequent ones, where f-scores are in the range 68.7783% – 87.4267%. When we annotated with all 5 most frequently-occurring functions, grammatical function accuracy remained high at 84.4528%. We also do reasonably well at detecting the most frequent lexical features, where f-scores are in the range 70.4782% – 92.1833%. However, it is questionable as to whether it is really useful to be able to detect such features – they are useful for helping us get better phrase-structure tree accuracy, but do not add much additional information to the output parse. Finally, we do well at detecting the more frequent atomic features, achieving f-scores in the range 81.0526% – 95.1879%.

7 Conclusions and Future Work

This paper proposed a new model – GF-DOP – which combines the robustness of the DOP model with some of the linguistic competence of LFG-DOP. This model incorporates more detailed linguistic information than Tree-DOP and, consequently, improves on the Tree-DOP model in that the output parses are more informative and the probability model is sensitive to this additional information. Although GF-DOP incorporates only some of the linguistic competence of the LFG-DOP model, it nevertheless constitues an improvement over LFG-DOP from both theoretical and practical perspectives: it maintains the integrity of the probability model because there is no 'leaked' probability mass and is more computationally tractable because the increase in grammar size induced by backing-off is not exponential.

We hypothesised that the GF-DOP model would (i) give us better phrase-structure tree parse ac-

	unlabelled			labelled			features			occ
	precision	recall	fscore	precision	recall	fscore	precision	recall	fscore	#
BASELINE	96.0619	96.3603	96.2109	92.6076	92.8953	92.7512	—	—	—	—
ABBREV	96.1300	96.4379	96.2837	92.7148	93.0117	92.8630	100.000	100.000	100.000	30
ACONSTR	96.0817	96.3894	96.2353	92.7148	93.0117	92.8630	100.000	100.000	100.000	0
ADEG-DIM	96.0720	96.3797	96.2256	92.7051	93.0020	92.8533	100.000	64.7059	78.5714	17
ADEGREE	96.1591	96.4670	96.3128	92.8019	93.0991	92.9502	87.9859	88.9286	88.4547	280
ADJUNCT-TYPE	96.2744	96.5641	96.4191	92.8876	93.1670	93.0271	87.5000	75.4902	81.0526	612
ADV-TYPE	96.1107	96.4185	96.2644	92.8212	93.1185	92.9696	85.5691	77.2477	81.1958	545
ANIM	96.2253	96.4962	96.3606	92.7023	92.9632	92.8326	92.7273	83.0018	87.5954	553
ATYPE	96.1304	96.4476	96.2888	92.8122	93.1185	92.9651	87.0370	87.3606	87.1985	269
CASE	95.9965	96.3506	96.1732	92.5829	92.9244	92.7533	89.5514	87.3345	88.4290	1737
DEIXIS	96.0430	96.3506	96.1965	92.6954	92.9923	92.8436	100.000	25.0000	40.0000	16
EMPH	96.0720	96.3797	96.2256	92.7148	93.0117	92.8630	100.000	100.000	100.000	0
EMPHASIS	96.0720	96.3797	96.2256	92.7148	93.0117	92.8630	100.000	100.000	100.000	0
GEND	96.1777	96.4670	96.3221	92.8005	93.0797	92.9399	100.000	73.5294	84.7458	34
GERUND	96.0433	96.3603	96.2016	92.8219	93.1282	92.9748	89.2857	92.5926	90.9091	108
GRAIN	96.0240	96.3409	96.1822	92.3672	92.6720	92.5194	92.3945	91.2306	91.8089	2064
LAYOUT-TYPE	95.9981	96.3894	96.1933	92.6728	93.0506	92.8613	91.3136	81.5516	86.1569	1057
MOOD	96.1505	96.4865	96.3182	92.7459	93.0700	92.9077	90.7631	85.3904	87.9948	794
NUM	95.9687	96.3506	96.1592	92.4207	92.7885	92.6043	94.3262	92.5641	93.4369	2730
NUMBER-TYPE	96.0720	96.3797	96.2256	92.7148	93.0117	92.8630	100.000	95.9596	97.9381	99
PASSIVE	96.2156	96.4865	96.3509	93.1572	93.4194	93.2881	94.8529	91.6519	93.2249	1126
PERF	96.2447	96.5156	96.3799	93.0314	93.2932	93.1621	96.1015	92.6573	94.3480	1144
PERS	96.0460	96.4282	96.2367	92.8171	93.1865	93.0014	95.9372	94.4538	95.1897	2975
POL	96.1297	96.4282	96.2787	92.6851	92.9729	92.8288	0.0000	0.0000	0.0000	6
PREDET-TYPE	96.1107	96.4185	96.2644	92.6858	92.9826	92.8340	100.000	100.000	100.000	12
PROG	96.0928	96.4379	96.2651	92.8820	93.2156	93.0485	96.2557	90.3171	93.1919	1167
PRON-TYPE	96.1788	96.4962	96.3372	92.9186	93.2253	93.0717	94.9664	90.2232	92.5341	941
PROPER	96.0902	96.3700	96.2299	92.5385	92.8079	92.6730	100.000	73.5294	84.7458	34
PSEM	95.9776	96.3409	96.1589	92.6030	92.9535	92.7779	93.2886	88.2540	90.7015	315
PTYPE	96.0913	96.3991	96.2450	92.7245	93.0215	92.8727	92.3810	92.6752	92.5278	314
SPEC-TYPE	96.0746	96.4476	96.2608	92.6327	92.9923	92.8122	94.6950	87.5000	90.9554	1632
STMT-TYPE	95.9675	96.3215	96.1442	92.5926	92.9341	92.7630	92.1453	87.1406	89.5731	1252
TEMPORAL	96.0913	96.3991	96.2450	92.7148	93.0117	92.8630	0.0000	0.0000	0.0000	2
TENSE	96.0925	96.4282	96.2601	92.8233	93.1476	92.9852	87.0968	76.5464	81.4815	388
TIME	96.1107	96.4185	96.2644	92.7148	93.0117	92.8630	0.0000	0.0000	0.0000	2
TYPE	96.0433	96.3603	96.2016	92.8219	93.1282	92.9748	89.2857	92.5926	90.9091	108
VTYPE	96.2350	96.5059	96.3702	93.0410	93.3029	93.1718	93.5426	89.6819	91.5716	1163

Table 5: Evaluation of the DOP model with training data annotated with atomic grammatical features.

curacy than the Tree-DOP model and (ii) allow us to learn grammatical features with a high degree of accuracy. In a number of experiments on the HomeCentre corpus, we investigated the veracity of this hypothesis. We generated a number of versions of this treebank with varying grammatical feature annotations, and trained and tested our model on these treebanks. We evaluated the output parses in terms of unlabelled parse accuracy, labelled parse accuracy and feature detection accuracy.

Our experiments show that annotating with multiple grammatical functions gives the greatest improvement in phrase-structure tree parse accuracy and that, overall, the GF-DOP model generally improves phrase-structure tree parse accuracy: 93.75% of the runs conducted gave improvements over the baseline in terms of labelled f-score. Our experiments also show that performance in terms of detecting grammatical features where feature occurrence is greater than 100 ranges between 68.7783% and 95.1879%, depending on the feature or group of features being tested and how often those features were seen in the training data.

In future work, we would like to incorporate available subcategorisation information into the model, perhaps by distinguishing between those functions which are subcategorised for and those which are not. We would like to scale also to larger corpora, in particular to be able to investigate features which were too infrequent in the data used here to be able to draw strong conclusions about their usefulness in this context. We intend to achieve this using the resources of Cahill et al. [2004]. Finally, we would like to investigate further the characteristics of the model's back-off element.

Acknowledgements

This work was generously supported by the Irish Research Council for Science and Technology (IRCSET) and Science Foundation Ireland (SFI).

References

Steven Abney. Stochastic Attribute-Value Grammars. *Computational Linguistics*, **23**(4):597–618, 1997.

Rens Bod. An Improved Parser for Data-Oriented Lexical-Functional Analysis. In *Proceedings of the 38th Conference of the Association for Computational Linguistics*, pages 61–68, Hong Kong, 2000a.

Rens Bod. An Empirical Evaluation of LFG-DOP. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 62–68, Saarbrucken, Germany, 2000b.

Rens Bod. *Enriching Linguistics with Statistics: Performance Models of Natural Language*. PhD thesis, Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands, 1995.

Rens Bod and Ronald Kaplan. A DOP model for Lexical-Functional Grammar. In Rens Bod, Remko Scha, and Khalil Sima'an, editors, *Data-Oriented Parsing*, pages 211–232. Stanford CA.: CSLI Publications, 2003.

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

Aoife Cahill, Michael Burke, Ruth O'Donovan, Josef van Genabith, and Andy Way. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. In *Proceedings of the 42th Conference of the Association for Computational Linguistics*, pages 320–327, Barcelona, Spain, 2004.

Joshua Goodman. Efficient Parsing of DOP with PCFG-Reductions. In Rens Bod, Remko Scha, and Khalil Sima'an, editors, *Data-Oriented Parsing*, pages 125–146. Stanford CA.: CSLI Publications, 2003.

Joshua Goodman. Efficient Algorithms for Parsing the DOP model. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing (EMNLP 1)*, pages 143–152, Philadelphia, PA., 1996.

Joshua Goodman. *Parsing inside-out*. PhD thesis, Harvard University, MA., 1998.

Víctor M. Jiménez and Andrés Marzal. Computation of the N Best Parse Trees for Weighted and Stochastic Context-Free Grammars. In *Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition*, pages 183–192, London, UK, 2000. Springer-Verlag.

Khalil Sima'an. An optimized algorithm for Data Oriented Parsing. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, 1995.

Khalil Sima'an. *Learning Efficient Disambiguation*. PhD thesis, University of Amsterdam, The Netherlands, 1999.

Andy Way. A Hybrid Architecture for Robust MT using LFG-DOP. *Journal of Experimental and Theoretical Artificial Intelligence*, **11**:441–471, 1999.

Andreas Zollmann and Khalil Sima'an. A Consistent and Efficient Estimator for Data-Oriented Parsing. *Journal of Automata, Languages and Combinatorics (JALC)*, **10**(2/3):367–388, 2005.

COMP IN (PARALLEL) GRAMMAR WRITING

Martin Forst

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Proceedings of the LFG06 Conference

Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

This paper is a grammar writer's reaction to the 'COMP debate', which has been going on in the LFG community for more than a decade now. Taking as a starting point the work by Dalrymple and Lødrup (2000), Alsina et al. (2005) and Berman (2006), I first consider the question with respect to a German large-coverage LFG. I show that, in addition to the reasons put forth by Alsina et al. (2005) and Berman (2006), there are further reasons to reinterpret as OBL$_\theta$S (or OBJ$_\theta$S) the arguments that Dalrymple and Lødrup (2000) analyze as COMPs in German, a language which they consider as 'mixed'. These have to do with COMPs subcategorized for by nouns and, to a lesser extent, with past participles of OBJ experiencer psych-verbs. I then present some data from Spanish, a 'non-mixed' language, and show that the distinction introduced in the reinterpretation of COMPs of German nouns has a correlate in Spanish. Furthermore, I point out how the reinterpretation of COMP can increase parallelism between grammars, an argument that needs to be taken with caution, of course, but in my opinion, does have its place in parallel grammar development. The final section explains why the linguistically more adequate analysis without COMP is also more attractive from the point of view of grammar engineering or, in other words, why the enhanced descriptive elegance of a grammar leads to improved efficiency in its processing. I report an 11% gain in processing time with a revised grammar compared to an equivalent grammar that makes use of COMP.

1 Introduction

The status of the grammatical function COMP (and – to a lesser extent – XCOMP) has been the subject of a considerable amount of work in theoretical LFG. However, the implementational efforts for various languages realized in the *ParGram* initiative do not reflect any of the results of this work so far, probably because grammar developers avoid the major effort of adapting their grammars as long as the controversy does not converge towards a consensus. This paper is an attempt to contribute to a possible consensus and to show that implemented grammars do benefit from insights from theoretical work, as a better understanding of the generalizations at work in the languages considered allows for improved lexicon acquisition and more general, and hence more efficient, grammars.

Before considering the linguistic data themselves and their modelling in the implemented grammars, let us recall the major steps of the 'COMP debate': In a contribution to the LFG List, Alsina et al. (1996) suggest reinterpreting COMPs as OBJs, arguing that the difference in category at the c-structure level between a nominal OBJ and an argument clause should not be reflected by a difference in grammatical function at the f-structure level if there were no further reasons to differentiate OBJs and COMPs. Dalrymple and Lødrup (2000) take up this argument and show that it holds for some argument clauses, but not for all. They thus propose to reinterpret some COMPs as OBJs, but keep COMP in the inventory of grammatical functions, even if, according to their terminology, COMP only exists in 'mixed' languages, whereas it does not in 'non-mixed' languages. Alsina et al. (2005), finally, revise their initial proposal of reinterpreting all COMPs as OBJs and suggest instead to reinterpet COMPs as OBJs, OBJ$_\theta$S or OBL$_\theta$S, depending on the subcategorizing element. One central argument of theirs is the alternation of non-OBJ argument clauses with different clitics in Catalan; another one is the parallelism between Catalan ('mixed') and Spanish ('non-mixed') translational equivalences. Interestingly, Berman (2006) comes to a similar conclusion, although she bases her argumentation on German ('mixed') facts only.

The remainder of this paper is organized as follows: In section 2, I show that, in addition to the reasons put forth by Alsina et al. (2005) and Berman (2006), there are further reasons to reinterpret as OBL$_\theta$S (or OBJ$_\theta$S) the arguments that Dalrymple and Lødrup (2000) analyze as COMPs in German, a language which they consider as 'mixed'. These have to do with COMPs subcategorized for by nouns and, to a lesser extent, with past participles of OBJ experiencer psych-verbs. In section 3, I then present some data from Spanish, a 'non-mixed' language, and show that the distinction introduced in the reinterpretation of COMPs of German nouns has a correlate in Spanish. Furthermore, I point out how the reinterpretation

of COMP can increase parallelism between grammars, an argument that needs to be taken with caution, of course, but in my opinion, does have its place in parallel grammar development. Finally, section 4 explains why the linguistically more adequate analysis without COMP is also more attractive from the point of view of grammar engineering or, in other words, why the enhanced descriptive elegance of the respective grammars leads to improved efficiency in their processing. This claim is sustained by the result of a small experiment with two grammar versions, one with and one without COMP.

Finally, it should be noted that my arguments with respect to the reinterpretation of COMP also apply to the arguments called VCOMPs in our grammar. These are infinitival arguments that are anaphorically controlled, i.e. arguments of equi verbs. I do not advocate, however, the reinterpretation of XCOMPs, which, in the German *ParGram* LFG, are functionally controlled arguments of modal, raising and AcI (accusativus cum infinitivo) verbs. Their behaviour is clearly different from the behaviour of VCOMPs with respect to passivization, the alternation with DPs[1] and control, so that I prefer maintaining XCOMP as a grammatical function, as long as no linguistically and technically adequate alternative is available.

2 The status of COMP in German and English

In our subcategorization lexicons for verbs and adjectives, we observe that almost all COMPs alternate with either OBJs or OBL$_\theta$s. (COMPs that alternate with OBJ$_\theta$s seem to be rare.) This redundancy seems undesirable to me, both for conceptual and for practical reasons; I will thus propose a reinterpretation of some COMPs as OBJs, as suggested by Dalrymple and Lødrup (2000), and then reinterpret the remaining COMPs as OBL$_\theta$s (and potentially OBJ$_\theta$s), along the lines of Alsina et al. (2005) and Berman (2006).

2.1 Uncontroversial OBJ clauses of verbs

In the theoretical literature, there seems to be a consensus now that certain COMPs should be reinterpreted as OBJs. The main criteria for distinguishing OBJ clauses from non-OBJ clauses are their alternation with DPs, their ability of being fronted and of being promoted to SUBJ status in passivized sentences. I will briefly go through these criteria again, although they have been discussed in the literature mentioned, because most *ParGram* grammars do not yet distinguish OBJ clauses from non-OBJ clauses and thus make wrong predictions with respect to the behaviour of either the OBJ clauses or the non-OBJ clauses.

2.1.1 Alternation with DPs

OBJ clauses subcategorized for by verbs alternate with DPs, as can be seen in (1) and (2). Non-OBJ clauses do not (see (7) and (8)). In the German *ParGram* LFG, as in most *ParGram* grammars, this alternation is stipulated through the presence of two unrelated subcategorization frames in the entry of the verbs concerned.

(1) *I believe [that the earth is round] / it / that.*

(2) *Ich glaube [, dass die Erde rund ist] / es / das.*
I believe that the earth round is / it / that.
'I believe that the earth is round / it / that.'

[1]The distinction between DPs and NPs is without importance for our argumentation. We use the term *DP* throughout this paper because there is a category DP in the German *ParGram* LFG. For grammars that do not have such a category or for readers that have reservations towards the notion of DP, the adequate term would be *NP*.

2.1.2 Fronting

OBJ clauses subcategorized for by verbs can be fronted, as in (3) and (4), whereas non-OBJ clauses cannot.

(3) *[That the earth is round] / That I believe.*

(4) *[Dass die Erde rund ist,] / Das wurde nicht geglaubt.*
That the earth round is / That was not believed.
'That the earth is round / That was not believed.'

2.1.3 Passivization

OBJ clauses subcategorized for by verbs can be promoted to SUBJ status in passivized sentences, as can be seen in (5) and (6). Non-OBJ clauses do not participate in passivization in the same way.

(5) *[That the earth is round] / That was not generally accepted.*

(6) *[Dass die Erde rund ist,] / Das glaube wurde nicht allgemein akzeptiert.*
That the earth round was / That was not generally accepted.
'That the earth is round / That was not generally accepted.'

2.2 Potential OBL_θ clauses of verbs

Argument clauses that are neither SUBJ nor OBJ are OBJ_θ or OBL_θ according to Alsina et al. (2005). In German (and English), OBJ_θ clauses seem to be rare. For the sake of simplicity, I thus talk about OBL_θ clauses here, although OBJ_θ clauses are expected to behave similarly.

2.2.1 Alternation with PPs, not DPs

OBL_θ clauses subcategorized for by verbs do not alternate with DPs, but with PPs, as can be seen in (7) and (8). In most *ParGram* grammars, this alternation is stipulated through the presence of two unrelated subcategorization frames in the entry of the verbs concerned.

(7) *The secretary has already insisted [that I have to fill in the form] / *it / [on it].*

(8) *Die Sekretärin passt auf [, dass ich das Formular ausfülle].*
The secretary pays attention that I the form fill in.
'The secretary is attentive that I fill in the form.'

2.2.2 Fronting

In English, OBL_θ clauses can only be fronted with a stranded preposition appearing after the verb. In German, they can only be fronted together with the corresponding pronominal adverb. Both the stranded preposition and the pronominal adverb indicate the type of OBL_θ function the fronted argument clause has; without this indication, the OBL_θ clauses, unlike their OBJ counterparts, cannot be fronted.

(9) *[That I have to fill in the form] the secretary has already insisted *(on).*

(10) **(Darauf,) [dass ich das Formular ausfülle,] passt die Sekretärin auf.*
On that that I the form fill in pays the secretary attention.
'The secretary is attentive that I fill in the form.'

As the German *ParGram* LFG, as it is, does not distinguish OBJ clauses from OBL_θ clauses, it wrongly parses (9). The non-distinction of OBJ clauses and OBL_θ clauses thus causes overgeneration in this case.

2.2.3 Passivization

In English, passivization is only possible with a stranded preposition appearing after the verb, and in German, the argument clause must be preceded by the corresponding pronominal adverb. For English, the explanation is that not only *Obj*s are promoted; in the German example, the argument clause is clearly not the SUBJ of the sentence (since PPs never are SUBJs), so that the construction has to be analyzed as an impersonal passive.

(11) *[That I have to fill in the form] has already been insisted *(on).*

(12) **(Darauf,) [dass ich das Formular ausfülle,] wird aufgepasst.*
On that that I the form fill in is paid attention.
'They are / Someone is attentive that I fill in the form. (impersonal passive)'

Again, the German *ParGram* LFG overgenerates due to the non-distinction of OBJ clauses and OBL_θ clauses, by wrongly parsing the unacceptable version of (12).

2.3 OBJ clauses of adjectives

Although adjectives are often believed not to take OBJs, a small number of German adjectives, like *gewohnt* ('used to') and *wert* ('worth'), do.

2.3.1 Alternation with DPs

Interestingly, the OBJ clauses and infinitives subcategorized for by adjectives alternate with DPs, just like OBJ clauses and infinitives subcategorized for by verbs. But again, just like in the lexical entries of verbs, this alternation is stipulated by two seemingly unrelated subcategorization frames.

(13) a. *Wir sind bei diesen Themen ja gewohnt [, dass die Damen*
We are with these topics indeed used that the ladies among
unter sich sind].
themselves are.
'With respect to these topics, we are indeed used to the fact that the ladies stick to themselves.'[2]

b. *Wir sind es / das bei diesen Themen ja gewohnt.*
We are it / that with these topics indeed used.
'With respect to these topics, we are indeed used to it / that.'

(14) a. *Die Begründung ist ?(es) wert [, im Wortlaut wiedergegeben zu werden]:*
The justification is it worth in the wording reproduced to be:
'The justification is worth being reproduced in its exact wording:'

b. *Die Begründung ist es / das wert.*
The justification is it / that worth.
'The justification is worth it / that.'

[2]This example, as most of the following examples, is an edited version of a corpus sentence. The corpora consulted were the TIGER Corpus, the Huge German Corpus (HGC) and the Europarl Corpus.

2.3.2 Fronting

Just like OBJ clauses and infinitives subcategorized for by verbs, OBJ clauses and infinitives subcategorized for by adjectives can be fronted. The German *ParGram* LFG, however, does not provide the necessary functional uncertainty path in the annotation of the fronted clausal or infinitival constituent, so that it cannot parse (15) and (16). Here, the grammar thus undergenerates.

(15) *[Dass die Damen unter sich sind,] sind wir bei diesen Themen ja gewohnt.*
That the ladies among themselves are are we with these topics indeed used.
'With respect to these topics, we are indeed used to the fact that the ladies stick to themselves.'

(16) *[Im Wortlaut wiedergegeben zu werden,] ist die Begründung nicht wert.*
In the wording reproduced to be is the justification not worth.
'The justification is not worth being reproduced in its exact wording.'

2.4 OBL_θ clauses of adjectives

In my view, most clausal and infinitival arguments subcategorized for by adjectives are OBL_θs. This is confirmed by the criteria that I have applied to OBL_θ clauses subcategorized for by verbs above.

2.4.1 Alternation with PPs, not DPs

OBL_θ clauses subcategorized for by adjectives alternate with PPs, not with DPs.

(17) *Ich bin froh, dass es alle geschafft haben.*
I am glad that it all made have.
'I am glad that they all made it.'

(18) *Ich bin *es / *das / darüber froh.*
I am *it / *that / about that glad.
'I am glad about that.'

2.4.2 Fronting

OBL_θ clauses subcategorized for by adjectives cannot be fronted without the corresponding pronominal adverb, whereas OBJ clauses subcategorized for by adjectives can, as we have seen above.

(19) **(Darüber,) [dass es alle geschafft haben,] bin ich froh.*
About that that it all made have am I glad.
'I am glad that they all made it.'

2.5 Why can OBL_θ clauses not be fronted (or, at least, only exceptionally)?

We have seen above that OBL_θ clauses cannot be fronted with the remainder of the sentence staying unchanged. Dalrymple and Lødrup (2000) take this observation as an argument for the existence of a distinct grammatical function COMP, which cannot be fronted in German and English. (They note, however, that in earlier stages of the German language, non-OBJ argument clauses could be topicalized.)

Berman (2006), who is in favour of the reinterpretation of COMP, gives a relatively complicated explanation for the fact that, in modern German, non-OBJ argument clauses cannot appear in SpecCP and she makes claims with respect to the ability of non-OBJ argument clauses to appear in topicalized partial VPs that, to me, seem to complicate the picture artificially.

What I believe is active in English and modern German is a constraint on the linear order of the subcat-frame-evoking element and the OBL_θ (or OBJ_θ) clause, which states that a (morphologically unmarked) *that*/*dass* clause can only function as an OBL_θ (or OBJ_θ) if it appears to the right of the verb, adjective or noun that subcategorizes for it. OBL_θ PPs can be fronted without problems because this constraint simply does not apply to them. With respect to argument clauses, this constraint explains the relevant data,[3] and I think this is a plausible constraint, since OBL_θ is a more marked grammatical function than SUBJ and OBJ, and morphologically unmarked constituents such as clauses can only be interpreted as such if the subcat-frame-evoking element prepares the hearer to do so.

In older German, this constraint apparently was weaker than today, but even in modern corpora we can find examples where, like in (20), a fronted *dass* clause or a fronted infinitival VP functions as an OBL_θ.

(20) *Sie zu ächten und zu verabscheuen gibt es gute Gründe;*
Them to ostracise and to loathe gives it good reasons;
'There are good reasons to ostracise and loathe them;'

2.6 COMPs subcategorized for by nouns

As for nouns, none of those from our subcategorization lexicon that subcategorize for a COMP can alternatively subcategorize for an OBJ, which is not surprising, as nouns are known to be intransitive. However, a large proportion of these nouns can alternatively subcategorize for an OBL_θ. I will show that the COMPs subcategorized for by these nouns can safely be reinterpreted as OBL_θs, in the very same way as many COMPs subcategorized for by verbs and adjectives, and the same restrictions on unbounded dependencies apply for all clausal OBL_θs, as example (21) illustrates.

(21) a. *Es gibt keinen Zweifel (daran), dass hier eine höhere Summe stehen sollte.*
It gives no doubt at this that here a higher sum stand should.

b. **(Daran,) dass hier eine höhere Summe stehen sollte, gibt es keinen Zweifel.*
At this that here a higher sum stand should gives it no doubt.
'There is no doubt that there should be a higher sum here.'

But what about the COMPs that cannot be reinterpreted as OBL_θs, like the one in (22)?

(22) a. *Es gibt den Vorwurf (*dafür/dazu/...), dass sich die DDR-Journalisten moralisch diskreditiert hätten.*
It gives the reproach that themselves the GDR journalists morally discredited had.
'There is the reproach that the GDR journalists had discredited themselves morally.'

b. ** (Dafür/Dazu/...) Dass sich die DDR-Journalisten moralisch diskreditiert hätten, gibt es den Vorwurf.*
That themselves the GDR journalists morally discredited had gives it the reproach.

[3] Example (25) in Berman (2006) is not relevant in my view, since the *dass* clause there is a SUBJ, and SUBJs are known to appear in topicalized partial VPs only with a very small number of verbs.

I propose to treat these as a kind of apposition or adjunct rather than an argument, a solution already hinted at in Dalrymple and Lødrup (2000). This treatment is motivated by semantic considerations, but also by the fact that none of these COMPs is obligatory, whereas at least some of the OBL$_\theta$ clauses subcategorized for by nouns are, and that the restrictions on unbounded dependencies that apply to appositive clauses are more strict than the ones that apply to OBL$_\theta$ clauses.

Interestingly, the nouns that can take clausal appositions are the very same ones that can subcategorize for a clausal SUBJ when used predicatively. This is illustrated in (23), which contains the same *dass* clause and the same noun, namely *Vorwurf* ('reproach'), as (22).

(23) *Dass sich die DDR-Journalisten moralisch diskreditiert hätten, ist ein schwerer Vorwurf.*
That themselves the GDR journalists morally discredited had is a serious reproach.
'That the GDR journalists had discredited themselves morally is a serious reproach.'

Nouns that subcategorize for OBL$_\theta$ clauses do not show this behaviour, as (24) illustrates.

(24) **Dass hier eine höhere Summe stehen sollte, ist ihr Zweifel.*
That here a higher sum stand should is her doubt.

Finally, the distinction between OBL$_\theta$ clauses subcategorized for by nouns and appositive clauses which accompany non-predicatively used nouns and which correspond to SUBJ clauses when the noun is used predicatively also allows us to analyze examples like (25) properly. Here, the noun *Beweis*, which is predicatively used, subcategorizes for a clausal SUBJ, which is instantiated by the first *dass* clause, and for a clausal OBL$_\theta$, which is the latter *dass* clause.

(25) *Dass inzwischen neun Prozent als politisch Verfolgte anerkannt werden, ist für Kanther Beweis, dass das neue Recht Schutz garantiert, ...*
That now nine percent as politically persecuted recognized are is for Kanther proof that the new legislation protection guarantees, ...
'That nine percent are now recognized as political refugees proves, for Kanther, that the new legislation guarantees protection, ...'

At the moment, it is not at all recorded in our subcategorization lexicon which ones are the nouns that can take clausal SUBJs. However, thanks to the knowledge about the relationship between appositive clauses of non-predicatively used nouns and clausal SUBJs of predicatively used nouns, it should be easy to acquire this knowledge by revisiting all lexical entries of nouns that subcategorize for COMPs at the moment.

2.7 Participles of OBJ experiencer psych-verbs

Further evidence for the ability of CPs to function as OBL$_\theta$s comes from the subcategorization behaviour of the participles of OBJ experiencer psych-verbs (e.g., *beruhigt* 'reassured', *beunruhigt* 'worried', *genervt* 'annoyed', *schockiert* 'shocked', *überrascht* 'surprised'). These participles are special because they seem to subcategorize for a COMP although the corresponding active forms clearly do not. As a temporary solution in order to analyze sentences like (26), where such a participles occurs, we entered them as 'lexicalized' participles in our lexicon. However, apart from their subcategorization behaviour, nothing indicates that they are lexicalized in any way.

(26) *Ich bin schockiert [, dass sich Bernard so positioniert hat.]*
I am shocked that himself Bernard so positioned has.
'I am shocked that Bernard positioned himself this way.'

By reinterpreting certain COMPs as OBL$_\theta$S and, hence, potentially as OBL-AGs, I will be able to account for the subcategorization behaviour of these participles with the standard lexical rule for passive.

3 COMP cross-linguistically

Unlike the other core grammatical functions, which seem to be present in all languages, COMP seems to be used only by the so-called 'mixed' languages (Dalrymple and Lødrup 2000). To me, this assumption seems somehow surprising and, moreover, it forces us to assume non-parallel analyses for translational equivalents that only differ in the presence or absence of a preposition. (See, e.g., Alsina et al. (2005) for translational equivalents from Spanish and Catalan or the examples below for translational equivalents from Spanish and French.) Since there seems to be consensus as to the non-use of COMP in 'non-mixed' languages, the question that needs to be clarified before COMP is abandoned as a grammatical function is whether the COMPs in 'mixed' languages can reasonably be reinterpreted as something else. In section 2, I have argued that they can in German and English; in the following, I will show that they can in French, yet another 'mixed' language, that French (just like Catalan) provides another argument for doing so, and that parallelism between closely related languages that differ with respect to their 'mixedness' is greatly improved.

3.1 OBJ clauses in French (a 'mixed' language) and Spanish (a 'non-mixed' language)

Here, I will briefly show that the distinction between OBJ clauses and non-OBJ clauses makes sense in French and Spanish and that the criteria for the distinction used in German and English can be applied in these two languages as well. French and Spanish (just like Catalan and Spanish in Alsina et al. (2005)) are an interesting language pair because they are closely related, both historically and typologically, but, according to Dalrymple and Lødrup (2000), French is a 'mixed' language, whereas Spanish is a 'non-mixed' language.

3.1.1 Alternation with direct object clitic

Both in French and in Spanish, OBJ clauses alternate with OBJ clitics.

(27) a. *Les gens ne croyaient pas que la terre était ronde.*
The people NE believed not that the earth was round.
'People did not believe that the earth was round.'

b. *Les gens ne le croyaient pas.*
The people NE it believed not.
'People did not believe it.'

(28) a. *La gente no creía que la tierra era redonda.*
The people not believed that the earth was round.
'People did not believe that the earth was round.'

b. *La gente no lo creía.*
The people not it believed.
'People did not believed it.'

3.1.2 Fronting

When fronted, OBJ clauses cooccur with a resumptive OBJ clitic in both French and Spanish.

(29) *Que la terre était ronde, les gens ne le croyaient pas.*
That the earth was round the people NE it believed not.
'That the earth was round, people did not believe.'

(30) *Que la tierra era redonda, la gente no lo creía.*
That the earth was round the people not it believed.
'That the earth was round, people did not believe.'

3.1.3 Passivization

In both French and Spanish, OBJ clauses can be promoted to SUBJ status in passivized sentences.

(31) *Que la terre était ronde n' était pas généralement accepté.*
That the earth is round NE was not generally accepted.
'That the earth is round was not generally accepted.'

(32) *Que la tierra era redonda no era generalmente aceptado.*
That the earth was round not was generally accepted.
'That the earth was round was not generally accepted.'

3.2 OBL_θ clauses in French (a 'mixed' language) and Spanish (a 'non-mixed' language)

In Spanish, *que* clauses can be preceded by prepositions that indicate their status as OBL_θs. In French, *que* clauses cannot be directly preceded by prepositions. Just like in Catalan (Alsina et al. 2005), there are good reasons, however, to suppose that many *que* clauses are OBL_θs.

3.2.1 Alternation with *both* adverbial clitics (French) or PPs (Spanish) respectively

The most important reason is that French non-OBJ clauses alternate with the two adverbial clitics available in the language, depending on the type of OBL_θ the verb (or adjective or noun) subcategorizes for. If these non-OBJ clauses were COMPs, as proposed in Dalrymple and Lødrup (2000), we could not explain why the argument clause of *insister* in (33) alternates with the clitic *y*, whereas the argument clause of *réjouir* in (35) alternates with the clitic *en*. In Spanish, the non-OBJ clauses all alternate with PPs.

(33) a. *La secrétaire a déjà insisté que je dois remplir le formulaire.*
The secretary has already insisted that I must fill in the form.
'The secretary has already insisted that I have to fill in the form.'

b. *La secrétaire y a déjà insisté.*
The secretary Y has already insisted.
'The secretary has already insisted on it.'

(34) a. *La secretaria ya ha insistido en que tengo que llenar el formulario.*
The secretary already has insisted in that I have to fill in the form.
'The secretary has already insisted that I have to fill in the form.'

b. *La secretaria ya ha insistido en eso.*
The secretary already has insisted in that.
'The secretary has already insisted on that.'

(35) a. *Je me réjouis beaucoup que mes parents viennent pour Noël.*
I myself am glad much that my parents come for Christmas.
'I am very glad that my parents are coming for Christmas.'

b. *Je m' en réjouis beaucoup.*
I myself EN am glad much.
'I am very glad about that.'

(36) a. *Me alegro mucho de que mis padres vengan para Navidad.*
Myself am glad much about that my parents come for Christmas.
'I am very glad that my parents are coming for Christmas.'

b. *Me alegro mucho de eso.*
Myself am glad much about that.
'I am very glad about that.'

3.2.2 Fronting

In French, non-OBJ clauses can be fronted, but must then cooccur with the corresponding adverbial clitic, which is *y* in (37) and *en* in (39). In Spanish, non-OBJ clauses can only be fronted together with the preposition that precedes them.

(37) *Que je dois remplir le formulaire, la secrétaire y a déjà insisté.*
That I must fill in the form the secretary Y has already insisted.
'That I have to fill in the form, the secreaty has already insisted on.'

(38) *En que tengo que llenar el formulario la secretaria ya ha insistido.*
In that I have to fill in the form the secretary already has insisted.
'That I have to fill in the form, the secretary has already insisted on.'

(39) *Que mes parents viennent pour Noël, je m' en réjouis beaucoup.*
That my parents come for Christmas I myself EN am glad much.
'That my parents are coming for Christmas I am very glad about.'

(40) *De que mis padres vengan para Navidad me alegro mucho.*
About that my parents come for Christmas myself am glad much.
'That my parents are coming for Christmas I am very glad about.'

3.2.3 Passivization

Non-OBJ clauses cannot be promoted to SUBJ status in either French or Spanish. I just give a French example here because only in the 'mixed' language French is there a danger of overgeneration due to the non-distinction of OBJ and non-OBJ clauses.

(41) **Que je dois remplir le formulaire a déjá été insisté.*
That I must fill in the form has already been insisted.
'That I have to fill in the form has already been insisted on.'

3.3 COMPs subcategorized for by nouns

Let us now consider COMPs that seem to be subcategorized for by nouns in a crosslinguistic perspective. I have argued above that *dass* clauses like the one in (42) are OBL$_\theta$s, whereas clauses like the one in (44) are appositions. I will argue that the same holds true for the *que* clauses in (43) and (45) respectively. My main arguments are that, in Spanish, OBL$_\theta$ *que* clauses can be preceded by basically any preposition that can introduce OBL$_\theta$s, whereas clausal appositions are always introduced by the preposition *de*, and that basically the same restrictions as to unbounded dependencies apply to *que* clauses as to *dass* clauses. (See subsection 2.6.)

(42) … *[$_{DP}$ das Vertrauen, dass es auch in Zukunft ein Land Bosnien-Herzegowina*
… the confidence that it also in future a country Bosnia-Herzegovina
gibt] …
gives …
'… confidence that the country of Bosnia-Herzegovina will continue to exist in the future …'

(43) … *[$_{DP}$ la confianza en que en el futuro exista también un país como B-H]*
… the confidence in that in the future exist also a country like B-H
…
…
'… confidence that the country of Bosnia-Herzegovina will continue to exist in the future …'

(44) *[$_{DP}$ Die Tatsache, dass diese Misshandlung durch andere Muslime ausgeführt wurde,]*
The fact that this mistreatment by other Muslims carried out was
…
…
'The fact that this abuse was perpetrated by other Muslims …'

(45) *[$_{DP}$ El hecho de que los malos tratos fueran infligidos por otros musulmanes]*
The fact of that the bad treatments were inflicted by other Muslims
…
…
'The fact that this abuse was perpetrated by other Muslims …'

In this context, it is interesting to note that in the French translation of the two sentences above, which are from the Europarl Corpus, we find a construction consisting of the preposition *en*, the pronoun *ce* and the *que* clause in the case of the OBL$_\theta$ clause, whereas the appositive *que* clause directly follows the noun *fait*, on which it depends. This does not mean that all OBL$_\theta$ clauses subcategorized for by nouns are preceded by a preposition and the pronoun *ce* in French, but only OBL$_\theta$ clauses can be constructed this way. Appositive clauses always directly follow their governing noun in French.

(46) … *la confiance en ce qu' à l' avenir, la Bosnie-Herzégovine demeure aussi*
… the confidence in it that to the future the Bosnia-Herzegovina stays also
un pays …
a country …
'… confidence that the country of Bosnia-Herzegovina will continue to exist in the future …'

(47) *Le fait que ces actes de violence aient été perpétrés par d'*
The fact that these acts of violence have been perpetrated by Article.Indefinite.Pl
autres musulmans …
other Muslims …
'The fact that this abuse was perpetrated by other Muslims …'

A further interesting observation is that the generalization stating that nouns that can head an appositive clause when used non-predicatively are the ones that can take a clausal SUBJ when used predicatively carries over to French and Spanish.

(48) *[Que ces actes de violence aient été perpétrés par d' autres*
That these acts of violence have been perpetrated by Article.Indefinite.Pl other
musulmans] est un fait.
Muslims is a fact.
'That this abuse was perpetrated by other Muslims is a fact.'

(49) *[Que los malos tratos fueron infligidos por otros musulmanes] es un hecho.*
That the bad treatments were inflicted by other Muslims is a fact.
'That this abuse was perpetrated by other Muslims is a fact.'

3.4 Parallelism

The *ParGram* grammars are regularly checked for parallelism among them, parallelism referring mainly to f-structures as the level of representation that is used for applications that build on top of the parser output. Whenever translational equivalents in two *ParGram* languages are structurally similar, the f-structures associated with these translational equivalents are supposed to differ only in the values of the PRED features and perhaps minor morphosyntactic features.

3.4.1 Parallelism within a ('mixed') language

Although parallelism is generally viewed as a criterion for analyses across languages, it can also be applied as a criterion for analyses of related sentences within a language. In 'mixed' languages, the criterion of parallelism is interesting with respect to the alternation of argument clauses with DPs or PPs. The following two f-structures, associated with (33a) and (33b) after the reinterpretation of COMP as OBJ, OBJ_θ or OBL_θ, are parallel with respect to the grammatical functions subcategorized for by *insister*, whereas the f-structures currently produced by the French *ParGram* LFG are not.

$$\begin{bmatrix} \text{PRED} & \text{'insister}\langle\text{secrétaire, devoir}\rangle\text{'} \\ \text{OBL} & \begin{bmatrix} \text{PRED} & \text{'devoir}\langle\text{remplir}\rangle\text{pro'} \\ \text{SUBJ} & \boxed{1}\ \begin{bmatrix} \text{PRED} & \text{'pro'} \end{bmatrix} \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{'remplir}\langle\text{pro, formulaire}\rangle\text{'} \\ \text{SUBJ} & \boxed{1} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'formulaire'} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'secrétaire'} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{PRED} & \text{'insister}\langle\text{secrétaire, pro}\rangle\text{'} \\ \text{OBL} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PFORM} & \text{sur} \end{bmatrix} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'secrétaire'} \end{bmatrix} \end{bmatrix}$$

3.4.2 Parallelism across languages

Parallelism across languages, in particular between 'mixed' and 'non-mixed' languages, also greatly benefits from the reinterpretation of COMP. The two following f-structures, associated to (33a) and (34a), which are translational equivalents in French and Spanish, are parallel. If COMP were maintained as a grammatical function in 'mixed' languages, they would diverge.

$$\begin{bmatrix}
\text{PRED} & \text{'insister}\langle\text{secrétaire, devoir}\rangle\text{'} \\
\text{OBL} & \begin{bmatrix}
\text{PRED} & \text{'devoir}\langle\text{remplir}\rangle\text{pro'} \\
\text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{'pro'}\end{bmatrix} \\
\text{XCOMP} & \begin{bmatrix}
\text{PRED} & \text{'remplir}\langle\text{pro, formulaire}\rangle\text{'} \\
\text{SUBJ} & \\
\text{OBJ} & \begin{bmatrix}\text{PRED} & \text{'formulaire'}\end{bmatrix}
\end{bmatrix}
\end{bmatrix} \\
\text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{'secrétaire'}\end{bmatrix}
\end{bmatrix}$$

$$\begin{bmatrix}
\text{PRED} & \text{'insistir}\langle\text{secretaria, tener-que}\rangle\text{'} \\
\text{OBL} & \begin{bmatrix}
\text{PRED} & \text{'tener-que}\langle\text{llenar}\rangle\text{pro'} \\
\text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{'pro'}\end{bmatrix} \\
\text{XCOMP} & \begin{bmatrix}
\text{PRED} & \text{'llenar}\langle\text{pro, formulario}\rangle\text{'} \\
\text{SUBJ} & \\
\text{OBJ} & \begin{bmatrix}\text{PRED} & \text{'formulario'}\end{bmatrix}
\end{bmatrix} \\
\text{PFORM} & \text{en}
\end{bmatrix} \\
\text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{'secretaria'}\end{bmatrix}
\end{bmatrix}$$

4 Engineering advantages

4.1 Simplification of subcategorization lexicons

In section 2, I mentioned the huge redundancy that exists in our subcategorization lexicons for verbs and adjectives. I believe that this redundancy is harmful in several ways, not only conceptually but also in terms of grammar efficiency. In addition to the grammatical functions a verb or an adjective can take, our subcategorization lexicons encode what categories can realize a given function. For example, thematic SUBJs can maximally be realized as DPs, *dass* CPs, declarative verb-second CPs, interrogative CPs or infinitival VPs. Although this is not yet done in practice, underspecification could be used in cases where all five categories are possible as the SUBJ of a lexical element. This possibility is not available, however, for non-SUBJ functions if DPs and PPs are analyzed as OBJs (or OBJ$_\theta$s) and OBL$_\theta$s respectively and CPs and VPs are analyzed as COMPs and VCOMPs respectively. The reinterpretation of COMPs and VCOMPs as OBJs and OBL$_\theta$s would allow the use of underspecification with respect to category for all grammatical functions and, hence, open up the way for a great simplification of our subcategorization lexicons. Apart from the conceptual advantage this represents, in my opinion, it is reasonable to expect a substantive gain in efficiency from this simplification, since it considerably reduces the number of disjuncts in the lexical entries of verbs and adjectives that have to be tested by the parser which processes the grammar.

The two following examples illustrate this point: *akzeptieren* ('to accept') has the following lexical entry in the original verb subcategorization lexicon of our grammar.

```
akzeptieren !V-S xle
        {@(DPnom-DPacc %stem) @(AUX-HABEN)
        |@(DPnom-Sdass_corr %stem) @(AUX-HABEN)
        |@(DPnom-Sv2_corr %stem) @(AUX-HABEN)
        |@(DPnom-Swh_corr %stem) @(AUX-HABEN)
        |@(DPnom-VPzuinf_corr %stem) @(AUX-HABEN)
        }; ETC.
```

The templates ending in `_corr` allow for a clausal or infinitival argument both with and without the correlative pronoun *es*. Since the functional interpretation of the clausal or infinitival argument changes, depending on the presence or absence of the correlative element, each of these templates involves a two-way disjunction, so that there are actually nine disjuncts in the lexical entry.

This number could be reduced to three, if we made maximal usage of underspecification in the lexical entry. This means that we would not specify the possible categorial realizations of a grammatical function if all categorial realizations permitted by the grammar are possible. We would then have something like the following:

```
akzeptieren !V-S xle
        {@(SUBJ_DPnom-OBJ %stem) @(AUX-HABEN)
        |@(SUBJ_DPnom-COMP %stem) @(AUX-HABEN)
        |@(SUBJ_DPnom-VCOMP %stem) @(AUX-HABEN)
        }; ETC.
```

A further reduction is not possible because each disjunct evokes a functionally distinct subcategorization frame. If, however, COMP and VCOMP are reinterpreted as OBJ in the case of *akzeptieren*, we could further simplify the lexical entry as follows:

```
akzeptieren !V-S xle @(SUBJ_DPnom-OBJ %stem) @(AUX-HABEN); ETC.
```

My second example is a verb whose COMP, in my view, is actually an OBL_θ, namely *drohen*. Its lexical entry in the original verb subcategorization lexicon looks as follows:

```
drohen !V-S xle
        {@(DPnom-PP %stem mit dat) @(AUX-HABEN)
        |@(DPnom-PPSdass %stem mit dat) @(AUX-HABEN)
        |@(DPnom-PPSv2 %stem mit dat) @(AUX-HABEN)
        |@(DPnom-PPVPzuinf %stem mit dat) @(AUX-HABEN)
        |@(DPnom-Sdass %stem) @(AUX-HABEN)
        |@(DPnom-Sv2 %stem) @(AUX-HABEN)
        |@(DPnom-VPzuinf %stem) @(AUX-HABEN)
        |...
        }; ETC.
```

These seven disjuncts can be reduced to three if maximal usage of underspecification is made.

```
drohen !V-S xle
        {@(DPnom-OBL_noInt %stem mit dat) @(AUX-HABEN)
        |@(DPnom-COMP_noInt %stem) @(AUX-HABEN)
        |@(DPnom-VCOMP %stem) @(AUX-HABEN)
        |...
        }; ETC.
```

But again, further simplification is made impossible by the distinction of OBL$_\theta$, COMP and VCOMP. Only by reinterpreting the COMP and the VCOMP of *drohen* as OBL$_\theta$s can we further simplify this lexical entry.

```
drohen !V-S xle
        {@(DPnom-OBL_noInt %stem mit dat) @(AUX-HABEN)
        |...
        }; ETC.
```

4.2 Simplified and more regular functional uncertainty paths

In the German *ParGram* LFG as it is, i.e. with COMPs and VCOMPs, there are functional uncertainty paths in the annotation of both topicalized and extraposed CPs and VPs that lead to both over- and undergeneration, as explained in section 2. Moreover, the functional uncertainty path in the annotation of extraposed CPs and VPs involves a high number of disjuncts due to the fact that extraposed CPs and VPs that are not preceded by a correlative pronoun or pronominal adverb are analyzed as SUBJs, COMPs or VCOMPs respectively, whereas those that are preceded by a correlative element are analyzed as APP-CLAUSEs of SUBJs, OBJs or OBL$_\theta$s. With COMPs and VCOMPs being reinterpreted as OBJs or OBL$_\theta$s, all extraposed CPs and VPs would be analyzed as (APP-CLAUSEs[4] of) SUBJs, OBJs or OBL$_\theta$s. The revised functional uncertainty path in the f-annotation of extraposed CPs and VPs then involves fewer disjuncts and exhibits more regularity than the original functional uncertainty path, as is illustrated here.

```
... "Nachfeld"
CPdep[std]: { (^ SUBJ (APP-CLAUSE)) = !
            | (^ VP-PATH { COMP | { OBJ | OBL } APP-CLAUSE } = !
            | (^ DP-PATH COMP) = !
            | ...
            }

... "Nachfeld"
CPdep[std]: { (^ { SUBJ | VP-PATH { OBJ | OBL } } (APP-CLAUSE)) = !
            | (^ DP-PATH { OBL (APP-CLAUSE) | APP }) = !
            | ...
            }
```

4.3 Simplified and more regular application of the lexical rule(s) for passive

In the original grammar, there are three templates that implement lexical rules for passive: `PASSIVE-OBJ-TO-SUBJ`, `PASSIVE-COMP-TO-SUBJ` and `PASSIVE-VCOMP-TO-SUBJ`. The first one can only promote nominal objects to subjects and applies to all subcategorization frames that

[4] Although I believe that the function APP-CLAUSE should be removed in order to simplify the functional uncertainty path under consideration even further, I think that this issue should be kept separate from the status of COMP and VCOMP.

involve a thematic SUBJ and a thematic OBJ; the second one can promote COMPs to SUBJ status, but, for reasons that have no independent motivation in the grammar, applies only to subcategorization frames that involve a COMP, but no OBJ, and the same applies to the last template with respect to VCOMPs.

Once COMPs and VCOMPs are reinterpreted as OBJs, OBJ_θs or OBL_θs, it is sufficient to keep the template `PASSIVE-OBJ-TO-SUBJ`, which allows for the promotion to SUBJ status of any type of OBJ and is systematically applied to all subcategorization frames that involve a thematic SUBJ and a thematic OBJ, which can be clausal or infinitival in this case. No longer are there lexical rules that apply to subcategorization frames in an unsystematic way.

4.4 Improved acquisition of subcategorization information from corpora

In the context of COMPs of nouns, I have stated above that a distinction is to be made between clauses that are actually OBL_θs of nouns and clauses that function as appositions to nouns and, more importantly, that this distinction was related to the subcategorization behaviour of nouns when they are used predicatively. Two properties in the subcategorization behaviour of those nouns which, at first glance, seem to be unrelated thus turn out to be one and the same property in fact.

I believe that there are more properties of this kind, which are recorded as separate pieces of information in our subcategorization lexicons, but are in fact related very regularly. Many of them have nothing or little to do with the grammatical function COMP, but the COMP does contribute to blur the picture that we have of subcategorization and on whose basis we develop the theory that underlies the way we record subcategorization behaviour. To name just two examples, the possibility of a correlative *es* to cooccur with an OBJ clause is independent of the exact nature (*dass*, verb-second declarative, interrogative) of this clause, and all verbs that can subcategorize for an OBL_θ clause without a correlate can equally subcategorize for such a clause with some correlative pronominal adverb. As long as we make use of COMP as a grammatical function, we are highly unlikely to discover this kind of regularity because the constituents are analyzed as having different grammatical functions (COMP vs. OBJ in the case of OBJ clauses (not) preceded by a correlative *es*; COMP vs. OBL_θ in the case of OBL_θ clauses (not) preceded by a correlative pronominal adverb).

For the acquisition of a subcategorization lexicon from corpora that aims at completeness and consistency, it is of utmost importance to have a good understanding of all regularities that are at work in subcategorization. No corpus will contain all realizational variants of a given subcategorization frame, but if the theory on which we build the representation in which the subcategorization information is recorded captures regularities, there is hope that, via these generalizations, the acquired subcategorization information also covers most unseen realizational variants.

4.5 Grammar efficiency

In order to verify my claim that the reorganization of the subcategorization lexicons made possible by the reinterpretation of COMP has a positive effect on grammar efficiency, I created two largely equivalent grammar versions and had them analyze 1,956 sentences from section 8,001 through 10,000 of the TIGER Corpus. The versions mainly differ in the verb subcategorization lexicon used. Further, rather minor, changes were made necessary by the reinterpretation of COMP in the new subcategorization lexicon, such as changes in the f-annotation of CPs and VPs and in treatment of the correlative pronoun *es* and correlative pronominal adverbs.

The comparison of the two runs shows that the original grammar version needs 11% more time to parse the 1,956 sentences than the version with the revised subcategorization lexicon. While this is not an enormous gain in efficiency, it does represent an improvement, which, moreover, reduces the number

of timeouts (sentences that cannot be associated with a full parse within a bounded amount of time, set to 100 seconds in both runs) by 13 to 181 out of the 1,956 sentences.

5 Conclusions

COMP seems to be redundant as a grammatical function, both for reasons internal to 'mixed' languages like German (or Catalan, English, French etc.) and for reasons of parallelism between closely related languages that, in spite of their close relationship, differ as to their alleged 'mixed' or 'non-mixed' status, as it is the case, e.g., for Catalan and Spanish and for French and Spanish. Furthermore, categorically restricted functions like COMP and VCOMP pose problems for the efficient and technically economic organization of subcategorization lexicons that, at least in principle, treat the functional status of arguments and their possible realizations in terms of syntactic category as disjunct pieces of information.

References

Alsina, Àlex, KP Mohanan, and Tara Mohanan. 1996. Untitled submission to the LFG List.

Alsina, Àlex, KP Mohanan, and Tara Mohanan. 2005. How to get rid of the COMP. In *Proceedings of the 10th International LFG Conference (LFG'05)*, Bergen, Norway. CSLI Publications.

Berman, Judith. 2006. Functional identification of complement clauses in German and the specification of COMP. In Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson, and Annie Zaenen (eds.), *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*. Stanford, California: CSLI Publications.

Bodomo, Adams B., and Y. M. Lee. 2003. On the function COMP in Cantonese. In A. B. Bodomo and K. K. Luke (eds.), *Lexical-Functional Grammar Analysis of Chinese*. Journal of Chinese Linguistics Monograph 19.

Dalrymple, Mary, and Helge Lødrup. 2000. The Grammatical Functions of Complement Clauses. In *Proceedings of the 5th International LFG Conference (LFG'00)*, Berkeley, California. CSLI Publications.

Lødrup, Helge. 2002. Infinitival complements in Norwegian and the form-function relation. In *Proceedings of the 7th International LFG Conference (LFG'02)*, Athens, Greece. CSLI Publications.

Lødrup, Helge. 2004. Clausal complementation in Norwegian. *Nordic Journal of Linguistics* 27(1): 61–95.

THE COMPLEMENT OF *verba dicendi* PARENTHETICALS

Christian Fortmann
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

Proceedings of the LFG06 Conference
Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract.[*] The topic of this article is a quite frequent parenthetical construction in German. The predicate of this type of parenthetical is constituted by a verb which governs a COMP function, in particular a *verbum dicendi* or *verbum sentiendi* as for instance in *Theo kam – sagt Paul – mit seinem Hund* (*Theo came – says Paul – with his dog*). The value of this COMP function is not projected from a constituent within the parenthetical. Due to the interpretation of the construction, the host provides the complement of the parenthetical verb. It is argued that the value of the COMP function is represented by an f-structure whose PRED value is specified as 'pro'. This pronominal PRED value is anaphorically linked to the f-structure of the host. After the exposition of this account some restrictions on the this construction concerning the lexical choice of the parenthetical verb and its coconstituents will be consdered.

1 Parentheticals with *verba dicendi*

The examples in (1) show a certain type of parenthetical constructions in German which contain a verbal predicate – mostly a *verbum dicendi* or *verbum sentiendi* – that subcategorizes for a propositional argument.[1]

(1) a. *Theo kam – sagt Paul – mit seinem Hund*
Theo came says Paul with his dog
b. *Theo kam – so sagt Paul – mit seinem Hund*
Theo came so says Paul with his dog
c. *Theo kam – wie Paul sagt – mit seinem Hund*
Theo came as Paul says with his dog

In the case of an ordinary complementation structure, the propositional argument of a verb like *sagen* (say) is expressed by a clausal CP as in *Paul sagt, daß Theo kommt* (Paul says that Theo comes). The grammatical function which is assigned to this argument is COMP. Anyway, in order to meet the requirements of the completeness condition, a clause whose predicate governs a COMP function must also contain a complement clause. This requirement also holds for clausal parentheticals. The constructions in (1), however, are peculiar because the complement of the verb *sagt* is not included in the parenthetical string which is also referred to as a *reduced parenthetical*.

Since the constructions in (1) are grammatical as a whole, the parenthetical's verb cannot suffer from a completeness condition violation despite the fact that there is no (clausal) complement of this verb located within the parenthetical string itself. There is also no doubt with respect to the content of the (missing) complement. In all three cases in (1) the host clause is interpreted as a statement made by the parenthetical's subject *Paul*. Hence, the complement of *sagt* is somehow linked to the host clause. The question, then, is: how is the complement of the verb represented and how is it linked to the host clause?

One might speculate that the sentences in (1) are somehow derived from a monoclausal construction with the parenthetical as its root and the host as a complement. But such an analysis is questionable for a number of reasons.

[*] I want to thank the participants of the LFG2006 conference in Konstanz for a number of instructive and helpful comments.

[1] There are a number of (quasi) defining criteria to determine a parenthetical string. Optionality and sepration from the surrounding string by intonational breaks are two at least sufficient conditions. In the following, parentheticals are marked by dashes.

The parenthetical in (1c) has the structure of a verb-final clause. Verb-final clauses constitute subordinate clauses in German but not root structures. Therefore, it would be quite unreasonable to assume that the parenthetical in (1c) forms the matrix clause of the whole construction. The parentheticals in (1a) and (1b) have an apparent verb-first structure and verb-second structure respectively. Both types occur as root clauses in German. Hence, a monoclausal complementation structure is not per se dubious. However, restrictions on the interpretation like the impossibility of variable binding from the host into the parenthetical and vice versa and the restriction on scope of negation to the parenthetical and the host respectively indicate that both clausal components of the construction are not functionally integrated as complementation structures normally are.[2]

On the other hand it might be argued that an account of the reduced parenthetical constructions in (1) should refrain from any syntactic consideration altogether. In this case the saturation of the verbs propositional argument slot would have to be transferred to some mechanism of post syntactic semantic interpretation. However, such a turn also faces a number of empirical and conceptual problems.

In general, verbs do not allow of any dispensation from their subcategorization requirements in German. An argument of a verb may remain implicit, an issue we will discuss immediately. But it is not possible to infer a missing argument from the discourse, say, from the preceding sentence even if some salient entity is available.

Moreover, such a mechanism would require some kind of syntactic argument reduction which cancels the COMP function from the predicate's semantic form. This device, however, would have to be distinct from other known mechanisms of argument structure modification like passivization. While the syntactic realization of the argument within the functional domain of the verb would have to be suppressed, the requirement of the semantic argument to get saturated by some overt syntactic material would have to be maintained, since this argument cannot be missing altogether. For these reasons, a syntactic account is worth considering.

Instead of a monoclausal complementation structure, there are three possible modes of syntactic explanation. According to the first one, the complement of the parenthetical's verb is represented as an implicit argument which has to be anaphorically linked to the host clause. Implicit arguments are common with verbs like *essen* (to eat) or *öffnen* (to open) in German.

The second possible account employs a phonologically unexpressed copy of the host clause within the parenthetical. Finally, the complement may also be conceived as a pronominal which is anaphorically linked to the host clause. In the absence of phonological realization, this pronominal has to be represented either by an empty element in the constituent structure of the parenthetical or as an f-structure value of the COMP function implemented in the parenthetical's f-structure which is not projected from a c-structure complement.

The first two alternatives are not suitable due to empirical reasons. An account in the sense of the third variant, however, is capable of explaining the construction.

---

[2] The examples in (i) illustrate the facts about scope of negation.

(i) a. *Theo kam nicht – sagt Paul – mit seinem Hund* — *Neg>sagt, sagt>Neg
Theo came not says Paul with his dog

b. *Theo kam – so sagt Paul – nicht mit seinem Hund* — *Neg>sagt, sagt>Neg
Theo came so says Paul not with his dog

c. *Theo kam nicht – wie Paul sagt – mit seinem Hund* — Neg>sagt, sagt>Neg
Theo came not as Paul says with his dog

For a detailed discussion of these aspects of the parenthetical construction cf. Fortmann (2005).

2 Implicit Argument Account (to be rejected)

The internal argument of verbs like *essen* (to eat), *öffnen* (to open), *helfen* (to help) and some others may be missing in a clause. But this argument is after all present in the interpretation of the predicate as in implicit argument. (3) shows the counterparts of the transitive verbs in (2).

(2) a. *Theo ißt mit Appetit eine Schweinshaxe*
Theo eats with appetite a knuckle of pork
b. *Theo hat mir die Tür öffnet*
Theo has me the door opened
c. *Theo hilft seinem Chef nur widerwillig*
Theo helps his chief only unwillingly

(3) a. *Theo ißt mit Appetit*
Theo eats with appetite
b. *Theo hat mir öffnet*
Theo has me opened
c. *Theo hilft nur widerwillig*
Theo helps only unwillingly

Although in principle the argument of the verbs in (3) may remain implicit, its interpretation is not free but subject to selectional restrictions. So for instance the verb *öffnen* in its syntactically intransitive use restricts its implicit argument to the entrance to a room or locality. While (3b) may be satisfactorily substituted for (2b), (4b) is not a possible paraphrase of (4a).

(4) a. *Theo öffnet gerade die Sardinenbüchse*
Theo opens just the sardine tin
b. *Theo öffnet gerade* ≠ (4a)

A statement like (4b) is even impossible in a context from which the content of the argument can be inferred as in (5). In this case an overt pronoun is required. This means that the implicit argument of the verb is not accessible for an anaphoric relation to some suitable antecendent in the discourse environment.

(5) *Theo hat eine Sardinendose gekauft. Er öffnet *(sie) gerade.*
Theo has a sardine tin bought He opens (it) just

Verbs which occur in reduced parentheticals may impose selectional restrictions on their clausal complement, too. These restrictions affect the determination of the sentence mood of the complement. The verbs *glauben*, *meinen*, (to believe) for instance, require a declarative complement and are incompatible with an interrogative.

(6) a. **Paul glaubt/meint wer mit seinem Hund kam*
Paul believes who with his dog came
b. *Paul glaubt/meint daß Karl mit seinem Hund kam*
Paul believes that Karl with his dog came

However, selectional restrictions by the verb do not apply in the case of a reduced parenthetical.

(7) a. *wer kam – glaubt/meint Paul – mit seinem Hund?*
who came believes Paul with his dog
b. *wer kam – so glaubt/meint Paul – mit seinem Hund?*
who came so believes Paul with his dog
c. *wer kam – wie Paul glaubt/meint – mit seinem Hund?*
who came as Paul believes with his dog

The ineffectualness of selectional restrictions raises doubts as to the representation of the verb's complement by an implicit argument. The fact that the host clause is anaphorically linked to the argument of the parenthetical's verb does not accord with the properties of an implicit argument, either.

3 Copy Account (to be rejected)

Let us next turn to the second possible account in terms of a phonologically unpronounced copy of the host clause contained in the parenthetical clause. (8) represents the string of terminal elements of the sentence in (1a)[3].

(8) *Theo kam – sagt Paul ~~Theo kam mit seinem Hund~~ – mit seinem Hund*
Theo came says Paul with his dog

Although the facts about the interpretation namely that the statement of the host is attributed to the parenthetical's subject are captured, this account faces the same objections concerning the selectional requirements by the verb as pointed out in the previous section. In (9) the interrogative complement clause does not meet the restriction imposed by the verb *meinen*.

(9) *wer kam – meint Paul ~~wer kam mit seinem Hund~~ – mit seinem Hund?*
who came believes Paul with his dog

Furthermore, the claim that the complete host clause is interpreted as the parenthetical verb's complement must be relativized. In cases like those in (1) this interpretation is most natural. In (1) each host clause contains only one parenthetical. However, multiple insertion of reduced parentheticals into one host is also possible. In this case the whole construction is interpreted as, for instance, a reumé of a number of assertions made by different speakers. These assertions need not be completely identical. It is only necessary that the speakers refer to an identical event. Hence, (10a) is possible in the face of statements like (10b-d).

(10) a. *Theo - sagt Paul - ist heute - sagt Fritz - mit seinem Hund - sagt Karl - gekommen*
Theo says Paul has today says Fritz with his dog says Karl come
b. Paul: *Theo ist gekommen*
Theo has come
c. Fritz: *ein Mann ist heute gekommen*
a man has today come
d. Karl: *jemand ist mit seinem Hund gekommen*
someone has with his dog come

[3] The unpronounced copy is crossed out in the following examples.

If, on the other hand, it is intended to express that an identical statement is made by three different individuals this is most naturally achieved by inserting one parenthetical with a coordinated subject into the host clause as in (11).

(11) *Theo ist – sagen Paul, Fritz und Karl – heute mit seinem Hund gekommen*
Theo has say Paul, Fred and Karl today with his dog come

It is obvious that the differing interpretations of the verb's complements in (10a) cannot emerge from an identical copy of the host clause inside the three parentheticals.

4 Empty/Incorporated Pronoun

Anaphoric relations across clause boundaries are regularly established by pronominal elements. Pronominals may also remain silent in certain contexts, as in pro-drop languages. Therefore the representation of the complement of the parenthetical verb by an empty pronominal is worth considering.
In the first place this account is justified by the fact that a reduced parenthetical may be freely substituted by a parenthetical with an overt pro-form. Apart from possible pragmatic effects, the interpretation of both variants is the same. The counterparts of (1) with an overt pronominal expressing the parenthetical verb's complement are listed in (12).

(12) a. *Theo kommt – Paul hat* **es** *gesagt – mit seinem Hund*
Theo comes Paul has it said with his dog
b. *Theo kommt – so hat Paul* **es** *gesagt – mit seinem Hund*
Theo comes so has Paul it said with his dog
c. *Theo kommt – wie Paul* **es** *gesagt hat – mit seinem Hund*
Theo comes as Paul it said has with his dog

In the previous section it is pointed out that there is some flexibility in the anaphoric relation of the complement to the host which is evident in multiple parenthetical constructions. The very same flexibility persists if the complement is realized by an overt pronominal.

(13) a. *Theo ist heute – Fritz sagt* **es** *– mit seinem Hund – Karl sagt* **es** *– gekommen*
Theo is today Fred says it with his dog Carl says it come
b. Fritz: *Theo ist heute gekommen*
Theo has today come
c. Karl: *jemand ist mit seinem Hund gekommen*
someone has with his dog come

In order to represent the complement, an empty pronoun within the c-structure representation of the parenthetical may be employed. LFG provides for an alternative representation at the level of f-structure alone, which will be elaborated in the following.
In the case of verb-first reduced parentheticals an alternative approach based on topic drop (Huang 1984, Sternefeld 1987) might be proposed. In German, topic drop is possible with subject and object functions.

(14) A: *was ist mit Theo?*
what is with Theo (what about Theo)
B: *ist gerade weggegangen* subject-drop
has just left
B: *habe ich gerade getroffen* object-drop
have I just met

Topic drop is also available with sentential complements alternating with a subject or an object.

(15) A: *daß Fritz kommt hat Theo überrascht*
that Fred comes has Theo surprised
B: *hat mich ebenfalls überrascht* subject-drop
has me also surprised

(16) A: *Theo hat gesagt daß Fritz kommt*
Theo has said that Fred comes
B: *hat Paul ebenfalls gesagt* object-drop
has Paul also said

However, it is impossible with other functions than subject and object. Namely, obliques are excluded from topic drop. This restriction also holds of sentential complements which alternate with an oblique function.

(17) A: *Theo hat Paul* (*darüber*) *informiert daß Fritz kommt*
Theo has Paul (correlative Prn) informed that Fred comes
B: ??*hat mich ebenfalls informiert*
has me also informed

(18) A: *Theo hat sich* (*darüber*) *beschwert daß Fritz kommt*
Theo has refl. (correlative Prn) complained that Fred comes
B: **habe ich mich gefreut*
have I refl. enjoyed

If an account of verb-first reduced parentheticals in terms of topic drop were suitable, verbs like *informieren, sich beschweren*, which either take an oblique PP or a clausal complement, would be expected to be incompatible with this construction. As the examples in (19) show this is not the case.

(19) a. *Theo kommt – informiert uns Paul – mit seinem Hund*
Theo comes informs us Paul with his dog
b. *Theo kommt – beschwert sich Paul – mit seinem Hund*
Theo comes complains refl. Paul with his dog
c. *Theo kommt – freut sich Paul – mit seinem Hund*
Theo comes enjoys refl. Paul with his dog

Apart from the fact that a topic drop analysis cannot be extended to verb-second and verb-final parentheticals since in both cases the SpecCP position is filled, it is not capable of covering the facts about verb-first parentheticals in a consistent way.

5 A Possible Objection against a Syntactic Representation

As pointed out by Jonas Kuhn (p.c.) a possible objection against a syntactic representation of the complement of the *verba dicendi et sentiendi* in reduced parentheticals may arise from certain parenthetical constructions in German whose predicate is formed by verbs that do not denote speech acts or thoughts at all. The parenthetical is functionally complete in these cases. So, for instance, the host clause of the parenthetical construction in (20) is interpreted as an utterance by the referent of the parenthetical's subject although the verb *hereinstürzen* (*to rush in*) is a verb of movement.

(20) *Theo kommt – stürzte Arthur zur Tür herein – mit seinem Hund!*
Theo comes rushed Arthur to the door in with his dog

It is obvious that in the case of (20) the attribution of the utterance of the host to *Arthur* cannot be mediated by the parenthetical's predicate. Instead, some other pragmatic advice has to be postulated in order to achieve this interpretation. If some non-syntactic account is necessary anyway then, one may argue, it should be possible to extend it to reduced parentheticals as well.
A common characteristic of (20) and (1a) obtains with respect to the structure of the parenthetical clauses. Both are apparent verb-first clauses, in both cases the sentence mood is declarative instead of interrogative. The latter fact, by the way, confirms the assumption that the sentence mood of the verb-first parenthetical is determined independently of the non overt representation of the verb's complement.
Nevertheless, constructions like (20) diverge from reduced parentheticals as in (1) to an extend that casts doubt on a unified analysis of both types. For example, multiple insertion which is possible with reduced parentheticals, do not seem as natural with functionally complete ones. (21) sounds a bit odd.

(21) ?*Theo will – erhob sich Paul vom Stuhl – heute – stürzte Arthur zur Tür herein*
Theo wants raised refl.Paul from the chair today rushed Arthur to the door in
mit seinem Hund kommen.
with his dog come

(21) becomes completely acceptable if one or the other parenthetical is cancelled.
A second more substantial divergence concerns the determination of sentence mood of the host clause. Subjunctive mood of the host is compatible with a parenthetical containing a *verbum dicendi*, but it is unsuitable with a functionally complete one.

(22) *Theo komme – sagt Paul – mit seinem Hund*
Theo $\text{comes}_{\text{subjunct}}$ says Paul with his dog

(23) ??*Theo komme – stürzte Arthur zur Tür herein – mit seinem Hund*
Theo $\text{comes}_{\text{subjunct}}$ rushed Arthur to the door in with his dog

Finally certain adverbs and focus particles which may occur freely within a reduced parenthetical are excluded from functionally complete ones.

(24) a. *Theo kommt – behauptet* sicherlich (auch) *Paul – mit seinem Hund*
Theo comes claims certainly also Paul with his dog
b. *Theo kommt – glaubt* vielleicht (sogar) *Paul – mit seinem Hund*
Theo comes believes perhaps even Paul with his dog

(25) a. ??*Theo kommt – stürzte* sicherlich (auch) *Paul zur Tür herein – mit seinem Hund*
Theo comes rushed certainly also Paul to the door in with his dog
b. ??*Theo kommt – erhebt sich* vielleicht (sogar) *Paul – mit seinem Hund*
Theo comes raises refl. perhaps even Paul with his dog

An adjunct to a functional complete parenthetical, if possible, does only modify the event denoted by the verb (*stürzte* in (26b)) but not the mode of utterance of the host.

(26) a. *Theo kommt – sagte Paul hastig – mit seinem Hund*
Theo comes said Paul hasty with his dog
b. *Theo kommt – stürzte Paul hastig zur Tür herein – mit seinem Hund*
Theo comes rushed Paul hasty to the door in with his dog

Functionally complete parentheticals a in (20) obviously lack properties of a complementation structure which, on the other hand, are common with parentheticals that contain a complement taking verb. Furthermore, the interpretation of the host in its relation to the parenthetical resembles adjunction much more than complementation. The example in (20), for instance, may be paraphrased by (27).

(27) *Mit dem Aufschrei: Theo kommt mit seinem Hund! stürzte Arthur zur Tür herein*
with the shout Theo comes with his dog rushed Arthur to the door in

The adjunct in (27), as well as the host clause in (20), modify the event denoted by the parenthetical. Adjunction may also account for the *reported speech* reading which is obligatory with functionally complete parentheticals. It is not the propositional content of the host but the act of uttering it which qualifies the Modification of the parenthetical event. Reduced parentheticals, however, like true complementation structures are not restricted to this reading.

6 Implementation

As argued in section 4, the complement of the parenthetical predicate equals a pronominal complement apart from phonological realization. This parallelism can be modelled by an empty pronominal element in the c-structure representation of the parenthetical. In an LFG mode of representation, however, it is more suitable to represent this pronominal solely in the parenthetical's f-structure representation. For certain cases of pro-drop languages, for instance, an account in terms of pronoun incorporation has been proposed by Bresnan (2001). According to this analysis, agreement morphology on the verb provides an f-structure value for the verb's SUBJ function.
In the absence of object-agreement in German, pronoun incorporation by the parenthetical's verb seems unavailable.[4] Since the non overt realization of the propositional argument de-

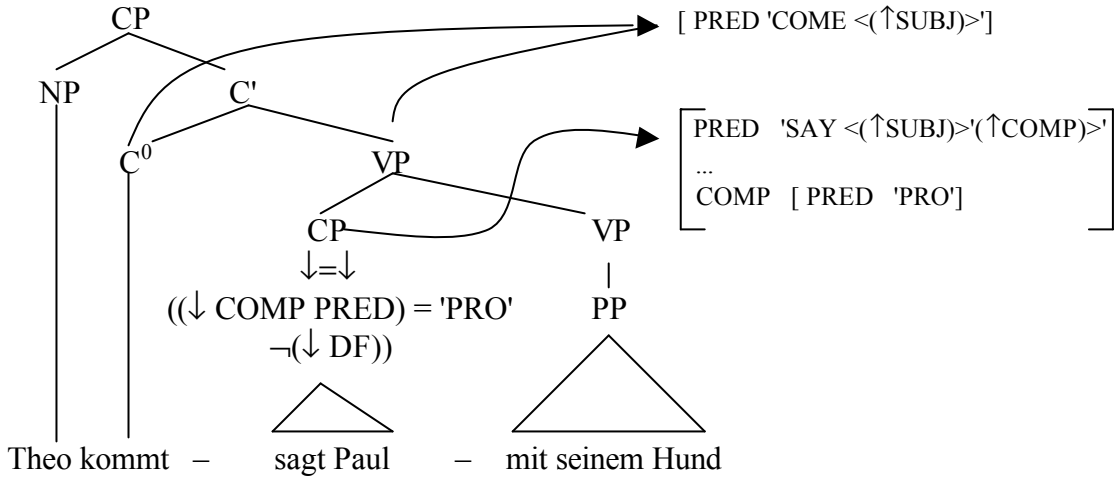
[4] But notice that an account in terms of pronoun incorporation might be pursued with reference to the morphological form of the pronominal *es*, which may occur in the parenthetical construction under consideration (cf. (12)). This pronominal element may also function as an expletive filling the SpecCP of a clause if no discourse function is defined or as a correlative element if a complement clause is extraposed (cf. Be-

pends on the specific construction, the value of the respective function has to be strusturally introduced by functional annotation of the c-structure node dominating the parenthetical.
Before going into details of the analysis proposed here, a remark on the structural relation of the parenthetical to the host is appropriate. As argued in Fortmann (2005) two types of clausal parentheticals have to be differentiated with respect to their structural integration into the host. Verb-first and verb-second parentheticals share a common c-structure representation with their host. Their f-structure representation, however, is not part of the host's f-structure. The functional dissociation is mediated by annotating a functional equation of the form ↓=↓ to the node dominating the parenthetical string. This annotation prevents the f-structure of the parenthetical from unification with the host's f-structure as well as from embedding it as the value of an f-structure attribute.
Verb-final parentheticals, on the other hand, are regular constituents, which constitute integral parts of the c-structure as well as the f-structure of the host. Their corresponding f-structure is embedded into the host's f-structure as a member of its ADJUNCT's set value.

6.1 Verb-first reduced Parentheticals

In the case of a verb-first parenthetical as in (1a) an optional annotation is added to the dominating CP-node in (28). This annotation has two components. There is a defining equation which defines the PRED value of the verbs COMP function. The restriction to verb-first structure in this type of parenthetical is captured by a negative constraint which excludes a discourse function in the parenthetical's f-structure and, as a consequence, prohibits the occurrence of any constituent in SpecCP.

(28)



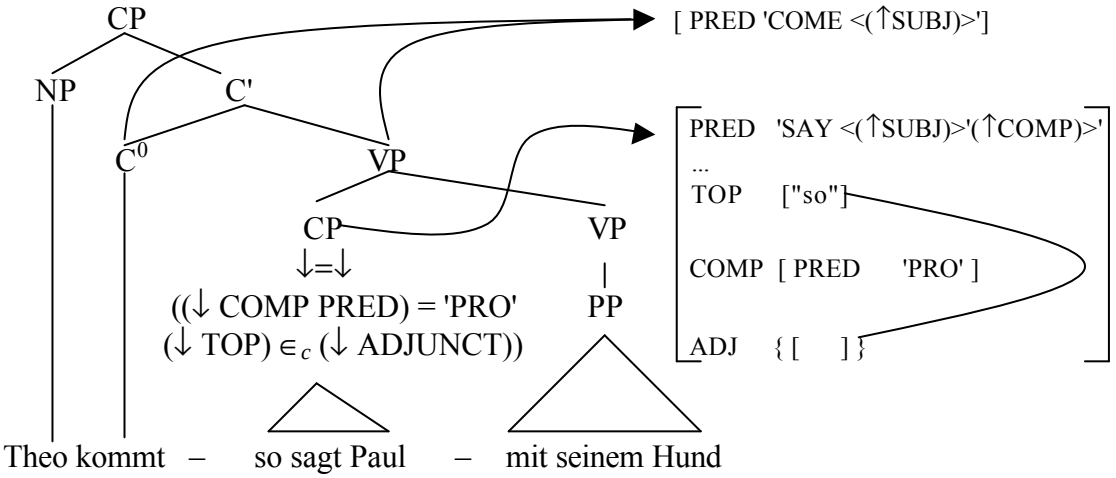
The interpretation of the host as the complement of the verb *sagt* (*say*) is mediated by the anaphoric relation of the COMP function's PRED value to the f-structure of the host clause. Since a deictic use of a pronominal is generally not possible with propositional arguments, the pronominal PRED value has to be linked to the next accessible f-structure.

man (2003)). *es* exhibits default specification of person (3. pers.) and number (sg.). It might be argued that a verb governing a COMP function is capable of defining the default agreement features of an incorporated pronominal.

6.2 Verb-second reduced Parentheticals

With respect to its structural relation to the host, a reduced verb-second parenthetical as in (1b) is on a par with a verb-first parenthetical. Its corresponding f-structure is not integrated into the f-structure of the host. The definition and the value of the verbs COMP function is likewise provided by the optional annotation of the parenthetical CP node. However, a distributional peculiarity of the reduced verb-second parenthetical has to be observed. Reduced verb-second parentheticals are only possible with a pronominal adverb *so* filling the preverbal SpecCP position.[5] Hence, the TOP function of a reduced parenthetical is excluded from unification with a governable grammatical function. Instead, this function has to be unified with a member of an ADJUNCT function. This is also justified by the interpretation. In the case of a *so*-parenthetical the literal utterance of the host is attributed to the parenthetical's subject. The proadverb *so*, which refers to the form of the host, simultaneously modifies the parenthetical's predicate.[6] The c- and f-structure representation is given in (29).

(29)



6.3 Verb-final reduced Parentheticals

As already mentioned, verb-final *wie*-parentheticals are functionally integrated into the host. In contrast to the verb-second *so*-parenthetical, no restrictions have to be imposed on the lexical choice of the adverb. Instead of *wie* (as), temporal and local adverbs may occur in the clause initial position of the parenthetical. The optional overt pronoun *es* in (30b/c) confirms the parallelism between reduced and functional complete parentheticals also in these cases.

---

[5] It is impossible to reverse the order of *so* and *Paul* in (1b):

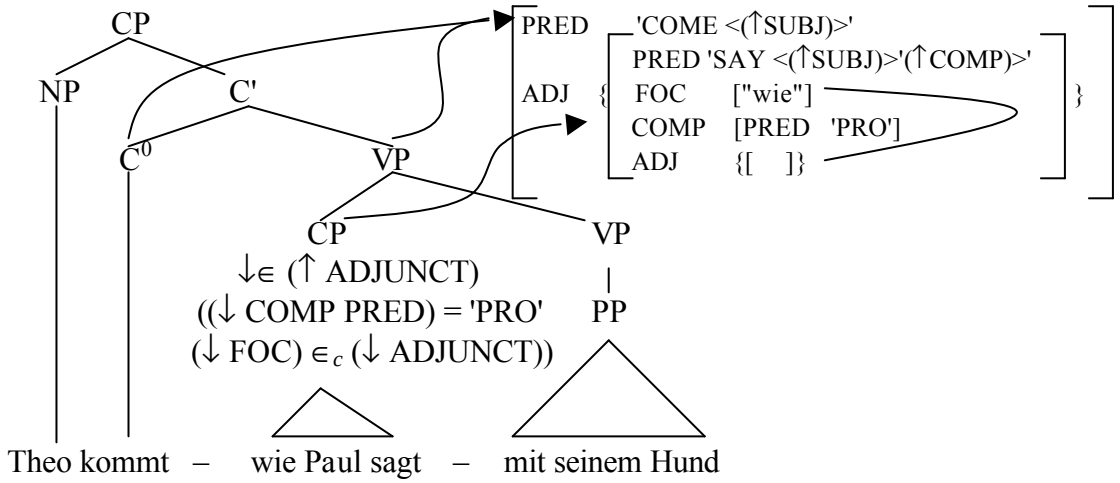
(i) **Theo kommt – Paul sagt so – mit seinem Hund*
Theo comes Paul says so with his dog

[6] The annotation proposed in (29) is necessary as far as the functional specification is concerned. It is not sufficient to determine the lexical choise of the proadverb *so*. This choice seems to depend on some general pragmatic conditions on the interpretation of this type of construction.

(30) a. *Theo kam – wie Paul sagt – mit seinem Hund*
Theo came as Paul says with his dog
b. *ich habe den Schlüssel – wo Paul* (*es*) *mir gesagt hatte – gefunden*
I have the key where Paul (it) me told had found
c. *der Zug ist – wann Paul* (*es*) *mir gesagt hatte – pünktlich angekommen*
the train has when Paul (it) me told had punctually arrived

All three parentheticals in (30) have in common that they function as free relative adjunct clauses of their respective host (cf. Desmets & Roussarie (2000) for an analogous analysis of french reportive *comme*-clauses).[7] The obligatory functional annotation of the parenthetical in (31) provides for a mapping of the CP onto an f-structure which is a member of the host's ADJUNCT 's set value. According to the matching condition on free relatives, the f-structure of the adverb *wie*, which is assigned the parenthetical's FOC function, has to be unified with a member of the parenthetical's ADJUNCT's set value. Finally, the CP node of the parenthetical has to be equipped with an optional definition of the COMP function, that is governed by the parenthetical predicate. The latter definition and the constraint on the on the unification of the FOC value are optional.

(31)



7 Restrictions

Parentheticals whose predicate is formed by verbs governing a COMP function exhibit a number of restrictions. Partly these restrictions are quite puzzling. They concern the lexical choice of the verb as well as the possibility of negation and their compatibility with certain adverbial modifiers. Some of these restrictions are independent of the structural encoding of the verb's COMP function. They obtain in reduced parentheticals as well as in parentheticals with an overt pronominal complement. Some restrictions rest on pragmatic conditions and some interact with the syntactic encoding of the COMP function.
In general a negated or negative predicate is excluded from a reduced verb-first parenthetical independent of the number specification of the subject.

[7] Notice that this account implies that the lexical items *wie, wann, wo* are categorized as Adverbs. They project maximal projections and occupy the SpecCP position of the parenthetical CP. In an analysis of *as*-parentheticals in English, Potts (2002) argues that *as* has to be categorized as a preposition which is complemented by a CP. He claims that this account holds of parallel construction in other languages as well. A uniform categorization of *wie, wann* and *wo* as preposition, however, would be rather idiosyncratic.

(32) a. **Theo kam – bestreite ich – mit seinem Hund* (subject: 1.pers)
Theo came deny I with his dog
b. **Theo kam – sage ich nicht – mit einem Hund*
Theo came say I not with his dog
c. **Theo kam – bestreitet Paul – mit seinem Hund* (subject: 3.pers)
Theo came denies Paul with his dog
d. **Theo kam – sagt Paul nicht – mit seinem Hund*
Theo came says Paul not with his dog

In (32) both the assertion of the host and the assertion of the parenthetical are attributed to the speaker. The assertion of the parenthetical, however, implies the refusal of (the truth) the host clause. The divergence of (32) seems to indicate that some condition on discourse coherence is offended. An assertion present in the discourse cannot be refused unless the refusal is explicitly marked.[8]
If the complement is expressed by an overt pronominal *es*, a negative predicate like *bestreiten* (*to deny*) is incompatible with a first person subject. Negative predicates with a second or third person subject and negated predicates in general are more acceptable than their counterparts in a reduced parenthetical. Yet a contrastive accent on either the verb or possibly some other constituent is necessary to make them fully acceptable.

(33) a. **Theo kam – ich bestreite **es** – mit seinem Hund*
Theo came I deny it with his dog
b. ??*Theo kam – ich sage **es** nicht – mit seinem Hund*
Theo came I say it not with his dog
c. ??*Theo kam – Paul bestreitet **es** – mit seinem Hund*
Theo came Paul denies it with his dog
d. ??*Theo kam – Paul sagt **es** nicht – mit seinem Hund*
Theo came Paul says it not with his dog

We will return to cases like (33b-d) immediately. Before, a second characteristic restriction has to be considered. Contrasting sentence adverbs like *allerdings*/*jedoch* (*however*) are incompatible with a reduced parenthetical.

(34) a. **Theo kommt – sage ich allerdings/jedoch – mit seinem Hund*
Theo comes say I however with his dog
b. **Theo kommt – sagt allerdings/jedoch Fritz – mit seinem Hund*
Theo comes says however Fred with his dog

Adverbs such as *jedenfalls* cannot occur in an isolated statement, anyway. They require some previous utterance in the discourse. They mark a contrast between these two statements. This contrast may result from the fact that the preceding statement is refused or otherwise modified. In (34), however, the statement attributed to the subject of the parenthetical and the statement of the host are identical. Likewise the parenthetical verb *sagen* denotes the same

[8] Such a requirement is not peculiarity of for parentheticals. If in a discourse a statement is negated by a following one at least contrastive stress on the verb or some other constituent of the following sentence is required.

action as is performed by uttering the host. Hence no contrast obtains between the parenthetical and its host.
The examples listed in (33b-d) with an overt pronominal complement improve considerably if a contrasting *jedoch* is inserted into the parenthetical and the main verb is stressed as indicated by small capitals in (35).

(35) a. *Theo kommt – ich* SAGE *es* ***jedoch*** *nicht – mit seinem Hund*
Theo comes I say it however not with his dog
b. *Theo kommt – Fritz* SAGT *es* ***jedoch*** *nicht – mit seinem Hund*
Theo comes Fred says it however not with his dog
c. *Theo kommt – Fritz* BESTREITET *es* ***jedoch*** *– mit seinem Hund*
Theo comes Fred denies it however with his dog

In all three cases the presence of the contrasting adverb is licensed by the negation or the negative predicate respectively and by the contrasting stress.
Furthermore, the interpretation of (35a/b) involves the disambiguation of the verb *sagen*. This verb means either *to utter a statement* or *to claim*. The subject of the verb is committed to the truth of the statement in the latter but not in the former case. The example in (35a) is only compatible with the first reading (*to utter*). In the case of (35b) this reading is at least preferred. The referents of the parenthetical's subjects do not deny the truth of the host (this would be contradictory in the case of (35a)), but they do not utter it. In the case of (35c) the refusal of the host, which is uttered by the speaker, is attributed to the parenthetical's subject. Since the referent of the subject and the speaker are not identical no contradiction arises.
Surprisingly the examples (32b-d) with a reduced parenthetical do not improve if *jedoch* is inserted and the main verb is stressed.

(36) a. ??*Theo kommt –* SAGE *ich* ***jedoch*** *nicht – mit seinem Hund*
Theo comes say I however not with his dog
b. **Theo kommt –* SAGT *Fritz* ***jedoch*** *nicht – mit seinem Hund*
Theo comes says Fred however not with his dog
c. **Theo kommt –* BESTREITET *Fritz* ***jedoch*** *– mit seinem Hund*
Theo comes denies Fred however with his dog

Prima facie (35) and (36) only differ with respect to the structural encoding of the parenthetical verb's complement, which is overt in (35) but not in (36).
But it is not the overt expression of the complement alone which distinguishes (35) from (36). In (35) the overt pronominal precedes the contrasting sentence adverb. As the examples in (37) demonstrate, this is indispensable. If the adverb precedes the pronominal, we obtain ungrammatical sentences.

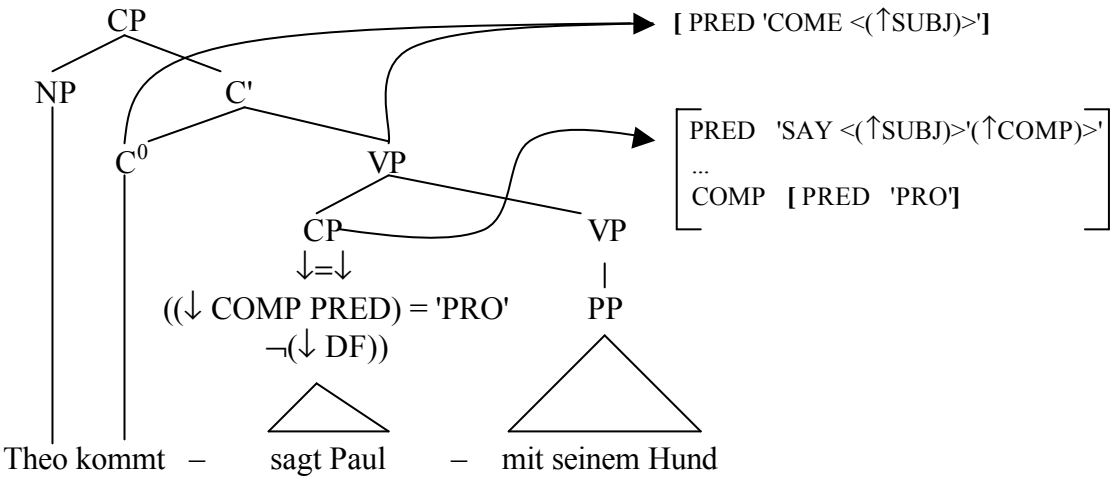
(37) a. **Theo kommt – ich* SAGE ***jedoch*** *es nicht – mit seinem Hund*
Theo comes I say however it not with his dog
b. **Theo kommt – Fritz* SAGT ***jedoch*** *es nicht – mit seinem Hund*
Theo comes Fred says however it not with his dog
c. **Theo kommt – Fritz* BESTREITET ***jedoch*** *es – mit seinem Hund*
Theo comes Fred denies however it with his dog

The divergence of (37) cannot emerge from a general ban on pronominal *es* in a position which is preceded by a sentence adverb. The second sentence in (38) is grammatical.

(38) der Bulle ist aus dem Stall ausgebrochen. Theo hat *freundlicherweise* es mitgeteilt
the bull has from the stable escaped. Theo has kindly it told

In order to account for the fact that the position of the pronominal *es* affects the grammaticality of the sentence, we have to determine a grammatical specification which depends on the position preceding the position of the sentence adverb.
The position preceding the sentence adverb has been identified as an *(aboutness) topic position* by Frey (2004). We may assume that the pronominal *es* can occupy this position in (35) but not in (36). Nonwithstanding the precise formulation of the structural conditions on the encoding of the associated topic function, it is clear that the complement of the parenthetical verb can only be associated with this function if the pronominal occupies the appropriate position. If the pronominal follows the adverb the topic function cannot be defined.
This reasoning also provides us with an explanation of the divergence of the examples in (36). On the one hand, in the absence of an overt pronominal in the appropriate position an aboutness topic cannot be defined as in the cases of (35). On the other hand, the optional annotation of the parenthetical in (28), repeated below, only provides an argument function, which is necessary to meet the completeness condition with respect to the parenthetical verb, but no additional discourse function.

(28)



After this sketch of an explanation of the structural conditions which differentiate the parenthetical constructions in (35) on the one hand, from constructions as in (36) and (37) on the other, we can turn to the question why the adverb *jedoch* requires the pronominal argument of the verb to be marked as a topic.
As pointed out above, this kind of adverb marks a contrast which obtains between two statements. In the case of (35) a contrast has to be established between the statement of the host and the statement of the parenthetical, which are both attributed to the speaker. The anaphoric relation of the pronominal complement of the parenthetical verb to the host by itself does only identify their respective content. It does not express that the host has actually been uttered in the discourse. However, by marking the pronominal as a topic of the parenthetical the content of the host is explicitly marked as the subject at issue and it can be inferred that a respective statement is already present in the discourse. Thereby the pragmatic licensing conditions on the contrasting adverb are optimally met.
The preceding discussion only considers restrictions on verb-first parentheticals. Similar restrictions can be observed with the two other types, verb-second and verb-final parentheticals.

They await further investigation. But an analysis seems promising which takes into account the structural conditions on the construction and the lexical semantics of the chosen lexical elements, in particular of the adverbs *so* and *wie*, and pragmatic conditions.

8 Summary

Based on parallels between reduced parentheticals and their counterparts containing an overt prononinal complement a syntactic representation of the missing complement of reduced *verba dicendi* parentheticals is proposed. Analogous to the case of pronoun incorporation, the locus of representation is the f-structure which corresponds to the parenthetical CP. The definition of the verb's COMP function value is provided by an optional annotation of the parenthetical CP. Structural peculiarities of the three types of reduced parentheticals – verb-first, verb-second and verb-final parentheticals – concerning the specifier position of the CP are captured by additional constraints. Certain restrictions on the choice of the parenthetical verb and its coconstituents are attributed to the interaction of pragmatic conditions on this type of parenthetical constructions and their syntactic representation.

References

Altmann, P., (1981) *Formen der 'Herausstellung' im Deutschen*. Tübingen: Niemeyer.

Berman, J., (2003) *Clausal Syntax of German*. Stanford, California: CSLI.

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

Desmets, M., L. Roussarie (2000) French Reportive *Comme* Clauses: a case of parenthetical adjunction, in: D. Flickinger, A. Kathol (eds.) *Proceedings of the 7th International HPSG Conference*. UC Berkeley.
http://csli-publications.stanford.edu/

Emonds, J., (1976) *A Transformational Approach to English Syntax*. Academic Press.

Emonds. J., (1979) Appositive Relatives Have no Properties. *Linguistic Inquiry* 10: 211–243.

Espinal, M.T. (1991) The Representation of Disjunct Constituents. *Language* 67(4): 726–762.

Fortmann, C., (2005) On Parentheticals (in German), in: M. Butt, T. Holloway King (eds.) *Proceedings of the LFG05 Conference*. Bergen.
http://csli-publications.stanford.edu/

Frey, W., (2004) A Medial Topic Position in German. *Linguistische Berichte* 198: 153-190.

Grewendorf, G., (1988) *Aspekte der Deutschen Syntax*. Tübingen: Narr.

Haider, H., (1993) ECP-Etüden. *Linguistische Berichte* 145: 185–203.

Huang, C.-J.T. (1984) On the Distribution and Reference of Empty Categories. *Linguistic Inquiry* 15(4): 531–574.

McCawley, J., (1982) Parentheticals and Discontinuous Constituent Structure, *Linguistic Inquiry* 13(1): 91–106.

Potts, C., (2002) The Syntax and Semantics of *As*-Parentheticals. *Natural Language and Linguistic Theory* 20(3): 623–689.

Reis, M., (1995): *wer glaubst du hat recht?* on So-called Extractions from Verb-Second Clauses. *Sprache & Pragmatik* 36: 27–83.

Ross., J., (1973) Slifting, in: M. Gross e.a. (eds.) *The Formal Analysis of Natural Languages*. 133–169.

Sternefeld, W., (1985) Deutsch ohne Grammatische Funktionen. *Linguistische Berichte* 99: 394-439.

THE SYNTAX OF THE MALAGASY RECIPROCAL CONSTRUCTION: AN LFG ACCOUNT

Peter Hurst
University of Melbourne

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu/

ABSTRACT
The verbal reciprocal construction in Malagasy is formed by a reciprocal morpheme prefixing on the main verb with a corresponding loss of an overt argument in c-structure. Analyses of similar constructions in Chichewa and Catalan both treat the reciprocalized verb's argument structure as undergoing an alteration whereby one of its thematic roles is either suppressed or two thematic arguments are mapped to one grammatical function. In this paper I propose that the reciprocal morpheme in Malagasy creates a reciprocal pronoun in f-structure - thus maintaining its valency and leaving the argument structure of the verb unchanged, while at the same time losing an argument at the level of c-structure.

1. INTRODUCTION

Malagasy is an Austronesian language and is the dominant language of Madagascar. The Malagasy sentences used in the analysis below are from the literature - in particular from a paper by Keenan and Razafimamonjy (2001) titled “Reciprocals in Malagasy” whose examples are based on the official dialect of Malagasy as spoken in and around the capital city Antananarivo.

The Malagasy reciprocal construction is formed by the addition of a prefix *-if-* or *-ifamp-* to the stem of the verb accompanied by the loss of an overt argument in object position. Compare sentence (1a) below with its reciprocated equivalent (1b):

(1) Malagasy
a. *N-an-daka an-dRabe Rakoto*
pst-act-kick acc.Rabe Rakoto
V O S
'Rakoto kicked Rabe'

b. *N-if-an-daka Rabe sy Rakoto*
pst-rec-act-kick Rabe and Rakoto
V S
'Rabe and Rakoto kicked each other' (Keenan & Razafimamonjy 2001:47)

Like Malagasy, the verbal reciprocal constructions in Chichewa and Catalan are similarly formed by a reciprocal morpheme attaching to the verb with a corresponding loss of an overt argument in c-structure. Furthermore, all the participants involved in the reciprocal relation are grouped into a plural NP:

(2) Chichewa (for same gender nouns)
a. *Galimoto inagunda njinga*
car it-past-hit-fv bicycle (where car and bicycle have the same gender)
'The car hit a bicycle'

b. *Njinga inagunda galimoto*
bicycle it-past-hit-fv car
'A bicycle hit a car'

c. *Galimoto ndi njinga zinagundana*
car and bicycle it-past-hit-rec-fv
'A car and a bicycle hit each other' (Mchombo & Ngalande 1980)

(3) Catalan

Les ballarines s'han vist

the ballerinas rec/refl'have seen

'The ballerinas have seen each other/themselves' (Alsina 1996:95)

The analyses of the verbal reciprocal constructions in Chichewa (Mchombo & Ngalande 1980, Mchombo 1991 and Dalrymple *et al.* 1994) and Catalan (Alsina 1996) both treat the reciprocalized verb's argument structure as undergoing an alteration whereby one of its thematic roles is either suppressed or two thematic arguments are mapped to one grammatical function. This is illustrated in figure 1 below:

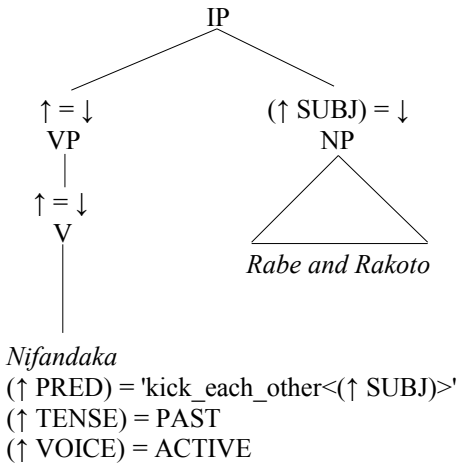
Figure 1.

Analysis of the Formation of Reciprocal Expressions in Chichewa (see Mchombo 1991:16)	**Analysis of the Formation of Reciprocal Expressions in Catalan (see Alsina 1996:260-263)**
A-structure: *respect*<Agent, Patient> ø *F-structure: respect_each_other*<SUBJ>	*A-structure:* *respect*<Agent, Patient> *F-structure:* *respect_each_other*<SUBJ>

In Mchombo's analysis of the reciprocal construction in Chichewa, he uses LMT (Lexical Mapping Theory) to suppress the patient argument of the verb in a manner very similar to the analysis of the passive construction in LFG. This results in the verb at the level of f-structure having only one argument. Alsina's analysis of Catalan is similar insofar as he maps both the agent and patient thematic roles to the subject of the verb at the level of f-structure. I call either of these analyses "valency reducing" because the number of arguments that the verb selects at the level of f-structure is reduced when the verb is reciprocated.

Under this valency reducing analysis, the f- and c-structures for sentence (1b) are represented by figure 2 below:

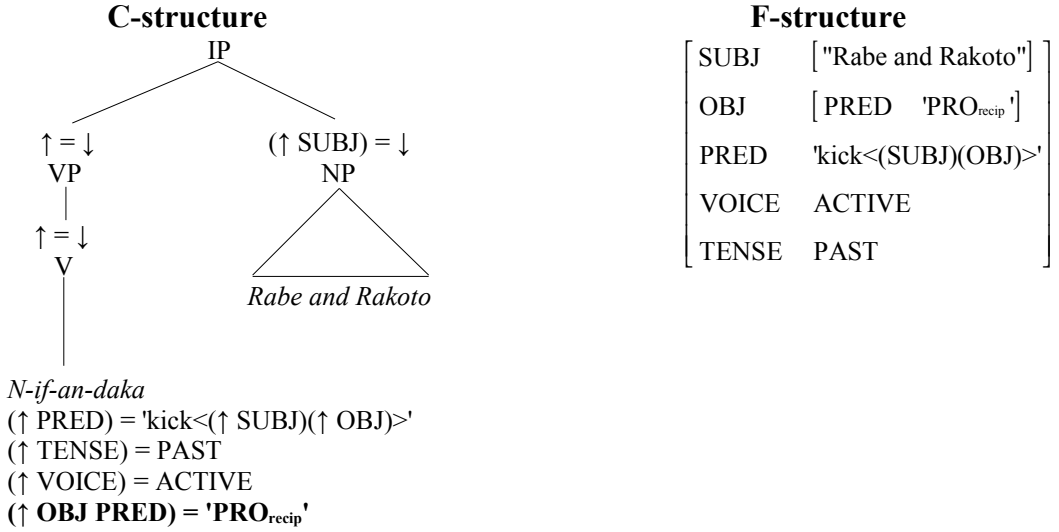
Figure 2. The Valency Reducing Analysis of 'Rabe and Rakoto kicked each other'



$$\begin{bmatrix} \text{SUBJ} & [\text{"Rabe and Rakoto"}] \\ \text{PRED} & \text{'kick\_each\_other<(SUBJ)>'} \\ \text{VOICE} & \text{ACTIVE} \\ \text{TENSE} & \text{PAST} \end{bmatrix}$$

However, the architecture of LFG allows for another way of representing reciprocal constructions. Under this representation, the a- and f-structures of the verb remain unchanged when it is reciprocated and instead, a reciprocal pronoun sits in the object position of the f-structure of the clause. Under this analysis, sentence (1b) has the f- and c-structures given in figure 3:

Figure 3. The Valency Preserving Analysis of 'Rabe and Rakoto respect each other'



I call this analysis the "valency preserving analysis" because the verb's valency at the level of f-structure doesn't change when it's reciprocated. This is because the object position in the f-structure is filled by a reciprocal pronoun. This reciprocal pronoun is created by the reciprocal morpheme which has the definition given in (4):

(4) *-if- / -ifamp-* (↑ OBJ PRED) = 'PRO$_{recip}$'

The definition for this reciprocal morpheme can be found in the lexical entries associated with the verb in which it appears (for example, see the lexical entries for the verb *Nifandaka* in figure 3 above.)

The reciprocal pronoun is a place-holder for reciprocal semantics. The mechanics of how it's linked to reciprocal semantics are beyond the scope of this paper - however, this idea of a reciprocal pronoun is compatible with the work by Dalrymple, Kanazawa, Kim, Mchombo and Peters (1998) where they treat reciprocal expressions as quantifiers. For example, sentence (1b) expresses a proposition like (5) below which has a quantifier RECIP, a domain of *Rabe* and *Rakoto*, and a scope, or expression to which the quantifier is applied:

(5) RECIP({Rabe,Rakoto}, λxy.kick(x,y))

To conclude, the valency preserving analysis of reciprocal constructions exploits the parallel structures in LFG, allowing an object to be missing at the level of c-structure, yet present at the level of f-structure. This creates a mismatch in the valency of the verb - and in the following sections I explain how this mismatch can account for the behaviour of the Malagasy reciprocal construction.

2. THE MALAGASY RECIPROCAL CONSTRUCTION

In this section I examine the Malagasy reciprocal construction in more detail - in particular, the more complex syntactic environments which will prove instrumental in demonstrating the efficacy of the valency preserving analysis of the Malagasy reciprocal construction.

Sentence (6a) below shows a typical transitive sentence with VOS word order. The object, when it is a pronoun or a proper noun, receives an ACC case marker. The verb has two prefixes, a tense marker *N*, and a voice marker *-an-*. In general, verbal morphology can also indicate aspect, reciprocity and causality. Sentence (6b) shows the corresponding reciprocal construction and sentence (6c) shows a typical intransitive sentence:

(6) Transitive Verb

a. *N-an-daka an-dRabe Rakoto*
pst-act-kick acc.Rabe Rakoto
V O S
'Rakoto kicked Rabe'

b. *N-if-an-daka Rabe sy Rakoto*
pst-rec-act-kick Rabe and Rakoto
V S
'Rabe and Rakoto kicked each other'

Intransitive Verb
c. *M-i-jaly Rabe*
pres-act-suffer Rabe
V S
'Rabe suffers' (Keenan & Razafimamonjy 2001:47,70)

As can be seen, the reciprocal construction is formed by:

- Prefixing the *-if-* morpheme to the verb.
- Gathering the participants into a plural (usually subject) NP.
- The loss of an overt, non-subject argument.

It is clear when examining these simple sentences that the reciprocal construction resembles an intransitive construction. It should be noted that treating *Rabe* and *Rakoto* in (6b) as the subject of the sentence is uncontroversial and can be demonstrated by a variety of tests. For example, Keenan and Razafimamonjy (2001) demonstrate that the subject NP in (6b) can be relativized - an operation only available to subject NPs in Malagasy:

(7) *Ny olona (izay) n-if-an-daka*
the people rel pst-rec-act-kick
The people (that) were kicking each other
(Keenan & Razafimamonjy 2001:22)

Keenan & Razafimamonjy (2001:22) provide further evidence demonstrating that the remaining argument is in fact the subject of the clause. For example, *Rabe* and *Rakoto* can be substituted with the nominative pronoun *izy*, but not its accusative equivalent *azy*. Keenan & Razafimamonjy (2001:22) state that comparable claims regarding the subjecthood of the NP hold for the other reciprocal sentences they studied (and which appear below) such as those formed from ditransitive or semi-transitive verbs.

The reciprocal construction in Malagasy is very productive, appearing with verbs with a variety of arities and in a variety of constructions. The examples below are by no means exhaustive. In (8) below, the ditransitive verb *manome* 'give' is missing its indirect object rather than its direct object when reciprocated:

(8) Ditransitive Verb

a. *M-an-ome* *vola* *an-dRabe* *Rakoto*
pres-act-give money acc.Rabe Rakoto
V DO IDO S
'Rakoto gives money to Rabe'

b. *M-if-an-ome* *vola* *Rabe sy Rakoto*
pres-rec-act-give money Rabe and Rakoto
V DO S
'Rabe and Rakoto give money to each other' (Keenan & Razafimamonjy 2001:49)

In (9a) below the semi-transitive verb *mipetraka* 'sit' takes the prepositional phrase "near Ranaivo" as a complement. The reciprocal equivalent to (9a) retains the preposition, but the NP it selected is now missing:

(9) Semi-transitive Verb

a. *M-i-petraka* *akaikin-dRabe* *Ranaivo*
pres-act-sit near-Rabe Ranaivo
V OBL S
'Ranaivo is sitting near Rabe'

b. *M-ifamp-i-petraka* *akaikin* *Rabe sy Ranaivo*
pres-rec-act-sit near Rabe and Ranaivo
V Prep S
'Rabe and Ranaivo are sitting near each other' (Keenan & Razafimamonjy 2001:50)

Sentences (10) and (11) are examples of causative and circumstantial constructions - both of which can co-occur with the reciprocal construction:

(10) Causative Constructions

a. *N-if-an-daka* *Rabe sy Rakoto*
pst-rec-act-kick Rabe and Rakoto
'Rabe and Rakoto kicked each other'

b. *N-amp-if-an-daka* *an-dRabe sy Rakoto aho*
pst-caus-rec-act-kick acc.Rabe and Rakoto 1sg.nom
'I made Rabe and Rakoto kick each other'
(Keenan & Razafimamonjy 2001:67)

(11) Circumstantialization

a. *N-if-an-tao* *farafara* *amin'ity vy ity* *Rabe sy Rakoto*
pst-rec-act-do bed with.this metal this Rabe and Rakoto
'Rabe and Rakoto made each other beds from this metal'

b. *N-if-an-tao-van-dRabe sy Rakoto* *farafara* *ity* *vy* *ity*
pst-rec-act-do-circ.Rabe and Rakoto bed this metal this
lit. 'This metal was made by Rabe and Rakoto beds for each other'
'This metal was made into beds for each other by Rabe and Rakoto'
(Keenan & Razafimamonjy 2001:56)

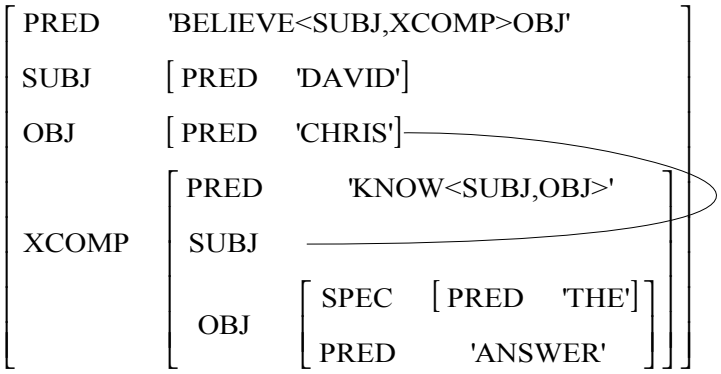
The analysis of reciprocalized circumstantial, nominalization and causative constructions will not form part of this paper.[1]

The first thing we see is that the reciprocal construction in Malagasy involves the loss of an overt argument. When simple transitive sentences are reciprocalized (*e.g.*, 6b), they look just like intransitive sentences (*e.g.*, 6c). In fact, if we were to examine only simple sentences, it would seem that the valency reducing analysis Mchombo made for Chichewa could be simply applied to Malagasy. What is needed are some syntactic environments which can tease apart these two analyses by showing that the valency of the verb has actually changed. I turn to two of these environments in sections 3 and 4 below.

3. FUNCTIONAL CONTROL

One environment which can be used to test the valency of a verb at the level of f-structure is functional control. In figure 4 below, I have reproduced the standard LFG analysis for the verb *believe* when used in a functional control sentence.

Figure 4. "David believed Chris to know the answer."



(Dalrymple 2001:314)

In the sentence *David believed Chris to know the answer*, *Chris* is acting as both the object of the main verb *believe* and as the subject of the lower verb *know*. Functional control constructions can be used to test the valency of the main verb because of the presence of the shared object. The valency preserving analysis of the Malagasy reciprocal construction predicts that the main verb in functional control sentences should be able to be reciprocated because the object is still present at the level of f-structure and is able to act as the subject of the lower clause. On the other hand, the valency reducing analysis of reciprocal constructions predicts the main verb should not be able to be reciprocated. This is because the object is no longer present at the level of f-structure - leaving the lower clause without a subject.

1 See Hurst (2003) for an account on how these constructions can be analysed in LFG.

3.1 Functional Control in Malagasy

I now turn to functional control in Malagasy. Among other features, Paul and Rabaovololona (1998) use the following criteria to identify a construction they call RTO: 'Raise to Object' in Malagasy:[2]

1. The embedded clause appears in typical object position (as opposed to other complements and sentential adjuncts which appear after the subject).
2. The embedded clause has atypical word order (SVO instead of VOS).
3. The complementizer *ho* is used instead of *fa*.

Compare (12a) and (12b) below to see the difference between a regular complement construction and an RTO construction:

(12)

a. *N-i-laza an-dRasoa ho namboly vary Ravelo*
pst-act-say acc.Rasoa Comp pst.cultivate rice Ravelo
[V O Comp V O] S]
lit. 'Ravelo said Rasoa to have cultivated rice'
'Ravelo said Rasoa has cultivated rice' -- 'RTO' Construction

b. *N-i-laza Ravelo fa namboly vary i Soa*
pst-act-say Ravelo Comp pst-cultivate rice Art Soa
[V Subj] [Comp V Obj Subj]
'Ravelo said that Soa cultivated rice' -- Complement Construction
(Keenan & Razafimamonjy 2001:50-51)

The RTO construction in (12a) is indicated by the embedded clause *ho namboly vary* - 'to have cultivated rice' between the object and the subject of the main clause. In contrast to this, the complement construction in (12b) has a complement clause appearing after the subject of the main clause. Furthermore, the complement clause is a complete clause with an overt subject and it is introduced by the complementizer *fa,* whereas the RTO construction in (12a) uses *ho.*

The construction which Paul and Rabaovololona identify as RTO has all the hallmarks of functional control - and this is seen in the unusual status of the NP *Rasoa* in (12a). *Rasoa* receives accusative case marking suggesting that it is the object of 'say'. However, although *Rasoa* in (12a) is marked as an object, it must also conform to the general requirement that all subjects in Malagasy be specific. This is demonstrated in (13) below, where the sentence *a child is reading a book* is grammatical with a subject NP that picks out a specific entity in the world, but non-grammatical with a non-specific subject:

(13) *mamaky boky i Bao/izy/ ny zaza /ilay zaza/*olona/*zaza*
pres-act-read book Bao/3SG/the child/that child/person/child
'Bao/(s)he/the child/that child is reading a book'
Some person/child is reading a book (Pearson 2001:19)

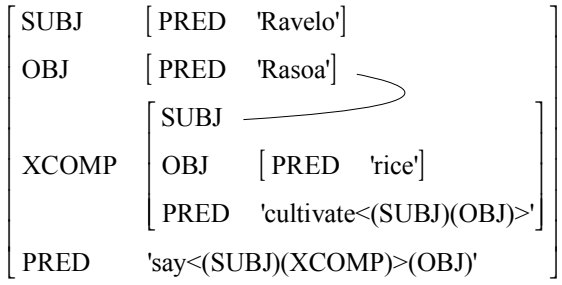
This specific subject requirement is also required of the NP marked in accusative case in RTO sentences. For example, in (14) a non-specific NP in the same position as *Rasoa* renders the sentence ungrammatical:

2 Paul and Rabaovololona use this term in a pre-theoretical sense - "we intend it (RTO) to describe a class of constructions without any implication as to the final analysis".

(14) *M-i-hevitra zanaka ho hendry aho
pres-act-think child comp wise 1sg(nom)
'I think some child is wise' (Paul & Rabaovololona 1998:55)

In other words, *Rasoa* in (12a) is acting as the object of main verb and so is marked with ACC case, and by acting as the subject of the lower clause, must also conform to the specific subject requirement. Figure 5 below shows the f-structure for sentence (12a) when analysed as functional control.

Figure 5. Functional Control: F-structure for (12a)



We now return to the issue raised at the start of this section - can functional control sentences in Malagasy be reciprocalized? As sentence (15b) demonstrates, the answer to this question is yes:[3]

(15)

a. *N-i-laza an-dRasoa ho namboly vary Ravelo*
pst-act-say acc. Rasoa Comp pst.cultivate rice Ravelo
[V O [Comp V O] S]
'Ravelo said Rasoa has cultivated rice'
lit. 'Ravelo said Rasoa to have cultivated rice'

b. *N-ifamp-i-laza ho namboly vary Rasoa sy Ravelo*
pst-rec-act-say Comp pst-cultivate rice Rasoa and Ravelo
[V [Comp V O] S]
lit. 'Rasoa and Ravelo said each other to have cultivated rice'
'Rasoa and Ravelo said of each other that s/he cultivated rice'
*'Rasoa and Ravelo said "we cultivated rice"'[4]
(Keenan & Razafimamonjy 2001:50-51)

In (15b) we see the reciprocal equivalent of (15a). An allomorph of the reciprocal morpheme, *ifamp,* has been prefixed to the main verb. The subject is now plural and the structure is accompanied by the lack of an overt NP in object position. As (15b) shows, the valency preserving analysis correctly predicts that functional control sentences can be reciprocalized. However, the grammaticality of this sentence creates problems if the valency reducing analysis is to be maintained. In section 3.2 below, I demonstrate precisely what these problems are, before demonstrating in section 3.3 how the valency preserving analysis of reciprocal constructions accounts for the grammaticality of sentences like (15b).

3 Sentence (12a) is reproduced here as (15a) for convenience
4 Keenan & Razafimamonjy (2001:50) specifically rule out the shared reciprocal reading in control verbs.

3.2 Analysis 1. The Valency Decreasing Analysis of the Malagasy Reciprocal Construction

In this section I demonstrate why the valency decreasing account of reciprocal constructions cannot account for the grammaticality of reciprocated control constructions.

For the sake of argument, assume that the reciprocated verb has lost an object at the level of f-structure. This means that its lexical entry would have to be modified from (16a), the typical definition of *Nilaza* - 'say', to the reciprocated version in (16b) - *Nifmpilaza* 'say_of_each_other':

(16a) *Nilaza* *v* (↑PRED) = 'say<(↑SUBJ)(↑XCOMP)>(↑OBJ)'
(↑XCOMP SUBJ) = (↑OBJ)
(↑TENSE) = PAST
(↑VOICE) = ACTIVE

(16b) *Nifampilaza* *v* (↑PRED) = 'say_each_other<(↑SUBJ)(↑XCOMP)>'
(↑XCOMP SUBJ) = ???
(↑TENSE) = PAST
(↑VOICE) = ACTIVE

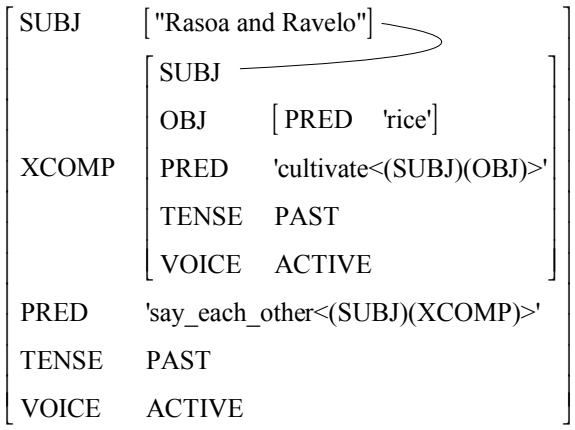
The resulting f-structure is no longer well-formed because the lower clause has a SUBJ feature with no attribute:

SUBJ	["Rasoa and Ravelo"]	
XCOMP	SUBJ	
	OBJ	[PRED 'rice']
	PRED	'cultivate<(SUBJ)(OBJ)>'
	TENSE	PAST
	VOICE	ACTIVE
PRED	'say_each_other<(SUBJ)(XCOMP)>'	
TENSE	PAST	
VOICE	ACTIVE	

To attempt to save this analysis, we could posit an additional change in the lexical entry of *Nifampilaza* so that the controlled verb's subject is linked to the subject of the main verb instead of the object. And so, the lexical entry for the reciprocated form of (16a) becomes (17):

(17) *Nifampilaza* *v* (↑PRED) = 'say_each_other<(↑SUBJ)(↑XCOMP)>'
(↑XCOMP SUBJ) = (↑SUBJ)
(↑TENSE) = PAST
(↑VOICE) = ACTIVE

This change to the lexical entry of *nifampilaza* at least produces an f-structure which is well formed:



However, this change cannot be maintained because reciprocalized control constructions can also be nominalized. In the case of nominalized reciprocated control expressions, the verb has lost both its subject and its object and so the lower clause is left without a subject.[5] To conclude, the valency reducing analysis is difficult to maintain in the face of reciprocated functional control expressions.

3.3 Analysis 2. The Valency Preserving Analysis of the Malagasy Reciprocal Construction

In this section I consider the valency preserving analysis where the valency of the verb doesn't change when it's reciprocated. Under this analysis the reciprocated verb's PRED does not change at all and instead, the reciprocal interpretation arises from the presence of the reciprocal morpheme *-if-* which creates a reciprocal pronoun in object position in the f-structure.

Examining the lexical entries in (19) for *nifampilaza* 'say', we see that the reciprocated verb has the same lexical entry as its non-reciprocated counterpart, apart from the additional information provided by the reciprocal morpheme which creates a reciprocal pronoun in object position:

(18) *-if-/-ifamp-* (↑OBJ PRED) = 'PRO_{recip}'

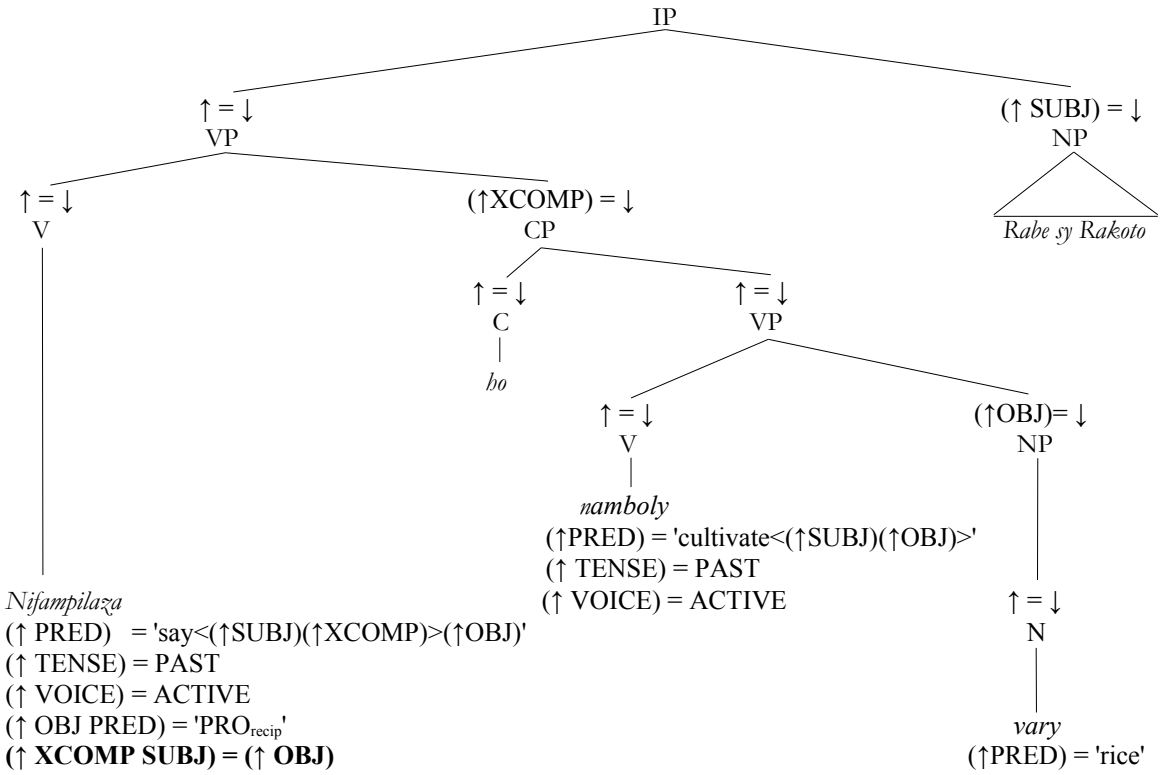
The lexical entries for (15b) are then:

(19)

N-ifamp-i-laza	*v*	(↑PRED) = 'say<(↑SUBJ)(↑XCOMP)>(↑OBJ)'
		(↑XCOMP SUBJ) = (↑OBJ)
		(↑OBJ PRED) = 'PRO_{recip}' (from the rec. morpheme)
		(↑VOICE) = ACTIVE
		(↑TENSE) = PAST
Namboly	*v*	(↑PRED) = 'cultivate<(↑SUBJ)(↑OBJ)>'
		(↑VOICE) = ACTIVE
		(↑TENSE) = PAST
Vary	*n*	(↑PRED) = 'rice'

5 For a detailed account of the reciprocalization of nominalized control expressions in Malagasy, see Hurst (2003).

Figure 6. C-structure for sentence (15b)



The c-structure above results in the following complete and coherent f-structure:

$$\begin{bmatrix} \text{SUBJ} & [\text{"Rasoa and Ravelo"}] \\ \text{OBJ} & [\text{PRED} \quad \text{'PRO}_{recip}\text{'}] \\ \text{XCOMP} & \begin{bmatrix} \text{SUBJ} & \\ \text{OBJ} & [\text{PRED} \quad \text{'rice'}] \\ \text{PRED} & \text{'cultivate<(SUBJ)(OBJ)>'} \\ \text{TENSE} & \text{PAST} \\ \text{VOICE} & \text{ACTIVE} \end{bmatrix} \\ \text{PRED} & \text{'say<(SUBJ)(XCOMP)>(OBJ)'} \\ \text{TENSE} & \text{PAST} \\ \text{VOICE} & \text{ACTIVE} \end{bmatrix}$$

This f-structure is both complete and coherent because the reciprocal pronoun is acting as both the subject of the XCOMP and the object of *say*. Because the valency preserving analysis does not change the valency of the reciprocated verb at the level of f-structure, the control construction does not require the invention of new syntactic rules to account for its reciprocal counterpart - reciprocal control constructions can be treated just like regular control clauses. In contrast to the valency decreasing analysis, the valency preserving analysis of the Malagasy reciprocal construction correctly predicts the grammaticality of reciprocated control expressions.

3.4 Summary

In section 3, I examined two possible analyses of the reciprocal construction in control sentences - one where the reciprocated verb has had its valency reduced in f-structure and another where the verb's valency remains unchanged and the reciprocal morpheme creates a reciprocal pronoun in the object position of the main verb.

The valency preserving analysis predicts that control constructions can participate in reciprocal relations - and this is what is observed. In contrast to this, the valency reducing analysis predicts that control sentences should be ungrammatical with a reciprocated verb. This analysis might be able to be saved by the questionable addition of a rule which changes the lexical entry of the verb so that the XCOMP's SUBJ no longer is linked with the object of the main verb, but with its subject - although for the reasons given above, I find this analysis unlikely.

4. POSSESSION CONSTRUCTIONS

Possession is another environment which helps us to understand the valency of a verb in f-structure. The valency preserving analysis of the Malagasy reciprocal construction predicts that sentences containing a possessor relation in object position should be able to be reciprocalized because the reciprocal pronoun can enter into a possession relation. In contrast to this, the valency decreasing analysis predicts that these sentences should not be able to be reciprocalized (unless entering into a new construction) because the resulting construction has no object. For example, the reciprocal construction in English, which uses an overt reciprocal pronoun, is able to enter into possessor relations:

(20) John and Mary saw each other's parents.

In (20) the reciprocal pronoun *each other* is acting as the possessor of *the parents*. Under a valency reducing analysis of reciprocity, the reciprocated verb loses an object - so it ought to be impossible to form reciprocated expressions of this type.

4.1 Possession in Malagasy

The Malagasy possession construction is fairly easy to form. However, the affected nouns do undergo some complex (but well understood) phonological changes (Paul 1996). The construction itself is formed by inserting the possessor noun to the right of the noun possessed with an attendant phonological change:

(21) *orona* + *olona* → *oron'olona*
nose person nose of a person / a person's nose (Paul 1996:77)

In (22) below, the possession construction occurs in the object of the transitive verb *maka* - 'ravish/take'. The resulting interpretation is that it is *Rabe's* spouse being ravished, not *Rabe*. Note that the *-d* in front of *Rabe* is not a case marker, but rather the result of the phonological change which comes about from merging *vady* and *Rabe*.

(22)

a. *M-aka* *ny* *vadin-dRabe* *Rakoto*
pres-take the spouse.of-Rabe Rakoto
V [Obj] [Subj]
'Rakoto ravishes the spouse of Rabe'

b. *M-ifamp-aka* *vady* *Rabe* *sy* *Rakoto*
pres-rec-take spouse Rabe and Rakoto
V Remnant N Conj N
'Rabe and Rakoto ravish each other's spouse' (Keenan & Razafimamonjy 2001:51)

Sentences such as (22a) can be reciprocalized productively as (22b) demonstrates. Examining (22b) we see that when reciprocating a sentence containing a possession construction in the object NP, the possessor disappears leaving, in this instance, just *vady* - the possessee. Following Keenan & Ralalaoherivony (1998), I will call what is left of the POSS construction after reciprocation the remnant (*e.g.*, in (22b) above, *vady* is the remnant). How to account for where the remnant belongs syntactically under the valency decreasing and preserving analyses is investigated in sections 4.2 and 4.3 below.

4.2 Analysis 1. The Valency Reducing Reciprocal Construction

Under the valency reducing analysis, it is possible that the reciprocal possession construction is a new construction and in this section I investigate this possibility. Under this analysis the verb *maka,* 'ravish', no longer selects an object when reciprocalized:

(23)
a. *maka* *v* (↑PRED) = 'take<(↑SUBJ)(↑OBJ)>'
b. *m-ifamp-aka* *v* (↑PRED) = 'take_each_other<(↑SUBJ)>'

Clearly, the remnant cannot belong to the object function under this analysis. So where can it belong? There are two possible candidates:

1. The Remnant Belongs to the Subject NP
We can immediately rule out this candidate. If *vady* were in a possessor relationship with the subject NP, it should have undergone the attendant phonological operation that binds it to the N or NP which possess it. For example compare (22b) with (24) where *vady* 'spouse' does belong to the subject NP:

(24) *N-if-an-lainga* *ny vadin-dRavelo* *sy* *Rasoa*
past-rec-act-lie [the spouse-of-Ravelo and Rasoa]
'The spouses of Ravelo and Rasoa lied to each other' (Keenan & Ralalaoherivony, 1998:84)

That *vady* is unchanged in (22b) indicates that it is not in a possessor relation with the subject NP.

2. The Remnant is part of an External Possession Construction
Could the remnant be analysed as being in an external possession construction with the reciprocated verb? Keenan and Ralalaoherivony (1998) detail several features which characterize the external possession construction in Malagasy. Among them are:

1. The loss of a determiner associated with the noun.
2. The newly formed verb+noun group acts as a prosodic word - question particles and adverbs can't be inserted between them.
3. There is a semantic shift where the subject of the verb is more involved in the event described.
4. Control verbs can't occur with the external possession construction (whether RTS or RTO).

Figure 7 below shows an example of a typical external possession construction.

Figure 7

Maty ny vadin-dRabe [died [the spouse-of-Rabe]] 'Rabe's spouse died/is dead'	↔	*Maty vady Rabe* [[died spouse] Rabe] 'Rabe was widowed' or 'Rabe underwent spouse death'
Regular Possession		**PossR / External Poss**

(Keenan and Ralalaoherivony 1998:65)

Compare the external possession construction (25a) with the reciprocal construction (25b):

(25)

a. *Maty vady Rabe*
dead spouse Rabe
'Rabe is widowed'

b. *M-ifamp-aka vady Rabe sy Rakoto*
pres-rec-take spouse Rabe and Rakoto
'Rabe and Rakoto ravish each other's spouse'

At first glance, it appears possible that external possession can account for the remnant. However, there are two insurmountable problems in treating the remnant as being part of an external possession construction. The first is that it is possible to get phonetic material between the remnant and the verb. For example, the semi-transitive verb 'lie' *mandainga* in (26a) uses a preposition *amin* to introduce the oblique phrase. When this sentence is reciprocated the possessee remains, but is preceded by the preposition:

(26)

a. *M-an-dainga amin'ny vadin-dRakoto Rabe*
pres-act-lie to.the spouse.of-Rakoto Rabe
V [OBL] S
'Rabe lies to Rakoto's spouse'

b. *M-if-an-dainga amin-bady Rakoto sy Rabe*
pres-rec-act-lie to.spouse Rakoto and Rabe
'Rakoto and Rabe lie to each other's spouse(s)' (Keenan & Razafimamonjy 2001:52)

Another example is in (27) where the possessor construction occurs in the indirect object. When reciprocated, the direct object is still between the remnant and the verb:

(27)

a. *M-an-ome vola ny zanan-dRavelo Rasoa*
pres-act-give money the child.of-Ravelo Rasoa
V O [IDO] S
'Rasoa gives money to the children of Ravelo'

b. *M-if-an-ome vola* ***zananaka*** *Rasoa sy Ravelo*
pres-rec-act-give money child Rasoa and Ravelo
V O remnant [S]
'Rasoa and Ravelo give money to each other's children'
(Keenan & Razafimamonjy 2001:52)

As well as syntactic difficulties in assigning a function to the remnant, there are also theoretical problems arising from the a-structure mapping. If we use the standard approach to argument suppression to account for the valence reduced reciprocal construction, then it is not clear why any remnant should be present at all. For example, to explain the loss of an argument in the Chichewa reciprocal construction, Mchombo (1991) suppresses the patient argument in the a-structure of the verb in a manner very similar to the analysis Bresnan (2001) uses for the passive construction (see figure 1). Likewise, Alsina's (1996) analysis of the reciprocal construction in Catalan is similar. He links both the agent and patient to just the SUBJ argument in the f-structure of the verb. Clearly Mchombo's analysis for Chichewa cannot be used for Malagasy because in a sentence like (25b), *vady* is part of the patient thematic role. For it to appear in a reciprocal sentence means that the patient hasn't been wholly suppressed. Likewise, under Alsina's analysis, if *vady* were going to appear anywhere, it would have to be part of the subject at the level of F-structure - and it's not. From this theory internal point of view then, it appears that the patient is not suppressed in reciprocal constructions - especially in conjunction with possession.

The valency reducing analysis failed to account for the reciprocation of possessor relations for the following reasons:

1. It could not assign a plausible function to the remnant.
2. The remnant is the possessee and belongs to the patient role of the verb. Contrary to the theoretical underpinnings of the valency reducing analysis, it appears that the patient is only partially suppressed in POSS constructions.

Analysis 2. The Valency Preserving Reciprocal Construction

The valency preserving analysis of reciprocal constructions predicts that these possession constructions should be grammatical - and the details of how it works are straightforward. As nouns in a possession relationship are modified by a lexical rule to allow them to select a POSS function, the lexical definition of the reciprocal morpheme must also be modified in a similar way so that the reciprocal pronoun can exist in a POSS function. This is accomplished by adding an optional POSS function to the lexical entry of the reciprocal morpheme:

(28) *-if-* (↑OBJ (POSS) PRED) = 'PRO_{recip}'

This is the only change required to account for the grammaticality of reciprocal constructions with objects containing a possession relation. For example, sentence (29) below can now be analysed straightforwardly using the standard analysis of control and possession in conjunction with the valency preserving analysis of the reciprocal construction:

(29)

M-ifamp-ilaza	*ray*	*aman-dreny*	*ho*	*mamboly*	*vary*	*Rabe sy Rakoto*
pres-rec-act-say	father	and mother.of	comp	pres-cultivate	rice	Rabe and Rakoto
V	[O	]	[XCOMP		]	[S]

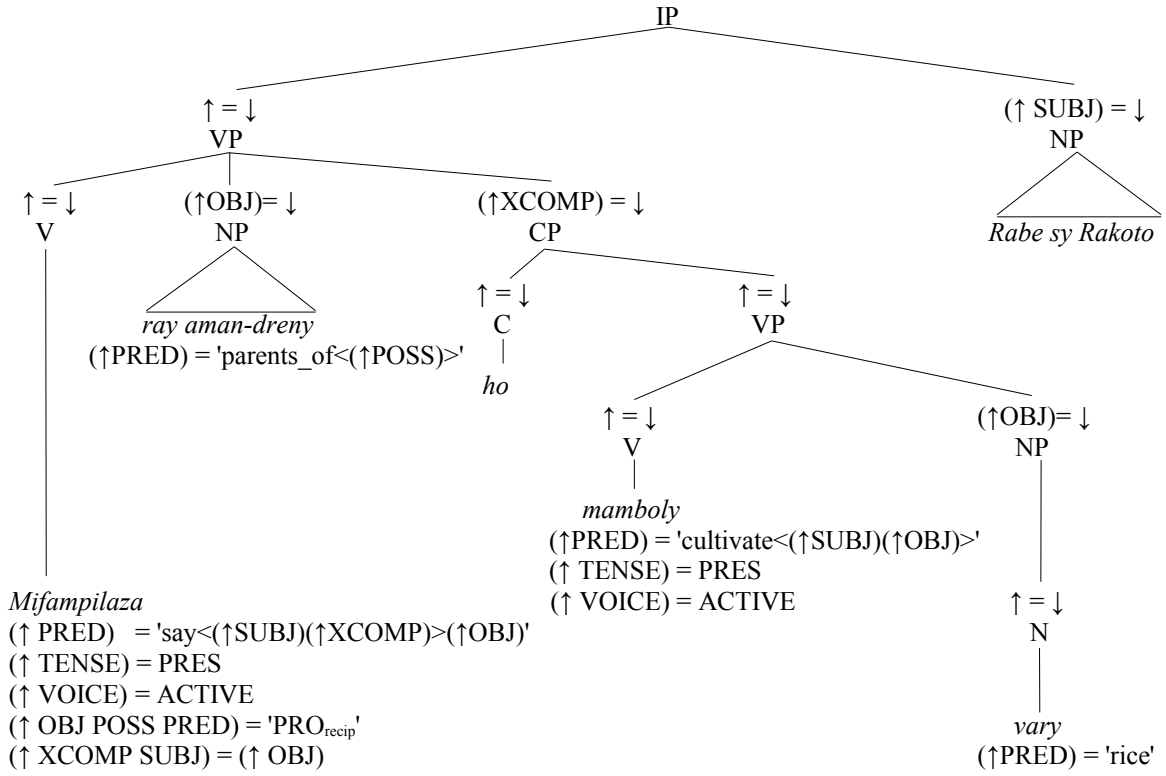
lit 'Rabe and Rakoto say each other's parents to be cultivating rice'
'Rabe and Rakoto say each other's parents are cultivating rice'

(Keenan & Razafimamonjy 2001:52-53)[6]

6 For clarity I have changed the names of the participants in this example.

M-ifamp-i-laza	*v*	(↑PRED) = 'say<(↑SUBJ)(↑XCOMP)>(↑OBJ)'
		(↑XCOMP SUBJ) = (↑OBJ)
		(↑TENSE) = PRES
		(↑VOICE) = ACTIVE
		(↑OBJ POSS PRED) = 'PRO$_{recip}$' (from rec. morpheme)
mamboly	*v*	(↑PRED) = 'cultivate<(↑SUBJ)(↑OBJ)>'
		(↑TENSE) = PRES
		(↑VOICE) = ACTIVE
ray aman-dreny	*NP*	(↑PRED) = 'parents_of<(↑POSS)>'
vary	*n*	(↑PRED) = 'rice'

C-structure



The lexical entries and c-structure (above) when combined give the complete and coherent f-structure below:

$$\begin{bmatrix} \text{SUBJ} & [\text{"Rabe and Rakoto"}] \\ \text{OBJ} & \begin{bmatrix} \text{POSS} & [\text{PRED} \quad \text{'PRO}_{recip}\text{'}] \\ \text{PRED} & \text{'parents\_of<(POSS)>'} \end{bmatrix} \\ \text{XCOMP} & \begin{bmatrix} \text{SUBJ} & \\ \text{OBJ} & [\text{'rice'}] \\ \text{TENSE} & \text{PRES} \\ \text{PRED} & \text{'cultivate<(SUBJ)(OBJ)>'} \\ \text{VOICE} & \text{ACTIVE} \end{bmatrix} \\ \text{TENSE} & \text{PRES} \\ \text{VOICE} & \text{ACTIVE} \\ \text{PRED} & \text{'say<(SUBJ)(XCOMP)>(OBJ)'} \end{bmatrix}$$

The advantage of the valency preserving analysis is that the f-structure of the clause remains structurally unchanged with the introduction of the reciprocal morpheme. In particular, this means that the possession and control constructions do not require the invention of new syntactic rules to account for their reciprocal counterparts: reciprocal control constructions can be treated just like regular control clauses.

The valency preserving analysis is able to account naturally for the remnant in reciprocal possessor constructions – it is assigned the function of object by the usual phrase structure rules and engages in a possessor relationship with the reciprocal pronoun. This account of the remnant correctly predicts that any Malagasy reciprocal construction can also engage in a possessor relation.

5. CONCLUSION

I looked at three languages with reciprocal constructions that appeared to be similar; Catalan, Chichewa and Malagasy. The analyses of the reciprocal construction put forward to account for the reciprocal construction in Chichewa (Mchombo & Ngalande 1980, Mchombo 1991 and Dalrymple et al 1994) and Catalan (Alsina 1996) both reduce the valency of the reciprocated verb in f-structure.

I propose an alternate approach allowed by the architecture of LFG where the verb's valency remains unchanged at the level of f-structure, and the reciprocal morpheme creates a reciprocal pronoun which sits in an internal argument position selected by the verb.

By allowing this mismatch between f- and c-structure, the valency preserving analysis of the Malagasy reciprocal construction is able to correctly predict the grammaticality of reciprocated functional control sentences and possession constructions.

Because the valency preserving analysis of the Malagasy reciprocal construction does not change the overall structure of the f-structure, this analysis predicts that the reciprocal construction should co-occur with other constructions in Malagasy. This is observed - the reciprocal morpheme can co-occur with causative, circumstantialization and nominalization constructions.

5. REFERENCES

Alsina, A. 1996. *The Role of Argument Structure in Grammar.* CSLI lecture notes 62. Stanford, California: CSLI Publications.

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

Dalrymple, M. Mchombo, S. & Peters, S. 1994. Semantic similarities and syntactic contrasts between Chichewa and English reciprocals. *Linguistic Inquiry*, Vol 25, Number 1, Winter 1994:145-163

Dalrymple, M. & Kanazawa, M. & Kim, Y. & Mchombo, S. & Peters, S. 1998. Reciprocal Expressions and the concept of reciprocity. *Linguistics and Philosophy* 21:159-210.

Dalrymple, M. 2001. *Syntax and Semantics: Lexical Functional Grammar*. Academic Press.

Keenan, E. L. & Ralalaoherivony, B. 1998. Raising from NP in Malagasy. In Paul, I. 1998. pp 65-93.

Keenan, E. L. & Razafimamonjy, J. P. 2001. Reciprocals in Malagasy. In Torrence, H. (ed.) 2001. *Papers in African Linguistics 1*, UCLA Department of Linguistics, Number 6, May 2001. UCLA working papers in linguistics.

Hurst, P. T. 2003. Syntactic Representations of the Malagasy Reciprocal Construction. Unpublished Post Graduate Thesis.

Mchombo, S. & Ngalande, R. M. 1980. *Reciprocal verbs in Chichewa: a Case for Lexical Derivation*. Bulletin of the school of Oriental and African studies, University of London, Vol 45. 1980

Mchombo, S. 1991. Reciprocalization in Chichewa: A lexical account. *Linguistic Analysis*, 21:3-22 1991.

Mchombo, S. & Ngunga, A. 1994. The syntax and semantics of the reciprocal construction in Ciyao. *Linguistic Analysis*, 24:3-31 (1994)

Paul, I. (ed.) 1998. *The Structure of Malagasy – Volume 2*. UCLA Occasional papers in linguistics, number 20.

Paul, I. 1996. The Malagasy genitive. In Pearson, M. & Paul, I. 1996.

Paul, I. & Rabaovololona, L. 1998. Raising to object in Malagasy. In Paul, I. 1998. pp 50-64.

Pearson, M. & Paul, I. (eds.) 1996. *The Structure of Malagasy – Volume 1*. UCLA Occasional papers in linguistics, number 17.

SPANISH *SE*-CONSTRUCTIONS: THE PASSIVE AND THE IMPERSONAL CONSTRUCTION

Carmen Kelling
Universität Konstanz

Proceedings of the LFG06 Conference

Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006

CSLI Publications

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

Abstract

Spanish *se*-constructions have many readings. The emphasis of this paper lies on the passive and the impersonal readings, which do not exist in German or English. I show that both constructions contain implicit information, contrary to previous analyses of *se* in the impersonal construction as an overt subject (e.g., Rivero 2002, D'Alessandro 2004). In order to account for the constructions, I assume systematic operations that concern different grammatical levels, actually lexical conceptual structure (LCS), argument structure, and functional structure. The correct analysis for the derivation of the *se*-constructions discussed here can be formulated adequately within Lexical Mapping Theory (LMT), the Lexical Functional Grammar linking module (Bresnan 2001).

1 Introduction

As is well-known, Spanish *se*-constructions have a variety of readings. The emphasis of this contribution lies on the passive (1) and the impersonal (2) readings, which have no equivalents in, e.g., German or English.

(1) Se firmaron los contratos. *passive*
REFL sign.PL the.PL contracts
'The contracts were signed.'

(2) Se admira a los futbolistas. *impersonal*
REFL admire.SG to the.PL soccer players
'One admires the soccer players.'

As Blevins (2003) states, first, the term 'passive' has been misapplied to a class of impersonal constructions, and, second, impersonal constructions have been almost entirely neglected in theoretical work.

The goal of this paper is to bring out the differences and similarities between the two constructions, showing that both constructions contain implicit information. In order to account for the structures in (1) and (2), I assume systematic operations that concern different grammatical levels, i.e. lexical conceptual structure (LCS), argument structure, and functional structure. A correct analysis for the derivation of the *se*-constructions discussed here can be formulated adequately within Lexical Mapping Theory (LMT), the Lexical Functional Grammar linking module (Bresnan 2001).

2 The theoretical framework: Lexical-Functional Grammar (LFG)

The theoretical framework underlying my analysis is Lexical-Functional Grammar (LFG: Bresnan 1982, Bresnan 2001), cf. (3). A-structure functions as the interface between lexical semantics and final syntactic structure. In LFG, two levels of syntactic structure (= final syntactic structure in (3)) are distinguished, i.e., constituent structure (c-structure), accounting for constituency, and functional structure (f-structure), which models relations among the grammatical functions like subject, object, etc.:

(3) lexical semantics
↓ Lexico-semantic projection
a-structure
↓ Lexico-syntactic projection
final syntactic structure

(Bresnan 2001:303)

For the analysis given here, I will need especially LFG's linking theory, i.e. Lexical Mapping Theory (LMT). In LMT, argument mapping is mediated by argument structure (a-structure), a level of representation in which argument positions are classified by a system of distinctive features for grammatical arguments: [±r] and [±o].
The feature [–r] refers to an unrestricted syntactic function, the kind of function which is not restricted as to its semantic role in the sense that it need not have any semantic role. The feature [–o] refers to a non-objective syntactic function.
The features constrain the mapping of thematic roles onto grammatical functions. (4) shows the intrinsic features of grammatical functions (GF), (5) shows the semantic classification of a-structure roles (see Bresnan 2001:309).

(4) Grammatical Functions (GF) classified by Features

GF	*Features*	
SUBJ	[–r, –o]	r: restricted
OBJ	[–r, +o]	o: objective
OBJθ	[+r, +o]	
OBLθ	[+r, –o]	

(5) Semantic Classification of A-Structure Roles for Function

patient-like roles	θ
	[–r]
secondary patient-like roles	θ
	[+o]

other semantic roles θ
[–o]

A mapping calculus can be constructed from features, a thematic role hierarchy as in (6), and mapping principles (7), that produces the appropriate mapping of thematic roles onto grammatical functions:

(6) Thematic Hierarchy
agent > beneficiary > experiencer/goal > instrument > patient/theme > locative

(7) Mapping Principles
a. Subject roles
The thematically most prominent role classified [–o] has to be mapped onto the subject function when initial in the a-structure. Otherwise a nonagentive, unrestricted role classified [–r] is mapped onto the subject function.
b. Other roles
All other roles are mapped onto the lowest compatible function in the partial ordering (8), where the subject is the least marked.

(8) Partial Ordering of Argument Functions
SUBJ > OBJ, OBL$_\theta$ > OBJ$_\theta$

Well-formedness constraints ensure that every sentence has a subject (9), and that two arguments cannot map onto the same grammatical function (10) (Bresnan 2001:311):

(9) The Subject Condition: Every predicator must have a subject.

(10) Function-Argument Bi-uniqueness
Each a-structure role must be associated with a unique function, and conversely.

The table in (11) shows the correct mapping of lexical conceptual structure (LCS) to functional structure (f-structure) for a transitvie verb like *firmar* 'sign'.

(11)

LCS		agent	theme
features	*firmar*	[–o]	[–r]
a-structure	'sign'	< x	y >
f-structure		SUBJ	OBJ

3 Classification of Spanish *se*-constructions

Before analyzing the passive and impersonal *se*-constructions in detail, I would like to give a list of the main uses of the Spanish reflexives. There are large differences in how reflexive constructions are classified, depending on the classification criteria as well as the theoretical frameworks. Subsequently, I will follow more or less the classification given in Kaufmann (2004).

The main uses are those in (12): the reflexive/reciprocal (a), the decausative (b), the middle (c), the causative (d), the passive (e), the aspectual (f), and the impersonal (g)

(12) a. Juan se afeita. / Juan y Pedro se afeitan. *reflexive/*
Juan REFL shaves.SG / Juan and Pedro REFL shave.PL *reciprocal*
'Juan shaves.' 'Juan and Pedro shave each other.'

b. El barco se hundió. *decausative*
The boat REFL sink.PAST
'The boat sank.'

c. Este libro se lee fácilmente. *middle*
This book REFL reads easily
'This book reads easily.'

d. Juan se afeita en la barbería. *causative*
Juan REFL shaves in the barber's
'Juan has himself shaved at the barber's.'

e. Se firmó la paz. ***passive***
REFL sign.PAST the peace
'The peace contract was signed.'

f. Juan se durmió. *aspectual*
Juan REFL sleep.PAST
'Juan fell asleep.'

g. Se invitó a todos los empleados. ***impersonal***
REFL invite.PAST to all the employees
'All employees were invited.'

The passive and impersonal constructions considered here are only two of a variety of interpretations, and not all of them encode implicit information The reflexive/reciprocal interpretation in (12a), e.g., does not contain implicit information. The agent argument is expressed by the subject *Juan*, and the theme argument by the reflexive clitic *se*. The decausative construction (12b) has no implicit argument. The agent argument was suppressed, there is no semantic agent argument at all (cf. Alencar and Kelling 2005).
Engelberg (2002) defines implicit arguments as in (13):

(13) Implicit Arguments
"I will assume that a verb's predicate constant has an implicit argument iff either (i) the verb has a variant with an explicit argument (i.e. an argument that gets syntactically realized) in the same semantic relation or (ii) there is a morphologically related verb with an explicit argument in the same semantic relation." (Engelberg 2002:375)

Condition (ii) holds for both constructions considered here, cf. the active sentences in (14) and (15) corresponding to the reflexive passive sentence in (1) and to the impersonal sentence in (2) respectively:

(14) El futbolista firmó los contratos. *active*
The president sign.PAST the contracts.
'The soccer player signed the contracts.'

(15) El jefe invitó a todos los empleados. *active*
The boss invite.PAST to all the employees
'The boss invited all the employees.'

In order to account for sentences like (1) and (2) or (12e) and (12g), I assume systematic operations that concern different grammatical levels, i.e. lexical conceptual structure (LCS), argument structure, and functional structure. In the case of the passive *se*-construction, I will assume an argument structure manipulation, and vor the impersonal *se*-construction, there is a change on the final syntactic structure level.

4 Analysis

An analysis for the derivation of the passive and the impersonal *se*-constructions will be formulated within Lexical Mapping Theory (LMT), the Lexical Functional Grammar linking module (Bresnan 2001).

4.1 The passive *se*-construction

The passive *se*-construction can only be derived from transitive verbs, and it is only available in the third person. In contrast to the periphrastic passive in (16a), the reflexive passive cannot be used when the agent of the action is mentioned (16b):

(16) a. Los contratos fueron firmados por el futbolista. *periphrastic*
The contracts were signed by the soccer player *passive*
'The contracts were signed by the soccer player.'
b. *Los contratos se firmaron por el futbolista. *reflexive*
The contracts REFL sign.PAST by the soccer player *passive*
'The contracts were signed by the soccer player.'

As in the periphrastic passive, the theme of the transitive verb is realized as a subject in the passive *se*-construction, see (17):

(17) a. El futbolista firmó los contratos. 'The soccer player signed the contracts.'
agent theme
SUBJ OBJ
b. Se firmaron los contratos. 'The contracts were signed.'
theme
SUBJ
c. Los contratos se firmaron. 'The contracts were signed.'
theme
SUBJ

The word order with the subject placed after the verb as in (17b) is less marked in passive *se*-constructions, but (17c) is also possible.
That the agent is present on the level of LCS in passive *se*-constructions can be shown by the classical agent diagnostics, for example, by adding a purpose clause (18) or an agentive adverb (19):

(18) Se firmaron los contratos para ganar más dinero.
REFL sign.PL the.PL contracts in order to earn more money
'The contracts were signed in order to earn more money.'

(19) Se retrasaron las reuniónes deliberadamente.
REFL delay.PL the.PL meetings deliberately
'The meetings were delayed deliberately.'

It follows from these facts that we need different passive rules for the periphrastic passive and for the reflexive passive, not only with respect to the morphological change, but also in order to account for different behaviors concerning the realization of the agent role. For the reflexive passive, I propose an operation, the Reflexive Passive Operation, that **suppresses** the [–o] feature of the agent argument, thus preventing it from being realized at functional structure. Applying the Reflexive Passive Operation gives the result in (20): the agent cannot be mapped onto functional structure. According to mapping principles, the y-argument is mapped onto the subject function.

(20)	LCS		agent	theme
	features	PRED	–	[–r]
	a-structure		–	< y >
	f-structure	REFL +	–	SUBJ

The structure shows that the agent argument is present at LCS as an implicit argument.
In contrast, the Periprastic Passive Operation **blocks** the realization of the agent argument, and it may be realized as oblique object, cf. (21):

(21)	LCS		agent	theme
	features	$PRED_{pass}$	[–o]	[–r]
	a-structure		< x	y >
	f-structure		(OBL)	SUBJ

Thus, the difference between the reflexive passive and the periphrastic passive comes out naturally by assuming suppression on the one hand, and blocking on the other hand. The effect is that with suppression, there is no mapping of the agent onto f-structure, whereas with blocking, the agent may be mapped onto an oblique function.

4.2 The impersonal *se*-construction

The impersonal *se*-construction can be used with many kinds of verbal predicates[1], as shown in (22). Examples include intransitive, unaccusative (22a) as well as unergative (22b), copulative (22c), and transitive (22d) predicates:

(22)	a.	Se	entra	por aquí.	*unaccusative*
		REFL	enter.PRES	by here	
		'One enters here.'			

[1] Examples are taken from Sánchez López (2002) and Butt and Benjamin (2000).

b. En este país se duerme mucho. *unergative*
In this country REFL sleep much
'People sleep a lot in this country.'

c. Se es feliz cuando se es honesto. *copulative*
REFL is happy when REFL is honest.
'One is happy when one is honest.'

d. Se encontró a los alpinistas desaparecidos. *transitive*
REFL found to the.PL mountaineer.PL disappeared.PL
'One has found the missed mountaineers.'

In contrast to the passive *se*-construction, impersonal reflexive constructions do not have an overt (theme) subject, as can be seen in the examples in (23): (23c) is ungrammatical because the verb *invitaron* 'invite.PL' neither agrees with the direct object *a todos los empleados* 'all employees' nor with a possibly existing null subject.

(23) a. El jefe invitó a todos los empleados. *active*
the boss invite.SG.PAST to all.PL the.PL employees
'The boss invited all the employees.'

b. Se invitó a todos los empleados. *impersonal*
REFL invite.SG.PAST to all.PL the.PL employees
'All employees were invited.'

c. *Se invitaron a todos los empleados.
REFL invite.PL.PAST to all.PL the.PL employees

Some linguists treat the *se* of the impersonal construction as subject (for example Oesterreicher 1992, Rivero 2002 or D'Alessandro 2004), equivalent to German *man* or French *on*. However, this is in contradiction with the distributional facts shown in (24) and (25) (cf. Mendikoetxea 1999, Sánchez López 2002, Suñer 1976; 1983).

(24) a. Ella siempre habla mucho. *active*
she always talks much
'She always talks a lot.'

b. *Se siempre habla mucho.
REFL always talk much

c. Siempre se habla mucho. *impersonal*
always REFL talks much
'One doesn't talk a lot.'

(25) a. Ella no habla mucho. *active*
She not talk much.
'She doesn't talk a lot.'

b. *Se no habla mucho.
REFL not talk much.

c. No se habla mucho. *impersonal*
not REFL talks much
'One does not talk a lot.'

Se does not have the distribution of subject pronouns in Spanish, neither with adverbs (24) nor with negation (25). Therefore, I assume that the subject is implicit, see (26) and (27).

(26) PRO siempre se habla mucho.
'One always talks a lot.'

(27) PRO no se habla mucho.
'One does not talk a lot.'

I do not assume an explicit subject argument. This is in accordance with the analysis of, e.g., Otero (1986) or Campos (1989) who analyzes the implicit subject of the impersonal *se*-constructions as an empty indefinite pronoun (PRO_{indef}).
In LFG, the PRO is accounted for by the interaction between constituent structure and functional structure. The empty element is not present at c-structure, but is there as PRO in the f-structure, see (28) and (29).

(28)

$$\text{SUBJ} \left[\begin{array}{ll} \text{PRED} & \text{PRO} \\ \text{NUM} & \text{SG} \\ \text{PERS} & 3 \end{array} \right]$$

(29) Impersonal *se*-construction

LCS	(...)	(...)
features	(...)	(...)
a-structure	(...)	(...)
f-structure	SUBJ	(...)
c-structure	Ø	

So there is no suppression or blocking in this case. However, the realization of the thematic argument is limited to a PRO.

For a transitive predicate like *invitar* 'invite' in a sentence like (30a) (= 23a), the active/transitive mapping structure is indicated in (31), and the impersonal mapping structure of (30b) (= 23b) can be seen in (32):

(30) a. El jefe invitó a todos los empleados. *active*
the boss invite.SG.PAST to all.PL the.PL employees
'The boss invited all the employees.'

b. PRO Se invitó a todos los empleados. *impersonal*
REFL invite.SG.PAST to all.PL the.PL employees
'One invited all the employees. / All employees were invited.'

(31)

LCS		agent	theme
features	*invitar*	[–o]	[–r]
a-structure	'invite'	< x	y >
f-structure		SUBJ	OBJ

(32)

LCS		(agent)	theme
features	*invitar*	[–o]	[–r]
a-structure	'invite'	< x	y >
f-structure	REFL +	SUBJ=PRO	OBJ

The subject of the impersonal sentence is PRO, and this is the agent role.

For an unaccusative verb like *entrar* 'to enter', for example in (33a) with an overt subject, and in (33b) with an implicit subject, we get the mapping structures in (34) and (35), respectively.

(33) a. Juan entra por aquí. *unaccusative*
Juan enter.PRES by here
'Juan enters here.'

b. PRO Se entra por aquí. *impersonal*
REFL enter.PRES by here
'One enters here.'

(34)

LCS		theme
features	*entrar*	[–r]
a-structure	'enter'	< x >
f-structure		SUBJ

(35)

LCS		(theme)
features	*entrar*	[–r]
a-structure	'enter'	< x >
f-structure	REFL +	SUBJ=PRO

In the case of an unaccusative verb, there is only a theme argument. This argument is realized as an overt subject in the unaccusative construction, whereas it is a PRO in the impersonal construction. The *se* indicates the change of the construction.

To sum up, the interpretation of the implicit argument in the passive and impersonal *se*-constructions result from different operations or conditions on different levels of the grammar.

5 Conclusion

Consider the sentences in (36):

(36) a. Es difícil vender periódicos en un país donde se leen poco.
is difficult sell newspapers in a country where REFL read.PL little
'It is difficult to sell newspapers in a country where they aren't read much.'

b. Es difícil vender periódicos en un país donde se lee poco.
is difficult sell newspapers in a country where REFL read.SG little
'It is difficult to sell newspapers in a country where people don't read much.'
(Butt and Benjamin 2000)

Both sentences, (36a) and (36b), contain implicit information. However, in (36a) we have a reflexive passive with a blocked agent argument. In this case, the theme argument of transitive *leer* 'read' is realized as subject (*periódicos* 'newspapers'). In (36b), there is no agreement between *periódicos* 'newspapers' and *lee* 'read.SG', so an implicit PRO-subject must be assumed.

The two different *se*-construction readings are produced on different levels of the grammar. In the case of the **passive *se*-construction** (36a), the agent argument's [–o] feature is suppressed, thus preventing it to be mapped onto functional structure.

In the **impersonal *se*-construction** (36b), the subject is there at the f-structure level. However, it is not realized at c-structure.

References

Alencar, Leonel Figureido de and Carmen Kelling. Are reflexive constructions transitive or intransitive? Evidence from German and Romance. In: Miriam Butt and Tracy Holloway King (ed.): *Proceedings LFG 2005*. Stanford: CSLI Publications (http://csli-publications.standford.edu). 1-20.

Blevins, James P. 2003. Passives and impersonals. *Journal of Linguistics* 39. 473-520.

Bresnan, Joan. ed. 1982. *The Mental Representation of Grammatical Relations*. Cambridge, Mass/London: MIT Press.

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

Butt, John and Benjamin, Carmen. ^{3}2000. *A New Reference Grammar of Modern Spanish*. London: Arnold.

Campos, Hector. 1989. Impersonal passive "se" in Spanish. *Lingvisticæ Investigationes* XIII:1. 1-21.

D'Alessandro, Roberta Anna Grazia. 2004. *Impersonal* si *constructions. Agreement and interpretation. Doctoral Dissertation.* URL: http://elib.uni-stuttgart.de/opus/volltexte/2004/1630/.

Engelberg, Stefan. 2002. Intransitive accomplishments and the lexicon: The role of implicit arguments, definiteness, and reflexivity in aspectual composition. *Journal of Semantics* 19. 369-416.

Kaufmann, Ingrid. 2004. *Medium und Reflexiv: eine Studie zur Verbsemantik*. Tübingen: Niemeyer.

Mendikoetxea, Amaya. 1999. Construcciones con *se*: medias, pasivas e impersonales. In *Gramática descriptiva de la lengua española*, ed. Ignacio Bosque/Violeta Demonte. Madrid: Espasa. 1631-1722.

Oesterreicher, Wulf. 1992. *SE* im Spanischen. Pseudoreflexivität, Diathese und Prototypikalität von semantischen Rollen. *Romanistisches Jahrbuch* 43. 237-260.

Otero, Carlos P. 1986. Arbitrary subjects in finite clauses. In *Generative Studies in Spanish Syntax*, ed. Ivonne Bordelois, Heles Contreras, and Karen Zagona. Dordrecht: Foris. 81-109.

Rivero, María Luisa. 2002. On impersonal reflexives in Romance and Slavic and Semantic variation. In *Romance Syntax, Semantics and L2 Acquisition. Selected papers from the 30th Linguistic Symposium on Romance Languages*, *Gainesville, Florida, February 2000,* ed. J. Camps and C. Wiltshire. John Benjamins: Amsterdam and Philadelphia. 169-195.

Sánchez López, Cristina. 2002. Las construcciones con *se*. Estado de la cuestión. In *Las construcciones con se*, ed. Cristina Sánchez López. Madrid: Visor Libros. 13-163.

Suñer, Margarita. 1976. Demythologizing the impersonal *se* in Spanish. *Hispania* 59. 268-275.

Suñer, Margarita. 1983. pro_{arb}. *Linguistic Inquiry* 14. 188-191.

ON THREE DIFFERENT TYPES OF SUBJECTLESSNESS AND HOW TO MODEL THEM IN LFG

Anna Kibort
Surrey Morphology Group, University of Surrey, UK

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

Outside LFG, the term 'subjectless' is found referring to a range of phenomena in which the expression of the predicate lacks an overt lexical item (a syntactic constituent) bearing the grammatical function of the subject. In some of these phenomena, for example in *pro*-drop, the architecture of LFG allows us to find the subject at the a-structure and f-structure levels despite there being no categorial element expressing the subject. There are, however, other subjectless constructions for which there are no readily available LFG accounts, and it is not always obvious how they could be analysed adequately. It is constructions of this type – often called 'impersonal' in traditional literature – that will be the focus of this paper, exemplified from Polish which is rich in impersonal forms.

I will begin with an overview of all Polish constructions which appear to be subjectless. I will identify three types of construction which lack subjects at some level of analysis: *pro*-drop constructions (including the so-called 'weather constructions' and 'adversity impersonals'), morpholexically derived impersonal constructions, and truly subjectless constructions. I will then demonstrate how they differ by highlighting their morphological and syntactic properties and suggest levels of representation at which the different types of 'null/missing subjects' can be captured theoretically.[1]

1 'Subjectless' constructions in Polish

Polish has a large number of different constructions that appear to be subjectless. As will be demonstrated in the further sections of this paper, their morphosyntactic properties allow them to be grouped into the following categories:

1. *pro*-drop constructions. These include clauses formed from personal predicates with a dropped personal pronoun (such as '[He] saw.3SG.MASC that the door was open and [he] went.3SG.MASC in'), and from personal predicates with a dropped indefinite pronoun, both the pronoun referring to humans ('[Someone] was-writing.3SG.MASC as if [he] wanted.3SG.MASC to warn us') and the pronoun referring to non-humans, as in 'weather constructions' and 'adversity impersonals' ('[Something] was-blowing.3SG.NEUT as if [it] wanted.3SG.NEUT to pull out trees with their roots', '[Something] threw.3SG.NEUT him to the side'). Contrary to the frequently found though unsubstantiated assumption, Polish weather constructions, adversity impersonals and other apparently subjectless clauses involving verbs of physical or psychological states do not lack a syntactic subject, nor do they have a suppressed or other empty category/zero subject. Instead, they result from subject ellipsis, with their omitted subject being the indefinite pronoun referring to non-humans – that is, they are instances of pro_{INDEF}-drop.

2. Morpholexically derived impersonal constructions. These include clauses formed from personal predicates whose fully operational (binding, controlling, available for raising) and interpretable syntactic subject has been 'suppressed' by a morpholexical operation and is not allowed to appear as a constituent in surface syntax. This category includes the so-called '*-no/-to* impersonal' (*Bito Piotra* 'Beat.IMPERS Peter(MASC).ACC' meaning '[They] beat Peter') and the reflexive impersonal (*Biło się Piotra* 'Beat.3SG.NEUT REFL Peter(MASC).ACC' meaning '[One] beat Peter').

[1] I gratefully acknowledge the current ESRC grant RES-051-27-0122. This paper draws from my PhD thesis on passive and passive-like constructions in English and Polish (Kibort 2004).

3. Subjectless constructions. These are formed from predicates without either an overt or an omitted/covert syntactic subject which could participate in syntactic operations such as control or raising. This category includes inherently impersonal predicates (a small class of defective, non-inflecting verbs as in: *Słychać ją* '[One] hear.NON-PERSONAL her.ACC', or '*Było widać łąkę* '[One] was.3SG.NEUT see.NON-PERSONAL meadow(FEM).ACC') and predicates which have lost their subjects as a result of derivation. The latter occurs, for instance, when passivisation is applied to an intransitive predicate.

4. Constructions with non-agreeing subjects. These include predicative adverbial constructions (e.g., *Miło z tobą podróżować* 'Nicely with you travel.INF' meaning 'It is nice travelling with you') or nominativeless clauses with predicates requiring a genitive argument (e.g., *Przybywa wody* 'Becomes-more water(FEM).GEN'). Clauses of this type complete the typology of Polish 'subjectless' constructions, but it is important to realise that their subjectlessness is only apparent. They do not lack subjects, but simply have non-agreeing subjects. Thus, they pattern with other clauses whose subjects have some nominal properties but are nevertheless not appropriate agreement controllers. Such subjects are prepositional phrases, quantifier phrases (with quantifiers requiring their complements to appear in genitive case), clausal subjects (including infinitival subjects), and certain indeclinable subjects such as acronyms and foreign place names. This type of construction will not be taken up further in this paper; for some more discussion and analysis see Kibort (2004:320-340).

The first three types of construction all lack overt subjects but each has different morphosyntactic properties, which will be exemplified in the sections below. I will show that the architecture of LFG makes it possible to identify these different types of subjectlessness at different levels of representation of the predicate, even though the new analyses may require revising some elements of LFG's theory of argument structure. Constructions of **TYPE 1 (*pro*-drop)** fall under the standard LFG analysis of unexpressed pronouns. Constructions of **TYPE 2 (morpholexical impersonals)** need a new analysis: they have an unaltered argument structure, but the categorial expression of their fully operational syntactic subject is suppressed. At functional structure level, the covert subject may be analysed as an obligatory PRO analogous to the PRO in arbitrary anaphoric control constructions. Finally, constructions of **TYPE 3 (truly subjectless predicates)**, which have no subject at any level of analysis (a-structure, f-structure, or c-structure), provide a strong argument against LFG's Subject Condition ('Every predicator must have a subject'; Bresnan 2001:311).

2 TYPE 1: *pro*-drop constructions

The following Polish sentences are not usually associated with the *pro*-drop phenomenon. They exemplify predicates denoting natural or supernatural phenomena including weather phenomena (1-2), the so-called 'adversity impersonals' (3), and predicates denoting physical or psychological states (4):

(1) *Pada/Świta.*
rains/dawns
'It is raining/dawning.'

(2) *W tym domu straszy.*
in this house spooks

'It haunts in this house.' (meaning: 'This house is haunted')

(3) *Wyrzuciło łódkę na brzeg.*
threw-out.3SG.NEUT boat(FEM).ACC onto shore
'The boat got thrown onto the shore.'

(4) a. *Mdli/Dusi/Skręca/Ciągnie/Boli/Swędzi/Kłuje mnie.*
nauseates/chokes/convulses/pulls/aches/itches/stabs me.ACC
'[Something] makes me nauseous/choke/convulse/contract my muscles/painful/itch/gives me shooting pains.'

b. *Mdli/Dusi/Skręca mnie od tego zapachu.*
nauseates/chokes/convulses me.ACC from this smell
'This smell makes me nauseous/choke/convulse.'

c. *Mdli/Dusi/Skręca mnie z bólu/zazdrości.*
nauseates/chokes/convulses me.ACC from pain/envy
'The pain/envy makes me nauseous/choke/convulse.'

Clauses of this type commonly appear without an overt nominative subject and use verbal forms displaying 'default' agreement. They are often treated as impersonal active clauses with covert inanimate subjects – that is, they are taken to contain an empty or 'zero' subject. The existence of the 'zero' subject is taken to trigger 'default' 3SG.NEUT agreement in the verb and impose on the construction an 'inherent inanimate force' interpretation.

However, it is straightforward to demonstrate that all predicates used in these constructions can easily appear with an overt nominative subject, whether in the singular or in the plural:

(5) a. *Padało. ~ Deszcz padał.*
rained.3SG.NEUT rain(MASC).NOM rained.3SG.MASC
'It was raining. ~ The rain was raining.'

b. *Świta. ~ Poranek świta.*
dawns morning(MASC).NOM dawns
'It is dawning. ~ The morning is dawning.'

c. *Często padają tu ulewne deszcze.*
often rain.3PL here torrential.NONVIR.NOM rains(NONVIR).NOM
'Torrential rains often rain here.'

(6) a. *W tym domu straszy.*
in this house spooks
'It haunts in this house.' (meaning: 'This house is haunted')

b. *W tym domu coś straszy.*
in this house something(NEUT).NOM spooks
'Something haunts in this house.' (meaning: 'This house is haunted by something/some ghost')

c. *W tym domu straszy duch pradziadka.*
in this house spooks ghost(MASC).NOM great-grandfather(MASC).GEN
'This house is haunted by the ghost of the great grandfather.'

(7) a. *Morze wyrzuciło łódkę na brzeg.*
sea(NEUT).NOM threw-out.3SG.NEUT boat(FEM).ACC onto shore
'The sea threw the boat onto the shore.'

b. *Fale wyrzuciły łódkę na brzeg.*
waves(NONVIR).NOM threw-out.3PL.NONVIR boat(FEM).ACC onto shore
'The waves threw the boat onto the shore.'

(8) a. *Wszystkie zapachy mnie mdliły. Nawet*
all smells(NONVIR).NOM me.ACC nauseated.3PL.NONVIR even
zapach kawy mnie mdlił.
smell(MASC).NOM coffee(FEM).GEN me.ACC nauseated.3SG.MASC
'All smells made me nauseous. Even the smell of coffee made me nauseous.'

b. *Ból skręcał mnie niemiłosiernie.*
pain(MASC).NOM convulsed.3SG.MASC me.ACC mercilessly
'The pain convulsed me mercilessly.'

c. *Bolała/Swędziała mnie głowa.*
ached/itched.3SG.FEM me.ACC head(FEM).NOM
'My head ached/itched.'

d. *Coś mnie dusi. / Dusiły mnie*
something(NEUT).NOM me.ACC chokes choked.3PL.NONVIR me.ACC
te zapachy.
these.NONVIR.NOM smells(NONVIR).NOM
'Something makes me choke. / Those smells made me choke.'

Furthermore, there are no morphosyntactic restrictions on any of these verbs which would prevent them from agreeing with a subject in a person other than third, e.g.:

(9) *Głośno wiejesz, wietrze.*
loudly blow.2SG wind(MASC).VOC
'You are blowing loudly, wind.'

All this suggests that these constructions do not lack a subject at any level of abstract representation of the predicate. They are personal predicates and their superficial subjectlessness results from the familiar *pro*-drop phenomenon. Wierzbicka (1966) argued against a *pro*-drop analysis of Polish 'weather constructions' assuming that the dropped pronoun would have to be a personal pronoun corresponding in gender to the nominal denoting the particular natural phenomenon, that is: *on* 'he[MASC]' for *deszcz* 'rain(MASC)' or *wiatr* 'wind(MASC)'; *ono* 'it[NEUT]' for *niebo* 'sky(NEUT)' or *powietrze* 'air(NEUT)'. She assumed that, if the 'subjectless' weather sentences were a result of subject ellipsis, the verb would have to display gender agreement with the dropped pronoun corresponding to the nominal denoting the natural phenomenon. Such agreement is indeed not established. However, this hypothesis makes an incorrect assumption about the subject of the weather constructions: the dropped subject is not the personal pronoun, but the indefinite pronoun.

All nouns and pronouns in Polish, whether denoting or referring to people, objects, abstract notions or natural phenomena, bear the feature of inherent grammatical gender:

MASC, FEM or NEUT in the singular, and VIR (masculine human) or NONVIR (all other, i.e. non-masculine human and all non-human) in the plural.[2] The so-called indefinite pronouns *ktoś* 'somebody', referring to humans (HUM), and *coś* 'something', referring to non-humans (NON-HUM), bear the grammatical features MASC and NEUT, respectively, and these are also the gender agreements that they trigger in the verb.

The following is an example of a definite (and referential) use of the indefinite HUM pronoun *ktoś* which is employed here in order to avoid specifying the gender (and number) of the referent of the agent:

(10) *Ten ktoś pisał, jakby chciał nas ostrzec.*
this.MASC.NOM someone(MASC).NOM wrote.3SG.**MASC** as-if wanted.3SG.**MASC** us warn.INF
'This person was writing as if he/she wanted to warn us [of something].'

If the pronoun is dropped, as in any other familiar case of ellipsis, the resulting sentence is:

(11) *Pisał, jakby chciał nas ostrzec.*
wrote.3SG.**MASC** as-if wanted.3SG.**MASC** us warn.INF
'He/she was writing as if he/she wanted to warn us [of something].'

Although sentence (11) taken out of context is ambiguous between a gender non-specific ('he or she') and a gender specific ('he') interpretation of its agent, both examples (10) and (11) show that **3SG.MASC agreement** is used with **unspecified singular human** subjects, whether overt or dropped.

By analogy, the following sentence:

(12) *Wieje, jakby chciało powyrywać drzewa z korzeniami.*
blows[3SG].**NEUT** as-if wanted.3SG.**NEUT** pull-out.INF trees with roots
'[The wind] is blowing as if it wanted to pull out the trees with their roots.'

illustrates the use of **3SG.NEUT agreement** with an **unspecified non-human** subject. In the sentence above, the subject has remained unexpressed overtly, as in example (11).

If we choose to specify the number and gender of the agent of the event denoted by the verb, the number and gender agreement corresponding to the unspecified agent is replaced by verbal inflection corresponding to the grammatical number and gender of the subject nominal. Therefore, in case of human agents, we can have, for example:

(13) *Piotr pisał, jakby chciał nas ostrzec.*
Peter.**MASC**.NOM wrote.3SG.**MASC** as-if wanted.3SG.**MASC** us warn.INF
'Peter was writing as if he wanted to warn us [of something].'

(14) *Ta kobieta pisała, jakby chciała nas ostrzec.*
this.**FEM**.NOM woman.**FEM**.NOM wrote.3SG.**FEM** as-if wanted.3SG.**FEM** us warn.INF
'This woman was writing as if she wanted to warn us [of something].'

[2]This is a simplified view of Polish gender in its interaction with number, but it is sufficient to describe the phenomena discussed in this paper.

Moreover, in case of subject ellipsis, the verb retains its agreement with the 'dropped *pro*' denoting a human agent, because personal pronouns are specified for exactly the same features which trigger the agreement as the nominals they correspond to:

(15) *(On)* *pisał,* *jakby chciał* *nas ostrzec.*
(he[**MASC**].NOM) wrote.3SG.**MASC** as-if wanted.3SG.**MASC** us warn.INF
'He was writing as if he wanted to warn us [of something].'

(16) *(Ona)* *pisała,* *jakby chciała* *nas ostrzec.*
(she[**FEM**].NOM) wrote.3SG.**FEM** as-if wanted.3SG.**FEM** us warn.INF
'She was writing as if she wanted to warn us [of something].'

In case of overtly expressed non-human agents, the gender and number agreement also corresponds to the grammatical gender and number of the subject nominal, as was shown in sentences to the right of the hyphens in example (5). However, even though *deszcz* 'rain(MASC)' or *wiatr* 'wind(MASC)' are grammatically masculine, and *niebo* 'sky(NEUT)' or *powietrze* 'air(NEUT)' are grammatically neuter, it is not possible to replace these nominals with the personal pronouns *on* 'he[MASC]' or *ono* 'it[NEUT]' unless we personify the natural phenomena in question.

The unacceptability – or, more accurately, the infelicity – of sentences such as:

(17) a. *#On* *padał.*
he[MASC].NOM rained.3SG.MASC
'It [he=the rain] was raining.'
b. *#Ono* *się* *ochłodziło.*
it[NEUT].NOM REFL cooled-down.3SG.NEUT
'It [=the air] has become colder.'

follows from the fact that, in addition to being specified for number and gender, personal pronouns in Polish conventionally denote human (HUM) agents, while verbs such as 'rain', 'snow' or 'cloud over' imply a non-human (NON-HUM) 'agent' or cause.

Since weather verbs in Polish are not normally used with personal pronouns, it is, therefore, not plausible to suggest that weather constructions without an overt subject result from personal pronoun ellipsis. It is, however, reasonable to see them as resulting from the ellipsis of the indefinite pronoun *coś* 'something' which is used to achieve the 'unspecified agent' interpretation and which triggers 3SG.NEUT agreement. In case of subject ellipsis ('*pro*$_{\text{INDEF}}$-drop'), the 3SG.NEUT verbal agreement is retained. One of the conventional uses of the 'indefinite' pronouns, both HUM and NON-HUM, is with a **definite** referent whom/which the speaker chooses not to specify. By omitting the indefinite pronoun *coś* 'something', the identity of the 'agent' is not questioned, but left unspecified, since it is in most cases understood from the context.[3]

Clauses with *pro*$_{\text{INDEF}}$-drop do not present problems for LFG. They fall under the standard analysis of unexpressed pronouns, e.g., Bresnan (2001:144-177). She analyses *pro*-drop as the functional specification of a pronominal argument by the head to which the pronominal inflection is bound, which entails the absence of the structural expression of

[3]For more detailed discussion of the morphosyntax of *pro*$_{\text{INDEF}}$-drop constructions in Polish, see Kibort (2004:295-318).

the pronoun as a syntactic NP or DP when the optional semantic and binding features of the pronominal inflection are present.

3 TYPE 2: morpholexical impersonals

There are two constructions in Polish whose grammaticalised function is to despecify the principal participant of the predicate: the *-no/-to* impersonal and the reflexive impersonal. The principal participant in these constructions is interpreted as either an unspecified or a generic human agent or experiencer. The constructions have particular morphosyntactic properties and morphological marking. The *-no/-to* impersonal uses a dedicated, uninflecting verb form ending in *-no/-to*, and is restricted to past tense, while the reflexive impersonal uses 3SG.NEUT verb form and the reflexive marker *się*, and can be used in all tenses.

3.1 The *-no/-to* construction

The *-no/-to* construction is exemplified in (18) and (19):

(18) *Budowano szkołę.*
built.IMPERS school(FEM).ACC
'A/The school was built. / [They] were building a/the school.'

(19) *Tutaj tańczono.*
here danced.IMPERS
'There was dancing here. / [They] danced here.'

One of the key properties of this construction is that it can be used with both intransitive and transitive predicates, and in the case of transitives the accusative object is retained, as in (18). Another key property is that it can be formed from both unergative and unaccusative predicates, including the habitual/iterative form of the verb 'be'. It can be formed from passivised predicates, therefore it has to be treated as independent of passivisation. The following example contains an impersonal form of the auxiliary (*bywano*) in a periphrastic passive construction with a passive participle (*bitymi*):

(20) *Dostawano różne kary i*
received.IMPERS various.NONVIR.ACC punishments(NONVIR).ACC and
bywano bitymi.
wasITERATIVE.IMPERS beat.PART.PL.INSTR
'[They/One] received various punishments and were/was beaten.'

The Polish *-no/-to* construction does not, under any circumstances, accept the surface expression of a nominative subject (21-22), nor does it accept the expression of the agent in an oblique phrase as in the passive, (23-24):

(21) **Władze budowano szkołę.*
authorities(NONVIR).NOM built.IMPERS school(FEM).ACC
'(intended) The authorities were building a/the school.'

(22) **Uczniowie tutaj tańczono.*
pupils(VIR).NOM here danced.IMPERS
'(intended) The pupils were dancing here.'

(23) **Budowano szkołę przez władze.*
built.IMPERS school(FEM).ACC by authorities
'(intended) A/The school was built by the authorities.'

(24) **Tutaj tańczono przez uczniów.*
here danced.IMPERS by pupils
'(intended) The dancing was done here by pupils.'

However, despite being superficially subjectless, the *-no/-to* impersonal appears to have a syntactically active 'covert' subject which participates in syntactic control and binding. The *-no/-to* predicate can share its subject with infinitives (25), with deverbal adverbials (26), and in a subject-raising construction (27); the covert subject of *-no/-to* is also capable of binding reflexive and reflexive possessive pronouns when they need to be bound by the subject (28-29):

(25) *Chciano wyjechać.*
wanted.IMPERS leave.INF
'There was eagerness to leave.'

(26) *Wsiadając do autobusu pokazywano bilety.*
get-on.PART$_{CONTEMP}$ into bus showed.IMPERS tickets(NONVIR).ACC
'On getting on the bus [they]/one showed the tickets.'

(27) *Zdawano się tego nie dostrzegać.*
seemed.IMPERS REFL this.MASC.GEN NEG notice.INF
'[They] seemed not to notice this.'

(28) *Oglądano się/siebie w lustrze.*
looked-at.IMPERS REFL/self.ACC in mirror
'[They] looked at [them]selves in the mirror. / One looked at oneself in the mirror.'

(29) *Oglądano swoje zbiory.*
looked-at.IMPERS own[REFL].NONVIR.ACC collections(NONVIR).ACC
'[They] looked at [their] own collections. / One looked at one's collection.'

The *-no/-to* impersonal is not agentless, either. Its agent (or experiencer) licenses all sorts of agent-oriented adverbials (e.g., *celowo* 'on purpose') and is invariably interpreted as an unspecified but definite human. The human interpretation of the agent/experiencer, which has been grammaticalised in the usage of this construction, overrides any semantic implications to the contrary that may arise from the meaning of the lexical items used in the clause, or from the context. Therefore, the *-no/-to* forms of predicates such as 'bark' or 'build nests' can only be interpreted as involving human activity.

The covert subject of the *-no/-to* impersonal triggers virile (plural) marking in agreeing (adjectival and nominal) predicative complements. Examples (30) and (31) show that expressions whose referents are, inflectionally, other than virile (plural) are incompatible with the *-no/-to* form and produce ill-formed clauses:

(30) (example adapted from Dziwirek 1994:222)

a. **Pracowano jako nauczyciel.*
worked.IMPERS as teacher(MASC).NOM

b. **Pracowano jako nauczycielka.*
worked.IMPERS as teacher(FEM).NOM

c. **Pracowano jako nauczycielki.*
worked.IMPERS as teachers(NONVIR)[FEM].NOM

d. *Pracowano jako nauczyciele.*
worked.IMPERS as teachers(VIR).NOM

'[They] worked as teachers. / One worked as a teacher.'

(31) a. **Wyglądano na szczęśliwego.*
looked.IMPERS to happy.MASC.ACC

b. **Wyglądano na szczęśliwą.*
looked.IMPERS to happy.FEM.ACC

c. **Wyglądano na szczęśliwe.*
looked.IMPERS to happy.NONVIR.ACC

d. *Wyglądano na szczęśliwych.*
looked.IMPERS to happy.VIR.ACC

'[They/One] looked happy.'

There is no off-the-shelf LFG analysis of impersonals, and therefore none to fit the *-no/-to* construction. The *-no/-to* impersonal is not a syntactic variant of the passive: it is neither an ill-behaved passive of the transitive, nor equivalent to the passive of the intransitive. It retains the accusative object, can be applied to unaccusatives, and exists alongside the passive – as was shown in (20), it can be formed from a passivised transitive predicate if the passive subject can be interpreted as human. It is, therefore, a different morpholexical construction to the passive (for more detailed argumentation against a passive analysis of this construction, see Kibort 2001).

The fact that *-no/-to* impersonalisation preserves both the grammatical relations and the internal (lexical) semantic structure of the predicate, but only suppresses the surface realisation of the subject, means that, unlike valency-changing operations, it is argument-structure-neutral – the argument structure of an impersonalised verb is unaltered. Thus, the a-structure representation of the impersonalised transitive verb *czytano* 'read.IMPERS' is the same as that of a personal active verb:[4]

(32) **impersonal of the transitive**

⟨ x	y ⟩
|	|
SUBJ	OBJ

[4]In the a-structure representations of impersonalised predicates that I had hypothesised prior to this paper, I placed the symbol Ø under the SUBJ to indicate that this grammatical function was prevented from being mapped onto a categorial argument. Cf. the LMT 'suppression' rule which says: 'Do not map an argument to the syntax' (e.g., Bresnan 2001:21-22; Falk 2001:111) and is notated with Ø. I understand now that this notation was superfluous. If we accept that the subject of the impersonalised predicate is an (obligatory) PRO (i.e., the impersonal inflection provides the specification (↑ SUBJ PRED) = 'PRO' in the f-structure), there can be no other NP that could be the subject at the same time.

The covert subject is not a phonetically empty pronoun (*pro*). It is not a null expletive – Polish does not have expletives at all, and the covert subject has a thematic role. It is not a dropped pronominal subject (VIR), either – the *-no/-to* morphology is not equivalent to normal VIR morphology, and we would have no way of explaining what prohibits the overt expression of the pronominal subject.

However, it is possible to analyse the covert subject of the *-no/-to* impersonal as a pronominal anaphor analogous to the null, or shared, subject of non-finite clauses in syntactic control contexts (PRO). In constructions involving arbitrary anaphoric control the reference of the pronominal element in the clause is not determined syntactically, but the controlled argument finds it referent in a way similar to an ordinary pronoun. Thus, the f-structure of *czytano* 'read.IMPERS' could be represented as in (33), with the c-structure appropriately lacking the node for the categorial expression of the subject:

(33)

$$f\!:\begin{bmatrix} \text{PRED} & \text{'czytano } \langle (f\ \text{SUBJ})(f\ \text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{HUMAN} & + \\ \text{NUM} & \text{PL} \\ \text{GEND} & \text{VIR} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \vdots \end{bmatrix} \end{bmatrix}$$

The [PRED 'PRO'] subject is introduced by the impersonal *-no/-to* inflection. The impersonal PRO differs from the pronominal anaphor of non-finite clauses in that it is clearly finite, and never syntactically controlled. The subject of the *-no/-to* impersonal is always interpreted as an **unspecified** human, but it is by no means always arbitrary – it may have either an unspecified arbitrary referent, or, very commonly, an unspecified definite referent. Furthermore, the human interpretation of the subject of the *-no/-to* impersonal cannot be overridden as it can be in infinitival clauses with the 'optional control' of the PRO by a superordinate non-subject argument: e.g., English *It's all too common to bark (at your kids/*in the dogpound)*, but: *It's all too common for all the dogs to bark all at once in the dogpound*. Finally, when the PRO_{arb} in Polish uncontrolled (i.e., arbitrarily controlled) infinitivals has an adjectival complement, the adjective has to be masculine (singular), while the covert subject of the *-no/-to* impersonal is compatible only with predicate adjectives which are virile (plural); compare examples (31) and (34):

(34) (example from Lavine 2005:97, ft. 26)

Jest ważne *być* *szczęśliwym* / **szczęśliwymi.*
is important.NEUT be.INF happy.MASC.INSTR / happy.PL.INSTR

'It is important to be happy.'

I understand that it is possible to draw all these properties of the *-no/-to* subject from the fact that it is introduced in a different way to the subject of non-finite clauses: here, it is the impersonal inflection itself that provides the (obligatory) [PRED 'PRO'] for its subject together with any other gender and number specifications that are required.[5]

[5]I would like to thank the participants of the LFG06 conference for a helpful discussion of the options of how to analyse the *-no/-to* subject, and for inclining to adopt this one as the most promising. Special thanks

3.2 The reflexive impersonal

The morphosyntactic behaviour of the Polish reflexive impersonal mirrors that of the *-no/-to* impersonal. It can be used with both intransitive and transitive predicates, and it retains accusative objects:

(35) *Budowało się szkołę.*
built.3SG.NEUT REFL school(FEM).ACC
'A/The school was built. / One was building a/the school.'

(36) *Tańczyło się.*
danced.3SG.NEUT REFL
'One danced.'

It can also be formed from both unergative and unaccusative predicates, and from passivised predicates, for example:

(37) *Było się żebrakiem.*
was.3SG.NEUT REFL beggar(MASC).INSTR
'One was a beggar.'

(38) *Było się bitym przez kaprala.*
was.3SG.NEUT REFL beat.PART.MASC.INSTR by corporal
'One was beaten by the corporal.'

The Polish reflexive impersonal does not, under any circumstances, accept the surface expression of a nominative subject (39-40), nor does it accept the expression of the agent in an oblique phrase as in the passive, (41-42):

(39) **Władze budowało się szkołę.*
authorities(NONVIR).NOM built.3SG.NEUT REFL school(FEM).ACC
'(intended) The authorities were building a/the school.'

(40) **Uczniowie tańczyło się.*
pupils(VIR).NOM danced.3SG.NEUT REFL
'(intended) The pupils were dancing.'

(41) **Budowało się szkołę przez władze.*
built.3SG.NEUT REFL school(FEM).ACC by authorities
'(intended) A/The school was built by the authorities.'

(42) **Tańczyło się przez uczniów.*
danced.3SG.NEUT REFL by pupils
'(intended) The dancing was done by pupils.'

The covert subject of the reflexive impersonal is also syntactically active and participates in syntactic control and binding. The reflexive impersonal can share its subject with infinitives (43), with deverbal adverbials (44), and its covert subject is capable of binding reflexive and reflexive possessive pronouns when they need to be bound by the subject (45-46):

to Rachel Nordlinger for further clarifying some issues to me.

(43) *Chciało się wyjechać.*
wanted.3SG.NEUT REFL leave.INF
'There was eagerness to leave.'

(44) *Wsiadając do autobusu pokazuje się bilet.*
get-on.PART$_{CONTEMP}$ into bus shows REFL ticket(MASC).ACC
'On getting on the bus one shows the ticket.'

(45) *Maluje się całego siebie od stóp do głów.*
paints REFL whole.MASC.ACC self.ACC from feet to heads
'One covers oneself with paint from head to foot.'

(46) *Nie niszczyło się swoich dokumentów.*
NEG destroyed.3SG.NEUT REFL own[REFL].NONVIR.GEN documents(NONVIR).GEN
'One did not destroy one's documents.'

Like the *-no/-to* impersonal, the reflexive impersonal has an agent (or experiencer) which licenses agent-oriented adverbials (e.g., *celowo* 'on purpose') and which has a 'default' human interpretation. However, unlike in the *-no/-to* impersonal, this default interpretation can be exceptionally overridden by providing a different referent for the unspecified agent somewhere in the context, for example:

(47) *Gdy się jest bocianem, gniazdo buduje się wysoko.*
when REFL is stork(MASC).INSTR nest(NEUT).ACC builds REFL high-up
'When one is a stork, one builds the nest high up.'

Furthermore, the reflexive impersonal verb form does not seem to impose the same inflectional requirements on its predicative complements as the *-no/-to* form. That is, if the context provides a specific agent/undergoer as the referent of the covert subject, agreeing (nominal and adjectival) predicative complements of the reflexive impersonal may carry any number and person markers corresponding to the features of the referent of this covert subject:

(48) a. *Pracowało się jako nauczyciel /nauczycielka*
worked.3SG.NEUT REFL as teacher(MASC).NOM /teacher(FEM).NOM
/nauczyciele /nauczycielki.
/teachers(VIR).NOM /teachers(NONVIR).NOM
'One worked as a teacher. / [We] worked as teachers.'

b. *Wyglądało się na biednego studenta /biedną*
looked.3SG.NEUT REFL to poor.MASC.ACC student(MASC).ACC /poor.FEM.ACC
studentkę /biednych studentów /biedne
student(FEM).ACC /poor.VIR.ACC students(VIR).ACC /poor.NONVIR.ACC
studentki, to i wpuszczali za darmo.
students(NONVIR).ACC so and let-in.3PL.VIR for free
'One looked like a poor student, so one was let in for free. / [We] looked like poor students, so [we] were let in for free.'

c. *Było się często bitym /bitą*
was.3SG.NEUT REFL often beat.PART.MASC.INSTR /beat.PART.FEM.INSTR
/bitymi.
/beat.PART.PL.INSTR
'One was often beaten.'

d. *Było się kiedyś szczęśliwym /szczęśliwą*
was.3SG.NEUT REFL in-the-past happy.MASC.INSTR /happy.FEM.INSTR
/szczęśliwymi.
/happy.PL.INSTR
'Once, one was happy.'

Like *-no/-to* impersonalisation, reflexive impersonalisation also preserves both the syntactic and semantic valency of the predicate, but suppresses the surface realisation of the subject. Therefore, the reflexive impersonal *czytało się* 'read.3SG.NEUT REFL' can be represented with the same a-structure as the *-no/-to* impersonal *czytano* 'read.IMPERS':

(49) **impersonal of the transitive** $\langle$ x y $\rangle$
| |
SUBJ OBJ

The reflexive impersonal has the same morphosyntactic properties as the *-no/-to* impersonal, therefore it can also be analysed as having an obligatory [PRED 'PRO'] subject, but its subject has different inflectional properties. Instead of the specific number and gender features, the agreement features of the reflexive impersonal's subject could be represented by the metavariable [$agr\ \alpha$]. Thus, the f-structure of *czytało się* 'read.3SG.NEUT REFL' could be represented as in:

(50)

$$f\!:\begin{bmatrix} \text{PRED} & \text{'czytało-się } \langle (f\ \text{SUBJ})(f\ \text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ agr & \alpha \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \vdots \end{bmatrix} \end{bmatrix}$$

Note that the exponent of the impersonal inflection introducing the 'PRO' subject in this construction is analytic, consisting of '3SG.NEUT marker + *się*'.

4 TYPE 3: truly subjectless predicates

There are two types of Polish nominativeless clauses which genuinely do not have syntactic subjects – that is, do not contain elements omitted only from surface syntax, whether due to ellipsis (*pro*-drop) or suppression (as in impersonalisation). They can be formed with two types of predicates which do not have subjects at a-structure as well as at f-structure and c-structure: a small class of defective (non-inflecting) verbs, and passives of intransitives. The existence of these predicates calls into question LFG's Subject Condition and similar principles expressed in other syntactic frameworks, such as GB's 'Extended

Projection Principle' and RG's 'The Final 1 Law'. Because the subject function is assumed to be universally required in clauses, subjects – including null or shared subjects – are standardly considered obligatory, and truly impersonal predicates do not feature in any standard syntactic analyses.

4.1 Inherently impersonal predicates

These clauses do not result from any derivation, and do not contain elements omitted only from surface syntax. Predicates which make these clauses are inherently subjectless – that is, their argument structures inherently lack the first argument.

The class of Polish inherently impersonal predicates is very small and comprises only a few defective (non-inflecting) verbs such as *widać* 'see.[NON-PERSONAL]', *słychać* 'hear.[NON-PERSONAL]', *czuć* 'feel.[NON-PERSONAL]', *stać* 'afford.[NON-PERSONAL]', *znać* 'know.[NON-PERSONAL]'. The form of these verbs resembles the infinitive, but their distribution and morphosyntactic behaviour are not like those of infinitives – they function in the clause as main verbs, resembling personal predicates. Here are examples of typical clauses with these verbs:

(51) a. *Słychać ją / jakieś mruczenie.*
hear.[NON-PERSONAL] her.ACC some.NEUT.ACC murmuring(NEUT).ACC
'One can hear her/some murmuring.'

b. *Było widać łąkę.*
was.3SG.NEUT see.[NON-PERSONAL] meadow(FEM).ACC
'One could see a/the meadow.'

c. *Czuć, że się wygina.*
feel.[NON-PERSONAL] that REFL bends
'One can feel that it is bending.'

As exemplified in the sentences above, all these verbs take complements in the form of an accusative noun/pronoun, a gerund or a finite clause.

If a sentence with a defective verb is meant to refer to the present, the verb may be used with or without the present auxiliary (*jest* 'is'). In the past, as in sentence (b) above, all these verbs require the past auxiliary (*było* 'was.3SG.NEUT') which carries tense marking.

The fact that these predicates are truly impersonal does not seem to be contested in any Polish grammars since, as phrased by Fisiak et al. (1978:24), 'there is no reconstructable noun phrase which can be regarded as being the deleted subject of sentences [with these predicates]' (see also Nagórko 1998:267 for a similar remark).

I suggest that impersonal predicates formed with defective verbs have lexically impersonal argument structures which, in the intransitive variants, may be represented simply as empty argument frames:

(52) **inherently impersonal predicate**

⟨ ⟩

while in the transitive variant they additionally include an object argument (apart from the unoccupied first argument position):

(53) **inherently impersonal predicate with an object**

$$\langle \quad y \rangle$$
$$|$$
$$\text{OBJ}$$

In an argument structure like (53) it is normally expected that the first argument of the predicate is assigned the grammatical function of the subject (as in the canonical anticausative, for example). However, in defective verbs the underlying object ([–r]) is preserved as a syntactic object [+o][6], which makes these verbs somewhat similar to morpholexical impersonals.

In contrast with morpholexical impersonals, defective verbs do not have a covert syntactic subject which would participate in syntactic control and reflexive binding, nor do they have an active agent which would control agent-oriented adverbials. On the other hand, they use the same lexical roots as the corresponding personal verbs which have agents/experiencers: *słyszeć* 'hear', *widzieć* 'see', *czuć* 'feel', etc. For this reason, despite being 'impersonal' at every level of argument structure (i.e. despite being subjectless, argumentless, and agentless), they are used exclusively in situations which involve animate (typically human) participants as agents/experiencers and they are interpreted accordingly. This might be the reason why they are exceptionally allowed to preserve their structural objects. There does not seem to be any other motivation for such a mapping, and the construction does not result from a productive derivational rule. On the contrary, the class of defective verbs in Polish is indeed very small and their morphosyntactic behaviour seems to be unusual.

If the a-structures above are accepted, the f-structure representation of the inherently impersonal verb *widać* 'see.[NON-PERSONAL]' with an object could be:

(54)

$$f\colon \begin{bmatrix} \text{PRED} & \text{'widać } \langle (f\ \text{OBJ}) \rangle\text{'} \\ \text{TENSE} & \text{PRESENT} \\ \text{OBJ} & \begin{bmatrix} \vdots \end{bmatrix} \end{bmatrix}$$

4.2 Passives of intransitives

The impersonal variant of the periphrastic passive results from the application of the passive rule to an intransitive predicate regardless of whether the predicate originally subcategorised for one argument only, or whether it happened to be an intransitive use of a potentially transitive predicate. It is, therefore, a derived construction which does not have a subject (either overt or covert), though it does, arguably, still have the original agent which can be mapped onto an oblique (as in examples (56) and (57) below).

Below are some examples of Polish impersonal passives:

[6]This argument is associated with a *primary*, not secondary, patientlike role, therefore I hypothesise that instead of being pre-specified as [+o], it is pre-specified as [–r] and then allowed to increase in markedness ([+o]) in order to be linked to OBJ. This operation can be referred to as 'object preservation' and it may be found in other types of clauses, e.g. the common (personal) active with a subject instrument that may not be conceptualised as an agent (Kibort 2004:368-372).

(55) *Wchodzisz i czujesz, że było palone.*
come-in.2SG and feel/smell.2SG that was.3SG.NEUT smoke.PART.SG.NEUT
'You come in and you can smell that there has been smoking [here].'

(56) *Czy na tej ulicy już było sypane (przez kogokolwiek)?*
INTERROG on this street already was.3SG.NEUT throw/spread.PART.SG.NEUT (by anyone)
'Has there already been spreading [of grit] on this street (by anyone)?'

(57) *Nie widać, żeby tutaj było sprzątane przez firmę.*
NEG see.[NON-PERSONAL] COMPL.[3SG] here be.-Ł-PART.SG.NEUT tidy-up.PART.SG.NEUT by company
'It doesn't look as if this place was cleaned by a [professional] company.'

(58) *Będzie ci wybaczone, jeśli przeprosisz.*
be.FUT.3SG you.2SG.DAT forgive.PART.SG.NEUT if apologise.FUT.2SG
'[It] will be forgiven you if you apologise.'

The personal passive of the transitive can be represented in the following way (e.g., Bresnan 2001:26):

(59) **passive of the transitive**

⟨ x	y ⟩
|	|
(OBL)	SUBJ

When the passive operates on an intransitive predicate, the result can be diagrammed as follows:

(60) **passive of the intransitive**

⟨ x ⟩
|
(OBL)

If the predicate does not subcategorise for any other arguments apart from the one being downgraded to oblique, there is no possibility of promoting any other argument to the status of syntactic subject. On the other hand, the mere presence of another argument does not guarantee its promotion either. In Polish, only 'underlying' objects, expressing patients/themes, but not beneficiaries or locatives, can become subjects. Although the same general rule applies in English passives, in the appropriate syntactic circumstances, the downgrading of the first argument may result in an oblique location argument being mapped onto syntactic subject.[7]

Locative inversion, described at length particularly in Bresnan & Kanerva (1989) and Bresnan (1994) can be exemplified by the following pair of sentences in English:

[7]As for English beneficiaries, when they are mapped onto passive subjects they occupy the second, direct object, position in the argument structure not the third position of the indirect object. See Kibort (2004:78-90) for discussion.

(61) a. *Those visitors came to the village.*
b. *To the village came those visitors.*

Bresnan (1994) demonstrates that, despite lacking the nominal morphology (and hence the agreement features) of subjects, inverted locatives in English have the properties of syntactic subjects as grammatical relations. Therefore, in sentence (61b) the nominal denoting the 'village' is a syntactic subject, while the nominal denoting the 'visitors' is a syntactic object.

The final mappings of arguments after locative inversion in a predicate such as *come* can be represented as in:

(62) **locative inversion** ⟨ x z ⟩
| |
OBJ SUBJ

Viewing locative inversion as downgrading of the highest argument to a lower grammatical function (in a similar way to passivisation) predicts correctly that locative inversion may be found with predicates which subcategorise for only one argument:

(63) a. *And then, those visitors came.*
b. *And then – came those visitors.*

As in the passive, the downgrading of an argument in (locative) inversion involves a concomitant promotion of another (lower) argument only if there is something to be promoted. If there is no argument available to become subject, (locative) inversion results in another subjectless construction, analogous to the impersonal passive of the intransitive:

(64) **(locative) inversion** ⟨ x ⟩
|
OBJ

The demotional (rather than promotional) analysis of both passivisation and locative inversion reveals that, when the two constructions are considered together, they emerge as complementary processes which are part of a larger system of operations occurring in the argument structure of predicates. It has been observed that there are crosslinguistic restrictions on the applicability of both passivisation and locative inversion which are based on the distinction between unergative and unaccusative predicates: passivisation applies only to unergatives, while locative inversion only to unaccusatives. In this way, the two operations apply to two complementary classes of predicates, and they essentially serve the same purpose: they both target the highest argument of the predicate in order to downgrade it to a lower grammatical function (the oblique, and the object, respectively).[8]

It has been noted that the overt expression of the downgraded agent in impersonal passives in Polish is not as easily acceptable as in personal passives. This may be due to the fact that passivisation of intransitive predicates yields clauses which structurally resemble and functionally pattern with unspecified-agent constructions.[9] If a predicate has only one argument, the agent, it can either be specified and appear in a personal clause, or be unspecified through a variety of means. Some of the means of despecifying the agent do not

[8]For detailed discussion of these operations, and references, see Kibort (2001; 2004).

[9]Blevins (2003:489) remarks that '[s]ubjectless passives often have an implicitly human interpretation, which suggests that this interpretation is associated with subjectless forms of personal verbs, irrespective of the syntactic source of that subjectlessness'.

make the clause subjectless: these are the use of lexical items with unspecified/generic reference (e.g., the English *one*, *people*, or *they*), or the use of conventionally interpreted verbal agreement (e.g., 3PL in Polish). If, however, the agent is despecified through impersonalisation or passivisation, the clause lacks a surface subject. Reintroducing the downgraded agent into an impersonal passive would contradict the intention to despecify it in the first place. Although it is syntactically legitimate in Polish, it is often more readily acceptable if the agentive phrase is an afterthought or addition to the main utterance, as in:

(65) *Dzisiaj było już sprzątane –przez sprzątaczki.*
today was.3SG.NEUT already clean.PART.SG.NEUT by cleaners
'The cleaning has already been done today – by cleaners.'

Reintroducing the agent into surface syntax as an oblique evidently does not pose the same kind of problem in personal passives. This may be because the prime motivation behind personal passives may be the need to locate the syntactic pivot on the initial object of the predicate, and not the need to despecify the agent of the predicate. Impersonal passives do not have the capacity to provide a different syntactic pivot for the clause. Therefore, unless the agent of an intransitive predicate needs to be unspecified, it will simply be kept as the subject of the personal active sentence rather than downgraded from this position only to be reintroduced to surface syntax as an oblique constituent.

Due to the lack of a syntactic subject, impersonal passives show 'default' impersonal agreement – that is, the verbs appear in 3SG neuter form. Recall that the same inflectional form is used in the Polish reflexive impersonal which does not have its own dedicated verbal morphology (the personal verb form is simply accompanied by the multifunctional marker *się*).

If the a-structures of the subjectless variants of the passive (60) and the (locative) inversion construction (64) are accepted, the f-structure representations of the impersonal passive *sprzątane* 'cleaned, tidied-up' and the inverted impersonal *came* could be, respectively:

(66)

$$f: \begin{bmatrix} \text{PRED} & \text{`sprzątane } \langle (f\ \text{OBL}) \rangle\text{'} \\ \text{OBL} & \begin{bmatrix} \vdots \end{bmatrix} \end{bmatrix}$$

(67)

$$f: \begin{bmatrix} \text{PRED} & \text{`came } \langle (f\ \text{OBJ}) \rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{OBJ} & \begin{bmatrix} \vdots \end{bmatrix} \end{bmatrix}$$

5 Conclusions

The range of Polish subjectless constructions seems to constitute a useful set for testing any syntactic theory. There is large number of constructions which appear to lack the subject, which have distinct morphosyntactic properties, and whose subjectlessness can therefore be attributed to different phenomena identified at different levels of representation that should

be posited for a predicate. In order to be properly distinguished, the constructions need to be handled by a correspondence-based model, and it appears that LFG can successfully provide one.

I have shown that two types of Polish constructions: morpholexical impersonals (TYPE 2) and truly subjectless constructions (TYPE 3) require new analyses.[10] I have suggested that morpholexical impersonalisation could be analysed as the provision of an obligatory PRO for the predicate's subject. As for truly subjectless predicates, once it is accepted that they have a valency slot that is unavailable for grammatical function mappings, the behaviour of their remaining arguments is not unusual. However, this analysis indicates that the 'Subject Condition' should be eschewed.

This last suggestion may not be as drastic as it seems because the Subject Condition may, in fact, be seen as simply redundant. The Mapping Principles of LMT (Bresnan 2001:311) appeal to the markedness hierarchy in (69) derived from the grouping of the grammatical functions into natural classes based on their features, where the highest syntactic function is the least marked:

(68) MAPPING PRINCIPLES

(a) Subject roles:

(i) a [–o] argument is mapped onto SUBJ when initial in the argument structure;[11] 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.

(69) MARKEDNESS HIERARCHY OF SYNTACTIC FUNCTIONS

[–o]/[–r] SUBJ $\succ$ [–r]/[+o] OBJ, [–o]/[+r] OBL_θ $\succ$ [+o]/[+r] OBJ_θ

In order to make full use of the markedness hierarchy, the Mapping Principles could be reformulated to a single one as follows:

(70) MAPPING PRINCIPLE

The ordered arguments are mapped onto the highest (i.e. least marked) compatible function on the markedness hierarchy.

[10] The passive (not discussed here, but see Kibort 2004 for details) would also benefit from a slightly revised analysis, specifically that it is an instance of alternative (non-default) mapping of grammatical functions onto the arguments of the predicate by means of which the underlying subject is downgraded to an optional oblique.

[11] The actual LFG formulation of this mapping principle is as follows: '$\hat{\theta}_{-o}$ is mapped onto SUBJ when initial in the a-structure' (Bresnan 2001:311), where $\hat{\theta}_{-o}$, referred to as the 'logical subject', is defined as 'the most prominent semantic role of a predicator' (ibid.:307). However, this formulation seems to contain superfluous information. Specifically, due to the Subject Condition, LFG excludes the formation of predicates without any 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 (68a)(i) is in fact just a more concise, but still faithful, version of the LFG principle.

The new formulation derives the principles of argument to function mapping directly from the markedness hierarchy, without the in-built condition that the first encountered argument has to be pre-specified as either [–o] or [–r], and without having to resort to the Subject Condition at any point. In other words, it is now the markedness hierarchy itself which determines the default mapping of arguments to surface grammatical functions. Thus, with the unergative transitive verb *clean*, the Mapping Principle in (70) ensures that its first ([–o]) argument is linked to SUBJ, and its second ([–r]) argument is linked to OBJ. Similarly, with the unaccusative intransitive verb *come*, the Mapping Principle ensures that its first ([–r]) argument is linked to SUBJ because this is the grammatical function which is the highest compatible one on the markedness hierarchy in (69).

The new formulation achieves correct mappings for various classes of predicates discussed in the literature (including unaccusatives and ditransitives, for example), but avoids stipulating 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. Since it makes redundant the Subject Condition, it enables LMT to handle inherently impersonal predicates and other constructions that may have posed problems of analysis due to their non-standard behaviour with respect to the subject.

References

Blevins, J. P. (2003). Passives and impersonals. *Journal of Linguistics* **39**. 473–520.

Bresnan, J. (1994). Locative inversion and the architecture of universal grammar. *Language* **70(1)**. 2–131.

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

Bresnan, J. & Kanerva, J. M. (1989). Locative inversion in Chichewa: a case study of factorization in grammar. *Linguistic Inquiry* **20(1)**. 1–50. Reprinted in Stowell, T. & Wehrli, E. (eds.), *Syntax and Semantics* 26: *Syntax and the Lexicon*, pp. 53–101. New York: Academic Press.

Dziwirek, K. (1994). *Polish Subjects*. New York: Garland.

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

Fisiak, J., Lipińska-Grzegorek, M. & Zabrocki, T. (1978). *An Introductory English-Polish Contrastive Grammar*. Warszawa: PWN.

Kibort, A. (2001). The Polish passive and impersonal in Lexical Mapping Theory. In Butt, M. & King, T. H. (eds.), *Proceedings of the LFG01 Conference*. Stanford, CA: CSLI Publications. 163–183. Available at http://csli-publications.stanford.edu/LFG/6/lfg01.html.

Kibort, A. (2004). *Passive and Passive-like Constructions in English and Polish*. Ph.D. thesis, University of Cambridge, Cambridge. Available online.

Lavine, J. E. (2005). The morphosyntax of Polish and Ukrainian *-no/-to*. *Journal of Slavic Linguistics* **13(1)**. 75–117.

Nagórko, A. (1998). *Zarys gramatyki polskiej (ze słowotwórstwem)*. Warszawa: PWN.

Wierzbicka, A. (1966). Czy istnieją zdania bezpodmiotowe. *Język polski* **46(3)**. 177–196.

OPTIONAL DIRECT OBJECT CLITIC DOUBLING IN LIMEÑO SPANISH

Elisabeth Mayer
Australian National University

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)

2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

In some Spanish dialects, direct object arguments of transitive clauses under certain conditions allow co-occurrence of a pronominal clitic with a coindexed lexical NP (direct object clitic doubling). In Limeño, as well as in Standard Spanish, in accordance with Kayne's Generalization, direct object clitic doubling obtains only under *a*-marking, by conveying animacy and specificity on direct objects. This paper explores the motivations and mechanisms of *a*-marking based on these referential categories with an emphasis on optional marking. It focuses on the resulting morphosyntactic reflexes: mood in relative clause, *a*-marking DOM, and clitic doubling. This analysis links Kayne's Generalization to topic marking in Spanish, which mostly seems to hold, by associating the semantic feature specificity with the discourse roles TOP and FOC. The evolution of clitics from marking agreement to marking a secondary topic is ascribed to a known grammaticalization process of the formative *a*.

1 Introduction*

Standard Spanish requires pronominal objects to be expressed by a clitic, allows optional doubling by a pronoun and rejects doubling of full NPs. All dialects require obligatory doubling of a pronominal direct object (DO) as demonstrated in (1).

(1) Pedro **lo** vió *a* él. All dialects
Peter DOCLMascSg saw-3Sg OM PROMasc3Sg
Peter saw him.

Standard Spanish rejects doubling of full lexical NPs as in (2a). All dialects abide by the principle of economy of expression [1] as in (2b): the referential PRO is supplied by anaphoric control.

(2) a. Invitaron *a* Beto y Carlos. Standard Spanish
They invited Beto and Carlos.

b. **Los** invitaron (*a* ellos). All dialects
DOCLMascPl invited-3P (OM PROMasc3Pl)
They invited them.

Direct object clitic doubling (DOCLD) is much more restricted than indirect object clitic doubling showing considerable cross-dialectal variation. In River Plate (RP) doubling extends optionally to animate full NP objects (Suñer 1988) and in Limeño to proper names and topics (Mayer 2003).

* I would like to thank all participants at the LFG06 conference who provided me with stimulating questions and constructive feedback. I am particularly grateful to Mary Dalrymple for pointing out to me the connection to FOC and TOP as well as to Miriam Butt and Tracy H. King for valuable comments on editing. I owe special thanks to Avery Andrews for extended discussions and insightful advice. All blunders are mine. Unless otherwise noted, the data discussed are drawn from fill-in questionnaires, interviews, and Limeño newspapers for my MLing thesis; the non Agr PRO contact data at the end are taken from my fieldwork in Lima for my ongoing PhD research.

[1] Economy of expression (Bresnan 2001b): the DO argument in cases like (2b) can be left out when it only supplies redundant information and when it is not needed for semantic expressivity.

(3) a. **(Los)** invitaron *a* Beto y Carlos. River Plate
(*DOCLMascPL) invited-3Pl OM Beto and Carlos*
They invited Beto and Carlos.

b. **(La)** nombraron *a* **Mara**. Limeño
(DOCLFemSg) nominated-3Pl OM Mara.
They nominated Mara.

Limeño also shows traits of leísmo similar to Quiteño[2], producing apparent indirect object constructions that can be doubled as in example (4a). These forms are ambiguous with respect to the grammatical function they are representing. Peninsular leísta dialects disallow doubling of these forms. River Plate does not show any leísmo effects, it unambiguously marks direct objects (4b).

(4)a. **Les** invitaron *a* Beto y Carlos nomás. Limeño Spanish
IOCLPl invited-3Pl OM Beto and Carlos only
They only invited Beto and Carlos.

b. **Los** invitaron *a* Beto y Carlos nomás. River Plate
DOCLMascPl invited-3Pl OM Beto and Carlos only
They only invited Beto and Carlos.

In following a standard LFG approach to clitics, I assume that DO clitics in a CLD construction as in examples (1), (3) and (4) are agreement markers co-occurring with an NP. Clitics occurring without the argument as in (2b) are thematic–role bearing pronouns. Optionality of this alternative function is specified in the lexical entries. For the thematic role bearing DO pronoun, the f-structure must show a referential feature [PRED 'PRO'] and agreement features such as number and gender as exemplified in (5a). The reduced f-structure in (5b) shows the entries for the clitic as agreement marker.

(5)a $\begin{bmatrix} \text{PRED} & \text{c 'PRO'} \\ \text{GEND} & \text{MASC} \mid \text{FEM} \\ \text{NUM} & \text{SG} \mid \text{PL} \end{bmatrix}$ (5)b $\begin{bmatrix} \text{GEND} & \text{MASC} \mid \text{FEM} \\ \text{NUM} & \text{SG} \mid \text{PL} \end{bmatrix}$

These notations are as in Andrews (1990:544) and should be read as a convenient informal notation for expressing equations, defining unless marked with c, in which case they are constraining.
In a clitic-doubling construction the exclusion of the co-occurrence of the [PRED 'PRO'] form is regulated by the Functional Uniqueness Principle. For DO clitics an optional constraint is needed to allow for the alternative function as pronouns or as agreement markers. The latter show animacy constraints to avoid conflicting semantic features from co-occurring in the functional head.

Morphological Blocking (MB) (Andrews 1990) does not quite cover specificity due to the fact that the *a* marker is not an inflection. Andrews assumes that doubling and non-doubling lexical entries require PRED=$_c$ PRO vs. PRED=PRO. Deal (2005) shows that the correlation between agreement and PRO drop is poor, casting severe doubt on this idea. PRED PRO seems to be supplied by anaphoric control rather than by the lexical entries.

[2] The Spanish dialect of Quito (Ecuador) simplifies the case paradigm in favor of the dative. This can also be found in Paraguayan Spanish and in Basque contact varieties. (cf. Yépez 1986, García 1990)

Therefore the entry for the DOclitics l(a/o)(s) in Standard Spanish is PRED=$_c$ PRO. MB then forces the clitic to be present when it can. For LS, the constraint on the clitic loosens to allow objects that are topics and an OPTIONAL feature when the DO is not a pronoun. This is illustrated in (6a). For RP the only constraint is optionality as in (6b), where the OPTIONAL here means that the entry does not trigger MB and thereby does not prevent a less specified form from being produced (Andrews 1990:543).

(6) a Limeño

$$\begin{bmatrix} \text{GEND} & \text{MASC} \mid \text{FEM} \\ \text{NUM} & \text{SG} \mid \text{PL} \\ \text{PRED} & = \text{c 'PRO'} \\ \begin{bmatrix} (\text{TOP}\uparrow) \\ \text{OPTIONAL} \end{bmatrix} & \end{bmatrix}$$

(6) b River Plate

$$\begin{bmatrix} \text{GEND} & \text{MASC} \mid \text{FEM} \\ \text{NUM} & \text{SG} \mid \text{PL} \\ \left\{ \begin{matrix} \text{PRED} = \text{c 'PRO'} \\ \text{OPTIONAL} \end{matrix} \right\} & \end{bmatrix}$$

2 Theoretical background

This proposal builds on "caseless" approaches such as Suñer's (1988) agreement approach which is based on the requirement of the "Matching Principle" and on the apparent loss of accusative case in favor of a binary [+DAT] and [-DAT] distinction as described in Alsina (1996). Spanish as a pro-drop language has only object clitics.[3] DO clitics as overt Agr PROs are inherently specified for person, gender and number, whereas IO clitics only show case and number. The case requirements imposed by the predicate on the f-structure will select the appropriate clitic as to ensure completeness and coherence. Subjects in Romance obey the Nondative Subject Constraint. Internal and external argument functions are not marked through abstract case distinctions but through a clear distinction of syntactic functions. For more evidence refer to Company (2001) arguing for a language shift in Spanish through multiple grammaticalization processes reinforcing dative marking and incorporating it into the clause as the prime object.
Consequently direct object clitics in clitic-doubled constructions are understood as caseless object markers obeying DOM as described in Bossong (1985) and Aissen (2003). Specificity is understood as intrasentential referential anchoring of an NP to another discourse object in the spirit of von Heusinger (2002). The definition of definiteness as a discourse pragmatic property, ensuring anaphoric linking, is based on Heim's Familiarity Principle (1988). Doubling of proper names will be analyzed relying on scopal specificity by Farkas (2002).

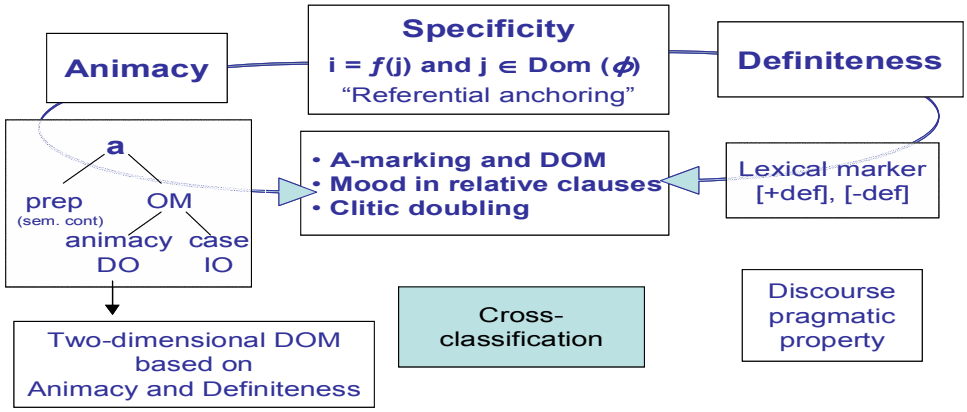
The aim of this paper is to show how the referential categories animacy, definiteness and specificity interact to license optional DOCLD in LS, a dialect that still exhibits agreeing clitics, as opposed to Limeño contact varieties, where clitics seem to have shifted from an agreement marker to a transitivity marker and/or topicality marker.

The paper is organized as follows: In section 3, I will proceed to define and show the scope of the three referential categories animacy, specificity and definiteness and provide a rather detailed analysis of the multifunctional formative *a*. Section 4 demonstrates the importance of *a*-marking DOM for specificity and discusses an interesting case of optional *a*. In section 5, I present various specificity effects on DO clitic doubling and link them to topic marking. In section 6, the role of a sole clitic in non Agr PRO Limeño contact clitics is analyzed as a topic marker. A short conclusion is given in section 7.

[3] The only SUBJ CL would be impersonal se ('one'). It cannot double an overt SUBJ.

3 Definition and scope of syntactic, semantic and pragmatic factors involved in DOCLD

The diagram in (7) shows the interaction of the three referential categories animacy, specificity and definiteness.



Animacy, definiteness and specificity are referential categories of different semantic and pragmatic natures, reflected in various morphosyntactic phenomena. In Spanish, animacy is encoded in the overt morphosyntactic marker *a* giving rise to DOM. Objects are marked for prominence on a "culture sensitive animacy Hierarchy" (Mohanan 1994), where Human > Animate > Inanimate. Definiteness is overtly encoded in a lexical marker; Spanish has a pair of number and gender distinctive definite and indefinite articles. Specificity in turn lacks such a lexical marker; it uses the formative *a* in its virtue as DOM, affecting definite as well as indefinite NPs. This gives rise to the assumption that specificity is not only a subcategory of indefinite NPs but an independent category that "can therefore form a cross-classification" (von Heusinger 2002:248). The term specificity corresponds roughly to identifiability as used by Bossong (1985). Specificity as a referential category shows the following morphosyntactic reflexes: i) mood in relative clause (with [+def] and [-def] nouns), ii) *a*-marking and DOM and iii) CLD.

3.1 Animacy

In Spanish animacy is encoded in the overt morphosyntactic marker *a*, broadly known as prepositional accusative. However, the multiple roles of the formative *a*[4] in Spanish resists a unified analysis in terms of a mere animacy marker. The one form **a** has three homophonously expressed functions. *A* is homophonous with a) the preposition **a** having semantic content (8); in this case **a** can be replaced by another preposition; with b) the dative case marker for indirect objects (9), and finally with c) the object marker (OM) marker for direct objects (10).

(8) Pasó el río con el agua **a** (hasta) la cintura.
crossed-3Sg DET river with DET water PREP (PREP) DET waist
He/she crossed the river with the water reaching to his/her waist.

(9) **Les** ofrecieron entradas gratis *a* **los visitantes** del hipódromo.
IOCLPl offered-3Pl tickets free to DET visitors of-DET racetrack
They offered free tickets to the visitors of the racetrack.

[4] In all examples the preposition **a** is marked bold and *a* as Case and object marker in italic.

Example (10) illustrates the complex use of the OM; it includes personal *a*[5] to mark human objects and DOM to mark personifying animal objects (pets) and inanimate and specific objects (as in examples (24-26) in section 5.1). Thus multifunctional *a* resists a unified account.

(10) Juan estima *a* Pedro.
Juan appreciates-3Sg OM Pedro
Juan appreciates Pedro.

3.1.1 True Animacy marker vs. specificity marker with quantified phrases

Relevance of *a* as an animacy marker is particularly obvious with quantified phrases. Example (11a) is understood as animate whereas (11b) is inanimate. Note that both can be CLD in LS. This is possible as *todas* in these examples is a quantification over a set, each *todas* refers to an imaginable subset of a superset, which can be *muchos* or *pocos*, a case of partitive specificity in the sense of Enç (1991).

(11) a. Ya **las** lavé ***a*** todas. [+animate]
Already DOCLFemPl washed-1Sg A all-(FemPl)
I already washed them all. (for example: the girls)

b. Ya **las** lavé todas. [-animate]
Already DOClFemPl washed-1Sg all-(FemPl)
I already washed everything. (for example: all dolls)
(Suñer1988:401)

3.1.2 Personal *a* versus specific indefinites

Definite articles presuppose that the common noun they modify is a singleton. Indefinite articles on the other hand, do not trigger presuppositionality even with a specific reading.
Leonetti (1999) claims that *a* in example (12) is a true animacy marker and does not convey specificity to the indefinite, animate NP as in the mood example (13b) below.

(12) a. Vimos ***(a)** unas mujeres en la plaza. [personal *a*]
saw-1Pl (A) DETindef women PREP DET market place*
We saw some women in the market place.
(Leonetti 1999:866)

From the previous examples it has become clear that the presence or absence of personifying ***a*** signals degree of animacy and/or distinctness. What is the difference then between (12a) and (13b)? Why is (13b) not simply another case of personal *a*-marking of an indefinite NP?

According to Luján (1987), Rivero (1975), and Torrego (1998) among others, there is a known correlation between subjunctive mood[6] and definiteness, and specificity. Despite the human DO, the use of the subjunctive in sentence (13a) is already enough to render the sentence unspecific, thus no *a*-marking.

[5] Personal *a* is relatively unknown in other Romance languages except for Sicilian and some other Southern Italian dialects.
[6] The aspect problem is another factor in CLD dealing with the impossibility of iteration of the VP predicate with DOCLD constructions. It does not seem to vary cross-dialectally.

(13) a. Fueron **a** buscar un médico experimentado que conociera bien las
went-3Pl PREP look-inf for DETindef doctor experienced that know-3SgSubj well DET
enfermedades del país.
diseases of-DET country
They went to look for an experienced doctor, who would know about the prevalent diseases in the country.

Version (13b) on the contrary, receives a specific reading through *a*-marking, which is indicated with the indicative in the subordinate clause. Specific in (13b) here means 'implied' existence, the doctor is known to the speaker. There is clearly established referential identity of the NP with a familiar entity in the sense of background information and the fact, that the doctor is well known, entails the existence of such a person.

(13) b. Fueron **a** buscar *a* un médico extranjero que gozaba de una
went-3Pl PREP find-INF OM DETindef doctor foreign that enjoyed-3Sg of DETindef
gran reputación.
great reputation
They went to look for a foreign doctor who enjoyed a great reputation.
(Bello 1984: 268)

Leonetti (1999:862) calls this an extensional argument. This is in contrast to example (13a) where novel information has been introduced; it is a purely intensional object: it does not relate to a fact or an accomplished situation. Intensional objects generally reject *a*-marking.

3.1.3 Word order

As Spanish, a pro-drop language, has relatively free word order,[7] *a*-marking is useful under certain conditions to distinguish between grammatical functions. Zubizarreta (1999) claims that Spanish, in contrast to Italian, shows overt morphological case that distinguishes objects from subjects. (14a) shows two postverbal inherently identical arguments, an overt lexical post verbal subject and object in canonical object position, both [+human]. The inflectional morphology of the sentence-initial verb would agree with either. As you can see in (14b) and (14c), a [+human] DO being highest on the Animacy scale must be *a*-marked to disambiguate the sentence.

(14) a. *Abrazó Juan María.
embraced-3Sg John Mary
He/she embraced Juan María. (Double name)

b. Abrazó Juan *a* María.
embraced-3Sg John OM Mary
John embraced Mary.

b'. Juan abrazó *a* María.
John embraced-3Sg OM Mary
John embraced Mary

c. Abrazó *a* Juan María.
embraced-3Sg OM John Mary
Mary embraced Juan.

c'. María abrazó *a* Juan.
Mary embraced-3Sg OM John
Mary embraced John

Note that personal *a* will be omitted in ditransitive constructions. A possible reason is that the topicality properties are being monopolized. Inanimate NPs in VSO and SVO constructions (15)

[7] As Spanish exhibits a well defined set of constraints on verbal complements, Demonte (1994) compares Spanish to English, German and Hindi in regard to object asymmetries and scrambling.

do not need personal *a*-marking for disambiguation, however, they can enter specificity relations by optional *a*-marking as we shall see in example (20) among others to come.

(15) a. Abrazó Juan el árbol.
embraced-3Sg John DET tree
John embraced the tree.

b. Juan abrazó el árbol.
John embrace-3Sg DET tree
John embraced the tree.

So far we can state that *a*-marking is a complex and multifunctional issue. Spanish uses the formative *a* as an obligatory animacy marker for [+human] DOs (personal *a*) and for [+animate] DOs for disambiguation. Optional *a*-marking as in (13b) allows for specificity effects on the DO argument. NPs without *a*-marking seem to have no or only the lowest form of specificity.

3.2 Specificity vs. definiteness

In the late 1960s the term specificity[8] was introduced to further describe a general phenomenon that attributes value to variables in a variety of ways. Indefinite and definite NPs can be distinguished semantically by the uniqueness condition. Some languages mark these differences morphologically or lexically, others don't. As we have seen, Spanish uses the animacy marker *a* as a morphosyntactic marker for specificity. The literature deals mainly with specific indefinite NPs and categorises them in various ways. The different kinds of specificity can be analysed from two main focus points: scope (Farkas 2002) and referentiality (Fodor & Sag 1982). See also Enç (1991) for another important distinction between relational and partitive specificity.

In the following I adopt von Heusinger's 'referential anchoring approach' meaning that specificity should be analysed in terms of the 'referential structure' of the text. The Specificity Condition (16) (von Heusinger 2002:269) restricts the linking of reference indices internally to the sentence, referentiality is thus sentence bound.

(16) An NP_i in a sentence ϕ with respect to a File *F* and the Domain of filenames DOM ϕ is [+specific] if there is a contextual salient function *f* such that $i = f(j)$ and $j \in DOM(\phi)$

In other words, the specificity of an NP is given if its file name (index) can be described as a contextually salient function of the file name of another NP within the same sentence (Domain of file names) or in short:

(17) $i = f(j)$ and $j \in DOM(\phi)$

The definition of definiteness on the other hand is a contentious issue. Givón (1979) defines a definite as an identifiable or referentially accessible existence within the Domain of the relevant discourse. Von Heusinger though claims that definiteness cannot solely be defined in terms of identifiability. Factors such as uniqueness, saliency, familiarity, functionality and many more play an important role and their importance is theory dependent. Definiteness always shows the functional connection of the new referent with a previously introduced item in the discourse. Thus definiteness is a discourse feature; its function crosses sentence boundaries. He defines definiteness after Heim's Familiarity Condition (in von Heusinger 2002:268) as follows:

[8] A. Martinet (1960) and G. Lakoff (1968) were probably among the first to mention specificity.

(18) An NP_i in a sentence ϕ with respect to a File *F* and the Domain of filenames DOM (*F*) is

i [+definite] if $i \in$ DOM (*F*), and it is

ii [-definite] if $i \notin$ DOM (*F*)

Meaning that any indexed NP that is part of the discourse is definite, but how exactly is the Domain *F* defined? Heim (1988:302) gives the following instructions for File-keeping: "For every indefinite, start a new card; for every definite, update a suitable old card." The Domain *F* is defined as "the set that contains every number which is the number of some card in *F*" (Heim 1988:304).

For my claim, that specificity and not definiteness is the licensing factor in DOCLD, it is fundamental to state that indefinites introduce a novelty into the Domain of discourse, whereas definite NPs must denote an entity familiar to the addressee. The chain of events is used to determine whether novelty has been brought into the Domain of discourse or not.

4 *A*-marking and DOM

As we have seen for Spanish, the semantic and pragmatic features of the DO decide whether it gets overtly marked for "case" or not. DOM (Bossong 1985, Aissen 2003) can account for cross-linguistic variation of these object alternations. "The higher in prominence a direct object, the more likely it is to be overtly case-marked." (Aissen 2003:436). Prominence[9] is determined by animacy and definiteness. Most languages use two-dimensional DOM, based on animacy and definiteness. Some languages change the scales to animacy and specificity. Persian and Turkish for example, case-mark all specific objects. In Turkish specificity marking includes cases where the speaker has a specific referent in mind, a parallel situation to Limeño Spanish and also River Plate.

Urdu and Spanish use the same case marker (*ko* in Urdu/*a* in Spanish)[10] for IO and DO. On DOs as in (19b) *ko* marks specificity/definiteness.

(19) a. nadya=ne jiraf dek^h-na hɛ
Nadya.F.Sg=Erg giraffe.M.Sg.Nom see-Inf.M.Sg be.Pres.3.Sg
'Nadya wants to see a giraffe/giraffes.'

b. nadya=ne jiraf =ko dek^h-na hɛ
Nadya.F.Sg=Erg giraffe.M.Sg=Acc see-Inf.M.Sg be.Pres.3.Sg
'Nadya wants to see the giraffe.'
(Butt 2005:143)

According to Butt, the speaker must have a specific giraffe in mind in (19b). The argument is considered to be a direct object in accusative case but on the level of s-structure, the NP should be interpreted as specific.

[9] Other factors like person (Silverstein 1976, Comrie 1989) and Topicality and Aspect (Kiparsky 1998 for Finnish and Torrego 1998 for Spanish) may also play a role cross-linguistically.
[10] The common genesis of Urdu *ko* and Spanish *a* as locative postposition and preposition respectively is also striking.

4.1 Optional *a*-marking with inherently identical arguments

The often cited controversially *a*-marked double inanimate sentence in (20) is a good example to show that optionality is a privative opposition. In *a*-marking this sentence, two interconnected issues have to be taken into account. A third issue, namely, disambiguation of subject and object, can be disregarded as the non *a*-marked version is equally accepted.

(20) El interruptor controla (*a*) la máquina. Limeño Spanish
DETMascSg switch controls-3Sg OM DETFemSg machine
The switch controls the machine.

The first issue deals with verbal preference for selecting a direct rather than indirect object. If this were the case, we could not have optional *a*-marking. However, we know that in contact varieties the distinction between DO and IO is fuzzy at times due to a known grammaticalization process of the formative *a.* Yet this seems to be far fetched. That leaves us with the last option: DOM to mark animacy and specificity.

I propose to call the *a*-marked version a familiar definite reference, in the sense of Enç (1991) "Having a specific referent in mind". The fact that a switch controls a machine is part of our daily life. The *a*-marked version reportedly sounds specific to a Limeño speaker. The marked version yields a marked meaning and allows identification of a certain machine.

Wh-questions show clearly the distribution of animacy: as expected animate subjects (21a) are felicitous with animate and inanimate objects. Question (21b) shows the optional *a*-marking. Inanimate subjects as in (21c) are not felicitous without personal *a* for animate/specific objects. (21d) allows for inherently identical inanimate objects without *a*-marking.

(21) a. ¿Quién controla *a* quién? * ¿Quién controla quién?
Who controls-3Sg OM whom

b. ¿Quién controla qué? ¿Quién controla *a* qué?
Who controls-3Sg what

c. ¿Qué controla *a* quién? *¿Qué controla quién?
d. ¿Qué controla qué?

Contrary to Jaeggli's argument, that strong PROs are favorably interpreted as animate and thus cannot refer to inanimate DOs as in example (22), there was no objection to the strong PRO *ella* for the inanimate object in the clitic doubled version. The doubled argument in (22) is considered redundant by most informants; it falls under the economy principle.

(22) El interruptor **la** controla (***a*** **ella**). Limeño Spanish
DETMascSg switch DOCLFemSg controls-3Sg (OM PROFem3Sg)
(Montalbetti in Andrews 1990 :541)

Aissen's analysis predicts that if strong personal pronouns occurred in direct object position in lieu of an inanimate object, they would be case-marked. (Aissen 2003:462-463).

For an analysis of *a*-marking with inanimate SUBJ and inanimate OBJ in bi-directional OT see de Swart (2003). For another account of *a*-marking the DO argument in the case of inherently identical arguments see Hanssen (1945).[11]

[11] In his analysis of the double inanimate sentence "El adjetivo modifica al sustantivo" (The adjective modifies the noun) *a*-marking is analyzed as marking the unique DO argument.

5 Clitic doubling and specificity effects

Cross-linguistic evidence for the participation of a formativ in CLD is an observable fact not only in Romance languages, mainly in Spanish and Romanian (23), but also in genetically unrelated languages, such as Hebrew, Swahili and Chicheŵa as described by Bresnan (1987), similar to the effects of the Animacy Hierarchy.

(23) L- am vizitat *pe* bunicul nostru.
DOCLMAscSg-TM visited-1Pl prep grandfather-DetMascAcc POSS1PlMasc
We visited our grandfather.
(Daniliuc & Daniliuc 2000:282)

In Limeño as well as in Standard Spanish, DOCLD obtains only under *a*-marking, observing Kayne's Generalization, by conveying specificity and animacy on DOs. It is generally assumed that CLD is related to referential problems of small clauses. Specifics, definites, demonstratives and possessives can take the position of head of a doubled NP, non-specific NPs, such as bare plurals, cannot and are therefore excluded from CLD. Obligatory DOCLD in all Spanish dialects only holds for pronominal arguments, all others are disallowed in Standard Spanish. Optionality in turn is a privative opposition and varies cross-dialectally.

5.1 Preposed arguments

In regard to CLD I distinguish between left dislocation and preposing. Left dislocation involves a pause, an intonational break and *a*-marking is optional. Left dislocated topicalized arguments cannot 'move' back into the original object position whereas preposed elements can. Left dislocated arguments will not be treated here.

In preposing, *a*-marking for DOs is subject to specificity and animacy restrictions. Preposed CLD objects are not argument functions but discourse functions in topic position (TOP) assuming simultaneously the in-clause function object and the discourse function TOP. The discourse function (preposed object) as well as the in-clause function (clitic) must be co-referential and show the same f-structure values. Fronted elements must have an appropriate relationship to a PRED, which means that if interpreted as FOC or TOP they have to be anaphorically linked with an integrated function and Functional Uniqueness has to be obeyed.[12] Focus arguments are usually associated with new, non- presupposed information, whereas topical arguments express what the sentence is "about", they are associated with presupposed material. CLD of preposed DOs is subject to specificity constraints in the first place and only secondly to animacy. If appearing in focus position, CLD preposed DOs have to be *a*-marked and are considered to be topics.

The examples (24)-(26) show various degrees of animacy, on "a culture sensitive animacy Hierarchy scale" as already mentioned in section 3. The proper name in (24) is "scopeless' like demonstratives, i.e. proper names always show widest scope, and are therefore assumed to be existentially presupposed.

(24) *A* Pablo **lo** escogieron para representar *a*l[13] colegio.
OM Pablo DOCLMascSg chose-3Pl for represent-INF OM-DET school
Pablo got chosen to represent the school. [+human, +spec]

[12] See Bresnan 1987 for a detailed analysis.
[13] Note that *a* + el (definite article, masculine) contracts to *a*l.

The animal in (25) is a specific and definite 'pet' and the fish in (26) are definite animals with no emphatic value but specific through the demonstrative *esos* which allows referential identification in situ.

(25) *A*l perro de mi vecino **lo** atropelló un carro.
OM-DET dog of POSS neighbour DOCLMascSg hit-3Sg DETIndef car
My neighbor's dog got hit by a car. [+anim, +spec]

(26) *A* esos peces hay que pescar**los** con anzuelos.
*OM DEM fish have-impersonal that catch-Inf-**DOCLMascPl** with hooks*
Those fish have to be caught with hooks. [+anim, +spec]

The non *a*-marked inanimate definite and specific NP in (27) is a case of a topicalized left dislocation with a resumptive pronoun.

(27) El compromiso de escribir, **lo** asumo totalmente.[14] [-anim,+def,+spec]
Det commitment of write-Inf DOCLMascSg assume-1Sg completely
I fully embrace the duty of writing.

Examples (28) and (29) show a semantic difference in meaning due to the verb's selection for direct or indirect object (dative-accusative alternation). Note in particular that (29) is **not** an instance of leísmo as described in (4a).

(28) No es malo que *a* un escritor **lo** silben de vez en cuando.
not is-3Sg bad that OM DETindef writer DOCLMascSg whistle-3Pl of time in time
It is not bad that a writer gets booed from time to time. [+anim,-def,+spec]

(29) No es malo que *a* un escritor **le** silben de vez en cuando.
not is-3Sg bad that OM DETindef writer IOCLSg whistle-3Pl of time in time
It is not bad that a writer gets whistled at from time to time. [+anim,-def,+spec]

Fodor and Sag (1982) distinguish between "speaker intent to refer" and reliance on other parts of the context for interpretation. That is the distinction is referential vs. non referential. The preposed transitive clauses (30d) and (30e) are obligatorily doubled and *a*-marked. But why is *a*-marking in the impersonal sentence in (30a) optional?

(30) a. (*A)* esa silla hay que poner**la** en otro sitio. [-anim,-spec,+def]
*(OM) THAT chair has that put-Inf-**DOCLFemSg** in other place*
That chair has to be put somewhere else.

b. Hay que poner esa silla en otro sitio. [-anim,+spec,+def]
That chair has to be put in another place.

c. *A* esta silla hay que poner**la** en otro sitio. [-anim,+spec,+def]
*OM THIS chair must-impers put-Inf-**DOCLFemSg** in other place*

[14] Examples (27)-(29) are taken from MarioVargas Llosa, 'Sólo miento en mis novelas' (my translation: I only lie in my novels). Interview in El Tiempo, May 6, 2003.

d. *A* esa silla **la** pongo en otro sitio. [-anim,+spec,+def]
OM DEM silla DOCLFemSg put-1Sg in other place.
I'll put that chair somewhere else.

e. *A* esa silla **la** quiero poner en otro sitio. [-anim,+spec,+def]
I want to put that chair somewhere else.

I propose to categorize specificity into ordinary and contrastive specificity in this case. Ordinary specificity would be the unmarked topic position (30a without *a*) and contrastive specificity the *a*-marked versions, showing a somewhat stronger form of specificity. This stronger form of specificity is directly dependent on the demonstrative denoting a specific chair the speaker has in mind in the sense of Fodor & Sag and on agentivity of the SUBJ (30d). Yet, there seem to be subtle differences as less doubt would even arise in (30c) with the demonstrative *esta*. No *a*-marking would be possible with a definite or indefinite determiner. In this corpus, it appears also that other inanimate preposed DOs of the patient type semantic role rejected *a*-marking as well, whereas all experiencer types were *a*-marked. The parallel to DOM and IO marking here is striking.

We have seen that the discourse functions FOC and TOP in preposed CLD objects carry a specific information load which is central to information structure. While topics are generally associated with non-focal and presupposed material, the focus position is the place for new non–presupposed information. The canonical direct object position is the preferred place for the latter.

The following analysis of three specificity effects on DOCLD in canonical position will shed more light on the previous discussion.

5.2. Scopal specificity

Proper names, as demonstratives, are 'scopeless' (Farkas 2002), i.e. they show widest scope as they introduce a new referent whose existence is presupposed. They do not depend on the context for reference like definite pronouns do. They introduce a unique reference and in this regard definite descriptions (definite lexical NPs) are closely related. In (31a and b) Mara gets singled out, is chosen above other candidates in a context where she is known by both speaker and hearer, and so Mara is the topic of the clause.

(31) a. La nombraron *a* Mara. (en especial) (LS)
DOCLFemSg called-3Pl OM Mara
They nominated Mara.

b. La nombraron *a* ella. (specifically her-instead of someone else)

Focus examples (31c and d) mention casually that a person called Mara got nominated. (31d) is puzzling in all dialects: a pronominal is required to be doubled by a clitic. The lack of the clitic is possibly due to the following complement.

(31) c. Nombraron *a* Mara. (como jefa del grupo – as group leader)

d. Nombraron *a* ella. (como gerente general – as general manager)
She got nominated.

In the example (32) below the clitic marks the proper name as the TOP of the clause ; without the clitic, *Grimanesa* is in FOC position.

(32) De repente (**la**) vió *a* Grimanesa bajando las escaleras.
Suddenly (DOCLFemSg) saw-3Sg OM Grimanesa coming down DET stairs
Suddenly he/she saw Grimanesa coming down the stairs.

The use of the definite article together with a proper name is redundant as the value condition contributed by the NP already satisfies the requirements. Some languages, e.g. German, allow co-occurrence of the definite article and a proper name under certain conditions. *Die Johanna hat angerufen* (The Johanna called). The person *Johanna* must be known to both the speaker and the addressee, it must be part of their known world or at least part of the discourse context. In Spanish the above-mentioned co-occurrence of article and proper name is also possible (*La Johanna que yo conozco* (The Johanna I know)) giving more evidence to the discussion above. In any case the naming function of N of a proper name is a fixed value (by convention) and cannot be modified by any other value function.

5.3 Partitive specificity

In Limeño, all indefinite, nonspecific and nonpronominal NPs are still barred from doubling despite the animate feature of the object. A specific clitic such as **lo** in (33b) and (33c) cannot co-occur with an unspecific argument in a reference chain, violating completeness and coherence. The indefinite *nadie* has no antecedent, *nadie* cannot be linked to another referent in the clause, so it cannot denote its topic. But in our world we know that expressions like *nadie, alguien* and *ninguno* (nobody, somebody and nobody (as adjective)) refer to a group of human beings.

(33) a. No veo *a* nadie. [+anim, -spec, -def]
Not see-1Sg OM nobody
I do not see anybody.

b. No ***lo** veo *a* nadie.
*Not*DOCLMascSg see-1Sg OM nobody*

c. No ***lo** vieron *a* nadie en la playa. [+anim, -spec,-def]
*Not *DOCLMascSg saw-3Pl OM nobody in DET beach*
They did not see anybody on the beach.

d. No **lo** vieron *a* nadie en **esta** playa. [+anim, +spec,-def]
Not DOCLMascSg saw-3Pl OM nobody in DEM beach
They did not see anybody on this beach.

Example (33d) has become marginally grammatical by adding the demonstrative *esta* to the locative PP, thus allowing a truth reading of the predicate argument. "The lexically specified evaluation parameter will ensure that the noun phrase will denote the individual the speaker has in mind."(Farkas 2002:5)

5.4 Referential anchoring

The definite, inanimate and unspecific argument in (34) is unavailable to clitic doubling. Clauses like (35) constitute clear evidence for the specificity effect: for the fact, that the

interaction of animacy plus specificity, that is two-dimensional DOM, licenses DO clitic doubling in Latin American Spanish dialects with agreeing clitics.

(34) a. *No **lo** vimos el bus. All dialects
*Not *it saw-1Pl Det bus*. [-anim, -spec, +def]
We did not see the bus.

b. ***Lo** vimos el bus.
**it saw-1Pl Det bus*
We saw the bus.

(35) (No) **lo** vimos *a*l bus (de la línea 38). Limeño Spanish
(no) DOCLMascSg saw-1Pl OM-DET car (of DET line 38) [-anim, +spec, +def]
We did not see the (route 38) bus.

This a classic example of Kayne's Generalization and argument for CLD clauses to be topics. The direct object position is the natural FOC position of a sentence and also the preferred place to introduce new referents or information. However, applying DOM to (35) scopal specificity is conveyed onto the inanimate NP leading to a specific interpretation by ensuring successful sentence-internal referential identification in Heusinger's sense. CLD obtains even in the scope of negation. The clitic is not redundant; it is needed for "semantic expressivity" as expressed in Bresnan (2001b). I assume that the clitic here in the CLD clause is marking a topical object and could be formalized in LFG as follows in (36).

The clitic **lo** has the (TOP↑) restriction. If a DO is a value of TOP then it must be *a*-marked.

(36) VP → V NP PP
↑= ↓ (↑OBJ)=↓ (↑OBJ)=↓
¬(TOP↑)

a: P, (TOP↑)

This formalization would account for the data presented in (34) as long as the non *a*-marked DO is not a topic and for the CLD argument in (35) as long as the absence of **lo** does not mean the NP is not a topic.

6 Clitics on the move

Limeño contact varieties display a hybrid clitic system showing either case or gender with an almost total lack of number agreement and also null direct objects.

According to Greenberg (1966:61) featurally unmarked forms can "act as a surrogate for the entire category." This seems to be the case in the archmorpheme **lo** as illustrated in the short discourse example (37) and the DOCLD example in (38), also called 'strange **lo'.** This is a well documented phenomenon apparently only in Peruvian contact varieties.[15]

(37) a. Yo **lo** ví *a* **la** **chica**. Allí estaban **ellas.**
*PRO1Sg **DOCLMascSg** saw-1Sg OM **DETFem girl**. There were-3Pl **PROFem3Pl***
I saw the girl. They were there.

[15] cf. Camacho and Sanchez (2002), Cerrón-Palomino (2003).

b. Los chicos **los** ignoraban.
DET boys ***DOCLMascPl*** *ignored-3Pl.*
The boys ignored them.

c. Y ahora en la mañana no **lo** ví *a* **ella**.
And now in the morning not ***DOCLMascSg*** *saw-1Sg OM* ***PROFem3Sg***
And this morning I did not see her.

(38) **Lo** frío *a* **la** **cebolla**.
DOCLMascSg *fry-1Sg OM* ***DETFemSg onion***
I fry the onion.

In Limeño contact non Agr PRO – as in this strange **lo** – occurs parallel with leísmo. Similar case paradigm variations can be found in Basque Spanish contact varieties.[16] It has also been reported for L2 English speakers of Hispanic background in the United States.[17] In Quiteño the strange **lo** is inexistent and the merging process of DO clitics and IO clitics has been almost completed in favor of the IO. Vincent (2001) calls this an extreme case of leísmo and argues that this loss may have given rise to null direct objects. Null direct objects as in (39b) seem to be constrained by definiteness and recoverability constraints.

(39) a. Recogiste **los** **documentos**?
picked up-2Sg ***DETMascPl documents***
Did you pick up the documents?

b. Ayer mismo **Ø** recogí.
Yesterday exactly **Ø** *picked up-1Sg*
I picked them up yesterday. (Data from 2nd fieldwork, LS contact)

The present analysis of specificity seen as referential anchoring and found to be licensing CLD in non-contact varieties is a small clause phenomenon and cannot be applied to contact varieties where a failure of coindexing produces ungrammatical results by failing the test for completeness and coherence as exemplified in (38). However, DOM in contact varieties seems to undergo a process of evolution due to contact and other factors like, for example, grammaticalization of the formative *a*. The evidence from examples (37)-(39) suggests that the reduction of the clitic paradigm and the loss of agreement features of clitics in contact varieties point towards a shift from an agreement marker to a transitivity marker and possibly discourse referent marking topicality as in (40) below. Following Bresnan (2001a) the evolution of Agr PRO to Non AgrPRO could be shown in the partial f-structures in (41).

(40) lo = (TOP↑)

(41) $\begin{bmatrix} \text{PRO} \\ \text{AGR} \end{bmatrix} \longrightarrow \begin{bmatrix} \text{TOP} \\ \text{PRO} \end{bmatrix}$

7 Conclusion

In this paper I discuss the complexity of optional DOCLD in Limeño Spanish. My primary concern here has been to show how the interaction of the referential categories animacy, definiteness and specificity accounts for optional DOCLD in LS. In a detailed discussion of the

[16] See Fernández Ordoñez 1994, Suñer 1989.
[17] See Luján and Parodi 1996.

formative *a,* I have argued against the simple term prepositional accusative and for DOM instead. The Limeño data showed that the scale for DOM has been pushed to include inanimate DOs for a specific reason. With the analysis presented here it is possible to link Kayne's Generalization (and DOCLD) to topic marking in Spanish. New data from contact Spanish corroborate the hypothesis that the clitic **lo** is shifting from an agreement marker to a secondary topic marker in the spirit of Dalrymple and Nikolaeva (2006). A plausible reason for this evolution process would be a known grammaticalization process of the formative *a*.

References

Aissen, Judith. 2003. 'Differential Object Marking: Iconicity vs. Economy'. *Natural Language & Linguistic Theory* **21**:435-483

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

Andrews, Avery. 1990. 'Unification and Morphological Blocking'. *Natural Language & Linguistic Theory* **8**:507-557

Bello, Andrés. 1984. *Gramática de la Lengua Castellana.* Madrid: EDAF, Ediciones-Distribuciones, S.A.

Bossong, Georg. 1985. *Empirische Universalienforschung, Differentielle Objekt –markierung in den neuiranischen Sprachen.* Tübingen: Narr

Bresnan, Joan. 2001a. The emergence of the unmarked pronoun. In *Optimality-Theoretic Syntax.* G. Legendre, J. Grimshaw and S. Vikner (eds). Cambridge, MA: MIT Press. 113-142

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

Bresnan, Joan and Sam Mchombo. 1987. Topic, pronoun and agreement in Chicheŵa. *Language* **63**:741-782

Butt, Miriam. 2005. *Theories of Case.* Cambridge: University Press

Camacho, José and Liliana Sanchez. 2002. Explaining Clitic Variation in Spanish. In *Language Universals and Variation.* M. Amberber and P. Collins (eds). Praeger Publishers: 21-41

Cerrón-Palomino, Rodolfo. 2003. *Castellano Andino. Aspectos sociolingüísticos, pedagógicos y gramáticales.* Lima: PUCP Fondo Editorial.

Company, Concepción. 2001. Multiple dative-marking grammaticalization. Spanish as a special kind of primary object language. *Studies in Language* **25** (1): 1-47

Comrie, Bernard. 1989. *Language Universals and Linguistic Typology.* Chicago: The University of Chicago Press

Dalrymple, Mary and Irina Nikolaeva. 2006. Topicality and nonsubject marking: Agreement, casemarking and grammatical function. Ms. Oxford University

Daniliuc, Laura and Radu Daniliuc. 2000. *Descriptive Romanian Grammar*. München: Lincom Europa

Deal, Amy Rose. 2005. *Pro-drop, topic-drop, and the functional lexicon. A constructional account of null arguments.* Honors Thesis. Brandeis University

Demonte, Violeta. 1994. On certain asymmetries between DOs and IOs. In *Paths Towards Universal Grammar: Studies in Honour of Richard S. Kayne.* G. Cinque, B. Koster, J.-Y. Pollock, L. Rizzi, R. Zanuttini (eds).Washington D.C.: Georgetown University Press: 111-120

de Swarts, Peter. 2003. The Case Mirror. MA thesis. University of Nijmegen

Enç, Mürvet. 1991. 'The Semantics of Specificity'. *Linguistic Inquiry* **22**(1):1-27

Farkas, Donka. 2002. 'Specificity Distinctions'. *Journal of Semantics* **19**(3):213-145

Fernández-Ordoñez. 1994. Isoglosas internas del castellano. El sistema referencial del pronombre átono de tercera persona. *Revista de Filología Española* **74**:71-125

Fodor, Janet & Ivan Sag. 1982. 'Referential and Quantificational Indefinites'. *Linguistics and Philosophy* **5**:355-398

García, Erica. 1990. 'Bilingüismo e interferéncia sintáctica'. *Lexis* **14**:151-195
Givón, Talmy. 1979. *On Understanding Grammar*. New York: Academic Press
Greenberg, Joseph.1966. *Language Universals. With Special Reference to Feature Hierarchies.* The Hague: Mouton
Hanssen, Federico. 1945. *Grámatica histórica de la lengua castellana.* Buenos Aires: Librería y Editorial "el Ateneo"
Heim, Irene. 1988. *The Semantics of Definite and Indefinite Noun Phrases.* New York & London: Garland Publishing Inc.
Kiparsky, Paul. 1998. Aspect and event structure in Vedic. *Yearbook of South Asian Languages and Linguistics.* 29-61
Lakoff, George. 1968. *Pronouns and Reference.* Bloomington: Indiana University Linguistics Club
Leonetti, Manuel. 1999. El Artículo. *Gramática Descriptiva de la Lengua Española.* I. B. Muñoz and V. Demonte Barreto (eds). Madrid: Espasa Calpe 2:787-890
Jaeggli, Osvaldo. 1982. *Topics in Romance Syntax.* Holland: Dordrecht. Foris
Luján, Marta. 1987. Clitic doubling in Andean Spanish and the theory of case absorption. L*anguage and Language Use. Studies in Spanish.* T. A. Morgan, J. F. Lee and B. van Patten (eds). Lanham: University Press of America. 109-121
Luján, Marta and Teresa Parodi. 1996. Clitic doubling and the acquisition of agreement in Spanish. In *Perspectives on Spanish Linguistics.* J. Gutiérrez-Rexach and L. Silva-Villar (eds). Los Angeles: UCLA .119-138
Martinet, André. 1960. *Éléments de linguistique générale.* Paris: Armand Colin
Mayer, Elisabeth. 2003. *Clitic Doubling in Limeño. A Case Study in LFG.* unpubl.Mling thesis. Canberra: Australian National University (http://anu.edu.au/languages/postgraduates/ elisabeth-mayer.asp)
Mohanan, Tara. 1994. *Argument Structure in Hindi.* Stanford: CSLI Publications
Rivero, María Luisa. 1975. 'Referential properties of Spanish Noun Phrases'. *Language* **51**: 32- 48
Silverstein, Michael. 1976. Hierarchy of features and ergativity. In *Grammatical Categories in Australian Languages.* R.M.W Dixon (ed). Canberra: Australian Institute for Aboriginal Studies. 112-171
Suñer, Margarita. 1988. 'The Role of Agreement in Clitic-Doubled Constructions'. N*atural Language and Linguistics Theory* **6** (3): 391-434
Suñer, Margarita. 1989. Dialectal Varation and Clitic-Doubled Direct Objects. In *Studies in Romance Linguistics.* C. Kirscher and J. Decesaris (eds). 377-397
Torrego, Esther. 1998. *The Dependencies of Objects.* Cambridge, Mass: MIT Press
Vincent, Nigel. 2001. LFG as a Model of Syntactic Change. *In Time over Matter: perspectives on morphosyntax.* M. Butt and T. King (eds). Stanford: CSLI Publications. 1-42
von Heusinger, Klaus. 2002. 'Specificity and Defniniteness in Sentence and Discourse Structure'.*Journal of Semantics* **19**(3):245-275
Yépez, María. 1986. *Direct Object Clitics in Quiteño.* MA thesis. Cornell: Ithaca New York
Zubizarreta, María L. 1999. Word order in Spanish and the Nature of Nominative Case. In *Beyond Prinicples and Parameters. Essays in Memory of Osvaldo Jaeggli.* K. Johnson and I. Roberts (eds). Dordrecht: Kluwer Academic Publishers. 45:223-251

A COMPUTATIONAL ARCHITECTURE FOR LEXICAL INSERTION OF COMPLEX NONCE WORDS

Bruce Mayo
Universität Konstanz

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

Derivational morphology has been conspicuously neglected in the LFG literature and elsewhere. Existing proposals in HPSG treat derivational patterns that are constrained and regular but sidestep problems raised by derivations that require knowledge-based, pragmatic evaluation, and most ignore the problem of nonce derivations, which must be processed on-line, i.e., concurrently with syntax. While practical implementation remains a formidable challenge, it is possible to specify a computational architecture that accounts for some 'worst case' patterns of productive nonce derivation in Italian that require pragmatic evaluations. This architecture factors lexical insertion into two functions, c- and m-insertion, for inflection and derivation. A buffer between these functions and the syntax component is shown to explain lexicalization phenomena, and it is argued that it may be one of the cognitive sources of Lexical Integrity.

1 Introduction

For computational and theoretical linguists, morphology has by and large meant inflectional morphology, and derivational morphology has been seen as a set of static, irregular and unpredictable relations within the lexicon that do not merit synchronic analysis. An important exception is work that has been done in the HPSG framework, where a number of researchers have shown that some, usually rather specific, derivational patterns can be modelled using that formalism's inheritance relations (Koenig and Jurafsky 2004, Koenig and Davis 2006). In LFG, the macro facility available in the extended formalism of the XLE system could, in principle, be used to derive related sets of words. In both cases, however, what results is little more than a kind of data compression: suitable roots are expanded to their derivational variants at compile time, sparing the lexicon writer the effort of creating explicit entries for them. More recent work in HPSG has explored the use of semi-productive lexical rules for inserting complex, transparently derivable words into syntax at run-time (Briscoe and Copestake 1999). As will be shown in this paper, however, the underlying nature of derivation inevitably makes it difficult to predict the semantics, mapping relations and other properties of all transparently derivable words without recourse to a level of general knowledge representation and conceptual operations. The task is not hopeless, however. A growing body of work, e.g., (Corbin 1990, Mayo et al. 1995, Stiebels 1996, Lieber 2004) demonstrates that the conceptual operations underlying derivation as well as the relationships between semantic representation and syntactic expression (Levin and Rappaport Hovav 2005) can be grasped. A computational implementation, however incomplete, can help to clarify what kinds of information must be available at each of the several interfaces that make up the linguistic system.

1.1 The KLU Computational Model

For many years, theoretical linguists were inclined to regard the derivational relationships among words as irregular and unpredictable, the product of historical processes lying outside the purview of grammar. However, experimental studies of the representation of lexical items, e. g., (Marslen-Wilson et al. 1994), reveal that some derivational roots and affixes seem to have independent mental representations, much like stems and inflectional affixes, suggesting that they ought to be freely combinable; and from corpus statistical studies (Baayen and Renoulf 1996) we know that certain derivational patterns are continually producing new words that no dictionary could hope to anticipate. Where words are produced freely, like sentences, it must be possible to identify and model the grammatical processes that create them. To this end, a small linguistic workbench for derivational morphology, called KLU, was developed at the University of Konstanz and was used to implement small grammars for word and sentence comprehension (Mayo 2000). Specifically, it was meant to provide formal solutions to the following problems:

- Complex and apparently idiosyncratic words can appear ‘out of the blue’. Hence, derivation must be possible concurrently with sentence analysis, i.e., within syntax.
- Derivation is not syntax, but its structures look much more like syntax than those of inflection: derived words do not fill out paradigms; like sentences, they exhibit structural embedding and name a limitless range of objects and events.
- Unlike inflectional attributes, the meanings that arise from derivation are subject to lexical shift, so that derived words often have competing transparent and opaque meanings.

To accommodate these requirements, the KLU program introduced three formal devices:

- **c-insertion**, which performs inflectional analysis and has tacitly always been a part of LFG
- **m-insertion**, which does derivational analysis, supplying newly derived stems to c-insertion, and
- a **morphological buffer**, which serves as an interface between the syntactic and the morphological components. The buffer is necessary for processing efficiency, and it helps to model the process of lexicalization.

The program was required to construct the semantics of sentences containing nonce derivations, mainly Italian. Its favourite derivation was *disiscrivere*, an invented Italian word meaning to ‘unregister’, as in

(1) La fata disiscrive il cavaliere dal castello.

‘The fairy unregisters the knight at the castle’

To allow the parser to process a sentence containing the non-lexical item, the program could create an ‘on-the-fly’ lexical entry for the unknown word, containing a lexical form and a semantic formula that, in the program’s output, looked like this:

```
/ dis iscriv e / ================================
§ Dis_iscrivere(S,O,L) =>accomplishment(v1)
                              agent(v1,S)
                              theme(v1,O)
                              localization(v1,L)
                              phase(v1,CAUSE(S,
                                    CHANGE(LISTED(O,L),NOT(LISTED(O,L)))))

TpLex:Verb -> /TpLex:Verb/   [PRED: dis_iscriv((ˆ SUBJ),(ˆ OBJ),{(ˆ OBL)}]
                              [AUX: AVERE]
                              [μˆ INFLCLASS: Vkere]
                              (ˆ OBL PCASE ) =/cc Da
```

Using these, the parser and the semantic analyser constructed a semantic analysis of the embedding sentence.

```
The relation dis_iscriv(Fata,Cavaliere,Castello) =>
     ACCOMPLISHMENT(411)
     AGENS(411,Fata)
     THEMA(411,Cavaliere)
     LOCUS(411,Castello)
     PHASE(411,CAUSE(Fata,
                 CHANGE(LISTED(Cavaliere,Castello),
                         NOT(LISTED(Cavaliere,Castello)))))
```

1.2 Computational Overview

While the KLU program was little more than a makeshift toy, it had to deal with a wide range of phenomena involved in derivational morphology, from lexical phonology to discourse structure, and it gave considerable insight into the kinds of problems that nonce words present for a lexically-oriented theory like LFG. To get a quick overview of how the functions listed above might deal with a sentence containing a new word, consider (2).

(2) Please open the wine bottles with those unflaskers.

Now *unflaskers* is not to be found even in the OED, but speakers have little difficulty understanding what is meant. Hence, some sort of analysis below the word level must be making the word available to syntax. Figure 1 is meant to give an impression of what happens; the details will be fleshed out in sections 2 to 4.

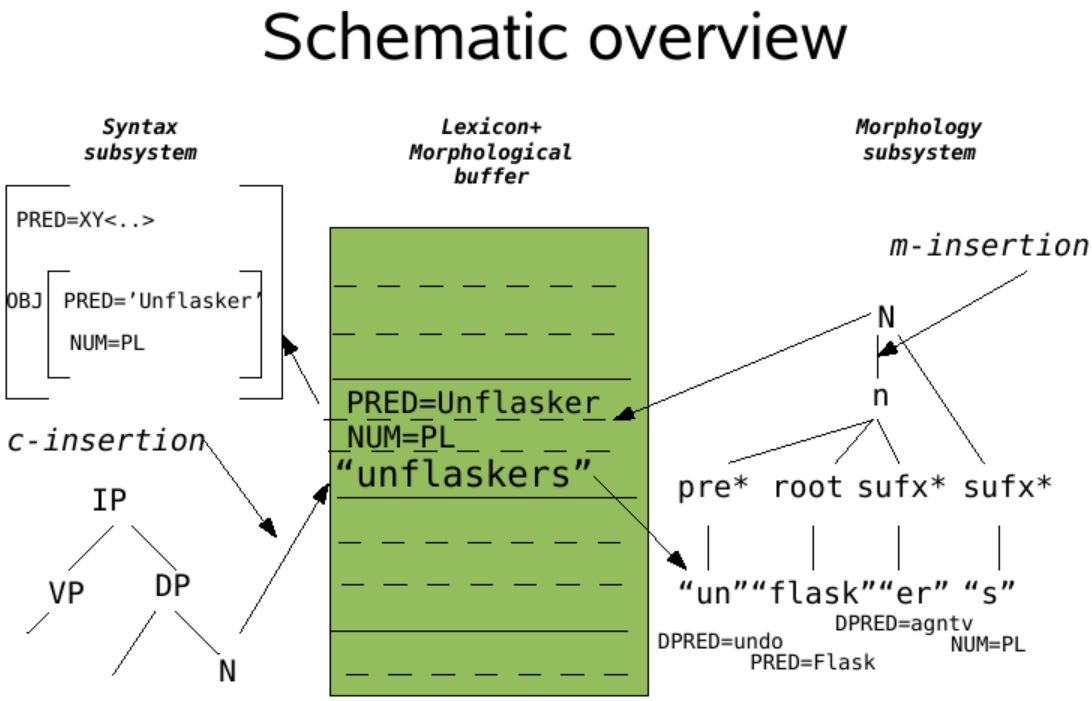


Figure 1. Schematic Overview of the KLU Model.

Consider a top-down parser working its way through the sentence, word by word. When it reaches a leaf node of the c-structure, in the proposed model it does not access the lexicon directly but calls the function **c-insertion** to obtain the lexical features of the item. For each word encountered, c-insertion places the surface form in a lexical buffer (shaded box), which queries the static lexicon. For words like *please*, *wine*, etc. it returns stored features without further ado. But when c-insertion encounters *unflaskers* there is no information in the lexicon to return. Thus, the buffer invokes morphology to decompose *unflaskers* into sublexical units (not really morphemes), the substrings "un" "flask", "er" and "s". Since these are not c-structure entities, their features cannot project to any c-structure nodes, and they do not even unify. Hence, they require a derivation. **c-insertion** calls a rather complex operation, called **m-insertion**, to answer a semantic riddle of the form "what object undoes something involving a flask". Presumably it finds a conceptual description of some sort of tool that is meant to open or empty flasks, or a relevant generic concept. Abstracting a semantic structure from this concept, Lexical Mapping Theory creates a corresponding lexical form and deposits a

PRED feature in the buffer. m-insertion also places a lexical-semantic formula in the buffer (not shown) as the referent of the PRED feature, so that the parser will be able to construct the semantics and discourse structure of the sentence. The morphological analyser then recognizes the segment "s" and deposits the feature Plural in the morphological buffer. This results in a morphologically complete lexical item. Now syntactic analysis can continue as if the form *unflaskers* had existed in the lexicon all along.

The following paragraphs describe these three components in more detail. For **c-insertion** (section 2) traditional LFG approaches to morphology are fine for inflection but not for derivation. For **m-insertion** (section 3), I shall fill in the outline just presented. Section 4 then sketches the entire process, showing how the Italian verb *sbobinare* 'unspool' would be inserted as a nonce derived word. For the **morphological buffer**, section 5 shows that, like c-insertion, it is an idea that has been around for a long time, but whose theoretical significance has never been comprehended. Computationally it is unavoidable, and it lets us model the effects of lexicalization and semantic drift of derived words. Its role in mediating between syntax and lexicon may be one of the reasons why we find a structural barrier between syntax and lexicon.

1.3 Lexical Integrity

It is evident that the model described in Figure 1 allows word analysis from within syntax but does not violate LFG's basic postulate of lexicalism. c-insertion mediates between syntax and morphology, constituting a kind of barrier between the two. The boundary between syntax and the lexicon is a rather complicated matter (Bresnan and Mchombo 1995); but in Bresnan's 2001 formulation it is simply the point where the structural principles of c-structure end and 'morphological completeness' begins:

> Morphologically complete words are leaves of the c-structure tree and each leaf corresponds to one and only one c-structure node (Bresnan 2001, 92).

Hence, one might see c-insertion as being the computational expression of this definition. But if it turned out to be necessary on other, independent grounds, we might want to see it as being a cognitive reason, or causal source, of Lexical Integrity. This is in fact a claim I want to make, but the justification will only unfold when we consider the last of the three proposed components, the lexical buffer, in section 5.

2 The c-insertion Function

Figure 2 describes c-insertion in more detail. It constitutes the interface from c-structure to the lexicon. At each c-structure leaf node, the syntactic parser passes the input string to c-insertion, which returns the string's lexical features if it can find or compute them. Seen from syntax, this is simply an access to the abstract lexicon. Whether the features are obtained from the static lexicon (upper shaded box) or from an analysis via the morphological buffer does not matter, and syntax cannot tell the difference. Formally, c-insertion can be described using an apparatus similar to that used in syntax (Börjars et al.1996) if the restrictions on word structure described in (Bresnan and Mchombo 1995) are born in mind.

c-insertion corresponds in the XLE system to the two-level morphology component, but in KLU its algorithms draw on ideas from lexical phonology. A significant feature is that segmentation is kept fully separate from morphological analysis proper, which deals only with the output from segmentation (details are discussed below and in more depth in Mayo 2000).

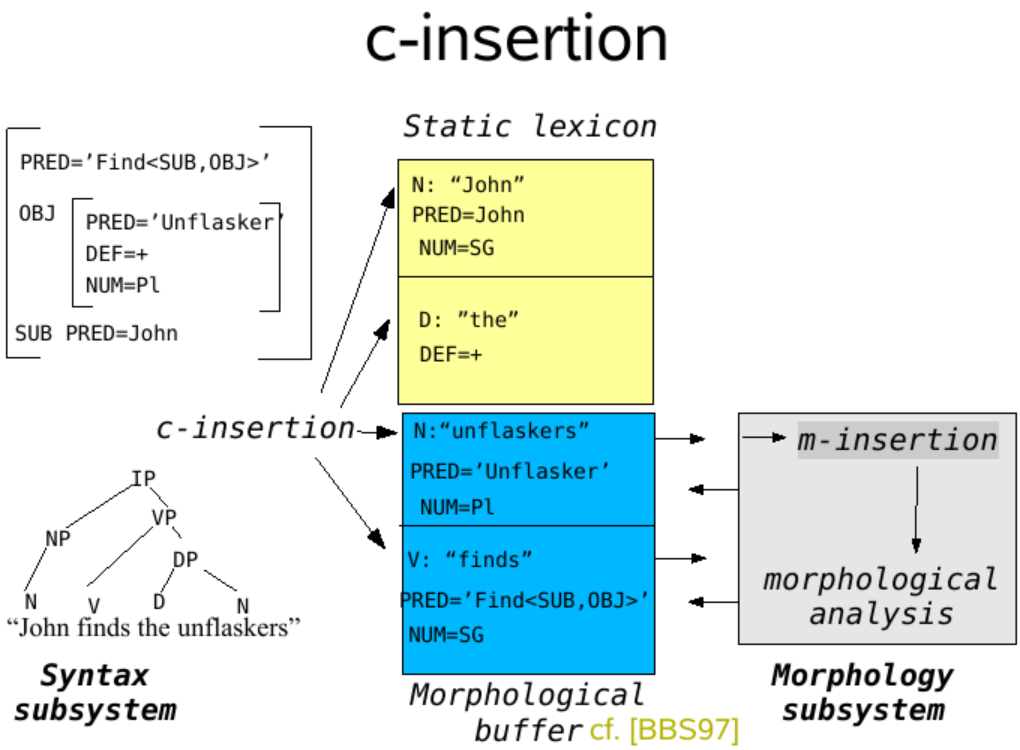


Figure 2. The c-insertion Function.

Monomorphemic words are fetched directly from the static lexicon; stems and inflectional segments unify locally within c-insertion and remain for a time as atomic morphemes in the buffer. If c-insertion cannot find a lexical stem for a string, it sends it to the derivational analyser (box on the right), which will try to return lexical features to the buffer.

In Figure 2, we see that the items "John" and "the" of sentence (2) are present with their lexical features in the upper shaded box, representing the static lexicon. In contrast, "finds" and "unflaskers" only appear in the buffer (lower box) after their constituents have undergone inflectional and derivational analysis. The unification takes place in local structures, and there is a test for morphological well-formedness (e.g, for the presence of inflection on categories that require it, and for the absence of features belonging to derivational morphemes). Once an analysed word is in the morphological buffer, it does not need to be recomputed, and it remains there for a while, indistinguishable from entries in the static lexicon, until it is eventually purged to make room for new computations. If a purged item is analysed again, it is purged more slowly the next time around. This suggests that frequently encountered inflected forms might stay in the buffer indefinitely and behave just as if they were monomorphemes in the static lexicon, in effect, as if they had no internal morphological structure. This is in fact what some experimental studies have found to be the case, e.g., (Baayen et al. 1997).

3 The m-insertion Function

The function m-insertion is actually only a sub-function of c-insertion, called when c-insertion encounters a derivational segment, whose features cannot be projected to syntax. Since derivation takes place inside inflection, many have been tempted to think of it as a sort of extension or elaboration of inflection, especially since inflection has proved to be computationally quite tractable. There are many kinds of derivational relations that seem to be regular and systematic, like passivization, causative formation, etc., especially in morphologically rich languages. Hence, one could think of derivation as filling out paradigmatic matrices, albeit

large ones, and therefore as being amenable to strategies that are effective for complex inflectional paradigms (cf. Karttunen 2003). If we want to be prepared for the worst cases, however, this would be much too simplistic. A representative 'worst case' is the denominal verb of removal in Italian. It creates not only a new semantic structure, more complex than that of its base, but also introduces a new argument structure and other morphological features like inflectional class. This is a type of derivation that is very productive and has been studied in detail, e.g., in (Mayo et al. 1995, von Heusinger and Schwarze 2006). An example is *sbonbinare*, from *bobina* 'spool'. It can be used transparently to mean 'pull wire from a spool' as in this example:

(3) *Un missile sbobina un filo* 'a missile unspools a wire' (Massari 2006)

Likewise, from *crema* 'cream' we can derive *scremare* 'to skim'; from *forno* 'oven' we get *sfornare* 'to take out of the oven'; *carta* 'paper' gives *scartare* 'unwrap' or 'remove from wrapping'. We can more or less get the compositional semantics of *sbobinare* using the word grammar shown below,

V → DPrefx Root
↑=↓ (↑Arg) = ↓

Root → N
↑=↓

s-, DPrefx (↑μ DPRED) = 'RemoveXfromY<-o,-r, (↑Arg)>
(↑μ CAT)= V
(↑μ CLASS)= -are

which unifies the constituents *s*- and *bobin(a)* to yield the following features at (local) f- and m-structure:

(↑μ DPRED) = 'RemoveXfromY<-o,-r, (↑Arg)>'

(ARG PRED)='Spool'
(↑μ CAT)= V
(↑μ CLASS)= -are

The base of the derivation, *bobin(a)*, has a PRED feature, which will be projected to a local f-structure so as to furnish the argument Arg, subcategorised by RemoveXfromY. The derivational morpheme, *s*-, has a special DPRED feature that is allowed to appear only in morphology. Unlike PREDs, DPREDs seem never take more than one argument, at least in West European languages, and there are reasons to think that this argument is too primitive to be considered a true grammatical function (e. g., its position in word structure is fixed; it is not assigned case; it has no anaphoric links). As suggested in section 1, Lexical Integrity implies a barrier between morphological constituents and sentence-global f- and m-structure, so that word-internal constituents do not project features directly to syntax but only as mediated by the morphological component. Morphology is non-monotonic insofar as it is permitted to perform operations like substituting the evaluation of a function — in this case, RemoveXfromY — for the function, while deleting the function's arguments. Hence, the DPRED can be deleted as well as the PRED feature of the base noun, 'Spool', and both will not appear in sentence-level f-structure. Instead, the interface projects a *new* PRED that is not a constituent of the word but is created by evaluation of the DPRED at run-time, and it creates a lexical semantic formula to which this new PRED refers.

Once the word grammar (via m-insertion) has eliminated the DPRED and its argument from *sbobin*- from the local f-structure, the only morphemic segment remaining is the verbal inflection, *-a.* It adds the features NUM, PERS, and TENSE monotonically to the newly derived stem, as shown below.

NUM=SG
PERS=3
TENSE=Pres
PRED=??? Lexical Semantics ???

Hence, inflection is relatively easy to handle. The task of m-insertion, on the other hand, is to obtain an f-structure PRED by substituting arguments to the DPRED's derivational function (RemoveXfromY<-o,-r,(↑Arg)>), so as to obtain a semantics, an argument structure, a lexical form, and other required attributes. A daunting job that, at present, has not been solved for the general case but can be solved for some specific but highly complex derivational patterns, like that of *sbobina*, as we shall see in the next section.

4 c-insertion in Detail: *sbobinare*

Figure 3 sketches how the derivation of a denominal verb like *sbobinare* takes place in the KLU program. At the time KLU was written, much less had been said about Lexical Mapping Theory and about conceptual unification than is now the case, but a genuine implementation of these functions would still be a large piece of work. In KLU they were implemented as very sketchy dummies. Nevertheless, my impression is that most of the pieces exist; someone with ample resources just needs to put them together.

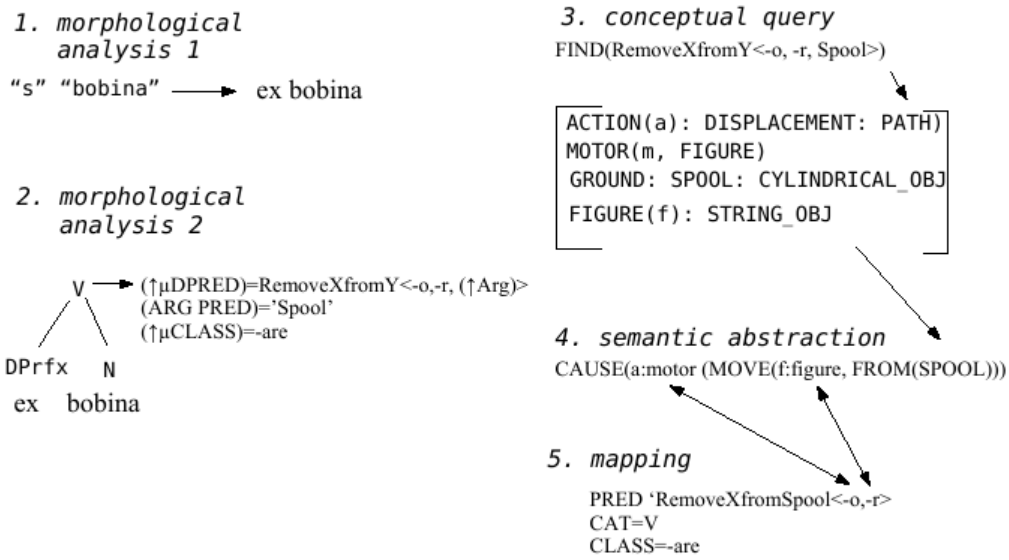


Figure 3. c-insertion Step-by-Step.

Apart from its job of filling the lexical buffer from the static lexicon and purging it, when it is called with an internally segmentable item, c-insertion can carry out the five steps **Morphological analysis 1 and 2, conceptual query, semantic abstraction** and **mapping**. For inflected stems, only Morphological analysis 1 and 2 are needed to add inflectional features to the stem. If a new stem needs to be derived, all steps must be carried out. At the end, m-insertion returns all required features, including a new lexical form and a semantic formula, to c-insertion, which copies the features relevant to syntax into the morphological buffer and finally returns them to the leaf node of c-structure that invoked c-insertion. Let me explain the steps one by one.

4.1 Segmentation (Morphological analyser 1)

To insert a nonce word to syntax, it is first necessary to obtain lexical features from each of the morphological constituents. This task, as has been mentioned, can be carried out by a conventional two-level analyser. KLU, however, was meant to explore ideas, not directly related to problems of computational morphology, about the overall structure of the mental lexicon. To this end, several two-level analysers were constructed in such a way as to mimic roughly the "domains" of lexical phonology, with separate analysers for the root and for derivational, inflectional, and clitic domains. (The nomenclature is a bit misleading because, during segmentation, nothing is known about the morphemic structure, but a rough correspondence exists, e. g., in the outer domain of the phonology we find mainly inflectional morpheme segments.This structure is probably reflected in the "continuation lexicons" of most two-level analysers). Segmentation thus enforces a number of lexical-phonological constraints on word structure, more or less prohibiting inflection inside derivation and the like, but it does not fetch or unify lexical features of the segments. This is the task of a separate word-grammar component, to which segmentation merely furnishes the input. Thus, segmentation and word grammar share the responsibility for enforcing the constraints on word structure that make it markedly different from sentence structure.

KLU's segmentation component uses hand-coded orthographic transducers to lemmatize words or parts of words ("surface strings") to strings found in the lexicon ("lexical strings"). Surface strings that are already present as lexical strings have precedence over strings that must be derived, long strings have precedence over short strings, and short derivations beat long derivations, roughly implementing the Panini or 'elsewhere' principle of lexical phonology (Kiparsky, 1982). In part 1 of Figure 3, "s" (an allomorph of "dis") is reduced to a lexical string, *ex*, while the surface string "bobina" remains the lexical string *bobina*. If a surface string cannot be lemmatized to any lexical string (neither as a full-form word nor as a known morpheme), segmentation fails. However, not knowing anything about true morphological structure, segmentation cannot distinguish between genuine derivational morphemes like *re-* in *retake* and pseudo-morphemes like the *re-* in words like *rejoice*. Since segmentation cannot distinguish pseudo from real morphemic segments, it is arranged that the segment list *re.joice* matches the lexical entry /re°joice/ immediately and blocks morphological decomposition.

A segment that can be found in the lexicon is not segmented further, or its segmentation is postponed. A motivation for this approach was the consideration that in speech or optical character recognition, the inputs can be many-ways ambiguous, and, computationally, morphological analysis is likely to be far more expensive than it is for computer-encoded input. It was supposed that limiting analysis to the word's periphery and giving the lexicon precedence over analysis might be nature's way of coping with this problem.

Unlike the well-known two-level transducer techniques, KLU works from both ends of the word toward the middle until it identifies a single root segment. Because affixes are separated without knowledge of the underlying morphology, segmentation can obtain misleading embeddings, as in *unhappier*. The suffix *-er* cannot be removed from *unhappy* because *unhappy* has three syllables, and *-er* attaches only to words of one or two syllables. This forces segmentation to first remove *un-* and then *-er* (both belonging to the derivational domain). This would give the segmental bracketing [un-[[happi]-er]], which would mean NOT(MORE(HAPPY)) instead of MORE(NOT(HAPPY)). Therefore segmentation returns only flattened, non-embedded lists of lexical strings to the word grammar (morphological analysis).

Compounds (as in German) are not accepted, as it was felt that they present a separate problem.

4.2 Word Grammar (Morphological analyser 2)

The input to the second stage of morphological analysis is the flattened (non-bracketed) list produced by segmentation, shown in part 2 of Figure 3. "ex bobina" looks like a tiny syntactic phrase, and the morphological parser looks like a miniature version of sentence analysis. It projects features of affixes and roots from

the lexicon to local f- and m-structures and unifies them. But there are important constraints. The phrase-structure rules must describe regular grammars and in general obey the principles outlined in (Bresnan and Mchombo 1995), although the compiler does not enforce most of these rules. But the range of computable forms is still immense.

As mentioned earlier, the resulting feature set is purely local, i.e., it does not unify immediately with sentence level f- or m-structure. If the resultant f-structure is well-formed (e.g., does not contain a DPRED), the results are deposited in the buffer and returned to the leaf node of c-structure that called c-insertion. If not, the following steps, 3 to 5, are taken to produce a new, derived stem. This stem can then be unified with any inflectional affixes, as if it had been drawn from the lexicon directly.

4.3 Conceptual query

After evaluation of the morphological structure, the DPRED's lexical form 'RemoveXfromY<-o,-r, (↑Arg)>' for *sbobinare* is presented as a query to a knowledge data base, illustrated by the function FIND in part 3 of Figure 3. Note that the argument list of RemoveXfromY does not assign thematic roles. The [-o] argument will probably be an agentive subject, but [-r] can be a theme or a localization. This leaves two interpretations open, one in which the spool (theme) is taken from something, and one in which something is taken from a spool. Hence, in this kind of derivation it sometimes appears that we get two readings back from the knowledge base. For example, in English *unhand* seems to mean 'take a hand away from something' or 'release the hand's grasp on something'. But such cases are fairly rare.

In a forthcoming article, von Heusinger and Schwarze (2006) show that the semantic ambiguity of Italian removal verbs must usually be resolved within the derivation, because it fails to carry over into the sentence semantics. In fact, derivational rules like those we see here usually produce predicate-argument structures that are very vague, such as 'something-typically-done-with-spaghetti' (for Italian *spaghettata*), but the meanings and argument structures that result tend to be very specific ('a meal with spaghetti'). This means that conceptual evaluation is an important part of derivation. In the case of *sbobinare*, we must expect the query function FIND to return something like the result shown in 3 of Figure 3.

4.4 Semantic abstraction

I assume that what the knowledge base returns is a purely conceptual, framelike structure. Using a logic of proto-roles, perhaps like that of Dowty (2001), it should be possible to obtain simplified semantic abstractions that can be the base for conventional mapping algorithms; cf. (Kelling 2001), which shows how this can be done for two classes of French nominalizations. For *sbobinare* the abstraction would produce a semantic formula like that of 4 in Figure 3, containing semantic relations and typed argument variables, some of which may be bound (as is the argument of FROM in this case, which is bound to the semantics of SPOOL).

Needless to say, the implementation of the required data base and abstraction rules would not be trivial, and most of the (very extensive) work in computational knowledge representation has been done without a clear idea of what outputs might be useful at the interface to lexical semantics and mapping. The KLU knowledge base was, of course, just a dummy that provided a few pre-arranged answers to conceptual queries.

4.5 Mapping

Mapping must create the lexical form (the PRED value) and its argument structure and establish the mappings from the lexical form's argument structure to the participants of the associated semantic representation, as shown by the arrows between 4 and 5 of Figure 3. In KLU the input to mapping is always a semantic structure, not a lexical form. However, an often-voiced opinion is that derived words inherit their argument structure directly from the argument structure of the base, not from the semantics. A point in favour of this

view is that nonsense words can be used as the bases of derivatives. Thus if I can *frobble* something, I can say that it is *frobblable*, without ever having found out what it is to *frobble*. It is conceivable, however, that the knowledge base has a default class of generic transitive actions and returns an abstract generic concept for *frobble* which could be the basis for mapping.

A study by Meinschaefer (to appear) shows that certain nominalizations from verbs derive their argument structures not from the argument structures of the base verb but from an underlying, semantic structure shared with the verb. Moreover, the often cited restrictions on passivization (e.g., *the hat fits you well* vs. **you are fitted well by the hat*) suggest that even in the very regular passive, more is involved than reorganizing the syntactically visible argument structure of the base. The task of formulating the semantic decompositions that the mapping algorithm from semantics will require is far from completion, as Levin and Rappaport Hovav (2005) point out. The value of attempting a computational implementation, however sketchy, along the lines indicated here is that it forces the contributing theories to pay attention to the entire gamut of interfaces involved.

In the KLU system, mapping was also made a catch-all for creating features like inflectional class and declension, aspect, gender, etc. Clearly some of these require access to information in the morphological analysis that gets lost in conceptual analysis. For example, some Italian diminutives take the grammatical gender of their bases, regardless of any physical gender properties of the base, but in other cases it is the affix which determines gender and inflectional class.

At this point we can account for what happens when the grammatical system encounters a spontaneous derivation like *sbobinare*. The syntactic parser does not concern itself with the word-internal structure but turns the job over to c-insertion. c-insertion can perform inflectional analysis in well-understood ways. What it cannot do is derivational analysis, but it contains a sub-function that can, at least in principle.

5 The Morphological Buffer

Now I turn to the last of the three components, the morphological buffer, the interface between syntax and morphology. Strictly speaking, the buffer is only a data structure within c-insertion, but c-insertion might not be necessary if there were no buffer to administer. To help explain why the buffer is there, let me again return to *sbobinare*, but in a different usage meaning 'transcribe' rather than 'pull from a spool'.

(4) La prego, mi dica: lo ha *sbobinato* soltanto, o lo *h*a scritto lei? (Scarpa 2004)
'I ask you, please, tell me: did you only *transcribe* it or did you write it [yourself]?'

Judging from my searches in Google, this is now by far the most common reading of *sbobinare*. It must have arisen at a time when wire or tape recorders were commonly used for taking dictation, and secretaries transferred the spoken texts to paper by 'pulling the speech' from the spool. At a time when the word was unlikely to be in the Italian mental lexicon, a derivation in the way described earlier would have been easy enough for most speakers because a typical action involving pulling something from a spool was transcription. Each time a speaker made up or heard this use of *sbobinare*, a lexical entry would appear in her morphological buffer, and the more often it happened, the longer it would stay there. The entry in the lexical buffer, however, is a pure Saussurian sign. It has no internal structure; it is simply a pairing of a surface form with a semantic item. At some point, it will be learned and used by other, younger speakers, and will become a part of the Italian lexicon. At the same time, wire recorders and tape recorders will give way to cassette recorders and MP3 players, and the conceptual connection between pulling something from a spool and transcription disappears. The lexical data repeatedly created in the lower shaded box of Figure 2 gradually move to the upper shaded box, the static lexicon. (Computationally, this is the same sort of process that takes place in memory management systems when chunks or pages of memory become increasingly 'non-purgable', i.e.,

permanent parts of the loaded system.) Young speakers who do not learn *sbobinare* as an opaque sign will not produce it in this sense spontaneously, and they will probably have difficulty understanding sentences like (3).

Thus the morphological buffer would seem to be the mechanism by which lexicalization takes place. Its psychological reality has been confirmed in many experiments, and buffers of this form are an indispensable data structure in virtually all large-scale computer programs. That it has not been a part of computational LFG is a historically curious accident.

Interestingly enough, in the early days of LFG it was recognized that even where derivation is relatively simple and systematic, as it is with the passive, it can be computationally very expensive. It was apparently assumed that something like a morphological buffer must exist to retain the results of this expensive computation when it was unavoidable. The "Introduction" to *The Mental Representation of Grammatical Relations* Bresnan and Kaplan wrote

> ...lexical computations are not required in generating sentences, since ... lexical rules, as long as they have a finite output, can always be interpreted as redundancy rules ... As such, the rules could be applied to enter new lexical forms into the mental lexicon, and the derived lexical forms could subsequently be retrieved for lexical insertion rather than being re-derived (Bresnan and Kaplan 1982, xxxiii).

The picture we how have of lexical rules and derivation is, if anything, only more complex that the transformational accounts of 25 years ago. All the more reason why we should expect the synchronic grammar to use a buffering mechanism to avoid expensive computations as much as possible, even if it has access to the mechanisms that produce and analyse new words. Bresnan and Kaplan did not identify the lexical buffer as an entity distinct from the lexicon itself, with its own storage-managing regime. This seems to have led to endless misunderstandings and to a wide-spread impression that LFG and similar unification-oriented models of grammar cannot give a formal account of spontaneous word formation. Across research traditions, lexicalism has unfortunately been identified with a conception of the lexicon as a static set of well-formed words that cannot participate in the creative, spontaneous introduction of new forms.

To be sure, some computational projects in the 1990s did in fact consider, but did not implement, what would have been a morphological buffer. The authors of the Alvey Natural Language Tools, for example, thought a word-formation cache would speed up processing, but "would be of little linguistic interest" (Ritchie et al. 1992, 177).

The moral of the story is this: Computational models are not just the servants of theory; they also strongly influence how we think about theory. We should bear in mind that formal descriptions of computational systems are simplifications of underlying reality. Computational systems are full of buffer-like structures, and cognitive psychology tells us that our own brains are, too. The buffer between syntax and morphology allows the cognitive grammar to avoid the enormous costs that would result from continuously recomputing all complex words, and there is evidence that even very frequent inflected forms are stored rather than computed (Baayen et al. 1997). However, the inherent, context-dependent flexibility of conceptual interpretation can cause frequently buffered words to lose their connection to their compositional semantics, leading to lexicalized forms that cannot be reconstructed easily once the conceptual context that engendered them is not generally available.

6 An Afterthought: Separable Prefixes in Derivation

The model I have described depends crucially on the one-to-one correspondence of c-structure nodes to morphologically complete words, as required by Lexical Integrity. On a word-and-paradigm view of inflectional morphology, syntactic paraphrases are part of the inflectional system, so that this correspondence is

not always given: the surface constituents realizing a cell of the paradigm lie under different c-structure nodes, in apparent violation of Lexical Integrity. The same can be found in derivation. In Germanic languages, the so-called particle or separable prefix verbs defy analysis in the model I have shown. Consider (5), a non-lexicalized but semantically transparent derivation.

(5) Max wird seinen Mitgleidsbeitrag fur den Alpenverein abwandern.
'Max will hike off his dues to the Alpine Association' (Stiebels 1996, 143)

In the framework I have described, we can account for this derivation by associating the particle *ab* with a DPRED Decrement<-o,-r, (↑Arg)>. This requires, however, that (↑Arg) be within the boundaries of the morphologically complete word given to c-insertion. This is not the case when the particle is separated, as in (6).

(6) Max wandert seinen Beitrag zum Alpenverein ab.
'Max is walking off his dues to the Alpine Club.'

A possible solution might be to loosen the barrier imposed by c-insertion in the following sense: Following the suggestion of Frank and Zaenen (2004), we assume a projection logic that carries purely morphological features beyond the lexicon-syntax barrier, but assembles them in a sentence level m-structure. This would let *ab* find its base *wandern* in f-structure by specifying the path to its argument with functional uncertainty, i.e.,writing (↑ X* Arg) insead of (↑ Arg) in the DPRED of *ab*. An unpleasant consequence of this strategy is that, after the derivation, the lexical form of the base must be replaced in f-structure by the newly derived lexical form, and the derivation itself, as we have seen, cannot be fully accomplished within the existing formal apparatus of LFG. A call to m-insertion, with affix and base, is necessary from some point above lexical insertion. What's worse, the replacement must happen prior to the tests for completeness and coherence, because *wandern*, which is syntactically like English *wander*, cannot take a direct object. Conceivably, the derivation could be implemented in an extension to the constraint tests.

On the positive side, the length of the path from the base to its affix might provide a measure of grammaticality. This is useful for the following reason: Distributionally, the particle is similar to an adjunct, but in an interesting study Jochen Zeller (2003) shows that the position of the derivational particle in German is actually more restricted than that of an adjunct. Where the first sentence is fully acceptable, the second is judged as marginal.

(7) Laut quietschte die Ziehharmonika 'Loudly screeched the accordion'

(8) ?*Auf schrie die Zieharmonika ' The accordion shrieked' (Zeller 2003, 188)

From a statistical study of grammaticality judgments Zeller concludes that the particle must "be strictly head-governed by the verb" (p. 199), while admitting that it's difficult to give a precise movement analysis that would predict the degree of ungrammaticality. It would be interesting to see if path length in f-structure might provide the required quantitative measure.

References

Baayen, Harald; Cristina Burani; and Robert Schreuder (1997). Effects of semantic markedness in the processing of regular nominal singulars and plurals in Italian. In *Yearbook of Morphology* 1997, pp. 13-33.

Baayen, R. Harald and Antoinette Renoulf (1996). Chronicling the times: Productive lexical innovation in an English newspaper. *Language*, 72:1, pp. 69–96.

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

Bresnan, Joan (1982). *Lexical Functional Syntax*. Malden, MA: Blackwell, 2001

Bresnan, Joan and Ronald M. Kaplan (1982). Introduction: Grammars as mental representations of language. In (Bresnan, ed., 1982) pp. xvii–lii.

Bresnan, Joan and Samuel A. Mchombo (1995). The lexical integrity principle. Evidence from Bantu. *Natural Language and Linguistic Theory* 13, pp. 181-254.

Briscoe, Ted and Ann Copestake (1999). Lexical Rules in Constraint-based Grammars. *Computational Linugistics* 25:5, pp. 487-526.

Börjars, Kersti, Nigel Vincent, and Carol Chapman (1996). Paradigms, periphrases and pronominal inflection: a feature-based account. In *Yearbook of Morphology*, pp. 155–180.

Corbin, Danielle (1990). *Associativité et stratification dans la represéntation des mots construits*. Contemporary Morphology. Berlin: de Gruyter.

Dowty, David (1991). Semantic Proto-roles and Argument Selection. *Language* 67:3, pp. 547-619.

Frank, Anette and Annie Zaenen (2004). Tense in LFG: Syntax and Morphology. In *Projecting Morphology*. Stanford: CSLI, pp. 23-65.

von Heusinger, Klaus and Christoph Schwarze (2006). Underspecification in the Semantics of Word-Formation. The Case of Denominal Verbs of Removal in Italian. *Linguistics* 44:6.

Karttunen, Lauri (2003). Computing with Realizational Morphology. In: Computational Linguistics and Intelligent Text Processing, Proceedings of the 4th International CICLing2003. Lecture Notes in Computer Science 2588. pp. 203-214. Berlin: Springer.

Kelling, Carmen (2001). Agentivity and suffix selection. In: Butt, Miriam and Tracey Holloway King, eds., Proceedings of LFG'01. University of Hong Kong. Stanford: CLSI, 147-162 (http://csli-publications.stanford.edu).

Kiparsky, Paul (1982). Word-formation and the lexicon. In: Frances Ingemann, editor. 1982 Mid-America Linguistics Conference, number 1982, pp. 3-29. Lawrence, KS, 1982. Linguistics Dept., Univ. of Kansas.

Koenig, Jean-Pierre and Anthony Davis (2006). The KEY to lexical semantic representations. *Journal of Linguistics*. 42, pp. 71-108.

Koenig, Jean-Pierre and Daniel Jurafsky (1994). Type underspecification and on-line type construction in the lexicon. In Raul Aranovich, Willian Byrne, Susanne Preuss, and Martha Senturia, editors, Thir-

teenth West-Coast Conference on Formal Linguistics, number 1994, pp. 270–285, University of Califorina at San Diego. Stanford: CLSI.

Levin, Beth and Malka Rappaport Hovav (2005). *Argument Realization*. Cambridge: Cambridge University Press.

Lieber, Rochelle (2004). *Morphology and Lexical Semantics*. Cambridge: Cambridge University Press.

Marslen-Wilson, William and Lorraine Komisarjevksy Tyler, Rachelle Waksler, and Lianne Older (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, 101:1, pp. 3–33.

Massari, Alessandro (2006). Il Great Barracuda. http://www.seaspin.com/magazine/articolo.php?idart=42

Mayo, Bruce (2000). A Computational Model of Derivational Morphology. Dissertation, University of Hamburg,. http://deposit.ddb.de/cgi-bin/dokserv?idn=961769629

Mayo, Bruce, Marie-Theres Schepping, Christoph Schwarze, and Angela Zaffanella (1995). Semantics in the derivational morphology of Italian: Implications for the structure of the lexicon. *Linguistics*, 33, pp. 883–938.

Meinschaefer, Judith (to appear). The syntax and argument structure of deverbal nouns from the point of view of a theory of argument linking. In: Dal, G.; Miller, P.; Tovena, L.; Van de Velde, D. (eds.) *Deverbal nouns*. Amsterdam: Benjamins.

Ritchie, Graeme D. and Graham J. Russell, Alan W. Black, and Stephen G. Pulman, editors (1992). *Computational Morphology: Practical Mechanisms for the English Lexicon*. ACL-MIT Press Series in Natural Language Processing. Cambridge, MA: MIT Press.

Scarpa, Tiziano (2004). Posted in vasicomunicanti on April 19th, 2004. http://www.nazioneindiana.com/2004/04/19/i-narrificatori/

Stiebels, Barbara (1996). *Lexikalische Argumente und Adjunkte*. studia grammatica 39. Berlin: Akademie-Verlag.

Zeller, Jochen (2003). Moved preverbs in German: Displaced or misplaced? In *Yearbook of Morphology* 2003, pp. 179-212.

Univ. Konstanz until 1998. Author's current e-mail address: bruce.mayo@arcor.de

A TREATMENT OF WELSH INITIAL MUTATION

Ingo Mittendorf and Louisa Sadler
University of Essex

Proceedings of the LFG06 Conference

Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006
CSLI Publications
http://www-csli.stanford.edu/

Abstract

A peculiarity of Welsh and the other Celtic languages is their system of Initial Mutations. These are regular alternations of word-initial phonemes triggered by a variety of lexical and syntactic triggering contexts. This feature of the Celtic languages poses a number of challenges to grammatical description, not least because it requires direct reference to adjacency relations in the linear string. We describe here an approach which covers the full range of mutation processes and their distribution in Welsh using the XLE grammar development environment and the associated finite state and tokenisation tools (Crouch et al., 2006).

1 Introduction

A peculiarity of Welsh and the other Celtic languages is their system of Initial Mutations. These are regular alternations of word-initial phonemes triggered by a variety of lexical and syntactic triggering contexts. This feature of the Celtic languages poses a number of challenges to grammatical description, not least because it requires direct reference to adjacency relations in the linear string. We describe here an approach which covers the full range of mutation processes and their distribution in Welsh using the XLE grammar development environment and the associated finite state and tokenisations tools (Crouch et al., 2006).

The rest of this paper is structured as follows. Section 2 provides some basic background on the system of initial mutations and a brief introduction to the types of conditioning environments. In section 3 we present a word-and-paradigm based view of initial mutation as a morphosyntactic phenomenon, a view which underlies our morphological approach. Following this, section 4 shows how a multiword transducer is defined to account for the distribution of initial mutations determined by specific lexical items. Section 5 then turns to cases of syntactically conditioned mutation, and outlines the c-structure approach to this phenomenon.

2 Initial mutations

A peculiarity of Welsh and the other Celtic languages is their system of Initial Mutations. These are regular alternations of word-initial phonemes such that, under the appropriate circumstances, a word like Welsh *tad* ‘father’ appears as *dad*, *thad* or *nhad*. (1) shows the possible range of alternations in initial consonant phonemes. These alternations can be arranged into different sets, for which the traditional terms are Radical (the citation form), Soft Mutation, Nasal Mutation and Aspirate Mutation.

(1) Welsh Initial Consonant Mutations

Radical	p	t	c /k/	b	d	g	m	ll /ɬ/	rh /r̥ʰ/
Soft Mut	b	d	g	f /v/	dd /ð/	Ø	f /v/	l	r
Nas Mut	mh /m̥ʰ/	nh /n̥ʰ/	ngh /ŋ̊ʰ/	m	n	ng /ŋ/			
Asp Mut	ph /f/	th /θ/	ch /χ/						

As can be seen, there is a wider range of alternations if the initial phoneme is a voiceless consonant (/p/, /t/, /k/) than if it is a voiced or other consonant. Consonants not listed (/n/, /s/, /f/, etc.) show no alternations. Basically two different types of environment can be distinguished in which these mutation forms appear. First, initial mutation can be *triggered* by a range of lexical items including proclitic pronouns, prepositions, determiners and numerals. Each trigger is followed by a specific, lexically determined mutation. The *target* of these mutation triggers, that is, the word that shows the requisite initial mutation, is the word directly following the trigger: a lexical mutation trigger and its mutation target are always adjacent. This means that the target is to some degree unpredictable. For example, in (2) the clitic pronoun *fy* 'my' in (2) triggers NM; in (2a) the target of this mutation is the noun *diddordebau* 'interests'; in (2b), the pre-nominal adjective *prif* 'main'; and in (2c), the numeral *tri* 'three'.[1] The pre-nominal adjective and the numeral in turn trigger their own mutations, SM and AM respectively.

(2) a. *fy niddordebau*
(fy) (NM.diddordebau)
my interests

b. *fy mhrif ddiddordebau*
(fy) (NM.prif) (SM.diddordebau)
my main interests

c. *fy nhri phrif ddiddordeb*
(fy) (NM.tri) (AM.prif) (SM.diddordeb)
my three main interest(s)

There is no connection between the category of the trigger and the triggered mutation: Different prepositions trigger different mutations; different clitic pronouns also trigger different mutations; and so on. (3a) shows the 1SG clitic *fy* 'my' triggering NM; (3b) the 3SG MASC clitic *ei* 'his' triggering SM; and (3c) the 3SG F clitic *ei* 'her' triggering AM.[2] As the

[1]Cardinal numerals are followed by the singular form of nouns in Welsh. This has no bearing on the issue here.

[2]The analysis of *ei thad* in (3c) has been slightly simplified at this point. For a more accurate analysis see (11 b) .

last two examples with the homophonous triggers *ei* ‘his’ and *ei* ‘her’ also illustrate, there is no connection between the phonological makeup of the trigger and the triggered mutation. Initial mutation is not a sandhi-phenomenon.

(3) a. *fy nhad*
(fy) (NM.tad)
my father

b. *ei dad*
(ei) (SM.tad)
his father

c. *ei thad*
(ei) (AM.tad)
her father

Second, initial mutations can be syntactically conditioned, that is, triggered by a syntactic environment. For example, attributive APs, which by default appear in post-nominal position, are subject to Soft Mutation if the head noun is FEM SG; otherwise (with MASC SG nouns or PL nouns of either gender, MASC or FEM), the AP appears in the radical form; cf. (4).

In such syntactic environments it is the first word in the relevant domain which is subject to mutation. In attributive APs this will usually be the adjective, but if the adjective is preceded by an adverb, it will be the adverb; cf. (6). (The adverb in turn triggers its own mutation.) A comparison between the examples in (6) incidentally shows that it would be wrong to view soft-mutated *bwysig* as an (attributive) FEM SG form of the adjective *pwysig*.

(4) *ci mawr* — *cath fawr*
(ci) (RAD.mawr) — (cath) (SM.mawr)
dog.M.SG big — cat.F.SG big

(5) *cath ddu fawr*
(cath) (SM.du) (SM.mawr)
cat.F.SG black big

(6) *agwedd bwysig* — *agwedd dra phwysig*
(agwedd) (SM.pwysig) — (agwedd) (SM.tra) (AM.pwysig)
aspect.F.SG important — aspect.F.SG very important
‘(an) important aspect’ — ‘(a) very important aspect’

Attributive AP mutation illustrates why syntactically conditioned mutation should be distinguished from lexically conditioned mutation. Attributive AP mutation is not subject to lexical idiosyncracy. Moreover, as (4c) illustrates, when the FEM SG noun is followed by two APs, each of these is subject to SM independently, and furthermore the trigger (noun) and target (second AP) are not adjacent, which is uncharacteristic of lexical mutation triggers and targets. Note also that lexical triggers are always followed by a target, whereas, of course, attributive APs are optional so that a FEM SG noun is only a trigger when a post-nominal AP is present.

3 Regular Mutation Paradigms

There are a number of different (and sometimes partial) approaches to Celtic initial mutations in the theoretical literature. In some analyses, initial mutations are viewed essentially as phonological processes triggered by syntactic environments, in a framework in which a direct interface between syntax and phonology is assumed (Ball and Müller, 1992). Our approach takes the alternative view that initial mutation is close to inflection in nature and is essentially a morphosyntactic phenomenon. Our approach has much in common with the view of initial mutation in the Goidelic languages proposed in Green (2003) and Stewart (1992): for detailed discussion of this position and criticisms of the phonological view, see those references.

(1) gave an overview over the possible mutation *forms*. These forms, however, cannot simply be equated with what could be call mutation *functions* or *mutation states*. Mutation forms are the morphological exponents of mutation functions with their different values. We assume that each word has a mutation paradigm with different cells filled with the possible mutation forms. There is not necessarily a one-to-one relationship between forms and functions: the paradigmatic nature of mutation forms establishes the different values of the mutation functions.

We illustrate this with a close look at AM in (1). Special AM forms exist for words with an initial voiceless consonant. There are no special forms beginning with other phonemes. This does not of course mean that words with a non-voiceless-stop initial are barred from those syntactic environments where the AM form of voiceless-stop initials is called for. Rather, what happens in such cases is that the radical form "stands in" for the non-existent discrete AM form (the radical is thus the morphological default).

Whatever applies to Aspirate Mutation also applies to Nasal Mutation: here, words with initial /m/, /ɬ/ <ll> and /r̥ʰ/ <rh> have no discrete forms; again the radical stands in. And with words which start with a "non-mutatable" phoneme such as /s/, the radical appears in all mutation environments. A first version of a mutation paradigm could therefore look as in (7).

(7)

	Vl stops			**Vd stops**			**m**	**ll / rh**		**Other C**
Rad	p-	t-	c-	b-	d-	g-	m-	ll-	rh-	s- *etc.*
AM	ph-	th-	ch-	b-	d-	g-	m-	ll-	rh-	s-
SM	b-	d-	g-	f-	dd-	Ø-	f-	l-	r-	s-
NM	mh-	nh-	ngh-	m-	n-	ng-	m-	ll-	rh-	s-

We now turn to the question of the number of mutation functions or states, which we have so far taken to be four (as in (7)), and show that the picture is actually slightly more complicated. There are mutation environments which straightforwardly require the set of Soft Mutation forms, or the Aspirate Mutation set. But in other mutation environments a mixture of such forms appears. First, there are environments which select only a subset of the SM forms. In these environments initial voiceless and voiced stops and /m/ undergo

SM, but if the initial phoneme is <ll> /ɬ/ or <rh> /r̥ʰ/, the radical form is required. This "Restricted Soft Mutation" (SMR) applies, for example, to FEM SG nouns following the definite article; cf. (8).

(8) *y gath* / *faner* / *ferch* / *llinell* / *rhwyd*
(y) (SMR.cath) / (SMR.baner) / (SMR.merch) / (SMR.llinell) / (SMR.rhwyd)
the cat / flag / girl / line / net

Second, a further group of triggers (negation particles mostly) is followed by AM forms if the initial phoneme is a voiceless stop, but by SM forms, if available, otherwise; this mutation is usually called Mixed Mutation (MM); cf (9).

(9) *ni chanodd* / *ddaeth* / *fudodd* / *lwyddodd* / *redodd*
(ni) (MM.canodd) / (MM.daeth) / (MM.mudodd) / (MM.llwyddodd) / (MM.rhedodd)
not sang / came / moved / succeeded / ran

These cases motivate distinguishing two further mutation states or functions (10).

(10)

	Vl stops			Vd stops			m	ll / rh		Other C
Rad	p-	t-	c-	b-	d-	g-	m-	ll-	rh-	s- *etc.*
AM	ph-	th-	ch-	b-	d-	g-	m-	ll-	rh-	s-
MM	ph-	th-	ch-	f-	dd-	Ø-	f-	l-	r-	s-
SM	b-	d-	g-	f-	dd-	Ø-	f-	l-	r-	s-
SMR	b-	d-	g-	f-	dd-	Ø-	f-	ll-	rh-	s-
NM	mh-	nh-	ngh-	m-	n-	ng-	m-	ll-	rh-	s-

A further complication is introduced when we consider the form of words with an initial vowel, which sometimes occur with an initial /h/. This prevocalic aspiration appears with some (but not all) mutation triggers which require the radical or AM on consonants. If these vocalic alternations are taken to be part of the mutation system, two additional mutation functions must be assumed, RAD-H and AM-H, which differ from plain RAD and AM only where words with a vocalic initial are concerned. The examples in (11) contrast plain AM with AM-H, and those in (12) plain RAD with RAD-H.

(11) a. *tri chi* / *tri afal*
(tri) (AM.ci) / (tri) (AM.afal)
three dog(s) / three apple(s)

b. *ei chi* / *ei hafal*
(ei) (AM-H.ci) / (ei) (AM-H.afal)
her dog / her apple

(12) a. *eich* *ci* — *eich* *afal*
(eich) (RAD.ci) — (eich) (RAD.afal)
your.PL dog — your.PL apple

b. *eu* *ci* — *eu* *hafal*
(eu) (RAD-H.ci) — (eu) (RAD-H.afal)
their dog — their apple

The inclusion of prevocalic aspiration in the system of Welsh initial phoneme alternations is, incidentally, the reason why we use the term 'Initial Mutation' for this phenomenon, and not 'Initial Consonant Mutation' that can often be found instead.[3]

The paradigmatic distribution of mutation forms, including pre-vocalic aspiration, leads us to assume the following system of mutation functions and regular mutation forms as their morphological exponents:[4]

(13)

	Vl stops			**Vd stops**			**m**	**ll / rh**		**Other C**	**V**
Rad	p-	t-	c-	b-	d-	g-	m-	ll-	rh-	C-	V-
Rad-H	p-	t-	c-	b-	d-	g-	m-	ll-	rh-	C-	hV-
AM-H	ph-	th-	ch-	b-	d-	g-	m-	ll-	rh-	C-	hV-
AM	ph-	th-	ch-	b-	d-	g-	m-	ll-	rh-	C-	V-
MM	ph-	th-	ch-	f-	dd-	Ø-	f-	l-	r-	C-	d_V-
SM	b-	d-	g-	f-	dd-	Ø-	f-	l-	r-	C-	V-
SMR	b-	d-	g-	f-	dd-	Ø-	f-	ll-	rh-	C-	V-
NM	mh-	nh-	ngh-	m-	n-	ng-	m-	ll-	rh-	C-	V-

A final complication, which is simple to deal with, is that there are some lexical exceptions to the regular patterns shown in (13). These irregularities mostly concern the fact that some words do not have a discrete SM form and use the radical form instead. Among these is a group of recent English loanwords with initial *g-* such as *gêm* 'game'. In the regular case, words whose initial phoneme is /g/ have SM forms where this phoneme is missing (*gardd* ∼ *ardd* 'garden'). Words like *gêm* are exceptional in that radical and SM forms are identical (*gêm*). Another word with a rather unusual mutation paradigm is the interrogative *ble* 'where'. The table in (14) shows a comparison between the regular and the irregular paradigms of *gardd* and *gêm*, and *bardd* 'bard' and *ble*, respectively. (Because of the nature of its triggers – negation particles – MM only applies to verbs and has been omitted here; only some mutations apply to *ble*.) Such irregularities (and others) are no problem for our word-and-paradigm based approach.

[3] A different approach to prevocalic aspiration would be to view prefixation of /h/ as a kind of liaison phenomenon: /h/ in fact would "belong" to the preceding word; if the following word starts with a consonant, then /h/ is lost; if, on the other hand, it starts with a vowel, it moves to this word. We do not discuss this alternative further here.

[4] There is a further minor complication concerning vowel initial words undergoing the mixed mutation MM: for reasons of space we suppress discussion of the details of this matter here.

(14)

	gardd	*gêm*	*bardd*	*ble*
Rad	gardd	gêm	bardd	ble
Rad-H	gardd	gêm	(bardd)	*n/a*
AM-H	gardd	gêm	(bardd)	*n/a*
AM	gardd	gêm	bardd	ble
SM	ardd	gêm [!]	fardd	ble [!]
SMR	ardd	gêm [!]	(fardd)	*n/a*
NM	ngardd	ngêm	mardd	mhle [!]

4 Lexical Mutations: The Multiword Transducer

The basic challenge in providing a treatment of lexically conditioned mutations is that the triggering relation is adjacent in the linear string, rather than any more abstract syntactic relation. Within XLE (Crouch et al., 2006) access to the linear relation between strings is possible using a user-defined `MULTIWORD` transducer within the `MORPHOLOGY` component. For those not familiar with XLE, we first give a brief overview of XLE's architecture, before describing our approach using the multiword transducer.

4.1 XLE

An XLE grammar contains a number of different sections, including:

- a RULES section that contains phrase structure rules that are functionally annotated;
- a LEXICON section that lists lexical entries with their c-structure categories and associated constraints;
- and a MORPHOLOGY section that specifies the transducers (finite state or other) used for morphological analysis (in the wider sense of the word).

The main components of interest for the treatment of initial mutations are the MORPHOLOGY section, the LEXICON section in which the tags used in the morphological analysis are mapped to syntactic terminals, and the sublexical rules, from the RULES section, which interface the morphological analysis with the syntactic terminals.

Within the MORPHOLOGY section, a string passes through several sequenced components (here described from the persective of analysis but fully reversible):

- The TOKENIZE section specifies the transducer whose main task it is to break up a parse string into individual words (or, properly speaking, tokens). (This is also the place to deal with sandhi phenomena.)
- The next section lists the transducers that are used for the morphological analysis proper and that map surface strings on to lexical strings and vice versa, i.e. they

pair tokens with morphological analyses which consist of a stem (usually the citation form is chosen) and a number of tags that encode any morpho-syntactically relevant information about a lexical item. In morphological analysis tokens are analysed individually one after the other, and thus at this stage adjacent tokens are not accessible to each other.

– The third section makes provision for the processing of multiword units. At this stage, adjacent tokens are once again accessible, and morphological analyses of individual words are concatenated. Multiword expressions may be built from the morphology and the lexicon via a built-in transducer and, if desired, marked with a special tag. In addition a user-defined multiword transducer can be specified to manipulate the concatenated string.

The morphologically analysed parse string is then passed to the grammar proper where it receives its c- and f-structure analysis. The grammar LEXICON lists all the stems[5] and tags used in the morphological analysis. The tag entries take the form of ordinary lexical entries, that is, in an XLE lexicon they consist of the stem (= the tag), a category label, a morphcode signalling whether to use the output of the morphological analyser (*XLE*, always required for tags) or not (*), and associated constraints if any. Some simple (and simplified) examples, stem and tags for *cath* 'cat' specifying that this is a FEM SG noun, are given in (16),[6] based on the morphological analysis in (15).

(15)

Surface	**Lexical**
cath	`cath +Noun +F +Sg`

(16)

```
cath    N      XLE   (^ PRED)='cath'.
+Noun   NSFX   XLE.
+F      DGEND  XLE   (^ GEND)=fem.
+Sg     DNBR   XLE   (^ NUM)=sg.
```

Finally, (sublexical) c-structure rules (Kaplan et al., 2004) describe the possible constituents (stem and tags sequence) of a category, N in the case of *cath*. (All sublexical constituents are appended with _BASE in these rules.) The sublexical rule for a noun is given in (17), again somewhat simplified.[7]

(17)

N −− >	N_BASE	*cath*
	NSFX_BASE	*+Noun*
	DGEND_BASE	*+F*
	DNBR_BASE.	*+Sg*

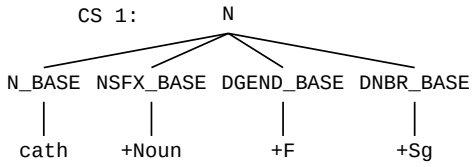
[5]It is actually not necessary to list every single lexical item/stem in the XLE lexicon. XLE allows the use of blanket entries, which is especially useful for open word classes such as nouns, adjectives, etc (Kaplan et al., 2004).

[6]^ equals ↑ in XLE notation, and ! equals ↓. It is ParGram policy to use lower case for atomic values (Butt et al., 2002).

[7]This rule also shows an XLE simplification of the usual LFG functional annotations in that a c-structure category without annotations is understood to be annotated with ↑=↓. (Some restrictions apply.)

Given as input the morphological analysis shown in (15), the tag entries in (16) and the (sublexical) rule in (17), XLE produces the c-structure in (18) and the f-structure in (19) for the string *cath*.

(18)



(19) `"cath"`

```
  ⎡PRED 'cath'     ⎤
1 ⎣GEND fem, NUM sg⎦
```

4.2 The Multiword Transducer

Before we can outline how our multiword transducer works and fits into the XLE architecture described in the previous section, and how we use it to deal with lexical mutations, we need to look at the (now "real" and unsimplified) morphological analyses of a lexical mutation trigger (the personal pronoun clitic 3SG MASC *ei* 'his') and of a mutation target (the FEM SG noun *cath* 'cat'). Both are shown in (20). We give two different mutation forms for the noun so that we can examine both a grammatical and an ungrammatical construction below.

(20)

Surface	**Lexical**
ei	`+Rad+ ei +Pron +Pers +Proclit +3Sg +M` **`+SM+`**
cath	**`+Rad+`** `cath +Noun +F +Sg`
gath	**`+SM+`** `cath +Noun +F +Sg`

The very last tag in the morphological analysis of the mutation trigger *ei* is a tag that encodes the initial mutation that this trigger governs, that is, the mutation state that the target must be in. For *ei* this would be Soft Mutation (`+SM+`, boxed in the example).

The very first tag in the morphological analysis of each and every word is a tag that encodes the mutation state of this word. One possible analysis for the mutation form *cath* would be Radical (`+Rad+`, boxed); one possible analysis for the mutation form *gath* would be Soft Mutation (`+SM+`, boxed). The mutation trigger *ei* also starts with such a tag, but this is immaterial in this context. Please note that we use the same set of tags for "mutation state" and "mutation governed", this difference being reflected solely in terms of position in the lexical string (start/end).

After each word (or rather token) has been morphologically analysed, XLE concatenates all the morphological analyses of the parse string. Given the two mutation forms of *cath* listed in (20), we might arrive at the two concatenated strings shown in (21).

(21) +Rad+ ei +Pron ... | +SM+ +Rad+ | cath +Noun ...
+Rad+ ei +Pron ... | +SM+ +SM+ | cath +Noun ...

As can be seen, the final mutation tag of the trigger *ei*, which constrains the mutation state of the target, and the initial mutation tag of the target encoding its mutation state are now adjacent. And only in the second of these concatenated strings do the two tags match, whereas in the first they differ. This first string (representing **ei cath*) is, in fact, ungrammatical because the target *cath* would show the wrong mutation.

It is at this point that our multiword transducer comes into action, checking that lexical mutation requirements have been satisfied by performing a test that checks whether the two mutation tags do in fact match.

There are several ways in which this test for matching mutation tags can be implemented. The way the test is performed now consists of two separate replacement operations. (The reason for keeping these two operations separate will become clear below in section 4.3):

1. Renaming the first of two matching mutation tags (the one originating from the mutation trigger): `+SM+` is replaced with `+T-SM`, `+Rad+` with `+T-Rad`, etc.[8] (22 b) shows a couple of lines from this section of the transducer.
2. Deleting the second of two matching mutation tags (the tag originating from the mutation target). The reason for this deletion has to do with syntactic mutations, which we will examine below in section 5. A couple of lines from this section of the transducer are shown in (22 a).

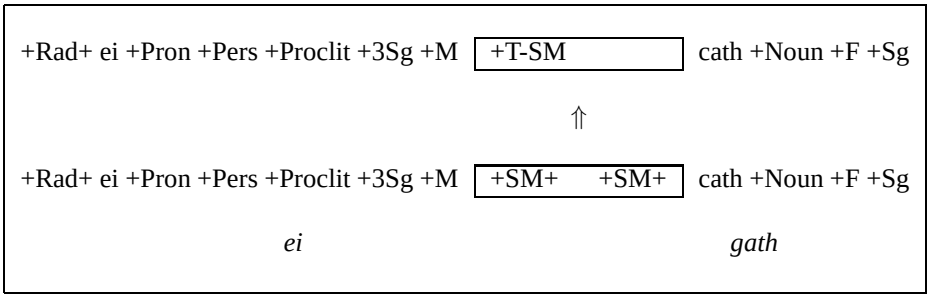
(22) a.
```
[..]  <- "+Rad+" || "+T-Rad" _
.o.
[..]  <- "+SM+" || "+T-SM" _
.o.
...
```

b.
```
"+T-Rad" <- "+Rad+" || _ "+Rad+"
.o.
"+T-SM" <- "+SM+" || _ "+SM+"
.o.
...
```

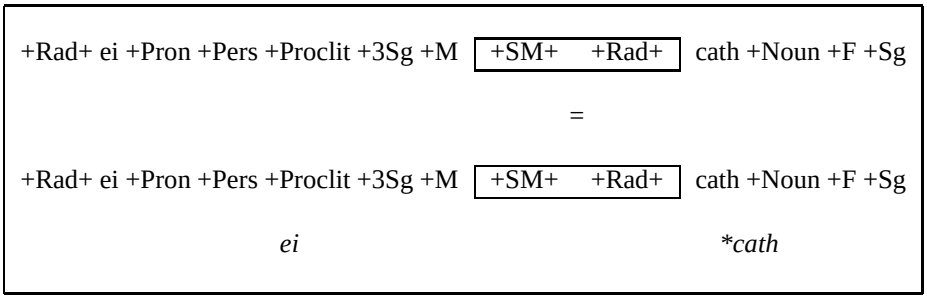
(23) shows the successful transformation of the concatenated grammatical string. If the two mutation tags do not match, no replacement takes place – and the test has failed; see (24).

[8] Instead of different replacement tags, one single blanket tag could be used instead (`+MutOK`, for instance). But because of the second replacement below this would mean that no indication whatsoever of the specific mutation triggered would remain in the Grammar proper. At least as a check some record should survive, if only at the sublexical level.

(23)



(24)



The final component of the treatment of Welsh lexical initial mutations involves the sublexical rules for mutation triggers in the grammar.

From the above it should have become clear that – provided the multiword test was successful – the lexical string of a mutation trigger will end in a renamed mutation tag (`+T-Rad`, `+T-SM`, etc.) instead of the mutation tag of its original morphological analysis (`+Rad+`, `+SM+`, etc.). These renamed tags have lexical entries in our grammar as shown in (25). Their (sublexical) category is `TMUT(_BASE)`, and there are no other tags of this category. There are no further constraints associated with them.

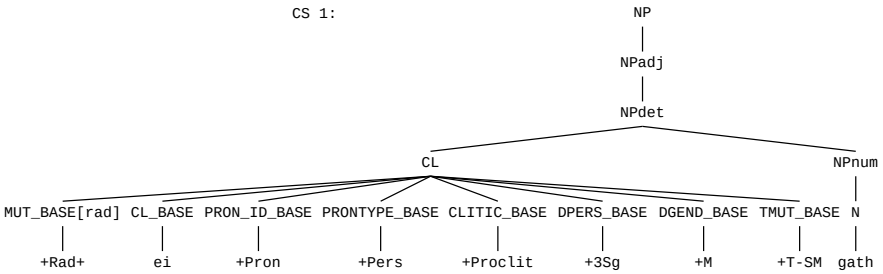
(25)
```
+T-Rad  TMUT  XLE.
+T-SM   TMUT  XLE.
...
```

If we now include this category in the sublexical rules for lexical mutation triggers, we have ensured that these are in fact followed by the correct mutation forms. (26) shows the sublexical rule for pronoun clitics like *ei* ‘his’ and (27) shows the c-structure for *ei gath*, with morphemes shown (split in two parts because of the size of the tree).

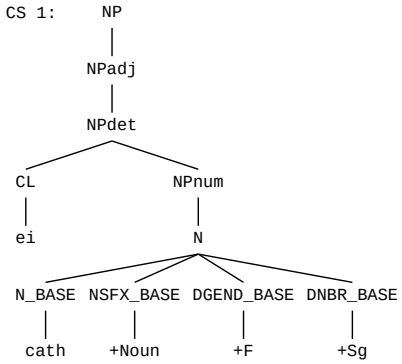
(26) CL -->

(MUT_BASE[rad])	*+Rad+*
CL_BASE	*ei*
PRON_ID_BASE	*+Pron*
PRONTYPE_BASE	*+Pers*
CLITIC_BASE	*+Proclit*
{ DPERS_BASE | CPERS_BASE }	*+3Sg*
(DGEND_BASE: (^ INDEX)=!)	*+M*
TMUT_BASE	*+T-SM*
MWE_BASE*.	*+MWE*

(27) a.



b.



4.3 Multiword Mutation Triggers

In the example above, the lexical mutation trigger was a single word. There are, however, some mutation triggers which are themselves multiword expressions (MWE) in our grammar. This introduces a slight complication into our own multiword transducer and makes necessary a minor adjustment.

One such multiword mutation trigger is the preposition *ar gyfer* 'for', which is followed by the radical of the mutation target. An example is given in (28). The morphological analyses for *ar gyfer* and the noun *cath* are shown in (29). Note the final +Rad+ tag in the analysis for *ar gyfer* (boxed) that specifies the mutation governed by this preposition.

(28) *ar gyfer cath*
for cat
'for a cat'

(29)

Surface	**Lexical**
ar gyfer	+Rad+ ar% gyfer +Prep +Nom +Rad+
cath	+Rad+ cath +Noun +F +Sg

In the MORPHOLOGY section we specify that multiword expressions should be built from the Morphology (and the Lexicon) and should receive the tag `+MWE` (30).[9]

(30)
```
BuildMultiwordsFromMorphology:
      Tag = +MWE
```

XLE will then attach this tag to the multiword analysis of *ar gyfer*:

(31)

Surface	**Lexical**
ar gyfer	+Rad+ ar% gyfer +Prep +Nom +Rad+ +MWE
cath	+Rad+ cath +Noun +F +Sg

This tag will be attached *before* the multiword transducer for lexical mutation checking comes into operation, that is, the architecture is as in (32).

(32)
```
welsh-multiword.fst
        .o.
  BuildMultiwords
        .o.
   MORPHOLOGY
```

The mutations transducer should then work with the output of `BuildMultiwords-FromMorphology` and has to accommodate one (or several) `+MWE` tags across which the check should be performed. (33) shows the modified version of the transducer. Note the addition of `["+MWE"]*` (= any number of `+MWE` tags including none) in the replacement context vis-à-vis the version in (22) without it.

(33)
```
[..]  <- "+Rad+" || "+T-Rad" ["+MWE"]* _
.o.
[..]  <- "+SM+" || "+T-SM" ["+MWE"]* _
```

[9]The purpose of this tag is to make it possible in the grammar proper to give MWEs preferential treatment over single word analyses via so-called OT marks (Frank et al., 2001). If the constraints associated with the tag `+MWE` include the appropriate OT mark, non-MWE analyses will be dispreferred.

```
.o.
...
"+T-Rad" <- "+Rad+" || _ ["+MWE"]* "+Rad+"
.o.
"+T-SM" <- "+SM+" || _ ["+MWE"]* "+SM+"
.o.
...
```

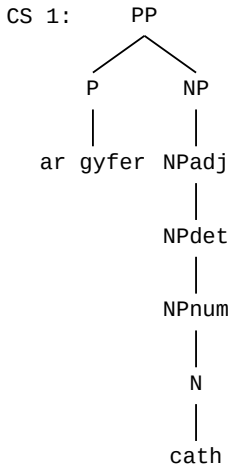
The presence of `["+MWE"]*` in these rules is the reason why there need to be two separate replacement operation sections in the multiword transducer, one for the first of the two matching tags and one for the second. These rules must work around `["+MWE"]*` (= any number of `+MWE` tags). They cannot be executed in one go because they would then have to include `["+MWE"]*` – which would imply replacing *any number of* `"+MWE"` with *any number of* `"+MWE"`, a result that we very definitely do not want.

(34) shows the successful transformation of the concatenated morphological analyses for *ar gyfer cath*, with the corresponding c-structure and sublexical analysis shown in (35) and (36).

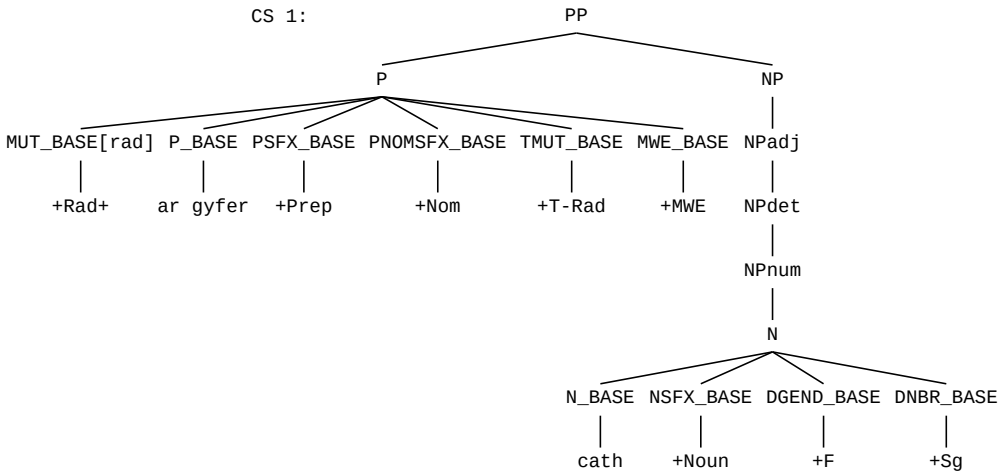
(34)

+Rad+ ar% gyfer +Prep +Nom	+T-Rad +MWE	cath +Noun +F +Sg
	⇑	
+Rad+ ar% gyfer +Prep +Nom	+Rad+ +MWE +Rad+	cath +Noun +F +Sg
ar gyfer		*cath*

(35)



(36)



5 Syntactic mutations

Recall that in addition to lexically triggered initial mutations, there are mutations which are triggered by syntactic environments. These two types of initial mutation have in common the fact that the exact mutation target is not predictable. Syntactic mutations apply to the first word in the relevant environment.

5.1 Syntactic Mutations as Categories

A (comparatively simple) example of a syntactic mutation is that governing (post-nominal) attributive APs, given above in (4)-(6) and repeated here as (37)-(39). The first word in a post-nominal AP appears in the Radical if the head noun is PL or MASC SG, but is soft-mutated if the head noun is FEM SG (37). All post-nominal APs are subject to this syntactic mutation (38). The mutation applies to the entire AP (i.e., the first word in the AP), not specifically to the adjective (see (39)).

(37) *ci* *mawr* *cath* *fawr*
(ci) (RAD.mawr) (cath) (SM.mawr)
dog.M.SG big cat.F.SG big

(38) *cath* *ddu* *fawr*
(cath) (SM.du) (SM.mawr)
cat.F.SG black big

(39) *agwedd* *bwysig* *agwedd* *dra* *phwysig*
(agwedd) (SM.pwysig) (agwedd) (SM.tra) (AM.pwysig)
aspect.F.SG important aspect.F.SG very important
‘(an) important aspect’ ‘(a) very important aspect’

(40) shows possible analyses for the two mutation forms *mawr* and *fawr* of the adjective 'big'. As with all morphological analyses, these start with a tag encoding the word's mutation state (boxed in the example).

(40)

Surface	**Lexical**	
mawr	+Rad+	mawr +Adj
fawr	+SM+	mawr +Adj

The tags, +Rad+, +SM+ etc. are in the lexicon, as shown in (41).[10]

(41)
```
+Rad+  MUT[rad]  XLE.
+SM+   MUT[sm]   XLE.
...
```

The easiest way to understand how we constrain syntactic mutations is through examination of the AP rule (42). The initial element of the right hand side of this rule is a disjunction of the relevant mutation categories. The associated (inside-out) constraints state that if the modified noun is FEM SG, the AP must be soft-mutated (MUT_BASE[sm]), that is, start with an initial soft mutation segment, and otherwise that it must begin with the radical (MUT_BASE[rad]). The mutation categories are then followed by the remaining constituents of the AP, whatever they are.

(42)

AP –>

{ MUT_BASE[sm]: ((ADJ ∈ ↑) GEND) =$_c$ fem
((ADJ ∈ ↑) NUM)=$_c$ sg
| MUT_BASE[rad]: { ((ADJ ∈ ↑) GEND)=$_c$ masc
((ADJ ∈ ↑) NUM)=$_c$ sg
| ((ADJ ∈ ↑) NUM)=$_c$ pl } }
... + remaining constituents of AP

That is, our treatment of syntactic mutations involves mutations mapping to syntactic categories which appear constituent-initially.

For this to work, it is crucial that the sublexical rules for an adjective, or a pre-adjectival adverb modifying the adjective, or indeed (almost) any other lexical category, do not start with a mutation category. If these rules did start with a (non-optional) mutation category, MUT_BASE[rad/sm] in (42) could not appear at the supralexical level in the c-structure

[10]The category names chosen currently have the format of a complex category MUT[*value*] where the value (in square brackets) can be passed to the left hand side of a rule in so-called parameterized rules (Crouch et al., 2006). This was necessary in an earlier approach to mutation in our grammar, but complex categories are no longer necessary in our current approach and could be replaced by simple categories such as MUTrad, MUTsm etc. We are, however, keeping them for the time being as they may become useful again for possible further improvements to the way we handle syntactic mutations.

because it would be associated with the lexical item it morphologically originates from. The sublexical rule for an adjective only contains the stem and any tags appearing after the stem as shown in (43), while the morphological analysis for adjectives as shown in (40) involves an initial mutation tag.

(43) A --> A_BASE *mawr*
ASFX_BASE *+Adj*
+ further (optional) sublexical constituents

Initial mutation tags are thus either consumed by lexical mutation triggers (i.e., deleted by our multiword transducer), in the case of lexically induced mutation, or they are treated supralexically as categories in c-structure rules in the case of syntactically conditioned mutation.

5.2 Syntactic Mutations as Edge Inflections

The treatment of syntactic mutations outlined in this section has the perhaps unexpected feature that it treats the initial mutation tag as mapping to a syntactic terminal in its own right, in apparent violation of lexical integrity. Within the context of our implemented grammar, the reasons for this treatment are largely of a practical nature, for this greatly simplifies the rule set required. Nonetheless, it is basically equivalent to treating syntactically conditioned initial mutation as a type of edge inflection, and an alternative direct encoding of an edge inflection approach is possible, though more complicated and less compact to state and more susceptible to coding error. We illustrate such a comparable approach as edge inflection with the rather simpler case of Basque case marking.

Although Basque is predominantly head-final, adjectival modifiers and demonstratives follow the noun. NPs (or perhaps DPs) in Basque can be inflected for case, number and determinedness (and a few other features). This inflection is marked on the last NP constituent only. Some examples are given in (44); case is always ABS[olutive].

(44) a. *zaldia*
horse.ABS.SG.DET
'(a/the) horse'

b. *zaldi txikia*
horse small.ABS.SG.DET
'(a/the) small horse'

c. *zaldi txiki hau*
horse small this.ABS.SG
'this small horse'

Clearly, whatever phrase structure rules and constraints we assume, we would have to ensure that only the last NP constituent is inflected, and that this inflection is passed up towards the top level of the NP. Respecting lexical integrity strictly, we could pass inflectional information upwards not only by using suitable f-structure annotations on sublexical constituents but also by using XLE's parameterized rules, which contain complex categories on both sides of the rule that hold a value (in square brackets) and whose value could be passed from the right hand side to the left hand side of the rule.

The morphological analysis for *txikia* in (44 b) would be as in (45), capturing the fact (*inter alia*) that the adjective *txikia* is inflected for case (+Abs). The lexical entry

for the tag `+Abs` is given in in (46) where the tag's category has the case value *abs*. The sublexical rule for adjectives, for instance, as shown in (47), where the case value of the `CASE(_BASE)[_case]` category, whichever this is, would be passed to the left hand side as value of `A[_case]` and all functional information would likewise be passed up via the annotations on the inflectional sublexical categories. This value could again be passed up to the AP as in (48) and from there to the NP as in (49). Similar rules can be written for nouns, NPs and demonstratives. But since non-final NP constituents appear uninflected, we would have to write additional rules for uninflected Ns, NPs, As, APs, etc.

(45) **Surface** **Lexical**
txikia `txiki +Adj +Sg +Art +Abs`

(46) `+Abs CASE[abs] XLE.`

(47)

A[_case] −− >	A_BASE	*txiki*
	ASFX_BASE	*+Adj*
	NUM_BASE: ((ADJUNCT $ ^)=!;	*+Sg*
	ART_BASE: ((ADJUNCT $ ^)=!;	*+Art*
	CASE_BASE[_case]: ((ADJUNCT $ ^)=!.	*+Abs*

(48) AP[_case] −− > A[_case].

(49) NP[_case] −− > N AP[_case].

In fact, since the case value is stored in the parameter, we could forego the functional annotation on `CASE_BASE[_case]` in (47) and provide this information wherever `NP[_case]` is instantiated as in (50).

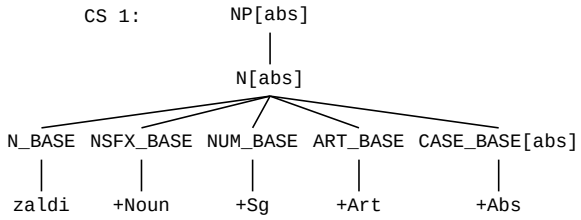
(50) { NP[abs]: (^ CASE)=abs
| NP[erg]: (^ CASE)=erg
| NP[dat]: (^ CASE)=dat *etc.* }

Turning to the NP rule, (51) encodes a flat NP analysis. What is crucial here is the number of disjuncts that take account of the requirement that the last NP constituent, whatever it is, is the locus of case inflection (i.e., a complex category). The distinction between complex (= inflected) and simple (= uninflected) categories in the rule ensures that inflection only appears where it is licensed in the c-structure. (52 a-c) shows the resulting trees for (44 a-c) with all morphemes displayed.

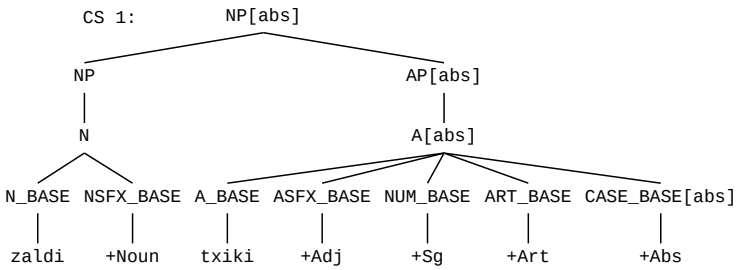
(51)

NP[_case] −− >	{ N[_case]
	| N AP* AP[_case]
	| N AP* DEM[_case] }.

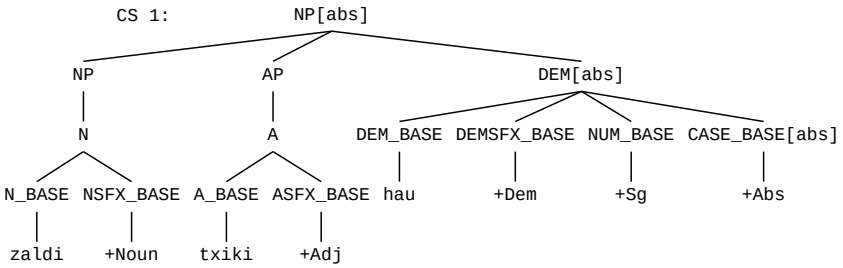
(52) a.



b.



c.



This Basque example shows how an inflectional value originating from the sublexical level can be passed up without necessarily passing up any functional information alongside. It seems to us that a similar approach to syntactic mutation in Welsh is possible, but would be extremely complex: while in Basque the inflection in question appears only on the final word in the NP, in Welsh the nature and distribution of syntactic mutations is much wider, inducing serious complications into the c-structure.

5.3 Default Mutation

One last point remains to be explained: why we deleted the second of two matching mutation tags in our multiword transducer dealing with lexical mutations. This is, in fact, not strictly necessary, but it gives us the considerable practical advantage of being able to specify *one* syntactic mutation as a default. As mentioned above, (almost) all sublexical rules end up without an initial mutation category. If syntactically governed, the mutation category appears in the supralexical c-structure rules; if lexically governed the corresponding tags are deleted even before they can enter the grammar proper.

Syntactic mutations almost always involve either Soft Mutation or the Radical. If we now include the mutation category corresponding to, say, the `+Rad+` tag in the sublexical

rules, and make this category optional, we only have to specify those syntactic mutations that do not involve the radical. This means that if we choose not to specify a syntactic mutation in the c-structure rules (and no lexical mutation applies), the mutation category/tag can remain with the lexical item it originates from, and if its only possible value is *rad*, the (overtly mutationally ungoverned) lexical item will default to its radical mutation state, but none other. (53) shows the slightly modified sublexical rule for an adjective vis-à-vis (43) above.

(53)
```
A -->  (MUT_BASE[rad])   +Rad+
       A_BASE            mawr
       ASFX_BASE         +Adj
       + further (optional) sublexical constituents
```


Acknowledgements

The work reported on here was carried in the project Verb Initial Grammars: a Multilingual, Parallel Approach: see `http://users.ox.ac.uk/~cpgl10015/pargram/index.html`. We are grateful for the financial support of the ESRC (research grant RES-000-23-0505, to Dalrymple and Sadler). We thank participants at LFG06 for useful comments, and especially John Maxwell and Ron Kaplan for much discussion of the multiword transducer.

References

Ball, Martin J and Nicole Müller. 1992. *Mutation in Welsh*. London: Routledge.

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*, pages 1–7. Online: http:www2.parc.com/isl/members/thking/coling02pg.pdf.

Crouch, Dick, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2006. XLE documentation. Tech. rep., Palo Alto Research Center, Palo Alto, CA.

Frank, Anette, Tracy Holloway King, Jonas Kuhn, and John Maxwell. 2001. Optimality Theory style constraint ranking in large-scale LFG grammars. In Peter Sells, ed., *Formal and Empirical Issues in Optimality Theory*, pages 367–97. Stanford, CA: CSLI Publications.

Green, Antony Dubach. 2003. The independence of phonology and morphology: the Celtic mutations. *ZAS Papers in Linguistics* 32:47–86. also online http://www.ling.uni-potsdam.de/ green/cv/independ.pdf.

Kaplan, Ron, John Maxwell, Tracy Holloway King, and Richard Crouch. 2004. Integrating Finite-state Technology with Deep LFG Grammars. In *Proceedings of the*

Workshop on Combining Shallow and Deep Processing for NLP (ESSLLI). Online: http://www2.parc.com/isl/groups/nltt/pargram/esslli04fst-xle.pdf.

Stewart, Thomas. 1992. *Mutation as Morphology: Bases, Stems and Shapes in Scottish Gaelic*. Ph.D. thesis, Ohio State University. Online http://www.ohiolink.edu/etd/send-pdf.cgi?osu1086046888.

•

THE EVOLUTION OF THE TENSE-ASPECT SYSTEM IN HINDI/URDU: THE STATUS OF THE ERGATIVE ALGNMENT

Annie Montaut

INALCO, Paris

Proceedings of the LFG06 Conference

Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006

CSLI Publications

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

Abstract

The paper deals with the diachrony of the past and perfect system in Indo-Aryan with special reference to Hindi/Urdu. Starting from the acknowledgement of ergativity as a typologically atypical feature among the family of Indo-European languages and as specific to the Western group of Indo-Aryan dialects, I first show that such an evolution has been central to the Romance languages too and that non ergative Indo-Aryan languages have not ignored the structure but at a certain point went further along the same historical logic as have Roman languages. I will then propose an analysis of the structure as a predication of localization similar to other stative predications (mainly with "dative" subjects) in Indo-Aryan, supporting this claim by an attempt of etymologic inquiry into the markers for "ergative" case in Indo-Aryan.

Introduction

When George Grierson, in the full rise of language classification at the turn of the last century, [1] classified the languages of India, he defined for Indo-Aryan an inner circle supposedly closer to the original Aryan stock, characterized by the lack of conjugation in the past. This inner circle included Hindi/Urdu and Eastern Panjabi, which indeed exhibit no personal endings in the definite past, but only gender-number agreement, therefore pertaining more to the adjectival/nominal class for their morphology (*calâ,* go-MSG "went", *kiyâ,* do-MSG "did", *bola*, speak-MSG "spoke"). The "outer circle" in contrast, including Marathi, Gujarati, Bengali, Oriya, Assamese, shows personal endings in every verb tense, therefore has a "conjugation", and should be sharply distinguished from the languages of the inner core, with intermediate languages arranged into a "middle circle" (Bhojpuri, Eastern Hindi).[2] What it means is that agreement only in gender-number, along with the ergative structure as we call it today, was supposed to be the mark of a truly authentic Indo-Aryan language. This theory was strongly criticized by Suniti Kumar Chatterji and later abandoned by Grierson, but it is still held that ergative Indo-Aryan languages (roughly speaking in the West) radically differ from non-ergative ones (in the East) and are extremely atypical within the wider Indo-European family. What is unique in fact is the modern development of a full fledged ergative structure out of the nominal predicates,[3] not the historical phase where participial predicates were used with instrumental agents, which in other languages got converted into a nominative structure. Both ergative and nominative patterns in Indo-European rather represent different stages of the same logic in renewing the system (section 2), both in the past and future (section 3). It will appear at the same time that the distinctiveness of the ergative alignment, at least in Indo-Aryan, does not consist in being an inverted mirror of the nominative alignment

[1] Grierson is the author of the *Linguistic Survey of India* (11 vol.), which is still a reference. The work represents the first attempt to group the Munda and Mon-Khmer languages as a distinct family (still called Austric or Austro-Asiatic) just after Dravidian languages had been separated as the second distinct Indian family, the first one being the Indo-Aryan family, identified right after the famous discovery by William Jones in 1786 that Sanskrit and Latin-Greek were sister languages. The first scholar who gave a scientific and wide description of the Indo-European family was Franz Bopp. For the description of the scientific and ideological context of these elaborations and their far reaching consequences in language classification, see A. Montaut 2005.

[2] His description, also based on a few phonetic features, like the alternation *s/sh*, supposedly a radical difference between both circles, was in conformity with the then theory of the settlement of the Aryan tribes in India, said to have come from the North-West in "concentric waves". The original, more ancient settlers occupied the nucleus around which circled those arrived later. Such a theory was no longer in fashion when the *Linguistic Survey of Inda* was completed. Moreover, the sharp critics of S.K. Chatterji modified Grierson's final presentation of the Indo-Aryan family.

[3] Which also occurred in some Iranian languages, like Pashto.

since it rather patterns with other localizing predications well established in the global economy of the system such as the experiential dative alignment (section 4). At the same time I will try to explore the main paths of grammaticization of aspect, tense and modality, starting with the non past system, which helps understand the evolution of the past system (section 1).

The aim of the paper is threefold: sketching the broad lines of the historical evolution of verb forms in Indo-Aryan and specially Hindi/Urdu; inquiring into the categories of aspect, tense, mood and the way they grammaticize; inquiring into the nature of the ergative alignment, along with other non-nominative alignments.

1. The non-perfect system

1.1. Generalities

The present shape of the Hindi/Urdu (HU) verbal paradigm may strike one as very bizarre: as opposed to most languages which have an unmarked indicative present, the two unmarked forms are the subjunctive (only personal endings) and the anterior or narrative past (only gender-number endings), and the present is marked (two words, 5 morphs). See table 1:

Tense Aspect - accomplished	Tense Aspect + accomplished	Mood	Mood	Tense
	calâ preterit went	*caltâ* counterfactual would go	*calûn* subjunctive	*calûngâ* future
caltâ hai present goes (*cal rahâ hai* prog) is going	*calâ hai* perfect has gone			
caltâ thâ imperfect walked (*cal rahâ thâ* prog) was going	*calâ tha* pluperfect had gone			

If we agree that unmarkedness is used by default and expresses the core meaning of a given sector of the mental map,[4] whereas marked forms express marked, less basic and less frequent meanings, the picture looks strange because we are not used to conceive of anteriority as the basic (core) meaning in tense, nor subjunctive in mood.[5] The basic oppositions (+/- progressive: *rah*/∅, +/- accomplished: ∅/*-t-*) only are represented below:[6] although present and past nicely parallel in the unaccomplished, as well as past and present in both accomplished and unaccomplished (last two lines), there is an asymmetry: whereas the simple form for the accomplished (- *t*) patterns with the two complex forms, structuring the whole of indicative forms, the simple form for the unaccomplished (+ *t*) does not pattern with the two complex forms and stands for a distinct mood. In these oppositions, the first is

[4] I am following Bybee (1994) in distinguishing unmarked from zero mark. Unmarked forms may have no overt mark and cover a wide (unspecified or with low specification) meaning, but the lack of overt marking may also refer to a specific meaning, therefore acquiring the status of zero.

[5] Tense can be defined as "the chronological situation of an event in relation with the speech act by which the speaker refers to that event", or time of utterance T0, pertaining then to the succession of events (anteriority, simultaneity or posteriority). Aspect on the contrary pertains to the way of representation of the process as a predicate (Cohen 1989: 11, 42). It maps the three basic notions of state, process and event into three types of topological intervals which can be represented with open of closed boundaries (Desclès 1980, 1992).

[6] The +/- accomplished is found in the other moods too.

expected (marked progressive) and parallels English translations, but the second does not (marked unaccomplished).

A word on terminology: perfective is the most frequent label used to design the simple form ("I went", *calâ* in Hindi) representing past events. It is named aorist in Nepali, simple past in French, preterit in English, and has received various names in the Indo-Aryan traditions, including indefinite perfect.[7] Given the very specific meaning of "perfective" in all the languages which oppose perfective to imperfective like Russian,[8] I will avoid the term and use the term preterit (referring to anterior events), leaving aside for the purpose of this paper the well-known non anterior meanings of the form (Montaut 2004, 2006). Since perfect is used by many as referring to present perfect ("I have gone", *cala hûn*) I will use perfect to refer to the whole system derived from *calâ*, present perfect, preterit and pluperfect, rather than accomplished, which has no currency in IA linguistics.

History only can make this paradigm understandable. The major event in verbal morphology was the drastic impoverishment in MIA of the rich ancient paradigm: whereas OIA had some forty synthetic forms for tense-aspect, mood, voice, MIA maintained very few finite forms, and in some regions only the present in the indicative (imperative was maintained everywhere. Some dialects and languages also maintained the old synthetic sigmatic future in *–Sya* (> *s* > *h*). All of them used the past participle to represent past events. Out of this extremely reduced paradigm of synthetic forms, a number of compound forms with auxiliaries developed, leading to the rich present analytical paradigm of HU: Nespital (1980) for instance registers 39 tense grams and Dymshits (1985), who, unlike Nespital, does not consider the vector verbs as aspect markers, registers about 20.

1.2. The non-past (non- perfect) system

If we start from the standard situation in MIA (deliberately simplified in order to account for Hindi/Urdu), we could expect that the Sanskrit indicative present in *–ati* remains a present throughout the period up to now. The form has indeed survived but is no longer an indicative present:

calati (go-PRS.3M.SG) > *calaï* > *calai* > *cale*

calanti (go-PRS.3M.PL) > *calaiN* > *caleN*[9]

for the base *cal-* "go/walk" is now interpreted as a subjunctive (with optative meaning in independent clauses and various meanings including non specific and virtual in dependant clauses)..

The reason why the old present did not retain its meaning in modern HU is that, among other factors, the synthetic future was ultimately not retained, but the old present form had still a present meaning up to 19th century. Simply, the synthetic form had a wider meaning, covering the all non-past area (future, eventual, present, both actual and habitual), since no other form was available. This wide meaning can be designed as an open meaning (Garcia & Putte 1989, Bybee 1994), embracing several restricted meanings later to be distinguished:

(1)a. *ve kheleN* they play

3M.PL play-3M.PL

---

[7] Kellogg (1875: 234), who opposes this simple form to the present perfect (with "be" auxiliary in the present) and past perfect (with "be" auxiliary in the past).

[8] A distinction which Indo-Aryan largely displays by means of vector auxiliaries, semi-lexicalized items similar to the semi-lexicalized opposition between perfective and imperfective in Slavic languages (Nespital 1997).

[9] The second person, originally ending in *–asi*, also evolved into a final *–e* (*-asi* > *-ai* > *-e*), like the third person. The first one *–âmi*, has a more complicated story. Analogy reshaped the plural modern forms where original endings disappeared except 3PL *anti* > *aiN* > *eN.* Gloss is according to the Leipzig glossing rules, with the addition of CP: conjunct participle, OVA: obligative verbal adjective, POT: potential, OBLIG: obligation, H: honorific.

is still described as a subjunctive sometimes used as a general present in Kellogg (1875) and is still found in the literature of the time with a present meaning, although rarely in the texts written by the language teachers of the Fort William College during the first decade of the 19th century (1800-1810), who were supposed to set the modern grammatical standards. It is still used today with this meaning in proverbs, expressions well-known for retaining archaic forms (*jaisâ kare/karai vaisâ bhare/bharai* "you (will) reap what you sow", *koî kare/karai, koî bhare/bharai* "someone does and another one benefits").[10] Along with this open meaning of the old synthetic form, the first periphrastic form in *–tâ hai* (lit." is … -ing") was, still in Kellogg's times, used as a form restricted to present, that is, not future and not subjunctive. In the 19th century the modern contrast between habitual/generic and progressive was still not well established, since the first form is glossed by both present meanings in Kellogg whereas the second longer form in *rah hai* (lit. is stayed) is glossed by a more expressive periphrastic turn ("be engaged in"):

(1)b. *ve* *khelte* *haiN* they play/ they are playing
3PL play-*ing* be-3M.PL

(1)c. *ve* *khel rahe* *haiN* they are engaged in playing
3MPL play stay-PP be-3M.PL

This means that the "is …-ing" form, today a "general present tense" (*sâmânya vârtamân kâl*) had still in the middle of the 19th century its expected progressive meaning, along with the general/habitual meaning. Texts from the Fort William College illustrate this situation, where the *rah* form is only in the process of being grammaticized as a progressive, still retaining a stronger emphatic and literal meaning ("engaged"), still used as a stylistic optional or disambiguizing device.[11] When the *rah* form lost its literal meaning and came to be required for the expression of an actual specific process, then no longer perceived as an expressive device, the other form restricted its meaning to the expression of habits and genericity, losing its open meaning. Such a process of restricting the open meaning of an old simple (unmarked) form because a new marked form became obligatory for the expression of the other (restricted) meaning has been well documented: the English simple present for instance was according to Garcia & Putte (1989) originally an open present (such as the French unmarked present still is) with both meanings, and the marked periphrasis gradually emerged as an expressive optional device used for stylistic emphasis or to prevent ambiguities. One can regard this process as a conventionalization of the inference which, in conformity with conversational rules, constrains the listener, in the absence of the periphrasis, to rule out the marked meaning (Carey 1994). When it generalized, the unmarked form retained only part of its earlier meanings and the marked one lost its expressive strength and got grammaticized. The unmarked form can be said to have grammaticized a zero mark for the new meaning ruling out progessive.

In HU the simple form indeed underwent such a process (open non-past > non past restricted to the potential: non-future, non-present), but the newly grammaticized "is –ing", originally a progressive present, in its turn underwent the same process (open present > non

---

[10] With future meaning too: *jaise kâlî kâmrî caRhai na dujo rang*, black will take no other hue

[11] Lalluji (*Premsâgar*, 4th chapter) for instance uses in the same context a *-tâ hai* (a) form and a *rahâ ha*i (b) form for describing an obviously actual and not habitual process, with ostensive indications, when Krishna suggests the cowherd to visit the nearby Brahmans from Mathura presently celebrating a sacrifice right now (the smoke is visible)

a *dekho jo dhuâN dikhâî detâ hai tahâN mathuriye kans ke Dar se <u>yagya karte haiN</u>*
look where smoke is visible, Brahmins from Mathura <u>are celebrating a sacrifice</u> for fear of Kansa

b *vahâN gae jahâN mathur baiThe <u>yagya kar rahe the</u>*
we went where the Brahmins from Mathura <u>were celebrating a sacrifice</u>

The present expression for Kellogg's translation of (1c) would be something like : *ve khelne meN lage haiN* (they play-INF in stuck are).

progressive). If the marked form in "is –ing" (*-tâ hai*) has already become an open present in the 19th century, it is because it was probably created for contrasting the actual process with the then open present expressed throughout ancient Hindi by the *–tâ* form.[12] This nominal form originates from the Sanskrit present active participle in *–anta* (> *ant* > *at*), later on suffixed with gender-number endings, and has been used as a predicate expressing general present from Chand Bardai's *Prithvirâj Rasau*, 12-13th century, in Old Rajasthani (2a) and Old Marathi to Kabir, in the 13-14th century, and to Tulsidas's *Ramayan*, 16th century, in Old Awadhi (2b):

(2)a. *kârtik karat pahukar sanân*
kartik do-*at* Pahukar bath
he takes his ritual bath in Pahukar (Beames: 130)

(2)b. *sab sant sukhi vicarant mahî*
all saint happy walk-*ant* earth-LOC
all the saints walked happily on the earth

(2)c. *puruS kahte*
man say-*t*-M.PL men say

The subsequent disappearance of the *–tâ* form from the domain of present left room for the *–tâ hai* form to occupy the entire space of present. Why did this participial form not retain its present meaning in the modern standard language,[13] why did it instead specialize in the expression of counterfactual? Bloch (1906) hypothesis that the predicative use of the nominal sentence dominated only in the accomplished (past) system, because of the resilience in the non-past domain of the old synthetic present, which indeed seems to have at least partly preserved its general present meaning since we still find it centuries later as in (1a).

To sum up, the non-past system between Old Indo-Aryan (OIA) and new Indo-Aryan (NIA) illustrates a cyclic process of widening>narrowing>widening in the meaning of certain forms. The Middle Indo-Aryan (MIA) opening of the meaning of the old Sanskrit present, due to the tumbling down of the whole finite paradigm, shows that a form tends to occupy the whole notional domain in the absence of other competing forms. The further gradual restriction of its meaning between MIA and modern NIA (non-past > future/optative > optative) was due to the emergence of new forms, first optional and emphasizing a contextual or stylistic meaning, then obligatory.

In contemporary Hindi/Urdu, the proliferation of new auxiliaries for habitual (frequentative), durative and progressive (V*â kar*-, V*tâ rah*-, V*tâ jâ*, V*tâ calâ jâ*-, etc.), still optional and commutable with adverbs, has not yet restricted the meaning of the older less complex forms (for that matter the V*tâ hai* or V *rahâ hai* forms), which can then been considered as open: the new forms are not fully grammaticized, hence not in a real complementary distribution with the less marked forms.

Open meaning and unmarkedness are relative. For instance, in a given portion of a notional domain (of the semantic map) like present, the "general present" (V*tâ hai*) is relatively unmarked compared to the progressive present (V *rahâ hai*). Unmarkedness, if associated with the defect meaning and core value of the notional domain (here, present), will tell us that habitual and generic is the basic meaning of present, not specific (actual,

---

[12] Old Hindi covers a wide number of regional dialects which are today distinct languages for some of them at least and for all of them with very distinct verbal paradigms. However, since no particular language can be identified as the direct exclusive ancestor of Standard Hindi or Kharî bolî, referring to this disparate or syncretic ancestry is a necessity if we want to go beyond the 18th century in the history of the language.

[13] As it did for instance in Garhwali still today, with personal endings suffixed:

mi *baccon thainki* *paRhâNo* *ku* *kâm* *kardûN*
1SG children DAT/ACC teach-INF GEN work do-PTCP.1SG
I do the job of teaching children

In Sindhi the *–ta/to* form is used as a future (Beames 1871: 126, Trumpp 1872).

progressive). This is confirmed by many other languages. But it would be at least awkward to conclude that counterfactual is the basic meaning and core value in the wider domain of non past because its form is unmarked (V*tâ*) compared with the general present (V*tâ hai*). (Un)markedness is also the product of history and can enlighten the linguistic mapping of cognitive realities only to a certain point.

2. The "past" system: the nominal sentence as an expression for perfect or accomplished

2.1. The problem

As opposed to the present system, in a similarly impoverished verbal paradigm, the past (accomplished) system was quite early dominated by the passive past participle in *–ita* (> *iya* > *ya* > *a*). Originally used for transitive processes, the participle expressed the result of the event, somewhat in the same way as we today can say "understood" for "I have understood". In classical Sanskrit already, the canonical expression of 'X had done /did Y' is 'by-X Y done', with the agent in the instrumental case (or genitive for pronouns) and the predicative participle agreeing in gender and number with the patient:

(3) *mayâ / mana tat kRtam*
I-instr / I-GEN this-NOM.N.SG done-NOM.N.SG
I did/have done that

As is well known, this is the pattern inherited by the present HU ergative structure (4a) in the perfect as opposed to the nominative structure in the present or future:

(4)a. *laRke ne /maiNne kitâb paRhî*
boy-OBL ERG /1.SG-ERG book-F.SG read-F.SG
the boy / I read the book

(4)b. *laRkâ kitâb paRh rahâ hai / maiN kitâbeN paRhtâ hûN*
boy-M.SG book-F.SG read stayed-M.SG is 1SG book-F.PL reading-M.SG be-1SG
the boy is reading a book I read books

Given the fact that Sanskrit gave birth to all modern Indo-Aryan languages, we may wonder why only some (roughly speaking Western) of the Indo-Aryan languages developed the aspectually split ergative structure. Bengali for instance is a consistently nominative language, with nominative subjects and verb agreeing in person with the subject at all tense-aspects (5):

(5)a. *âmi boi.Ta por.l.âm* b. *tui boi.Ta por.l.is*
1SG book-DEF read-PST-1SG 2SG book-DEF read-PST-2SG
I read the book you read the book

The question is all the more puzzling since a similar pre-ergative structure prevailed in all the Asoka Prakrits, in the East as well as in the West: (6a) is from Girnar in the North-Western region, whereas (6b), with the same structure as (3), is from Jaugada in the Magadhean region, presently Bengal-Orissa-Assam-Bihar. Since (a) and (b) have the same meaning and gloss, except for the verb base, causative in (a) and simple transitive in (b), I give them only once:

(6)a. *iyam dhammalipi devânâmpriyena priyadassina ranna lekhapita*
this law-scripture of-gods-friend friendly-looking king inscribed
NOM-F.SG NOM-F.SG INSTR-M.SG INSTR-M.SG INSTR-M.SG NOM-F.SG

(6)b. *iyam dhammalipi devanampiyena piyadassina [Iajina] lekhita*
the friendly looking king beloved of gods has (made) engraved this law-edict

Present predicates contrast with this structure in the same way as (4a) with (4b), as shown in (7), from Kâlidâsa's *Vikramorvasiya*, where the pronominal subject is in the nominative (*hau* <*aham*) whereas in the past it is in the oblique (*pai* < ?? *âtman*??) already used as a syncretic marker for several oblique cases):[14]

[14] Quoted by Bubenik & Paranjape (1996: 112). Cf. *tai rasau hinduâN* [you-OBL protected-M.PL Hindus-M.PL], "you protected Hindus", vs *hau acchari nâhi* [I-NOM apsara NEG] "I am not a celestial woman".

(7) *hau* *pai* *pucchimi … diTThî* *pia* *pai* *sâmuha* *jantî*
I-NOM you-OBL ask-PRS-1SG seen-F.SG loved-F.SG you-OBL in-front passing-NOM-F.SG
I ask you… Did you see (my) beloved passing in front (of you)?

This opposition between past and present systems started prevailing as soon as classical Sanskrit, and Bloch noticed the wide generalization of the nominal statements for expressing past: in *Vetâla* (10th century) 1115 expressions of past are of that type against 38 for finite verb forms (1906: 60). Predicative passive past participles were then used to express "various nuances of past tense and modality", but this dominance does not mean that no other form existed: various finite forms were still in use, but none prevailed on others, and they became less and less frequent in texts, almost disappearing in MIA (Bloch 1906: 47-48).

What we still find in ancient NIA (the earliest phase of modern Indo-Aryan from 12th to 16th century) is the same nominal structure for past / accomplished statements, that is to say a pre-ergative structure. The only difference with Asoka's statements in (6) is that the instrumental (or genitive) is no longer a distinct case since it got fused with other oblique cases, except with the locative which remained distinct in many languages. Old Bengali (8a-b), Old Awadhi (8c), which are Eastern dialects considered to derive from Magadhean Prakrit, present the same structure as Old Braj (9a), Old Panjabi (9b) and Old Marathi (9c) which are Western dialects considered to derive from Saurasenic or Marashtri Prakrits:

(8)a. *kona purane,* *Kanhâ,* *hena sunili* *kâhini*
which purana-LOC Krishna, so heared-PST-F.SG story-F.SG
in which Purana, Krishna, did (you/one) hear this story? /was the story told?

(8)b. *ebeN maï* *bujhila*
now 1SG-OBL understood-ø now I have understood

(8)c. *maï* *pâi* vs. *hau* *manuSa*
1SG-OBL obtained 1SG-NOM man (Jayasi)
I obtained (it) vs. I am a man

(9)a. *susai* *[bat]* *kahî*
hare-OBL (speech-F.SG) said-fs the hare said

(9)b. *guri* *dânu* *ditta*
guru-LOC gift-M.SG given-M.SG the guru gave the gift (Guru Granth Sahib)

(9)c. *aiseN* *myâ* *pahileN*
this-N.SG I-INSTR seen-N-SG I have seen this (Jnanesvari)

In (8) as well as in (9), the predicate is a nominal form agreeing in gender and number with the patient, whereas the agent, if expressed, is in the oblique form and does not control verb agreement. This series shows that up to a certain point the expression of past was general, and bifurcated later, between 14th and 16th c., since the first Eastern statements (from Chatterji 1926) are from 14th century caryâs.

2.2. The nature of the divergence : semantics and syntax of aspect

2.2.1. Evolution of aspectual semantics

As the structure in (6) got generalized, it started to loose its expressive meaning, originally emphasizing the result and not the process, so that it soon acquired an open meaning, encompassing process and result (cf. supra, Bloch's quote). The original restricted meaning of the passive past participle, a state, can be represented as an open space, not taking into accounts any boundary, as opposed to the anterior which only takes into account the bound interval (event) in disjunction from the utterance time, and in contrast to the perfect, which represents the adjacency of the resulting state with the event which produced it, allowing for the topological representation below (from Desclès 1992):

state ---[////]$_{T1}$-----$_{To}$
anterior ---[////]$_{T1}$-----$_{To}$
perfect ----[---]/////[$_{T1}$----- $_{To}$

When the participles generalized with open meaning (anterior event / preterit, resulting state / perfect), they got more and more perceived as an active predication since there was no other option, and lost the passive meaning of the patient orientation. As the need was felt in certain statements to avoid ambiguity or to emphasize the resulting state, a new form was created by the adjunction of a copula, originally expressive then grammaticizing in the meaning of resulting state. Initially the copula occurred in the first and second person to prevent agent ambiguity (Bloch):

10a *kenâsy* *abhihatah*
who-INSTR-be-2SG beaten-NOM-M.SG
by whom have you (not he, not we, not she etc.) been beaten

10b *tenâsmi* *sopacaram* *uktah*
3SG-INSTR-be-1SG respectfully said-NOM-M.SG
I (not you, not they) have been told this by him = he told me

The copula later helped emphasizing stativity (to prevent another kind of ambiguity, event or state) or simply introducing stylistic variation according to Breunis and Bloch, and from the moment this originally stylistic variant became more expressive of state or "condition" it was no longer a stylistic variation but a grammaticized expression of perfect or resulting state of an event Breunis (1990: 141). At the same time, the simple form restricted its previously "open" meaning to the expression of anteriority (event: preterit). This echoes the story of the renewal of the present (first competing with a new progressive marker in the specific meaning then retaining only the other meaning). The situation found in early NIA similarly shows an open meaning, which was probably in the process of getting restricted in front of the competing copula form, whereas the contemporary situation clearly shows a strictly complementary distribution of both forms. If we agree with Bybee (1994) we may analyse this as an emergence of a zero mark with the meaning of anterior, whereas previously the unmarked form had unspecified meaning in the whole perfect system.

Obviously when the former participle is used as a predicate for representing events, even if the agent remains in an oblique case as in passive sentences, the emphasis is more on the process (source oriented) than on its result and the whole statement gets more and more perceived as active and no longer passive. Besides, it was the only expression for past processes. This is expressed by Nespital (1986: 145) as the emergence of a "Neuer Proto-aktiv Satz", which he observes since the pali stage in *Milindapanha*.

2.2.2. Changing morpho-syntactic patterns

This active transformation was implemented differently in the East and in the West, and here lies the today opposition in the syntactic alignments. In the East, the active renewal was radical, and the pre-ergative structure was de-ergatived so to speak, between the 14th and 16th century. Chatterji (1926) calls the process an active conversion, comparing the form, not the meaning, with the medieval structures (8). The agent, in conformity with the linguistic perception (active process) became expressed in the nominative or unmarked case, whereas new personal endings were affixed to the verbal form. What is interesting is that these affixes are still now clearly distinct from the older endings of the present.

(11) *âmi boi.Ta* *por.l.âm*
1SG book-DEF read-PST-1SG I read the book (present *âmi por-i*)
tui por.l.i: 2-nonH read-2nonH, "you read" (present *tui por-is*)
tumi por.l.e 2 read-2 "you read" (present *tumi por-o*)

The transformation then ends up providing a nominative alignment with standard personal predicates with a standard past marker *–l-*, as rightly today analysed. But its origin denotes no trace of anteriority marking, since this suffix is widely found throughout the nominal class, mostly with the meaning of a "diminutive" affix (*rangilâ* "coloured" from *rang*

"colour", *kanTilâ* "thorny", from *kânT* "thorn"). It also behaved more or less like the so-called "enlargement" suffix *–k-* extensively added to nominal bases in late OIA.[15]

The same transformation happened in Bhojpuri and to a lesser degree in Awadhi: "when the original passive construction was lost in Bhojpuri as in other Magadhean dialects, the Prakritic constructions with the passive participle became a regular verb in Bhojpuri, and it began to be conjugated by adding personal terminations which came from the radical tense as well as from the *s/h* future" (Tiwari 1966: 171).

Western languages on the contrary, instead of re-aligning the morpho-syntactic pattern on the nominative model fit for action processes, reinforced the oblique marking of the agent by using a postposition, either specific (HU) or not (Marathi), and so developed the full fledged ergative structure for the perfect system (anterior, present perfect, pluperfect).[16] Only some modern IA languages retain the old oblique agent (Jaisalmeri, Western Rajasthani dialects). But this recent re-characterization of the old instrumental does not make the structure more passive and its "perception" as an active structure shows in the various subject properties attached to the marked agent, who has now most of the control properties (reflexivation, conjunctive participle), but still never controls agreement, even with a marked patient.[17] Bubenik & Paranjape (116-7) suggests that the placing of the agent in the first position in late MIA correlates with the linguistic perception of the oblique noun as a semantic subject. Breunis (1990) in his chapter on word order (chapter 6) suggests that the fronting of the agent is earlier, which is confirmed by many of the examples from Bloch (1906). The fronting of the marked agent amounts to treat it as a topic, which is a first step on the way to shifting it to the subject status.

We can then summarize this general evolution by saying that Eastern languages have simply gone one step further than Western ones in the same logic, they have fully endowed the agent with subject properties, whereas the Western languages have gone a step further in the ergative pattern but still have endowed the agent only with the semantic, syntactic and to a certain extent pragmatic properties.

Bengali is a good example of the full cycle from a nominative language (Sanskrit) to a pre-ergative one (Old Bengali) and back to a nominative one, and Hindi/Urdu of the first part of the cycle (from a nominative to an ergative one. This cyclic evolution has of course been gradual and is still in process, and the occurrence of personal endings in Marathi at the second person, as well as the use of nominative agents for first and second person in Marathi and Panjabi, [18] may be interpreted as a sign of a transitional stage towards a nominative patterning. For instance, (12a) in Marathi and (12b) in Panjabi exactly structured as (4a) in Hindi/Urdu, show a marked agent, only gender-number agreement with the patient on the participle-like predicate, but (13a) in the second person shows, after the gender-number agreement with the patient, a *–s* which is a personal ending referring to the agent and (13b) in Panjabi shows unmarked agent at the first and second person:

(12)a. *tyânî* *pothiâ* *lihiliâ*
3M.SG book-F.PL read-PST-F.PL
he read the book

[15] That is why *–l-* is observable in other tenses in Bhojpuri (present, past) and Pahari (future, past). Although Tiwari traces the origin of the future/present *–l-* in *lag* "touch", it is generally considered as a diminutive (*laghutâvacak*: Chatak), cf. Tessitori (*l* < *ll* < *ill*). Tiwari relates the past *-l-* to the one in *tonaïla* (< *tunda* + *illa*) "pot bellied man".

[16] In the various moods too.

[17] In which case a default agreement occurs (M.SG in Hindi/Urdu and Panjabi, N.SG in Marathi).

[18] The Marathi past ending always differs from the present one (*-s* also) since in the present, *–s* follows a vowel which varies according to the subject gender (*tu topi kâDh-t-os* : « you-M.SG take off the hat », *tu topi kâDh-t-es* : « you-F.SG take off the hat »), whereas in the anterior it follows a vowel referring to the patient (*tu topi kaDh-l-i-s* « you-M.SG took off the hat »). In the first person, Marathi like Panjabi in both first and second person, has unmarked agent and agreement with the unmarked patient.

(12)b. *one* *sanun* *tîn* *botlâ* *dîttiyâ*
3SG-ERG 1PL-DAT three bottle-F.PL give-F.PL
he gave us three bottles

(13)a. *tu* *kâm* *keleNs*
2SG work-N.SG do-PST-S.SG-2SG you worked/did the work
tu *pothî* *lihi.l.î.s* *tu* *pothiâ* *lihi.l.iâ.s*
2SG book-F.SG read-PST-F.SG-2SG 2SG book-F.PL read-PST-F.PL-2SG
you read the book you read the books

(13)b. *main (tû, tusî) ih* *kamîzâN* *kharîdîâN*
I (you) this shirt-F.PL buy-F.PL
I (you) bought these shirts

2.3. A similar shift in other Indo-European languages: from passive to active?

A very similar evolution has been studied by Kurylowicz for Persian (1953) and French (1931, 1965), and by Benveniste (1952, 1960, 1965), also for Persian and French. Like late Sanskrit, late Latin substituted to the old synthetic perfect a new periphrastic expression with the agent in the dative case (*dativus auctoris*), the patient unmarked and a passive past participle as a predicate (often followed by the copula).[19] The forms in Persian (14) are exactly similar to (3) in Sanskrit, including the lexical bases, except that the instrumental is not an option for the agent, always in the genitive case, and the Latin (15) is similar morpho-syntactically:

(14)a. *mana* *kardam*
I-GEN done-N.SG I have done [that]

(15)a. *mihi* *id* *factum*
I-DAT this-NOM-N.SG done-N.SG I have done that

Table 2 summarizes the analogies of the periphrastic perfects (I did / have done this) in the three ancient languages, which still accounts for the present state of HU:

	marked agent	unmarked patient	Verb-^{Patient}
	N1-oblique	N2-nom	verbal adjectif^{N2}
OIA	*mayâ*	*tat*	*kRtam*
OPer	*manâ*	*tya*	*krtam*
Latin	*mihi*	*id*	*factum*
NIA (W)	S-ergative	O-absolutive	Verb-^{OD}
	maiNne	*yah*	*kiyâ*

Whereas Persian later undergone the same evolution as Bengali, shifting the agent to the nominative case while adding new personal endings to the old participle (14b), Latin realised the same syntactic and semantic shift by using the "have" auxiliary, lacking in IA (15b):

(14)b. *man kardam* [I-NOM did-1SG], *to kardi* [2SG-NOM did-2SG], etc.

(15)b. *ego* *id* *habeo* *factum*
I-NOM this-N.SG have-ISG done-N.SG I have done this

(15b) is the structure now inherited by the modern Romance languages, such as French, with "have" verb conjugated in the present as an auxiliary for the present perfect and agreeing with the subject, before the participle, the latter still agreeing with the object in some cases:

[19] What is generally meant by perfect in the traditional grammar of Latin is the–*(v)i* form, usually translated by either as an anterior (*amavi* "I loved") or a present perfect ("I have loved"). The difference in both IA and Romance languages is that the old synthetic form was maintained and is still living as the simple past or aorist or definite past (various terminologies according to languages and centuries), only written in contemporary French but very common in spoken Spanish and Italian.

(15)c. *j'ai fait (cela), tu as fait (cela), nous avons fait*
I have-1SG (this) 2SG have-2SG done (this), 1PL have-1PL done[20]

Kurylowicz as most of the then scholars admitted the "passive" origin of the modern perfects derived from the passive past participle: "In the evolution that we consider, the decisive step is in the replacement of the dative + *esse* [be] + nominative by nominative + *habere* [have] + accusative. The passive construction has been transformed into an active one" (1931: 107). This is also the implicit assumption of Chatterji and Tiwari when they interpret the periphrastic renewal (nominative pattern) as an active conversion. Benveniste on the contrary argued for a "possessive" meaning of the perfect, aiming at both the ancient periphrastic expression and the present meaning ("le sens possessif du parfait"). One of his argument is casual: the genitive case used to represent the agent of the Latin or Persian perfect is also the possessive marker in both languages, distinct from the case used in Old Persian for the agent of passive verbs (*hacâma* in Old Persian, *a me* in Latin).[21] For instance *mihi filius est* (I-dat son-nom-ms is) "I have a son" or *mihi pecunia est* (I-dat money-nom-fs is) "I have money" is structured in the same way as "I did this" in (15a) and has been renewed in the same way as (15b) by the use of "have" verb, nominative subject and accusative object: *ego pecuniam habeo* (I-nom money-acc-fs have-1s). His other argument for the possessive reading is that the auxiliary "have" is also the stative verb which forms possessive statements: the older dative "possessor" has simply been transformed into a nominative possessor. That obviously the casual argument does not really hold for Sanskrit and Prakrits (instrumental is the agent in passive statements, and never expresses a possessor), does not entail that the general hypothesis is wrong. We come back to these problems and to the notions of possession and stativity later (section 4).

3. The modal future: a similar development

3.1. Parallel historical facts

But Kurylowicz's theory of the passive meaning of the old periphrastic passive allows him to grasp a very interesting analogy between perfect and future in the Romance languages. The development of the modern future in Romance languages also stems from a periphrastic renewal of the older synthetic Latine future (*amabo* "I will love"). This renewal occurred in Late Latin at the same time as the periphrastic perfect and on the same pattern: *mihi cantandum est* (Kurylowicz 1965) parallels *mihi factum est*, with a dative "subject", a passive verbal adjective or gerund, originally meaning obligation in *–nd-* (glossed OVA for obligative verbal adjective), agreeing with the patient if any (16a) or else in the neuter *–um* (16b).

(16)a. *mihi virtus colenda est* *mihi id faciendum est*
I-DAT virtue-F.SG cultivate-OVA-F.SG be-3SG I-DAT this-N.SG do-OVA-N.SG be-3SG
I shall/have to cultivate virtue I shall/have to do this

(16)b. *Carthago delenda est*
Carthago-F.SG delete-OVA-F.SG be-3SG
Carthago is to be destroyed
Carthago should/will be destroyed, (we) shall destroy Carthago

[20] With a preposed object: *les choses que j'ai faites, je les ai faites* (the things-F.PL which I have done-F.PL, them I have done-F.PL).
[21] "This difference in the casual form shows of the pronoun *manâ* on one hand, *hacâma* on the other hand, shows that the perfect must be interpreted as a category in its own right, altogether distinct from passive, it is an active perfect with possessive expression". (PLG1 179-80).

The Indo-Aryan data developed a strikingly similar structure, since in Asoka's times the obligative future (then the future) is expressed by an obligative passive participle in *–tavya* agreeing with the patient. (17) is the second part of example (6), again with a Western expression in Girnar (17a) and an Eastern expression in Jaugada (17b), identically patterned:

(17)a. *idha na kimci jîvam arâbhitpâ prajuhitavyam na ca samâjo kattavyo*
here no some living kill sacrifice. no and assembly do
NOM-N.SG CP OVA-NOM-N.SG NOM-M.SG OVA-NOM-M.SG

(17)b. *hida no kimci jive alabhitu pajohitavye no pi ca samâje kattavye*
one should not sacrifice by killing a living creature nor hold a meeting
(it should not be sacrificed by killing a living being nor a meeting should be held)

Table 3 summarizes these analogies in IE periphrastic forms for future and perfect:

		Marked Agent N1-instr/gen/dat	Unmarked Patient N2-nom	Verbe-^{Patient} verbal adjective
OIA	perfect	*mayâ*	*tat*	*kRtam*
P		*manâ*	*tya*	*krtam*
LATIN	perfect	*mihi*	*id*	*factum*
		mihi	*filius* (*liber*)	*est*
	future	*mihi*	*id*	*faciendum*
OIA		*mayâ*	*tat*	*kartavyam*

The last part of the story is exactly similar to what happened with the perfect: this passive (according to Kurylowicz) structure got transformed into an active one by shifting the dative/instrumental agent to the nominative, the patient to the accusative and using the auxiliary "have" (*habere*) after the infinitive:

(18)a. *ego cantare habeo*
I-NOM sing-INF have-1SG I shall sing
I have to do that/I will do that

(18)b. *ego id facere habeo*
I-NOM this-ACC-N.SG do-INF have-1SG I shall do it

And the modern future, although written today in one word, is clearly derived from the have construction since the personal endings paradigm of future in French for instance is the present of "have" verb:

(18)c. *je chanter-ai, tu chanter-as, nous ferons*
I sing-have-1SG, you sing-have-2SG, we sing-have-1PL
I will sing, you will sing, we will sing

The old system of (17) prevailed in the Magadhean languages up to around the 16th century. Transitive as well as intransitive have for their future, the old verbal adjective of obligation (OVA) in *–tavya* > *-abba* > *ab* > *b*) with an instrumental agent.[22] But the old modal meaning, quite perceptible in Late Sanskrit (19) is gradually lost in NIA and replaced by a temporal meaning of future as shown in Old Bengali (20a-b) or Old Awadhi (20b-c):

(19)a. *tribhir yâtavyam*
three-INSTR go-OVA-NOM-N.SG the three have to go

(19)b. *na kSeptavyâ brahma-vâdinâ na câvamânyâh*
neg neglect-OVA-NOM-M.PL Brahman-knower-M.PL neg contempt-OVA-NOM-M.PL
(you) should not neglect nor contempt those who know the Vedic word

(20)a. *maï dibi piricha (SK mayâ dattavyâ pRcchâ)*
I-INSTR give-*b*-F.PL question-F.PL 'I will ask questions" (Chatterji)

[22] The *–tavya* form is still present in some Hindi *tatsam* words, with its modal meaning, usually as nouns (*kartavya* "what has to be done, duty").

(20)b. *Thakiba, khaïba maï*
stay-*b*-∅ eat-*b*-∅ I-INSTR I will stay, eat
(20)c. *ghar kaise paithaba maï*
house how enter-*b*-∅ I-INSTR how shall I enter?
(20)d. *sukh lahab râm vaidehî*
bliss get-*b*-∅ Ram Vaidehi Ram and Sita will find happiness

The later evolution of these *–b-* futures has been similar to the evolution of perfects in the East: personal endings were added to the participle, similar to the perfect endings and distinct from the present ones in Bengali, in parallel with the shifting to a nominative structure:

(21) *âmi boiTâ porbâm, tu porbi, tumi porbe*, etc.
1s book-DEF read-*b*-1SG, 2nonH read-*b*-2nonH, 2 read-*b*-2
I will read the book, you will read, etc.

In Bhojpuri too and Awadhi, Saxena (1937: 261) notes that the *–b-* future was generalized in ancient NIA in the region, before the re-introduction in Western Awadhi of the sigmatic forms for the 1st and 2nd persons.[23]

The Western dialects in contrast either maintained or re-acquired the sigmatic synthetic future or developed a periphrastic future with a "go" verb (*gachati*), added to the synthetic open form for non past, like Hindi/Urdu *calegâ* "he will go".

3.2. Modalities and the non-nominative pattern

This striking parallel in Bengali between past and future shows, as already argued by Kurylowicz, that perfect and future share a common evolution which suits a common meaning. Benveniste opposed this view and denied any relation at the semantic level between future and the obligative participle.[24] But many various languages show a possible grammaticization of an obligative form in the meaning of a future (Heine 1993), an the IA data is a particularly clear evidence of such a development. Kurylowicz (1965) maintained that both future and perfect evolved on similar lines from passive nominal structures (X been done, X to be done) to auxiliated active structures with "have" (have this done, have this to do) because they are both views over the process from the present utterance time: "future and past structures are originally forms of present, they are related to the time and situation of utterance. They do not express action, but the need or intention to act, and the present result of an action which has already been accomplished".

The link between the old nominal obligative structure and perfect is confirmed by the Marathi data in a different way, since Marathi does not exhibit a future of the Bengali type. But it maintained the old obligative verbal adjective, in modal structures closer to the original than in the Magadhean modern languages : potential and obligation not only maintain the *–âv/av-* morphology inherited from the *–tavya* verbal adjective, they also maintain the old syntax with an instrumental subject (Joshi 1900: 468) and they also allow interesting case alternations. Bloch already noted that the "the use of these forms is similar to that of the form for past" (1935: 264), on the basis of the obligative statement borrowed from Joshi, where the "logical subject" *ahmî* is instrumental and *karâveN* agrees in the neuter-singular:

(22) *ahmî kây karâveN*
I-INSTR what do-*âv*-N.SG what should I do?

The pair in (23), from Joshi (1900), with obligative meaning, shows the "active conversion" of this "passive" structure in a way very similar to what happened in Bengali. (23a) is a quasi

[23] The sigmatic Sanskrit future (*Sy.ati* > *-s-* > *-h*) was retained in some Western languages like Western Rajasthani, but also in Awadhi at certain persons.

[24] According to him, *habere* in Late Latin future was only used in the past, with passive infinitive, to express a predictive meaning, specially in the Christian predication; the meaning "have to" could in no way produce a future meaning and was never confused, and still today is never.

ergative alignment and *neN* marker although the verb is intransitive, agreeing in the neuter whereas (23b), still competing in the 19th century, shows a nominative alignment with a verb agreeing with its nominative subject:[25]

(23)a. *tyâneN* *ghariN* *yâveN*
3M.SG-ERG home-LOC come-OBLIG-N.SG he should come home

(23)b. *to* *ghariN* *yâva*
3M.SG-NOM home-LOC come-OBLIG-3M.SG
he should /may he come home

In contemporary Marathi, although according to Pandharipande, ergative (agent) case can also have the optative meaning ("he may go home" is the translation she gives for *tyâne gharî dzâwe*), according to other modern writers there is now a difference in meaning, the ergative pattern being obligative while the nominative one is "optative" (Wali 2004: 31), "may he come home". The next series in (24) illustrates the potential modality, also derived from the obligative verbal adjective, also allowing casual alternation. The alternation here is between two oblique forms within the same syntactic pattern, the dative and the "instrumental", according to Joshi and Pandharipande, who however glosses the same *ne* as agent in obligative statements (1997: 438, 434):

(24)a *majhyâneN* */malâ* *câlavleN*
I-INSTR / I-INSTR go-POT-PST-N.SG I could/was able to go

(24)b. *majhyâneN* / *malâ* *dhadâ* *sikhavlâ*
I-INSTR / I-DAT lesson-M.SG learn-POT-PST-M.SG
I was able to learn the lesson

(24)c. *titSyâne* / *tilâ* *bharbhar* *tsâlvât* *nâhi*
3F.SG-ERG / 3F.SG fast walk-POT NEG
she cannot walk fast

It is however remarkable that *neN*, whether identically glossed or not, a single morphological unit with a single origin (see infra), alternates with both dative and nominative markers for the main participant. Examples (23) and (24) are a further argument to regard the modal system originated from the *–tavya* verbal adjective as a parallel structure to the perfect pre-ergative or ergative structures, a fact clearly captured by Bloch in the early 20th century (1920). At the same time, they are a further argument, too, to consider the ergative IA pattern as part of a larger way of mapping non action, instead of viewing it as an aspectual split.

4. Place of these evolutions within the global economy of the NIA system

4.1. Parallel patterns for what is aimed at, accomplished, experienced

Benveniste, who also claimed that future and past do not represent tense but "views on time from the present" (1965), is however only concerned with perfect since he does not recognize any deep or interesting analogy with the development of futures. But he clearly states that the "so-called" passive structure, in fact according to him a possessive structure with its *dativus auctoris*, is a stative one. Instead of viewing the "avoir/have" conversion as a converting device from passive to active (as did Kurylowicz), he regards it as a device for "inversion". The idea stems from the possessive statement which in Latin patterns as the periphrastic future (table 3): "*avoir* is nothing else than a "be-to" inverted (*mihi est pecunia-money = habeo pecuniam*). The nominative is not an agent but the localizer of a state,[26] seemingly transitive but in reality intransitive and stative". Similarly when used as auxiliaries as in the perfect "I have done" (Benveniste 1960: 197).

[25] Significantly, as in the past, the verb adds a –s personal ending for the second person.
[26] In French, "un siège d'état".

The above formulation makes the expression of perfect one among other stative predications of localization. Viewed under this light, the term of "possessive" applied to perfect is understandable, providing we do not over-semanticize it and read it as a label for "have" sentences in general, most of them are indeed stative and only some possessive. The 'be' to 'have' "inversion" which transforms a dative alignment into a nominative alignment retains the static feature and the semantic role of localizer of the first nominal (in the dative or nominative). Adapted to the ergative IA pattern which is the continuation of the 'be' structure, the periphrastic perfect commented by Benveniste as a stative, not passive structure, such an analysis suggests that the *ne* sentences too are localizing predications,[27] similar to (25a) for obligative predicates with verb "be", perception or cognitive predicates (25b) and more generally experiential statements, transitive and intransitive (25c):

(25)a. *mujhko jute kharîdne hoNge*
I-DAT shoe-M.PL buy-INF-M.PL be-FUT-M.PL
I will have to buy shoes

(25)b. *mujhko choTe choTe ghar dîkh rahe the*
I-DAT small-M.PL small-M.PL house-M.PL appear PROG-M.PL be-PAST-M.PL
I saw (was discovering) houses

(25)c. *mujhko Thand hai*
I-DAT cold-F.SG be-PRS-3.SG
I am cold (French "j'ai froid")

The series (25) morpho-syntactically patterns exactly as (4a) and (6), even when the predicate is a single participant one since in HU such predicates usually consist in verbo-nominal expression (NV) and the verb agrees with N. Similarly, possessive statements (with locatives) present a stative verb, mostly "be", which agrees with the object possessed, and the possessor, although the main participant in the first position, is marked (*ke pâs* "near", *meN* "in") and does not control agreement. Significantly, the equivalent of type (25) statements in Romance languages involves the verb 'have' more often than in English and Benveniste includes these statements too in his analysis of the "possessive perfect".
Table 4 summarizes the analogies between the various types of predications of localization:

OIA	agent-INSTR	patient-NOM	verbal.adjective^{patient}
	mayâ	*tat*	*kRtam/ kartavyam*
Latin	*mihi*	*id*	*factum /faciendum*
NIA (W)	agent-*ne*	patient-NOM	Verb^{patient}
	maiNne	*yah*	*kiyâ*
	experiencer-*ko*	theme-NOM	Verb^{theme}
	mujhe	*yah*	*dîkhâ*

4.2. The cognitive scenarios of non-transitive processes

This suggests more affinity with an intransitive model than with a transitive one. If we come back to the aspectual semantics of perfect (emphasis on result), it is a well-known fact since DeLancey (1981), who first associated both ergative and dative experiential statements, that aspectual semantics requires the viewpoint to be associated with the result (goal) and not with the source at the " natural " origin of the process, which is encountered secondarily (hence marked), upstream so to speak. In this logic, the source no longer retains the same relation with the process and its goal: in the standard transitive model, the source is the natural start-point of a process ending on the goal (endpoint), whereas in the ergative pattern the

[27] More details in Montaut 2004b.

source is outside the predication, which has the goal as its start point. This means that the ergative case is not a simple grammatical marker used to reverse the same trajectory, within the same cognitive scenario. The trajectory itself is a different cognitive scenario. As Langacker (1999: 35) puts it, ERG encodes an altogether different relation, involving a different perceptive strategy, thus being rather a semantically significant case and " only incidentally associated with grammatical relations " (cf. section 4.3). It only profiles the last part of the clause as "onstage" (the " trajector " and main figure being the patient), in an autonomous way (not dependant on the source), whereas a nominative transitive alignment profiles the full path (the "trajector" and main figure being the agent) and builds the relation as dependant on the source. The ergative pattern is then more like an intransitive structure, corresponding to what Langacker calls a thematic relation ('the ice melted', profiling only the end part of the action chain, whereas 'Bob melted the ice' profiles the whole chain). As a thematic relation, " it enjoys a certain autonomy vis-à-vis the agent and the flow of energy, even for inherently energetic processes", and is thus an "absolute construal" (Langacker 1990 : 245-8). The starting point has conceptual autonomy from the source, a reason why " the path involved is more abstract and of lesser cognitive salience ". Both structures are thus shown to differ deeply, and not only at the morphological level.

The affinity with intransitive patterns is evidenced by Hindi/Urdu examples such as (26), where 26b) in the ergative may give particular emphasis to the resulting state (26c) by adding the past participle of "be" to the predicate ('is having been done'), in a quasi equivalent meaning as the intransitive nominative pattern (26a):

(26)a. *maiN unse mitratâ banâe hue hûN*
I-NOM 3PL-with friendship-F.SG make-caus being be-1SG

(2)b. *maiNne unse mitratâ banâî hai*
I-ERG 3PL-with friendship-F.SG make-PRF-3F.SG
I have made friendship with him

(26)c. *sîtâ ne aTahârû pahne hue the / sâRî pahnî huî thî*
Sita ERG earing-M.PL wear-PP been be-M.PL / sari-F.SG wear-PP been be-F.SG
Sita was wearing (had put on) earings /a sari"[28]

Whereas table 4 showed tripartite models, things could then be reformulated in a binary model with the localizer outside the profiled relation, which itself is basically intransitive and mapped into an "absolute construal" (Langacker's terms) into table 5:

[agent-ne]	**patient-nom**	**Verb$^{patien\ t}$**
[experiencer-ko]	**theme-nom**	**Verb^{theme}**

4.3. Case semantics

Now, if the forms inherited from the *–tavya* participle may encode this localizer in the dative as well as in the ergative (Marathi data), the alternation makes it dubious that ergative is basically a marker for voluntary controlled action. The volition-control feature is certainly

[28] We may say that "*huâ*" is not a specific marker for stativity since we also find it with unaccomplished participles, as in *vah gâtâ huâ â rahâ thâ* (3s singing *huâ* come PROG PST) "he was coming (while) singing" where it simply marks concomitance. But the relation between resultant state (perfect) and concomitance is well known (Cohen 1992), both marking the link of the process with the situation of reference (set by utterance), either through a relation of inherence (progressive: being in the process) or by a relation of adjacency (perfect: being with or after the process). (26c), like (26b) can be substituted by the intransitive :
Sitâ aTahârû pahnî huî (pahne hue) thî
Sita-fs earing-mp worn been-fs (worn been-adv) was Sita was wearing earrings

present in a massive majority of ergative statements, but it is probably linked with the semantic class of transitive predicates, rather than with the case marker,[29] since transitive basis in HU are generally + volitional or + consciousness/awareness. In contrast, the use of dative refers to lack of conscious awareness, as shown in (27): the ergative/nominative statement only involves conscious awareness rather than a deliberate choice, whereas the dative statement rules it out:

(27)a. **us din maiNne tumse irSyâ kî thî par iskâ bodh nahîn thâ*
that day I-ERG 2-with jealousy do PPRF but this-of awareness NEG was

(27)b. *us din mujhe tumse irSyâ huî thî par iskâ bodh nahîn thâ*
that day I-DAT 2-with jealousy be-PPRF but this-of awareness NEG was
that day I felt jealous from you but I was not conscious of it

When alternating with nominative case as in Marathi (23a), ergative (glossed either as such or as instrumental by linguists) is obligative, whereas nominative is optative or epistemic (Wali 31), which refers to a "demand" or "wish" from the speaker and not from the subject in the non-first person. Here ergative appears less "volitional" than nominative. In Delhi Hindi (DH), Hindi the use of the ergative marker has developed for obligative statements as (28), supposedly under the influence of Panjabi (*ne* ergative, *nuN* dative), competing with the standard Hindi construction in the dative (25a).

(28) DH *maiNne jânâ hai*
I-ERG go-INF is I have to go
SH *mujhe jânâ hai*
I-DAT go-INF is

While it sometimes conveys a "conscious choice" (Butt 1994) as opposed to the standard dative construction, it has been proved (Bashir 1997) to also convey different meanings varying according to the person of the verb and to the context, including a "prospective, anticipated, injunctive" meaning, which is consistent with the modal nominal pattern of (x).

But the very fact that dative and ergative can alternate in patterns like (29) and that closely linked languages have either one or the other case for obligative statements suggests that there is a deep affinity between dative and ergative. For example, Pahari in both its regional variants Garwhwali (29a) and Kumaoni (29b) use only the ergative marker in the "obligative future", expressed by a bare infinitive, where standard Hindi/Urdu use the dative. Garhwali uses *na* or *la*, and Kumaoni uses *le*:[30]

(29)a. *maiNna /maiNla âj barat rakhNa* I have to fast today

(29)b. *maiNle âj barat rakhNa*
I-ERG today fast keep-INF

(29)c. *mujhe âj vrat rakhnâ hai*
I-DAT today fast keep-INF is

All these facts of alternation suggest that there is no polar opposition between *ne/le* and *ko/la*, the markers for ergative/dative, although in many contexts they convey distinct and even opposed meanings. The instrumental use of *ne/ni* in Marathi (for inanimate cause and instruments), hence the gloss, as well as the interpretation of the ergative structure as passive, with instrumental agent, wrongly represent the case marker as a source, opposed to the dative (goal). But the historical evidence for the origin of both tales a different tale, more in conformity with Benveniste's "possessive" reading and my own analysis as a localizer for stative predication.

---

[29] Ergative predicates like *maiNne dekhâ* "I saw" (aside with "I looked"), *maiNna pâyâ* "I found", *maiNne mahasûs kiyâ* "I felt" make it clear that ergativity in Hindi is not always associated with volitionality and control.

[30] Both languages are classified as belonging to the PahâRî Madhy BhâSâ, Garhwali probably more influenced by Hindi since the traditional ergative marker *la/le* tends to commute with *na* in urban places. The obligative future (*bhaviSyat kâl*) is considered by Juyal (1976) as passive in meaning *karNîy arth*).

4.4. Origin of the markers

First of all, it is obvious than the ergative *ne/ni* can in no way originate from the Sanskrit instrumental *–ena*, even reinforced: Hindi *main* may reasonably be assumed to derive from a reinforcing of the classical instrumental form *mayâ* via **mayena* (Chatterji: 744) and shows only a nasal ending vowel, as all forms derived from the Sanskrit *–ena*. It does not seem to have appeared before the end of 14th century (Namdev has *tâyaneN*) and was not generalized then. In the early century Konkan, the *n, na, nî* form means "to" and similarly *ne* in Bhili, *ne/nai* in Rajasthani has both meanings "by" and "to" (Grierson). Today *nûN* means "to" in Panjabi and *ne* is the agent marker. The etymology of this obviously single form has been extensively discussed and sometimes associated to *nyâya* (manner < rule), questioned by Bloch (1914) who does not suggest an alternative. The most convincing etymology is traced by Tessitori (1913; 1914: 226-7), according to whom *nain, naï, nî, ni, ne* is a shortening of *kanhaïN* found in Old Rajasthani texts. *KanhaïN* (<Apabramsha *kaNNahî*) comes from the reconstructed * *karNasmin* (< Sanskrit *karNe*), a locative form meaning "aside, near". Trumpp (1872: 401) also gives the original meaning "near" for *naï/ne*. This meaning, according to Tessitori, "may be understood either in the sense of the locative "Near to" or of the accusative-dative "Towards, to". The second meaning is the origin of the Western marker for goal (Panjabi *nûn*), and the first one of the ergative markers of the *ne* type, clearly a locative.

As for *le/la*, which in Pahari (and modern Nepali) is the agent marker and the instrumental (allomorphs *–l, al, lè*),[31] it is assumed by most to derive from *lagya* > *lege* >*laï*, *le* "having come in touch with", "for the sake of", "with the object of" (Juyal 1976). We may notice the similar origin for the dative marker *lâ, la* (Marathi), from *lag*, (> *lâgi*, "up to, for the sake of"), according to Turner (Old Marwari *lag* "up to, until":[32] it is obvious that both locative and dative, although quite distinct now in most IA languages, stem from a common notion of vicinity and adjacency, presented either as dynamic (entity aimed at: dative, goal or patient) or non dynamic (localizer of the process: ergative).

Originally, both *ne* and *la* markers are then semantically quite close, and these facts make the IA date even closer to the Latin data.

Conclusion

The above data for perfect and future compared with experiential patterning, do not of course amount to say that ergative statements are presently perceived as states, no more than was the Latin periphrastic perfect once grammaticized as a perfect. No more did Benveniste's "possessive" perfect really meant that perfect was perceived as the possession of a result by an agent. But it shows that a similar logic has restructured all predications that were not actual processes (such as processes aimed at or accomplished, or experienced states) into localizing predications. In NIA, most of the localizing predications with two participants came to be represented as non-nominative statements, a historical development which amounted to split grammatical subject properties and syntactic, semantic or pragmatic subject properties on two separate entities.[33] Whereas in Romance languages this gap has been overcome by the "have" restructuring, allowing topic, subject and agent to coincide in a grammatical subject, IA languages, lacking a "have" verb, still display a subjectless patterning for most of these

---

[31] *hamanle callo mâr cha* [1PL-ERG bird-M.SG strike PRF-M.SG] « we killed the bird »]
apnâ hâtel khan banuni [REFL-OBL hand-INSTR food make-PST]"(they) prepared food by their hand"

[32] Against Tiwari, who suggests a possible derivation from *labhati* "acquire, benefit".

[33] « Coding properties » in Li's (1976) terms (case marking, agreement), *vs* syntactic (control), semantic (agentivity, animacy) and pragmatic (topic) properties.

predications.[34] Western NIA is in this respect more "conservative" than Eastern NIA, which has differently restructured its modal and perfect statements into a nominative pattern. Given the historical evolutions above mentioned, useless to say that the relation between unmarkedness and core meaning is to be used cautiously: shall we say that in Hindi/Urdu the preterit is the unmarked form, then, anteriority is the basic meaning for tense, because there is no tense-aspect-person mark, as opposed to present for instance, whereas in Bengali, with a similar history of grammaticization upto the 16th century, perfect was already marked by *–l-* and personal endings got added to the form, hence marked more than optative? Still forms are indicative of paths of grammaticization, if not, at least not directly, of the cognitive domains they are supposed to map.

References

Bashir, E.. 1999. The Urdu and Hindi Ergative Postposition *ne*: its Changing Role in the Grammar. *South Asian Languages Yearbook* :11-36.

Beames, J. 1970 [1871]. *A Comparative Grammar of the Modern Aryan Languages of India*. Delhi: Munshiram Manoharlal.

Benveniste, E.[1952]. La construction passive du parfait transitif. [1960], Etre et avoir dans leurs fonctions linguistiques. Reprinted in 1966. *Problèmes de linguistique générale* 1. Paris : Gallimard (176-86; 187-207).

Benveniste, E.. 1965. Les transformations des catégories grammaticales. Reprinted in 1966. *Problèmes de linguistique générale* 2. Paris : Gallimard (127-136).

Bloch, J.. [1970] 1914. *The Formation of the Marathi Language*. Delhi: Motilal Banarsidass

Bloch, J.1906. *La Phrase nominale en sanscrit*. Paris: Champion

Bubenik, V. & Ch. Paranjape. 1996. Development of Pronominal Systems from Apabhramsha to New Indo-Aryan. *Indo-Iranian Journal* 39 (11-32).

Butt, M., 1993. Conscious Choice and Some Light Verbs in Urdu. *Complex Predicates in South Asian Languages*. Delhi: Manohar (31-46).

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

Bybee, J. 1994. The Grammaticalization of zero. *Current Issues in Linguistic Theory* (*Perspectives on Grammaticalization*) 109 (237-54).

Bybee, J., R.D. Perkins & W. Pagliuca (eds.). 1994. *The Evolution of Grammar, Tense, Aspect and Modality in the Languages of the World*. Chicago-London: Chicago Univ.Pr.

Cardona, G. 1970. The Indo-Iranian Construction *mana* (*mama*) *krtam*. *Language* 46 (1-12).

Carey, K. 1994. The Grammaticalisation of the Perfect in Old English. In Pagliuca N. (ed.), *Perspectives on Grammaticalization*. Amsterdam: Benjamins (103-116).

Chatterji, S.K. 1986 [1926]. *The Evolution of Bengali Language*. Delhi: Rupa (3vol.).

Cohen, D. 1989. *L'Aspect verbal*. Paris:PUF.

DeLancey, S. 1981. An interpretation of split ergativity and related patterns. *Language* 57.3 (626-57).

Desclès, J.P. 1980. Construction formelle de la catégorie d'aspect. In *La Notion d'aspect*, J. David et R. Martin (eds.). *Recherches Linguistiques* 5 (198-237).

Desclès, J.P. 1992. La prédication opérée par les langues. *Langages* 103 (83-96).

Dixon, R.M.W., 1994, *Ergativity*, Cambridge University Press.

Dymshits, Z. 1985. *Vyavahârik Hindî VyâkaraN*. Delhi: Rajpal.

Garcia, E. & F. van Putte. 1989. Forms are silver, Nothing is Gold. *Folia Linguistica Historica* VIII-1-2 (365-84).

Grierson, G.A. 1903-28. *Linguistic Survey of India*, Delhi (reprint): Motilal Banarsidass.

Heine, B. 1993. *Auxiliaries. Cognitive Forces and Grammaticalisation*: Oxford: OUP.

[34] For a discussion of « subjectless », see Kibrik (1997), and for a view of Hindi as a subjectless language according to these lines see Montaut 2004b.

Juyal, G. 1976. *Madhya Pahari Bhasha (Garhvali Kumaoni) ka anushilan aur uska hindi se sambandh*. Lucknow: Navyug Granthagar.
Kellogg, R. 1875 [1972]. *A Grammar of the Hindi Language*. Delhi: Oriental Book Reprints.
Kibrik, A. 1997. Beyond Subject and Object: towards a Comprehensive Relational Typology. *Linguistic Typology* 1 (279.346).
Kurylowicz, J., 1953 [1960]. Aspect et Temps dans l'histoire du persan. And 1931 [1960]. Les Temps composés du roman. *Esquisses Linguistiques*, Krakov: Polska Akademia Nauka (109-118, 104-108).
Kurylowicz, J. 1965. The Evolution of Grammatical Categories. *Diogène* 51 (51-71).
Langacker, R. 1999. *Grammar and conceptualization*, Berlin-New-York: Mouton de Gruyter.
Langacker, R. 1990. *Concept, image and symbol*, Berlin-New York: Mouton de Gruyter.
Montaut, A. 2006. Mirative Meanings as extensions of aorist in Hindi/Urdu. *Yearbook or South Asian Linguistics*. Amsterdam: Mouton (71-86).
Montaut, A. 2004b. Oblique main arguments in Hindi as localizing predications. In *Non nominative Subjects* (eds. Bhaskararao & Subbarao). Amsterdam: Benjamins (33-56).
Montaut, A. 2004. *Hindi Grammar*. Munchen: Lincom-Europa.
Montaut, A. 2005. Colonial language classification, postcolonial language movements and the grassroot multilingualism ethos in India. In Mushirul Hasan & Asim Roy (eds.). *Living together separately. On the historicity of India's composite culture*. Delhi: Oxford University Press (75-106).
Nespital, H. 1980. Zur Aufstellung eines Seminventars der Tempus Kategorie im Hindi und Urdu und zu seiner Charasteristik. *Zeitschrift der Deutschen Morgenländischer Geselschaft* 130-3 (490-521).
Nespital, H. 1986. Zum Verhältnis von Genus Verbi, Nominativ- und Ergativ- Konstruktionen im Hindoarischen aus synchroner und diachroner Sicht. *Müinchener Studien zur Spachwissenschaft* 47 (127-58).
Nespital, H. 1997. *Dictionary of Hindi Verbs*. Lokbharati Prakashan: Allahabad.
Ojha, T. 1987. *Pramukh Bihârî boliyon kî tulnâtmak adhyâyan*. Varanasi: Vishvavidyalay Prakashan.
Saxena, R.B., 1937, *Evolution of Awadhi*, Delhi (reprint), Motilal Banarsidass
Tessitori, L. 1913. On the Origin of the Dative and Genitive Postpositions in Gujarati and Marwari. *JRAS* (553-67).
Tessitori, L. 1914. On the Origin of the Perfect Participles in *1* in the Neo-Indian Vernaculars. *Indian Antiquary* (225-36).
Tiwari, U.N. 1966. *The Development of Bhojpuri*. Calcutta: The Asiatic Society vol. X.
Tiwari, U.N. 1961. *Hindi Bhasha ka udgam aur uska vikas*. Prayag: Bharati Bhandar.
Trumpp, E. 1872. *Grammar of the Sindhi Language*. London-Leipzig.
Wali, K. 2004. *Marathi*. Munchen: Lincom-Europa, Languages of the World/Materials.

CREATING RAISING VERBS
An LFG analysis of the Complex Passive in Danish

Bjarne Ørsnes
Copenhagen Business School

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)
2006
CSLI publications
http://csli-publications.stanford.edu/

Abstract

Raising configurations are sometimes described as the result of a diachronic process of semantic bleaching whereby an equi verb loses a semantic component of VOLITION. The verb is left with only a propositional argument in its argument structure and has to raise the subject of this embedded argument in order to fulfil a syntactic requirement that every predicator has a SUBJ ([Barron,1999]). Building on Barron's analysis, I argue that raising configurations may also arise as the result of argument structure operations and not just as the result of semantic changes. This means that verbs that are not otherwise raising verbs, may function as raising verbs if the right argument structure properties are present. Such a case is passivisation of verbs with propositional complements, where the most prominent argument is suppressed, leaving the verb with a propositional argument and no most prominent argument. The object of investigation is the Complex Passive in Danish.

1 Introduction

In Danish as in several other languages, verbs taking infinitival complements fall into two groups.[1] In the first group the subject of the matrix verb is not a thematic argument of the matrix verb, but a thematic argument of the embedded infinitival complement. These verbs are referred to as raising verbs. In the second group, the subject of the matrix verb is a thematic argument of the matrix verb as well as a thematic argument of the embedded infinitival complement. These verbs are referred to as control or equi verbs.[2] Several diagnostics distinguish these two groups: raising verbs may occur with an expletive subject while equi-verbs do not allow expletive subjects, equi-verbs license agent-oriented adverbials while raising verbs do not allow agent-oriented adverbs and raising verbs preserve the (non-compositional) meaning of idioms while equi-verbs crucially do not. In addition, the infinitival complements of these two classes of verbs also exhibit distinct syntactic behaviour as regards passivisation, topicalisation and substitutability with DP-objects.

As noted by several researchers the borders between raising and equi verbs are not clear. Verbs such as *begin* and *want* exhibit both raising and equi properties depending on the context. The topic of this paper is another case of fuzzy borders between raising and equi verbs: equi verbs that turn into raising verbs due to morpho-lexical operations altering the argument structure of the verbs.

While raising verbs do not passivise at all, equi-verbs in Danish allow two kinds of passives. Consider the examples below:

(1) Peter forsøger at reparere bilen
Peter is.trying to repair the.car

[1]For valuable comments and suggestions I wish to thank Line Mikkelsen, the anonymous reviewers of LFG06, the audience at LFG06 as well as the editors of the proceedings.

[2]In LFG the relation between an antecedent and a covert co-referential subject is termed control. This means that the relation between a raised constituent and a covert subject is also a relation of control. To avoid confusion, the group of verbs with infinitival complements and thematic subjects is referred to as equi-verbs and not control-verbs which is otherwise common in the literature.

(2) a. at reparere bilen forsøges / der forsøges at reparere bilen
to repair the.car is.tried / there is.tried to repair the.car

'as for repairing the car an attempt is made / an attempt is made to repair the car'[3]

b. bilen forsøges repareret
the.car is.tried repaired

'as for the car, an attempt is made to repair it'

Example (1) shows the equi verb *forsøge* in the active construction. When the verb is passivised as in (2) two different kinds of passives are observed. The passive in (2a) follows the pattern of passivisation of transitive verbs: the (infinitival) complement raises to subject (the personal passive) or the (infinitival) complement retains its grammatical function and the passive matrix verb occurs with an expletive *der* 'there' (the impersonal passive).[4]

In example (2b), however, the internal argument of the embedded verb is raised to subject of the passive matrix verb while the verbal complement surfaces as a (passive) past participle. The passive in (2b) is referred to as the Complex Passive.

An analysis of the Complex Passive should account for two crucial features of this construction. In the Complex Passive an internal argument of an embedded verb is raised to subject of the passivised matrix verb. As we will see in the next section there is no corresponding active construction where the internal argument of the embedded verb surfaces as the object of the matrix verb. The construction thus appears to violate a locality condition since heads generally do not "have access to" the internal structure of their complements, in this case the arguments of the embedded verb. Secondly the analysis should account for the fact that the infinite complement of the Complex Passive is realized as a passive past participle while the active form of the matrix verb requires an infinitival complement. Moreover, an analysis of this construction should uncover the syntactic and semantic constraints on this construction, i.e it should uncover whether any particular constraints pertain to the matrix-verb and to the embedded verb, and it should uncover to what extent these constraints follow from the specific properties of the construction.

2 A first characterisation of the Complex Passive

The Complex Passive is complex in the sense that it is composed of two passive verbs: a passive matrix verb followed by one or more passive past participles. The matrix verb must be a verb selecting a verbal complement, i.e. it must select a propositional argument. As shown in the examples below, the matrix verb may appear in both the synthetic and the periphrastic passive with

[3]Throughout the paper, the translation "as for X" is meant to indicate the topicality of a referential subject in the Complex Passive.

[4] Actually two kinds of impersonal passives are observed with either *der* 'there' or *det* 'it'. The exact analysis of these two kinds of impersonal passive is unclear. The passive with *there* appears to be an instance of a presentational *there*-sentence. Intransitive verbs may appear in presentational *there*-sentences with an indefinite postverbal complement. Passive verbs with sentential complements are intransitive and given that sentential complements do not carry definiteness, the verbs may appear in a presentational *there*-sentence. The impersonal passive with *det* 'it' on the other hand appears to be an instance of extraposition of a clausal complement. However, the exact analysis is not of immediate concern here.

the semantic and pragmatic differences observed for these two passive forms ([Engdahl, 1999]). Interestingly, if one of the embedded participles is itself a verb selecting a verbal complement as the verb *forsøge* 'to try' in (5) below, a further passive past participle may occur as its complement, thus giving rise to a (in principle infinite) recursive embedding.

(3) bilen forsøges repareret
the.car is.tried repaired
'as for the car an attempt is made to repair it'

(4) bilen blev forsøgt repareret
the.car was tried repaired
'as for the car, an attempt was made to repair it'

(5) bilen blev lovet forsøgt repareret
the.car was promised tried repaired
'as for the car, a promise was made to try to repair it'

On the basis of the examples in (3) through (5) the Complex Passive may be schematically depicted as in (6) below ([Hellan, 2001]): A passive verb followed by at least one passive past participle.

(6) SUBJECT V_{pass} (synthetic or periphrastic) $V_{pastpart\_passive}+$

Crucially, the Complex passive has to be distinguished from the superficially similar construction of a passivised ECM-construction (subject-to-object raising). Consider the example in (7). The construction conforms to the schematic characterisation of the Complex Passive above: a passive matrix verb followed by a passive past participle. However, the example in (7) does have an active counterpart where the subject of the passive verb *forvente* 'to expect' surfaces as the object of the matrix verb followed by a passive past participle. The relevant example of the active construction is given in (8).

(7) forslaget forventes vedtaget
the.proposal is.expected adopted
'the proposal is expected to be adopted'

(8) man forventer forslaget vedtaget
you expect the.proposal adopted
'everyone is expecting the proposal to be adopted'

The apparent Complex Passive in (7) may thus be derived by raising of the object to subject as in the canonical case of a passivised transitive verb. For this reason there is nothing special about the passivisation in this case. However, the true Complex Passive does not have an active counterpart where the subject of the Complex Passive surfaces as an object followed by a passive past participle as shown below.

(9) * Peter forsøger bilen repareret
Peter tries the.car repaired
'Peter is trying to repair the car'

Since verbs such as *forvente* 'to expect' do have an active counter-part, I do not consider examples such as (7) Complex Passives. A Complex Passive is thus a syntactic construction conforming to the schematic characterization given in (6) for which there is no active counterpart where the subject surfaces as an object of the active matrix verb followed by a passive past participle.

3 A closer look at the Complex Passive

The Norwegian Complex Passive has been extensively discussed in the literature ([Hellan, 2001, Engh, 1995, Nordgård and Johnsen, 2000]). It even occupies a whole section in the Norwegian Reference Grammar ([Faarlund et al., 1997]). For some reason the Complex Passive in Danish does not seem to have been discussed at all and it is not even mentioned in the traditional grammars of Danish ([Diderichsen, 1957, Robin Allan and Lundskær-Nielsen, 1995, Hansen, 1967]). Due to the lack of previous discussions of the Complex Passive in Danish, I will draw heavily on the accounts of the Norwegian Complex Passive. Interestingly there turn out to be significant differences between the construction in the two languages.

Also due to the lack of comprehensive studies of the Complex Passive in Danish, the present analysis is based on an extensive corpus investigation of the syntax of verbs selecting sentential complements. Approximately 125 verbs with sentential complements were randomly selected. All verbs were searched in Korpus2000[5] and on Danish web-pages (through Google) and analysed in their syntactic context. For each verb the following properties were recorded: complementation in the active, ability to occur with an expletive subject in the active, ability to form personal and impersonal passive, ability to occur with an ECM-construction and ability to form the Complex Passive. Also information on complementisers heading finite complements was recorded. All results were stored in a database and the following discussion is based on the main findings of this investigation as regards the Complex Passive.

3.1 Constraints on the matrix verb

Of the approximately 125 investigated verbs with sentential complements (supplemented with occasionally observed verbs forming Complex Passives), 12 verbs form the Complex Passive. The constraints on these verbs emerging from the data are presented below.

Obligatory control verbs The matrix verbs forming the Complex Passive are obligatory control verbs in the sense of [Culicover and Jackendoff, 2005]. The verbs select infinitival complements, denoting controlled actions. Both subject and object control verbs form the Complex Passive as shown below where representative samples of the Complex-Passive-forming verbs are given.

- Subject control verbs: *forsøge* 'to try, *agte* 'to intend', *simulere* 'to pretend' ...
- Object control verbs *bede* 'to ask to', *pålægge* 'to force to', *forbyde* 'to forbid' ...

[5]Korpus2000 is a corpus of contemporary Danish provided by *Det Danske Sprog- og Litteraturselskab*: http://korpus.dsl.dk.

[Holmberg, 2002] arrives at a different generalisation concerning the verbs forming Complex Passives in Norwegian. He claims that verbs forming the Complex Passive are restructuring verbs taking a somehow reduced or defective sentential complement. A property of this class of verbs is that they license non-thematic subjects also in the active as shown in (10) for the Danish verb *forsøge* 'to try'.

(10) der forsøgte at komme mange med bussen
there tried to come many with the.bus
'many people tried to get on the bus'

However, this property appears to be an exception, rather than a defining characteristic of the verbs forming the Complex Passive. The majority of the other verbs do not allow non-thematic subjects in the active.

(11) * der agter at oprette mange en ny forening
there intends to found many a new association
'many people intend to try to found a new association'

(12) * der pålægger regeringen at indføre nye afgifter
there forces the.government to introduce new taxes
'someone is forcing the government to introduce new taxes'

While it does seem to be the case that one of the most frequent verbs forming the Complex Passive, namely the verb *forsøge* 'to try' may exhibit a raising-like behaviour in the active, there is no indication that this property is somehow related to the ability to form Complex Passives. As a matter of fact it appears to be an exception.

Verbs selecting infinitival complements The verbs forming the Complex Passive select an infinitival complement headed by a *to*-infinitive. Only one verb *bede* 'to ask' combines with a bare infinitive. [6]

(13) de beder ham flytte bilen
they ask him to.remove the.car

(14) bilen bedes flyttet
the.car is.asked removed
'please remove the car'

Interestingly many obligatory control verbs in Danish select infinitival complements marked by a semantically vacuous preposition. Cf. the examples given below.

(15) der satses på at gennemføre konkurrencen
*there is.aimed at.*PREP *to complete the.contest*
'the intention is to complete the contest'

[6] Lars Heltoft (p.c.) points out that passive *bedes* 'is asked' does not behave as a passive full verb. Rather it appears to have developed into a kind of modal marker for politeness. On this analysis the verb may be excluded from the verbs forming the Complex Passive and the appropriate generalization is that verbs forming the Complex Passive combine with *to*-infinitives.

(16) de advarer mod at forsøge at reparere bilen
they warn against to try to repair the.car
'they warn against trying to repair the car'

Verbs selecting infinitival complements marked by prepositions are systematically excluded from occurring in the Complex Passive ([Christensen, 1986, Hellan, 2001]):

(17) * konkurrencen satses på gennemført
*the.contest is.aimed at.*PREP *completed*
'the intention is to complete the contest'

(18) * bilen advares mod forsøgt repareret
*the.car is.warned against.*PREP *tried repaired*
'there is a warning against trying to repair the car'

In section 4, I show how this falls out of the analysis of the Complex Passive as a raising construction.

Verbs with an agentive subject Verbs taking controlled actional complements and experiencer subjects do not form the Complex Passive, even though the verbs do passivize when combining with sentential or nominal complements.[7]

(19) a. de lokale helte blev glemt
the local heroes were forgotten

b. det blev glemt at checke motoren
it was forgotten to check the.engine
'As for the engine, it was forgotten to check it'

c. * bilen blev glemt repareret
the.car was forgotten repaired
'As for the car, it was forgotten to repair it'

3.2 Constraints on the verbal complement

After having uncovered the constraints on the matrix verb in the Complex Passive, we next turn to the constraints on the verbal complement.

Only embedded participles It is a defining characteristic of the Complex Passive that the second passive form is a past participle. Infinitival complements are excluded even though the active verb selects an infinitival complement, cf. the discussion in section 3.1.[8]

[7] [Holmberg, 2002] attributes this constraint to the fact that psychological state predicates are typical non-restructuring verbs. However, as shown above restructuring does not appear to be a necessary condition for the formation of Complex Passives in Danish. In addition, Holmberg uses the verb *hade* 'to hate' as illustration. But this verb is independently ruled out from occurring in the complex passive since it does not passivise at all when combining with a sentential complement. So even if there is a correlation between taking an experiencer subject and being a non-restructuring verb, it has not been established that this correlation has any influence on the ability to form the Complex Passive.

[8] The restriction against infinitival complements seems to be subject to some variation. This may be due to the fact that controlled verbal complements may be either infinitival or participial.

(20) ??/* bilen forsøges at blive repareret
the.car is.tried to be repaired
'as for the car, an attempt is made to repair it'

Only participles with a suppressed argument position In his discussion of the Complex Passive in Norwegian, [Hellan, 2001] notes that the class of participles occurring in the Complex Passive is co-extensive with the class of participles with unaccusative subjects, i.e. passivised transitive verbs and participles of unaccusative verbs. On Hellans account passivised unergative verbs are excluded from the Complex Passive.

A very different picture of the Complex Passive emerges from the Danish data. The Complex Passive is possible with passivised transitive verbs (21a), passivised transitive verbs with expletive subjects (21b), passive unergative verbs (21c) and passive unergative verbs with prepositional complements (21d).

(21) a. bilen forsøges repareret
the.car is.tried repaired
'as for the car an attempt is made to repair it'

b. der forsøges repareret en bil
there is.tried repaired a car
'an attempt is made to repair a car'

c. der forsøges løbet
there is.tried run
'an attempt is made to run'

d. der forsøges indrapporteret på en ikke-eksisterende medarbejder
there is.tried reported on a non-existing employee
'an attempt is made to report on a non-existing employee'

The fact that the Complex Passives allows intransitive unergatives shows that these participles are not instances of Adjectival Passive Formation ([Levin and Rappaport, 1986]). The participles in (21c) and (21d) may not occur prenominally as shown below. In addition, the participles observed in the Complex Passives generally do not allow *un*-prefigation (24).

(22) * en løbet mand
a run man

(23) * en indrapporteret på medarbejder
a reported on employee

(24) * en urepareret bil
an unrepaired car

Excluded from the Complex Passive are active past participles, i.e. the past participles used to form the perfect tense as in *han har læst bogen* 'he has read the book', example (25), and participles based on unaccusative verbs, example (26) (contrary to Norwegian).

(25) * Peter forsøges læst bogen
Peter is.tried read the.book
'Someone is trying to make Peter read the book'

(26) * Peter forsøges omkommet
Peter is.tried died
'someone is trying to make Peter die'

Participles based on unaccusative verb as in (26) may be excluded on semantic grounds. Unaccusative verbs generally do not denote controlled actions and consequently they cannot be embedded under verbs with obligatory control. However, the constraint may also be stated in purely syntactic terms to the effect that the participle must contain a suppressed argument position. In section 5, I show how this restriction follows from the interaction between the semantics of equi-verbs and raising.

3.3 Further properties of the Complex Passive

The subject of the matrix verb is the subject of the most embedded participle As already hinted at in the introduction, the subject of the Complex Passive is the subject of the embedded participle. A clear indication of this is that the subject must meet the selectional restrictions imposed by the embedded participle on its subject. A further indication is the behaviour of embedded postverbal objects in the presence of an expletive subject as in (27a) and (27b). In example (27a) the subject of the Complex Passive is the expletive *der* 'there'. Expletive subjects occur with passivised unergative verbs, and with passivised transitive verbs. However, in the presence of an expletive subject, the object of the passive verb has to be indefinite (the well-known definiteness effect of *there*-sentences). The same restriction is observed in the Complex Passive as shown in example (27b). Here the object has to be indefinite as expected if the expletive subject of the matrix verb is the subject of the embedded participle.

(27) a. der forsøges repareret en bil
there is.tried repaired a car
'an attempt is made to repair a car'

b. * der forsøges repareret bilen
there is.tried repaired the.car
'an attempt is made to repair the car'

The "raised" constituent cannot stop half-way The "raised" constituent has to raise to the subject of the top-most matrix verb. This property follows from the complementation properties of the matrix verb. In (28), the verb *love* 'to promise' does not subcategorise a non-thematic object in addition to an expletive subject, thus example (28) violates COHERENCE.

(28) * der loves en bil forsøgt repareret
there is.promised a car tried repaired
'a promise is made about the car to repair it'

The Complex Passive exhibits Unit Accentuation (destressing of main verb) An interesting characteristic of the Danish Complex Passive is that it exhibits Unit Accentuation, i.e. destressing of the main verb. In (29a) the main verb *forsøge* 'to try' carries stress on the second syllable. In (29b) the main verb is destressed and the two verbs form a single stress group with the main stress on the participle. Unit Accentuation is a phonological characteristic of syntactic noun incorporation ([Asudeh and Mikkelsen, 1999, Thomsen, 1992]) and seems to suggest that the Complex Passive is some kind of complex predicate. Cf. however the discussion in section 4.

(29) a. de $_o$for'søgte at $_o$re$_o$pa'rere bilen
they tried to repair the.car

b. bilen blev $_o$for$_o$søgt $_o$re$_o$pa'reret
the.car was tried repaired
'as for the car, an attempt was made to repair it'

The Complex Passive may (marginally) occur in prenominal position The accounts of the Complex Passive in Norwegian ([Hellan, 2001, Nordgård and Johnsen, 2000]) all note that the Complex Passive may occur in prenominal position as exemplified in (30). This is not attested in Danish and (30) appears to be marginal.

(30) ?? en forsøgt stjålet bil
a tried stolen car
'a car that someone had tried to steal'

4 The GF of the non-finite complements

Evidence from the separability of the verbs and scrambling shows that equi-verbs do not form complex predicates in Danish as opposed to the analysis of German equi-verbs in [Müller, 2002]. The same diagnostics apply to the Complex Passive. Further evidence pertaining solely to f-structure properties is the possibility of having two subcategorized GFs in equi-constructions and Complex Passives as shown below.

(31) dommen påbyder regeringen at give firmaerne pengene tilbage
the.verdict orders the.government to give the.companies the.money back
'the.verdict orders the government to pay back the money to the companies'

(32) den gule stjerne blev af nazisterne påbudt båret af alle jøder
the yellow star was by the.nazis ordered carried by all jews
'as for the yellow star, it was a requirement of the nazis that it be born by all jews'

Depending on the analysis of the first post-verbal DP, the f-structure of example (31) would contain two OBJs or two OBJ2s if the equi-verb formed a complex predicate with its verbal complement. Similiarly, the f-structure of (32) would violate uniqueness by containing two OBL_{ag} if it were a complex predicate. Furthermore, adjuncts may scope over the invidual parts of the complex passive which is unexpected if they formed a complex predicate at f-structure.

(33) patienten forsøges nu opereret i morgen
the.patient is.tried now operated tomorrow
'someone is now trying to have the patient operated tomorrow'

Examples such as (31), (32) and (33), however, are expected on an analysis of equi-verbs and the Complex Passive as bi-clausal structures. Thus, even though Unit Accentuation generally is taken to be an indication of syntactic incorporation e.g. in [Thomsen, 1992, Asudeh and Mikkelsen, 1999], there is no syntactic evidence to support an analysis of the Complex Passive as a complex predicate.

The next step is to determine the GF of the verbal complements of equi-verbs and the Complex Passive. Following [Dalrymple and Lødrup, 2000, Lødrup, 2004], I assume that sentential complements may be either OBJs or (X)COMPs. Turning first to active equi verbs, we see that the infinitival complement alternates with a DP-object (34b), it participates in Unbounded Dependency Constructions (34c) and it may raise to SUBJ in passives (34d). The infinitival complement of active equi-verbs is thus an OBJ on the diagnostics of [Dalrymple and Lødrup, 2000, Lødrup, 2004].

(34) a. de forsøger at gennemføre konkurrencen
they try to complete the.contest

b. de forsøger det
they try it

c. at gennemføre konkurrencen har de forsøgt
to complete the.contest have they tried
'as for completing the contest, they have tried to do so'

d. at gennemføre konkurrencen blev forsøgt
to complete the.contest was tried
'as for completing the contest, it was tried to do so'

Turning to the non-finite complement of the Complex Passive, the past participle, we see that it behaves as an XCOMP on the diagnostics of [Dalrymple and Lødrup, 2000, Lødrup, 2004]. It does not alternate with a DP-object (35b) and it does not participate in Unbounded Dependency Constructions (35c).[9]

(35) a. bilen forsøges repareret
the.car is.tried repaired
'as for the car, an attempt is made to repair it'

b. * bilen forsøges det
the.car is.tried it
'as for the car, an attempt is made to do so'

[9]The last diagnostic concerning the ability to occur as a SUBJ in passives, is not applicable since the verb is already in the passive.

c. * repareret forsøges bilen
repaired is.tried the.car
'as for repairing, an attempt is made concerning the car'

The analysis of the participle in the Complex Passive as an XCOMP straightforwardly accounts for the fact that equi-verbs verbs with prepositional complements do not form Complex Passives: prepositions take OBJ and not XCOMP. A similar idea is presented in [Christensen, 1986]. Working within a derivational framework, Koch Christensen assumes that prepositions have to be able to assign case and a past participle cannot receive case. On the account presented here example (36) is ruled out since the preposition selects an OBJ and not an XCOMP.

(36) * bilen satses på repareret
the.car is.aimed at repaired
'as for the car, efforts are made to repair it'

If the Complex Passive were a complex predicate, the impossibility of (36) would be unexplained. Actually (36) should be possible like other pseudo-passives.

(37) børnene passes på
the.children are.taken.care of

4.1 Functional and Anaphoric control

LFG recognizes two different kinds of control: anaphoric and functional control ([Bresnan, 1982, Bresnan, 2001, Dalrymple, 2001]). Anaphoric control is a relation of semantic co-indexing between a controller and a covert pronominal subject in the f-structure. Functional control, on the other hand, is a relation of token-identity between the controller and the controlled subject. In anaphoric control, the controller may be absent (38), the construction allows split antecedents (39) and the controller may be realized by a semantically restricted GF (in this case an OBL_{ag}) (40).

(38) der blev forsøgt at bygge en bro
there was tried to build a bridge
'an attempt was made to build a bridge'

(39) regeringen forsøger sammen med oppositionen at afvikle debatten
the.government tries together with the.opposition to conduct the.discussion
i fællesskab
together
'the governement tries to conduct a joint discussion with the opposition'

(40) ? det blev forsøgt af regeringen at få forslaget vedtaget
it was tried by the.government to have the.proposal adopted
'an attempt was made by the government to have the proposal adopted'

Languages with personal and impersonal passives, such as Danish, provide a further diagnostic for distinguishing anaphoric and functional control. Personal passives have thematic subjects and impersonal passives have non-thematic subjects. Since an anaphorically controlled infinite complement contains a PRED-bearing subject we predict that only personal passives are allowed in anaphorically controlled complements. If an anaphorically controlled complement contains an impersonal passive, the PRED-bearing SUBJ is not assigned a semantic role in violation of COHERENCE. This prediction is borne out as shown in the examples below along with the corresponding f-structures.

(41) a. det forsøges [at blive optaget i unionen]
it is.tried [to be admitted into the.union]
'an attempt is made to be admitted into the union'

$$\begin{bmatrix} \text{PRED } \text{'ADMIT}\langle \text{SUBJ}, \text{OBL}_{goal} \rangle\text{'} \\ \text{SUBJ} \left[\text{PRED 'PRO'}\right] \\ \text{OBL}_{goal} \left[\text{"UNION"}\right] \\ \text{VOICE PASS} \end{bmatrix}$$

b. * det forsøges [at arbejdes/blive arbejdet]
it is.tried [to be worked]

$$*\begin{bmatrix} \text{PRED 'WORK}\langle\rangle\text{SUBJ'} \\ \text{SUBJ} \left[\text{PRED 'PRO'}\right] \\ \text{VOICE PASS} \end{bmatrix}$$

In functional control, the controller is obligatory (42) and no split antecedents are allowed (43). In (42) the expletive *det* 'it' cannot be a controller of the SUBJ of *repareret* 'repaired' since *det* 'it' is used in extraposition constructions and not in presentational sentences.

(42) * det forsøges repareret
it is.tried repaired (only possible with it as a referential pronoun)
'an attempt is made to repair'

(43) ?? bilen forsøges sammen med cyklen repareret samtidigt
the.car is.tried together with the.bike repaired at.the.same.time
'someone is trying to repair the car and the bike at the same time'

Since in functional control the infinite complement does not contain a pronominal subject in the f-structure, impersonal passives are allowed (provided that the controller is an expletive pronominal). Example (44) shows an active raising verb with a verbal complement containing an impersonal passive, and example (45) shows a Complex Passive construction containing an impersonal passive in the complement.

(44) der plejer at blive gjort rent om mandagen
there uses to be cleaned on Monday
'usually cleaning is on Monday'

(45) der forsøges arbejdet
there is.tried worked
'an attempt is made to work'
$\begin{bmatrix} \text{PRED } \text{'WORK}\langle\rangle\text{SUBJ'} \\ \text{SUBJ}\begin{bmatrix}\ \end{bmatrix} \end{bmatrix}$ (simplified)

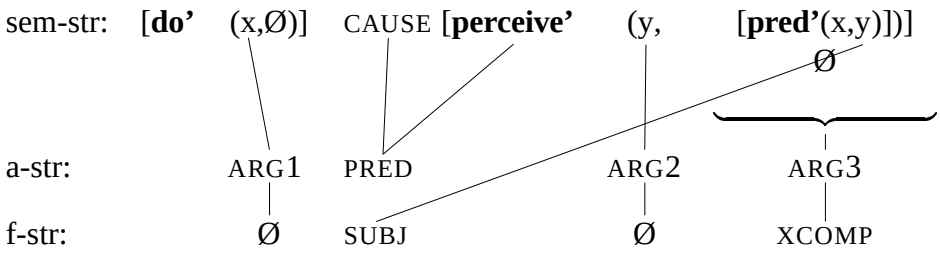
On the basis of the discussion above we may thus conclude that active equi verbs select anaphorically controlled objects, while the complex passive is a raising construction with a functionally controlled XCOMP.

5 Equi-verbs to raising verbs: an argument structure account

[Barron, 1999] presents a semantically based account of the development of the English raising verb *seem* from the Middle English equi-verb *semen* meaning 'to pretend'. On Barron's analysis the equi-verb is de-causativised, i.e. it looses the semantic component of "to do something" due to a process of semantic bleaching. Loosing this semantic component of volitional action, the verb also loses its most prominent argument and consequently its SUBJ. In order to fulfil the subject condition the verb raises the subject of the embedded complement giving rise to the raising verb *seem* of modern English.

The crucial claim here is that equi verbs may also turn into raising verbs due to operations on the argument structure without any shift in semantics. Cf. the example in (46) below. The semantic representation for the verb *simulere* 'to simulate' follows the account given in ([Barron, 1999]): someone performs some unspecified action which causes someone to perceive as if X performs some other action. The argument of the semantic component of DO maps to the ARG1 of the argument structure, while the unspecified receiver maps to ARG2. The embedded predicate in turn maps to ARG3. The ARG1 of the argument structure is suppressed due to passivisation, and the ARG2 is demoted or suppressed entirely ([Barron, 1999], p. 202). Given that there is no most prominent argument mapping to SUBJ, the verb raises the subject of the embedded predicate in order to fulfil the subject condition ([Bresnan, 2001]). The crucial claim is thus that suppression of the most prominent argument creates the canonical argument structure of a raising configuration, a verb with a propositional argument and no most prominent argument mapping to subject.

(46) han simuleres henrettet
he is.simulated executed
'As for him, his execution is simulated'



5.1 Linking the passives of equi-verbs

Recall from section 1 that equi verbs allow two kinds of passives, the canonical passive and the complex passive. In this section we will have a look at the linking of these various passives. The active equi verb is a verb with obligatory anaphoric control as shown above. The linking for the example in (47) is depicted below: The agent maps to SUBJ, the propositional argument to OBJ (cf. the discussion of the grammatical function of the verbal complements in section 4) giving rise to a lexical entry requiring a SUBJ and an OBJ where the PRED-value of the embedded SUBJ is specified as 'PRO' (anaphoric control).

a-str ⟨ [ag] [prop] ⟩ ⟹ ⟨ SUBJ OBJ ⟩
SUBJ OBJ (↑ OBJ SUBJ PRED) = 'PRO'

(47) Peter forsøger at reparere bilen
Peter tries to repair the.car
'Peter is trying to repair the car'

$$\begin{bmatrix} \text{PRED} & \text{'TRY}\langle \text{ SUBJ OBJ } \rangle\text{'} \\ \text{SUBJ} & \left[\text{"PETER"}\right] \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'REPAIR}\langle \text{ SUBJ OBJ } \rangle \\ \text{SUBJ} & \left[\text{PRED 'PRO'}\right] \\ \text{OBJ} & \left[\text{PRED 'CAR'}\right] \end{bmatrix} \\ \text{VOICE} & \text{ACT} \\ \text{TENSE} & \text{PRES} \end{bmatrix}$$

The linking of the canonical passive follows the general pattern of passivisation of transitive verbs. The most prominent argument is suppressed and the propositional complement may map to SUBJ. Alternatively the object retains its GF and the verb occurs with a non-thematic SUBJ (cf. the discussion in footnote 4). In both cases the passive exhibits anaphoric control giving rise to the lexical entries shown on the right-hand side of the arrow.

a-str: ⟨ [ag] [prop] ⟩ ⟹ { ⟨ SUBJ ⟩
Ø SUBJ/OBJ (↑ SUBJ SUBJ PRED) = 'PRO' |
⟨ OBJ ⟩ SUBJ
(↑ OBJ SUBJ PRED) = 'PRO' }

Examples of the canonical passive are given in (48)

(48) a. at finde en løsning forsøges
to find a solution is.tried
'as for finding a solution, an attempt is made to do so'

b. der forsøges at finde en løsning
there is.tried to find a solution
'an attempt is made to find a solution'

$$\begin{bmatrix} \text{PRED} & \text{'TRY}\langle\ \text{OBJ}\ \rangle\ \text{SUBJ'} \\ \text{SUBJ} & \left[\text{PRON-TYPE EXPLETIVE}\right] \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'FIND}\langle\ \text{SUBJ OBJ}\ \rangle \\ \text{SUBJ} & \left[\text{PRED 'PRO'}\right] \\ \text{OBJ} & \left[\text{PRED 'SOLUTION'}\right] \end{bmatrix} \\ \text{VOICE} & \text{PASS} \\ \text{TENSE} & \text{PRES} \end{bmatrix}$$

Finally we turn to the linking of the Complex Passive. Again the most prominent argument is suppressed creating a canonical argument structure of a raising construction. Raising configurations contain functionally controlled XCOMPs so the propositional argument maps to an XCOMP with functional control as shown below.

a-str:	⟨ [ag]	[prop] ⟩	⟹	⟨ XCOMP ⟩ SUBJ
	|	|	(↑SUBJ) = (↑XCOMP SUBJ)
	Ø	XCOMP		(↑XCOMP VFORM)=$_c$ PASTPART
				(↑XCOMP VOICE) =$_c$ PASS

An example of the complex passive is given in (49) below.

(49) konkurrencen forsøges gennemført
the.contest is.tried completed
'as for the contest, an attempt is made to complete it'

$$\begin{bmatrix} \text{PRED} & \text{'TRY}\langle\ \text{XCOMP}\ \rangle\ \text{SUBJ'} \\ \text{SUBJ} & \left[\text{PRED 'CONTEST'}\right] \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{'COMPLETE}\langle\ \text{SUBJ}\ \rangle\text{'} \\ \text{SUBJ} & [\ ] \\ \text{VFORM} & \text{PASTPART} \\ \text{VOICE} & \text{PASS} \end{bmatrix} \\ \text{VOICE} & \text{PASS} \\ \text{TENSE} & \text{PRES} \end{bmatrix}$$

However, the analysis of the complex passive as a raising construction does not explain the presence of the two constraining equations in the lexical entry above, i.e. that the XCOMP be a past participle, and that the XCOMP be passive. As a matter of fact, this latter fact is rather puzzling since raising constructions do not otherwise impose restrictions on the voice of their embedded complement.

5.2 Deriving the constraints on the XCOMP of the Complex Passive

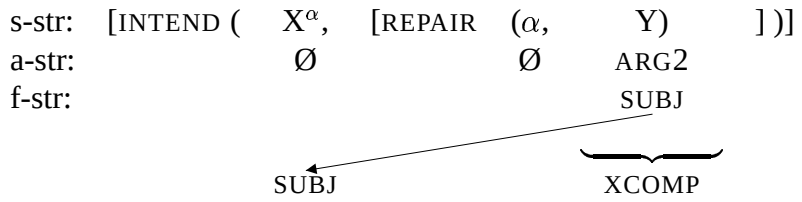
Passivisation avoids raising of a bound variable The Complex Passive derives its name from the fact that it obligatorily occurs with a passive past participle. In this section I show how this feature follows from the interaction between the semantics of equi and raising verbs.

Following ([Culicover and Jackendoff, 2005]) I assume the semantic structure of an obligatory control verb such as *try* as depicted below.

[INTEND(X^{α},[ACT(α,Y)]]

The semantics of the verb decomposes into a semantic component of intention and an actional complement (the controlled complement). Crucially the ACTOR of the intention-component (represented with X^{α}) is co-referential with the ACTOR (α) of the embedded actional complement (thus showing that it is a subject control verb). Cf. example (50) below with the semantic representation immediately below.

(50) bilen forsøges repareret
the.car is.tried repaired
'as for the car, an attempt is made to repair it'

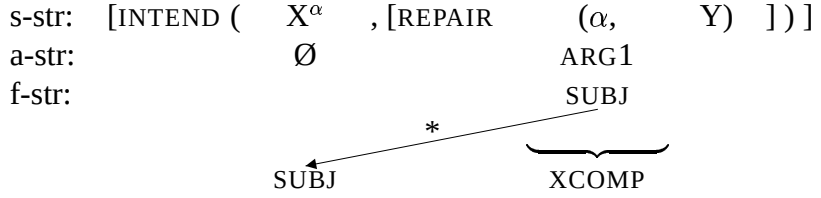


The most prominent argument of the matrix predicate has been suppressed, and also the most prominent argument of the embedded predicate. Consequently the argument Y may map to subject and since Y is not co-indexed with an argument of the matrix predicate, i.e. since it is not a semantic argument of the matrix predicate, it may raise to subject of the matrix predicate.

Consider next what happens if a complex passive is formed with an active past participle, i.e. the participle as used in *Peter har repareret bilen* 'Peter has repaired the car'.

(51) * Peter forsøges repareret bilen
Peter tries repaired the.car
'someone is trying to make Peter repair the car'

The semantic structure and the linking to syntax of the ill-formed example (51) is given below:



The argument linked to the ACTOR of the matrix verb has been suppressed being the most prominent argument. However, the co-referential subject argument (the bound variable) of the embedded predicate is raised to subject of the matrix verb. A raising verb is a verb with a syntactic complement which is not at the same time a semantic argument. But in this case the raised complement does count as a semantic argument of the matrix predicate due to the fact that it is inherently bound by an argument of the matrix predicate (indicated with α). In this way the subject is licensed by the argument structure of the matrix verb, so this cannot be a raising construction. In a raising construction the raised argument cannot be inherently bound by an argument of the matrix verb. Since the ACTOR of controlled actional complements is bound by an argument of the matrix verb, the ACTOR-argument of the controlled actional complement has to be suppressed so as to allow another argument to map to subject, the target of raising constructions.

To sum up - when the matrix verb functions as a raising verb, only an argument which is not inherently bound by an argument of the matrix verb is eligible. Since raising targets subjects of embedded complements, the most prominent argument of the verbal complement, the ACTOR has to be suppressed to allow another argument to map to subject, or to allow the subject not to be linked to argument structure at all. In this way the semantics of equi and raising verbs conspire to enforce passivisation of the embedded predicate.

The morpho-syntactic realisation of the verbal complement As far as the morpho-syntactic realisation of the infinite complement as a past participle and not as a passive infinitive, I can only offer a descriptive generalisation. Following [Lødrup, 2002], I assume that XCOMPs are canonically realised as non-finite VPs cross-linguistically. A non-finite VP can be either an infinitival or a participial clause and as regards the choice between these two forms, a clear pattern can be discerned: verbs selecting finite verbal complements in the active take infinitival XCOMPs and verbs selecting infinitival complements in the active take participial XCOMPs.

The Complex Passive is formed by verbs selecting infinitival complements, but also verbs selecting finite complements may function as raising verbs when passived. This is expected since the most prominent argument is suppressed and the verb has to raise the subject of an embedded complement to fulfil the subject condition. However these raising complements are realised as infinitives, and not past participles. Cf. the example in (52) below.

(52) a. de påstår at han er rejst
they claim that he has left

b. han påstås at være rejst
he is.claimed to have left

If, however, the verb selects an infinitival complement in the active, the raising complement of the passivised verb is realized as a participle as previously discussed. Thus there appears to be a kind of hierarchical ordering of the morpho-syntactic realisation of the raising complements as shown below:

FINITE CLAUSE << INFINITIVAL CLAUSE << PAST PARTICIPLE

If the verb selects a finite clause in the active, the raising complement is an infinitival clause, and if the verb selects an infinitival clause in the active, the raising complement is realised as a

past participle. However, an explanation of this generalisation must be left for future research.

6 Conclusion

This paper has provided an analysis of the Complex Passive in Danish. It was shown that the Complex Passive is a raising construction, and the difference between the kinds of passives of equi-verbs was shown to be a difference between a passive equi verb with anaphoric control and a passive raising verb with functional control. It was shown that the raising construction is triggered by the morpho-lexical operation of passivisation creating the right argument structure environment, i.e. a verb with a propositional complement and no most prominent argument. On this analysis, raising configurations may thus not only be triggered by semantics but also by argument structure properties without any shift in semantics. The constraint that the embedded complement be passive was shown to follow from the interaction between the semantics of equi and raising verbs.

References

[Asudeh and Mikkelsen, 1999] Asudeh, A. and Mikkelsen, L. H. (1999). Danish syntactic noun incorporation: A case study in grammatical interfaces. In HPSG99. Edinburgh.

[Barron, 1999] Barron, J. (1999). Perception, volition and reduced clausal complementation. PhD thesis, University of Manchester.

[Bresnan, 1982] Bresnan, J. (1982). Control and complementation. In Bresnan, J., editor, The Mental Representation of Grammatical Relations. Cambridge MA: MIT Press.

[Bresnan, 2001] Bresnan, J. (2001). Lexical-Functional Syntax. Blackwell Textbooks in Linguistics. Blackwell.

[Christensen, 1986] Christensen, K. K. (1986). Complex passive and conditions on reanalysis. Nordic Journal of Lingusitics, 9(2).

[Culicover and Jackendoff, 2005] Culicover, P. W. and Jackendoff, R. (2005). Simpler Syntax. Oxford Linguistics. Oxford University Press.

[Dalrymple, 2001] Dalrymple, M. (2001). Lexical-Functional Grammar, volume 34 of Syntax & Semantics. Academic Press.

[Dalrymple and Lødrup, 2000] Dalrymple, M. and Lødrup, H. (2000). The grammatical function of complement clauses. In Butt, M. and King, T. H., editors, Proceedings of the LFG00 Conference. University of Berkeley, CSLI Publications.

[Diderichsen, 1957] Diderichsen, P. (1957). Elementær dansk grammatik. København: Gyldendal.

[Engdahl, 1999] Engdahl, E. (1999). The choice between *bli*-passive and *s*-passive in danish, norwegian and swedish. NORDSEM Report 3, Göteborgs Universitet.

[Engh, 1995] Engh, J. (1995). Verb i passiv fulgt av perfektum partisipp. Betydning og bruk (1984). Oslo:Novus forlag.

[Faarlund et al., 1997] Faarlund, J. T., Lie, S., and Vannebo, K. I. (1997). Norsk referansegrammatik. Oslo: Universitetsforlaget.

[Hansen, 1967] Hansen, A. (1967). Moderne Dansk, I-III. København: Grafisk forlag.

[Hellan, 2001] Hellan, L. (2001). Complex Passive Constructions in Norwegian. NTNU, Trondheim/CSLI, Stanford.

[Holmberg, 2002] Holmberg, A. (2002). Expletives and Agreement in Scandinavian Passives. Journal of Comparative Germanic Linguistics, 4:85–128.

[Levin and Rappaport, 1986] Levin, B. and Rappaport, M. (1986). The formation of adjectical passives. Linguistic inquiry.

[Lødrup, 2002] Lødrup, H. (2002). Infinitival complements in Norwegian and the form-function relation. In Butt, M. and King, T. H., editors, Proceedings of the LFG02 Conference. National Technical University of Athens, CSLI Publications.

[Lødrup, 2004] Lødrup, H. (2004). Clausal complementation in Norwegian. Nordic Journal of Linguistics, 27(1):61–95.

[Müller, 2002] Müller, S. (2002). Complex Predicates. CSLI Publications.

[Nordgård and Johnsen, 2000] Nordgård, T. and Johnsen, L. (2000). Complex passives. a declarative analysis. In The 18th Scandinavian Conference of Linguistics. University of Lund.

[Robin Allan and Lundskær-Nielsen, 1995] Robin Allan, P. H. and Lundskær-Nielsen, T. (1995). Danish: A comprehensive grammar. Routledge Grammars. London: Routledge.

[Thomsen, 1992] Thomsen, O. N. (1992). Syntactic noun incorporation in Danish. In Fortescue, M., Harder, P., and Kristoffersen, L., editors, Layered Structure and Reference in a Functional Perspective, pages 173–231. John Benjamins Publishing Company.

PREPOSITION INCORPORATION IN MANDARIN: ECONOMY WITHIN VP

Jeeyoung Peck and Peter Sells

Stanford University

Proceedings of the LFG06 Conference

Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006

CSLI Publications

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

Abstract

In this paper, we examine the phenomenon of Preposition Incorporation (PI) in modern Mandarin. While the category PP is found in various positions within the clause, it is never found with VP. Instead, the P 'incorporates' into V, or else is absent. We argue that previous generative approaches have failed to provide a simple and consistent explanation which applies to all types of verbs.

We propose an OT analysis in which a given argument structure has different potential surface expressions, with interacting constraints to give the correct range of actual output forms. The constraints are of the familiar types: markedness, in particular, that the VP-internal structure be maximally simple, and faithfulness, in particular, that any non-core GFs are 'marked' by the presence of a P. Further, our analysis extends to different classes of ditransitive verbs; the 'put' class has a very limited range of surface expressions, a subset of those available to the 'send' class.

1. Introduction

Modern Mandarin has very strong restrictions on what XPs it allows VP-internally, following V. Typically, only subcategorized arguments are allowed (cf. Fang 2006); some verbs have two internal arguments and hence allow sequences such as V NP NP.

With intransitive verbs, the sequence V PP is never found, even though PPs are found in every other position in the clause. We call this hypothetical construction the 'full PP' construction. Instead of this, the P must be 'incorporated' into the verb, giving the sequence V+P NP. We call this 'PI' (preposition incorporation). As shown in the examples below, the aspect marker *-le* is a diagnostic for PI. And with some verbs, the incorporated P can be absent; we call this 'BVC' (bare verb construction).

The full PP construction, PI and BVC are all in principle expressions of the same argument-structure. We further develop the LMT/LFG analysis of Her (1999), using OT to account for:

1. the obligatory nature of PI (the reason for *V PP), and
2. the optionality of different surface representations.

We briefly compare our account with the previous analysis of Li (1990), Gao (2005), and Feng (2003), of which Gao's is the most thorough.

2. Two Types of Verb

2.1. The 'send' class

The relevant data for a verb like 'send', with a(rgument)-structure $<$ agent, goal, theme $>$, is as follows:

(1) a. *Ta song-le yibenshu **gei wo**. (*V NP PP)
he send-PERF one.CL.book to me
'He sent a book to me.'

b. *Ta song-le **gei wo** yibenshu. (*V PP NP)
he send-PERF to me one.CL.book

c. Ta **gei wo** song-le yibenshu. (PP_{go} V NP_{th})
he to me send-PERF one.CL.book

d. Ta song-**gei**-le **wo** yibenshu. (PI; V NP NP)
he send-to-PERF me one.CL.book

e. Ta song-le **wo** yibenshu. (BVC; V NP NP)
he send-PERF me one.CL.book

The key examples are (d) and (e). While these are grammatical with the structure V NP NP, examples (a) and (b) are not, with a PP in the structure.

Her (1999) presents an LMT analysis of simple ditransitives in Mandarin, which we follow here:

(2) *song(-gei)* 'send(-to)'

argument-structure:	<	ag,	go,	th	>
intrinsic:		[−o]	[+o]	[−r]	
GF:		SUBJ	OBJ_θ	OBJ	

The Goal argument of *song* is never expressed as an OBL, in a PP; and it does not passivize, hence it is categorized as [+o]. If this is correct, then the (a/b) examples above cannot be generated, as desired.

2.2. The 'put' class

The relevant data for a verb like 'put', with a-structure < agent, theme, location >, is as follows:

(3) a. *Wo fang-le nabenshu **zai zhuozishang**. (*V NP PP)
I put-PERF that.CL.book on desk.top
'I put that book on the table.'

b. *Wo fang-le **zai zhuozishang** nabenshu. (*V PP NP)
I put-PERF on desk.top that.CL.book

c. Wo **zai zhuozishang** fang-le nabenshu. (V NP)
I on desk.top put-PERF that.CL.book

d. *Nabenshu wo fang-le **zai zhuozi-shang**. (*V PP)
that.CL.book I put-PERF on desk.top

e. Nabenshu wo fang-**zai**-le **zhuozi-shang**. (PI; V NP)
that.CL.book I put.on-PERF desk.top
'I put that book on the table.'

f. Nabenshu wo fang-le **zhuozi-shang**. (BVC)
that.CL.book I put-PERF desk.top

Either the PP must appear external to VP, as in (c), or else the NP object must be external to VP, as in (e) and (f), with accompanying PI or BVC. In other words, topicalization of one complement commonly occurs when there is another complement in postverbal position as in (e) and (f), with this class of verb (Huang 1982). Topicalization is somewhat free in Mandarin, though we assume that such displacement always has some function for pragmatic or information-structure reasons. The discourse-related aspects of Mandarin constituent order are well-known (e.g., Li and Thompson 1981). The key point in our data is that while two postverbal complements are allowed in principle with 'send', only one postverbal complement is allowed with 'put'.

The a-structure of 'put' is as follows:

(4) *fang-(zai)* 'put(-on)'

a-str:	<	ag,	th,	loc	>
instrinsic:		[−o]	[−r]	[+r]	
GF:		SUBJ	OBJ	OBL_θ/OBJ_θ	

[+r] can be OBL_θ or OBJ_θ, and in Mandarin these can be expressed as PP or NP. The location argument is classified as [+r], and hence it could be a PP or an NP in c-structure. Note, however, that (5) is ungrammatical (cf. (3)e):

(5) *Wo fang-**zai**-le **zhuozi-shang** nabenshu. (*PI; V NP NP)
I put.on-PERF desk.top that.CL.book

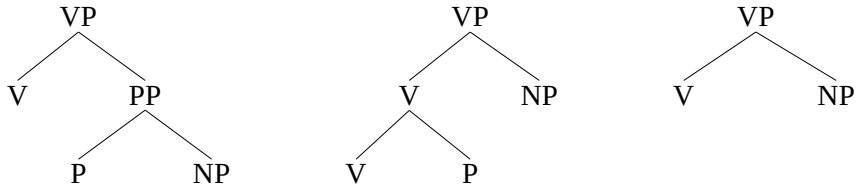
As seen in the contrast between (3e) and (5), two NPs may appear VP-internally with verbs of the 'send' class but not those of the 'put' class. Hence the expression of arguments within VP cannot solely be a fact of c-structure restrictions in Mandarin. Instead, we argue that the full account of which structures are grammatical also involves Optimality Theory-style competition between interacting constraints, which refer to the thematic hierarchy, for the contrast just mentioned.

3. Structures and Constraints

3.1. Competing Structures

In an OT analysis, structures compete as expressions of the same abstract information, and the structures that are relevant for Mandarin VPs are shown in (6), concentrating for now on structures with just one internal argument. For the (b) structure, we assume a lexical rule combining V and P as a complex V. As the aspect marker *-le* follows the sequence V+P, this is strong motivation that PI (verb) structures are formed lexically.

(6) (a) PP (b) PI (c) BVC



The (a) structure never surfaces in Mandarin. Gao (2005) adopts a movement analysis in which the (b) structure is derived from the (a) structure via 'Preposition Incorporation', and the (c) structures are derived from the (b) structures by 'Phonetic Suppression' of the P. This of course implicitly claims that the (a–c) structures share the same a-structure, for they are all derived from the same underlying structure.

In Gao's analysis, there is no simple mechanism which forces the P to incorporate if the PP is adjacent to V – the grammar can allow it as an option, but not force it as a necessary operation. Gao suggests that V and P assign different cases, and in situations of adjacency of V and PP, the V's case 'wins'. Note that this implicitly compares the favorability of V's case over P's case.

In our analysis, structures (a) and (b) are more FAITHFUL than (c), but (c) is more ECONOMICAL than (b) (and more than (a)).

3.2. Competing Constraints are Necessary

Gao's analysis suffers from a problem with intransitive Vs, which do not assign case (see (7)); therefore the motivation for PI cannot be that V's case and P's case clash, as V has no case. Instead, they show that the language prefers direct (NP) complements to V in favor to PP complements, regardless of the properties of the head V of the VP.

(7) a. Xiaotou pao **dao menkou**. (V PP)
small.thief run to entrance
'The thief ran to the entrance.'

b. *Xiaotou **dao menkou** pao-le. (*PP V)
small.thief to entrance run-PERF

c. *Xiaotou pao-le **dao menkou**. (*V PP)
small.thief run-PERF to entrance

d. Xiaotou pao-**dao**-le **menkou**. (PI)
small.thief run-to-PERF entrance

e. *Xiaotou pao-le **menkou**. (*BVC)
small.thief run-PERF entrance

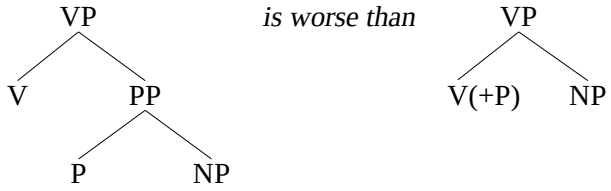
These examples illustrate the surface expression with an intransitive verb. (a) looks like a sequence of V and PP, out of which the PP cannot scramble (b).[1] However, the facts of aspectual *-le* in (c) and (d) show that the P is actually part of the verb (hence, in a PI structure), rather than heading a constituent PP. Finally, (e) shows that BVC is not possible with this verb.

3.3. Constraints

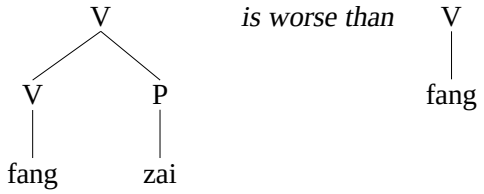
To account for the competitions between structures outlined above, we propose the following constraints. (8)–(10) are markedness constraints:

(8) θ(VP): Order in the VP obeys the thematic hierarchy $\ldots$ go $>$ th $>$ loc.
NP_θ precedes NP in Mandarin, so
NP(Loc) $\prec$ NP(Theme) *is worse than* NP(Theme) $\prec$ NP(Loc)

(9) ECON(VP): *XP within VP. (cf. Fang 2006)
V – PP violates this more than V – NP:



(10) ECON(V): V is mono-morphemic.
V+P violates this; V does not:



[1](7)b is grammatical, but with a different meaning, namely 'The thief came to the entrance and then ran away (from there)'.

And we have one faithfulness constraint, which requires marked values of the LMT features to have overt expression as a P:

(11) FAITH([+f]): An intrinsic [+o] or [+r] argument is marked by a P in c-structure.
(Only [$-f$] arguments are unmarked; the BVC violates the constraint.)

3.4. Constraint Ranking

The basic constraint ranking for Mandarin is that the first two constraints outrank the second two:

(12)

{ θ(VP), ECON(VP) }	$\gg$	{ ECON(V), FAITH([+f]) }
these restrict		these allow
the content of VP		PI or BVC

The key to our analysis is to restrict the internal contents of VP, and then let the other constraints arbitrate between the PI or BVC structures. In the full analysis, θ(VP) outranks ECON(VP), but the lowest two constraints are tied, which we take as an abbreviation for the idea that in a specific evaluation, one of the two possible rankings will taken; over multiple instances, the grammar will generate two outputs in 50-50 distribution, for a given relevant input (see e.g., Müller 1999).

4. The Analysis of PI

We illustrate our analysis first with some general considerations, and then go on to the 'send' and 'put' classes respectively.

4.1. Economy within VP

The motivation for PI and BVC is the pressure for Economy within VP, rather than Case Conflict, which cannot apply in examples like those in (7). The Mandarin VP is highly restricted in the post-verbal domain, allowing at most subcategorized arguments or a single adverbial, but not both (for detailed analysis, see Fang 2006).

In fact, if multiple VP-internal constituents are required, the V must be 'copied' (see Tai 1985, Li 1990, Sybesma 1999, Feng 2003):

(13) a. Zhangsan tan gangqin tan-le hen jiu.
Zhangsan play piano play-PERF very long
'Zhangsan played piano for a very long time.'

b. *Zhangsan tan-le gangqin hen jiu.
Zhangsan play-PERF piano very long

Fang (2006) presents an LFG analysis of examples like (13)a, with a modified coordinated VP structure. Each VP itself has minimal structure, as preferred by our high-ranking constraint ECON(VP).

4.2. The ‘send’ class

The ‘send’ class has the a-structure shown in (14). The Goal is intrinsically classified by LMT as [+o], so in the linking to GFs it be an OBJ of some kind. However, the Theme will be the unmarked choice for OBJ, by the usual principles of LMT.

(14) *song(-gei)* ‘send(-to)’

a-str:	< ag,	go,	th	>
intrinsic:		[+o]		
GF:	SUBJ	OBJ_θ	OBJ	
c-str:	NP	NP_θ / *PP_θ	NP	

The constraint ranking already disfavors expression of any PP within VP. To illustrate the analysis clearly, we present two tableaux, with the same inputs, with the two possible rankings of the lowest two constraints.

(15)

send<ag,go,th>	θ(VP)	EC(VP)	EC(V)	FTH([+f])	
[a] [V NP_{th} PP_{go}]	*	3	1		*NP PP
[b] [V+P NP_{go} NP_{th}]		2	*2		*PI
[c] [V NP_{go} NP_{th}]		2	1	*	BVC
[d] [V PP_{go} NP_{th}]		*3	1		*V PP

(16)

send<ag,go,th>	θ(VP)	EC(VP)	FTH([+f])	EC(V)	
[a] [V NP_{th} PP_{go}]	*	3		1	*NP PP
[b] [V+P NP_{go} NP_{th}]		2		2	PI
[c] [V NP_{go} NP_{th}]		2	*	1	*BVC
[d] [V PP_{go} NP_{th}]		*3		1	*V PP

We count violations numerically, so having a PP inside VP creates one more violation of *XP (EC(VP)) than having an NP there. The first two constraints are the most high-ranked, so the [a] and [d] candidates are eliminated, and the LMT principles will also guarantee that the Goal is linked as OBJ_θ. As in (1)d and (1)e, both [b] and [c] candidates are selected by the two rankings, each of which is shown in one tableau above.

The variation between [b] and [c] is determined by the relative ranking of ECON(V) and FAITH([+f]). These potential winning candidates correspond to (1)d and (1)e. It is worth noting that FAITH([+f]) is active even though the Goal argument of *song* is never expressed as a PP, as the incorporated P in PI also satisfies this constraint.

For verbs of this type, the apparent order NP – PP is possible in Mandarin. However Chao (1968) and Huang and Mo (1992) argue that in such a case *gei* is a co-verb heading a secondary VP. Each VP then obeys θ(VP), as well as the other constraints on VP structure:

(17) Ta [$_{VP}$ **song** nabenshu [$_{VP}$ **gei** wo]].
he send that.CL.book ‘give’ me

As shown by the bracketing, this structure has one VP as a complement inside another, and the structure is V – NP – VP, not V – NP – PP.

4.3. The 'put' class

The 'put' class has the a-structure shown in (18). The Loc role is instrinsically classified by LMT as [+r], so in the linking to GFs it must either be OBJ_θ or OBL_θ, which would correspond to expression as an NP or as a PP:

(18) *fang-(zai)* 'put(-on)'

a-str:	< ag,	th,	loc	>
intrinsic:			[+r]	
GF:	SUBJ	OBJ	OBL_θ/OBJ_θ	
c-str:	NP	NP	PP/NP_θ	

For this a-structure, θ(VP) eliminates candidates [b]/[c], for they do not follow the order in the thematic hierarchy. This renders irrelevant the ranking of the lowest two constraints, which we just show here in the same ranking order as (15). The only way out of the ineffability that θ(VP) triggers is that one argument of V must topicalize or otherwise be expressed external to VP:

(19)

put<ag,th,loc>	θ(VP)	EC(VP)	EC(V)	FTH([+f])	
[a] [V NP_{th} PP_{loc}]		3	1		*NP PP
[b] [V+P NP_{loc} NP_{th}]	*	2	2		*PI
[c] [V NP_{loc} NP_{th}]	*	2	1	*	*BVC
[d] [PP_{loc} [V NP_{th}]]		1	1		V NP

Due to ECON(VP), [d] always wins as it has the least structure in the lowest VP. This corresponds to (3)b.

Of course, fronting a category out of VP always involves some extra discourse information, as we noted in section 2.2. For this reason, the [d] candidate is technically unfaithful as an unmarked expression, even though it is the winner in the competition shown here. We assume that the few constraints that we have presented here outrank any constraints involving faithfulness to discourse-related information.

The [b/c] candidates above violate θ(VP). However, the [b′/c′] candidates shown in (20) with the NP_{th} in front of the V do not violate this constraint, and correspond to (3)e/f:

(20)

put<ag,th,loc>	θ(VP)	EC(VP)	EC(V)	FTH([+f])	
[b] [V+P NP_{loc} NP_{th}]	*	2	2		*PI
[c] [V NP_{loc} NP_{th}]	*	2	1	*	*BVC
[b′] [NP_{th} [V+P NP_{loc}]]		1	2		PI (see below)
[c′] [NP_{th} [V NP_{loc}]]		1	1	*	BVC

The contrast between [b] and [b′] is shown in (21). As the restricted object must always immediately follow the verb, the sequence NP NP with the verb 'put' always violates θ(VP).

(21) a. *Wo fang-**zai**-le **zhuozi-shang** nabenshu. (V+P NP NP)
I put.on-PERF desk.top that.CL.book

b. Nabenshu wo fang-**zai**-le **zhuozi-shang**. (V+P NP)
that.CL.book I put.on-PERF desk.top

The relative ranking of the lower two constraints will dictate whether the 'put'-type verbs surface with PI or BVC, as long as only one argument is VP-internal. The ranking as given in (20) is the same as in (15), and

BVC is the winning candidate. With the constraint ranking as in (16), PI would be the winner. Importantly, the 'put' class only allows one NP VP-internally, while the 'send' class allows two.

5. Conclusion

In conclusion, we have proposed this ranking for Mandarin:

(22)

{ θ(VP), ECON(VP) }	$\gg$	{ ECON(V), FAITH([+f]) }
these restrict		these allow
the content of VP		PI or BVC

The ranking of the first two constraints over the second two explains the preference for NP over PP within VP, and the restrictions on the order of constituents within VP, seen in the difference between 'send'-type and 'put'-type verbs. Specifically, due to θ(VP), a VP with the structure [V NP NP] is possible with 'send' but not 'put'. The high-ranking constraints also ensure that a constituent PP can never appear within VP, even though it can appear in other positions within the clause.

The lower-ranked constraints have effects as follows:

(23) a. FAITH([+f]) $\gg$ ECON(V) favors PI over BVC.

b. ECON(V) $\gg$ FAITH([+f]) favors BVC over PI.

There is of course further complexity in the data; here we have concentrated on illustrating the style of the analysis with ditransitive verbs. A fuller account of a wide range of Mandarin intransitive, transitive and ditransitive verbs can be found in Peck (2006).

References

Chao, Yuen-Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley, University of California Press.

Fang, Ji. 2006. *The Verb Copy Construction and the Post-Verbal Constraint In Chinese*. Doctoral dissertation, Stanford University.

Feng, Shengli. 2003. Prosodically constrained postverbal PPs in Mandarin Chinese. *Linguistics* 46, 1085–1122.

Gao, Man. 2005. Preposition incorporation in Mandarin. Paper presented at *NACCL-17*, DLI Foreign Language Center, Monterey.

Her, One-Soon. 1999. Interaction of thematic structure and syntactic structures: On Mandarin dative alternations. *Zhongguo Jingnei Yuyan ji Yuyanxue* 5, 373–412.

Huang, C.-T. James. 1982. *Logical Relations in Chinese and the Theory of Grammar*. Doctoral dissertation, MIT.

Huang, Chu-Ren, and Ruo-Ping Mo. 1992. Mandarin ditransitive constructions and the category of *gei*. *BLS* 18, 109–122.

Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese*. Berkeley and Los Angeles, California, University of California Press.

Li, Yen-hui Audrey. 1990. *Order and Constituency in Mandarin Chinese*. Dordrecht, Kluwer Academic Publishers.

Müller, Gereon. 1999. Optionality in Optimality-Theoretic Syntax. *GLOT International* 4(5), 3–8.

Peck, Jeeyoung. 2006. OT-LFG analysis of Preposition Incorporation in Modern Mandarin. Manuscript, Stanford University.

Sybesma, Rint. 1999. *The Mandarin VP*. Dordrecht, Kluwer Academic Publishers.

Tai, James H-Y. 1985. *Temporal Sequence and Chinese Word Order*. Amsterdam, John Benjamins Publishing Company.

jpeck@stanford.edu
sells@stanford.edu

ON THE NEED FOR A MORE REFINED APPROACH TO THE ARGUMENT-ADJUNCT DISTINCTION:

THE CASE OF DATIVE EXPERIENCERS IN HUNGARIAN

György Rákosi

University of Debrecen

Proceedings of the LFG06 Conference

Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006

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

Abstract

The so-called circumstantial PPs (instruments, benefactives, and the like) have traditionally been analyzed in terms of thematic roles, but it is generally recognized that they have a special syntactic status. In this paper, I argue that they can systematically be treated as adjuncts that bear a thematic role. I substantiate this claim through a case study of dative experiencers in Hungarian, which I analyze in terms of an LFG-theoretic application of the Theta System of Reinhart (2000, 2002). I distinguish between thematic arguments, thematic adjuncts and non-thematic adjuncts; and I argue that this threefold distinction is minimally needed to be able to account for the behavior of the three empirically distinct types of dative experiencers.

1. Introduction: the non-core thematic domain

The argument-adjunct distinction is a fundamental design feature of the standard LFG architecture. For expository purposes, I will be glossing over syntactic and semantic aspects of representation in the basic terminology that I use in this paper, and understand the terms *argument* and *adjunct* as follows. A typical *argument* is an element that is obligatory both syntactically and semantically, i.e. it is entailed by the predicate.[1] A typical *adjunct* is optional both syntactically and semantically, and acts as a semantic modifier. In LFG, arguments have a crucial syntactic influence as they provide the minimal information needed for the construction of the f-structure assigned to the predicate. Adjuncts are collected in sets and from a syntactic perspective, they are in general simply required to be integrated into the immediate f-structure containing a PRED feature. The two are further distinguished by the standard assumption that semantic arguments receive a thematic specification, whereas adjuncts do not.

There are certain expressions that are often discussed in terms of thematic roles but which are unlike typical arguments in being optional. In particular, most PPs described as *comitatives*, *instruments*, *benefactives*, (spatial or abstract) *sources* and *goals* are in this non-core thematic domain. The standard LFG approach is to regard these as "*possible grammatical arguments of the verb*" (Bresnan: 1982, 165). Bresnan (1982) assumes that the verbs in (1a) and (1b), for example, are two distinct lexical entries: the latter has been derived from the former by the lexical rule of *Instrumentalization*. This rule adds an additional *instrument argument* to the input agentive predicate.[2]

(1) a. *John assassinated the president*.
a.' *assassinate*$_1$ < agent, patient >

b. *John assassinated the president with the dynamite*.
b.' *assassinate*$_2$ < agent, patient, instrument >

Technically, the instrument is obligatory *qua* the argument of the predicate (1b') in this analysis. Its apparent optionality is the result of the fact that we are in general free to choose between a lexical entry with (1b') or without (1a') an instrument. An *optional argument* of a predicate P_2 is then such that it (i) bears a thematic role and (ii) there exist a

[1] I will not be concerned with non-semantic arguments (e.g.: *expletives*) in this paper.

[2] Here and throughout the paper I represent argument structure by inserting thematic roles directly into the argument list.

separate lexical entry P_1 in the lexicon which is minimally different from P_2 in not having this argument.[3]

This approach relies on a broad interpretation of argumenthood, which renders it difficult to capture more fine-grained distinctions between different uses of instruments, benefactives, and the like. Focusing on the example of instruments, there exist verbs which entail the existence of an instrument participant and which minimally contrast in this property with *assassinate* (*cut*, *peel*, *sow*, etc., see Levin & Rappaport 1995 and Reinhart 2002).

(2) *John peeled the potato with the knife.*

The conceptual structure of *assassinate* does not entail the existence of an instrument – one can in principle assassinate the president simply by jumping on him/her. One grammatical reflex of this conceptual/semantic difference between the two verbs is that the instrument PP in (2) can, but the one in (1b) cannot be mapped onto SUBJ. This is pointed out in Reinhart (2002), who also calls attention to the fact that *peel* cannot otherwise take cause subjects (4). The noun phrase *the knife* is licensed in (3b) as an instrument.

(3) a. **The dynamite assassinated the president*.
b. *The knife peeled the apple*.

(4) a. **The heat assassinated the president.*
b. **The heat peeled the apple.*

One possible explanation for the grammaticality difference between (3a) and (3b) is that only argument PPs can be re-linked to a term function as a result of a morpholexical operation. In other words, the data in (3) can be explained if we assume that the instrument phrase is an argument of *peel*, but it is an adjunct of *assassinate*.[4] This is in accord with the fact that *peel* entails the existence of an instrument, but *assassinate* does not. If, however, we assume the analysis in (1b'), then it is not possible at the level of argument structure (and, consequently, at f-structure) to differentiate *peel*-type verbs systematically from verbs that do not entail the existence of an instrument (like *assassinate*).

Considerations of this sort have led to proposals in which non-core thematic expressions are not analyzed on a par with regular arguments. Rather, they can be treated as event-

---

[3] If we assume that regular arguments can be reduced (as in the case of the formation of passives, middles or reflexive/reciprocal verbs), then condition (ii) is a necessary but not a sufficient condition for an expression to be an optional argument.

[4] Essentially the same argument is made in Grimshaw (2005:109, an updated version of Grimshaw 1989) to distinguish between two types of *for*-PPs. She claims that the PP in (ia) is a *possessor argument*, whereas the PP in (iia) is a benefactive and "*presumably an adjunct*". Consequently, only the former can undergo dative shift, i.e. be re-linked to a term function, cf. (ib) and (iib).

(i) a. *I'll fix*/*make a sandwich for the children*.
b. *I'll fix*/*make the children a sandwich*.

(ii) a. *I'll fix*/*mend the radiator for the children*.
b. **I'll fix*/*mend the children the radiator*.

internal participants which receive a special *circumstantial role* (Fillmore 1994, Cinque 1999, 2006), or a an *auxiliary theta role* (Ernst:2002, 65). It is important to emphasize that circumstantials receive a thematic role in their own right, unlike the *a-adjuncts* of Grimshaw (1990), which only gain their thematic specification by being related to a suppressed argument position, as is the case with passive *by*-phrases. The generally emerging view is that circumstantials are neither real arguments, nor real adjuncts, but are of a special intermediate sort. Cinque (2006), for example, builds up an analysis in which circumstantials have their own elaborate syntactic domain of insertion inside the VP.

I have started this paper with the point that the argument-adjunct dichotomy is fundamental in LFG. It seems desirable to maintain this design feature, which forces us to be strongly committed to categorizing circumstantials either as arguments or as adjuncts. As we have seen, Bresnan (1982) opts for treating them as arguments. Recently, Asudeh & Toivonen (2005, 2006) have proposed a different LFG-theoretic analysis of a subset of what I am referring to here as circumstantials. Let me concentrate on what they call the *Pgoal* (goal of perception), as this role type will directly be relevant to us. This role, which I regard to be a particular instance of experiencer, is borne by the optional *to*-PP of *seem*-type predicates, as in (5).

(5) *It seemed to Tom that Kate had won*.

Asudeh & Toivonen (2005, 2006) make two important claims concerning the status of the *to*-PP. First, it is claimed to be an adjunct. The most important reason for this is that it is optional syntactically. Second, they assume that it bears a semantic, rather than a thematic role (every thematic role is a semantic role, but not vice versa). For them, only syntactic arguments bear a thematic role. Non-thematic semantic roles are possibly assigned not only to what has been referred to above as circumstantials, but also to time, place and manner adjuncts in general. They only briefly mention that instruments and Pgoals differ from time, place and manner adjuncts in being lexically restricted, but the two types of adjuncts are assigned a semantic type from the same pool of semantic roles (2006:22).

In this paper, I carry out an investigation of dative experiencers in Hungarian, the counterparts of English *to*-experiencers. I show that these datives in Hungarian and the corresponding *to*-PPs in English fall into three empirical classes and I argue that these classes are best captured if we assume the following triadic classification.

(6) (a) *This appeals* ***to me***. [(thematic) argument]

(b) *This is important* ***to me***. [thematic adjunct]

(c) ***To me****, this is nice*. [(non-thematic) adjunct]

The *to*-PP in (6a) is a regular argument with a thematic role, which is selected by the predicate. The *to*-PP in (6c) is a regular event-external adjunct without a thematic role, and it is not selected by the predicate. I claim that the *to*-PP in (6b) is part of the non-core thematic domain, i.e. it is a type of circumstantial PP. I concur with Asudeh & Toivonen (2005, 2006) in categorizing this expression as an adjunct, rather than an argument. I argue nevertheless that these adjunct PPs receive the same sort of thematic specification as regular arguments. Accordingly, I refer to these objects of grammar as thematic adjuncts, or

ADJ_{Θ} for short.

The paper is organized as follows. I first discuss dative experiencer predicates in Hungarian (section 2). I show that dative arguments differ from dative thematic adjuncts in terms of their morphological encoding, their possible range of interpretations, and in terms of optionality (subsection 2.2.). Then I survey some empirical differences between thematic datives and non-thematic dative adjuncts (subsection 2.3.). I present an analysis in section 3. To represent thematic information, I assume a feature decomposition approach to thematic roles as in the Theta System of Reinhart (2000, 2002). I briefly show how Reinhart's Theta System can be accommodated in an LFG grammar (subsection 3.1.). Dative arguments (3.2.) and dative thematic adjuncts (3.3.) are then analyzed in this framework. Since the two thematic domains are formally kept separate along the lines of the argument-adjunct distinction, well-formedness conditions governing thematic entities can now be thought of as being applied to them in a distributive fashion. In particular, the uniqueness condition precludes the co-occurrence of two arguments or two thematic adjuncts of the same thematic type, but it is in principle not ruled out that an argument may co-occur with a thematic adjunct of the same thematic type. I show that this approach makes the right predictions with respect to how the thematic datives under investigation can be interpreted. I sum up in section 4.

2. Three types of dative experiencers in Hungarian: an empirical overview

2.1. *The predicate classes that take thematic datives*

Predicates that take dative experiencer arguments can be classified as falling into two groups in Hungarian. The first of these contains regular *piacere*-predicates (cf. Belletti & Rizzi 1988). I refer to the second group as *verbs of mental appearance*. This group includes psych-verbs which denote the emergence of a mental image or the resulting emotional state, and which are often metaphoric in nature. Respective examples are given in (8).

(7) **Verbs taking dative arguments**

(i) Group 1: *Piacere*-predicates

derogál	'it is beneath one's dignity to'
jól/ rosszul esik	'feels good/bad to' [*lit*. 'falls well/badly to sb]
sikerül	'succeeds, works well'
tetszik	'appeals to'

(ii) Group 2: *Verbs of mental appearance*

feltűnik	'appears, attracts attention'	
beugrik	'clicks, the recognition comes'	[*lit*. 'jumps in to sb']
bejön	'likes'	[*lit*. 'comes in to sb']
leesik	'gets it, picks it up'	[*lit*. 'falls down to sb']

(8) a. *Te tetsz-el János-nak.*
you appeal-2SG John-DAT
'You appeal to John.'

b. *Ez* *be-jön* *Kati-nak.*
this in-come Kate-DAT
'Kate likes this.' [*lit*. 'This comes in to Kate.']

The number of predicates that take dative thematic adjuncts is relatively large. The two most important groups are modal and evaluative predicates.[5] Besides, there exist a variety of verbs that take optional dative experiencers, some of which I list in Group 3 below.[6] I provide an example sentence with an evaluative adjective in (10).

(9) **Verbs taking dative thematic adjuncts**

(i) Group 1: Evaluative predicates

elég	'enough'
fontos	'important'
jó	'good'
kellemes	'pleasant'
korai	'early'

(ii) Group 2: Modal predicates

kell	'needs'
kötelező	'obligatory'
lehetséges	'possible'
szükséges	'necessary'

(iii) **Group 3: Miscellaneous verbs licensing dative thematic adjuncts**

számít	'matters, counts'
tűnik	'seems'
jelent	'means'
hiányzik	'is missing to'

(10) *Ez* *nagyon* *fontos* *nek-em*.
this very important DAT-1SG
'This is very important to me.'

In the next subsection (2.2), I substantiate first the empirical difference between dative experiencer arguments (7) and dative thematic adjuncts (9). Then in 2.3, I provide some arguments for why these two sorts of datives should be distinguished from non-thematic, event external dative adjuncts. As we will see, such dative adjuncts can be inserted freely into any sentence, and are not required to be selected by particular types of predicates.

2.2. *Dative arguments vs. dative thematic adjuncts*

There are three important properties in which dative arguments differ from the proposed

[5] See Dalmi (2005), Komlósy (1994), É. Kiss (2002) and Tóth (2000), among others, for a discussion of these predicate classes. These authors all assume that modal and evaluative predicates take dative arguments in Hungarian.

[6] A comprehensive taxonomy of the verbs in Group 3 can be found in Rákosi (2006).

dative thematic adjuncts. First, the morphology of dative arguments is always fixed in the lexicon, whereas dative thematic adjuncts can be coded by competing morphological markers. In English, experiencer arguments are always marked by the preposition *to* (11), whereas thematic adjuncts can generally appear as complements of either *to* or *for* (12).

(11) a. *You appeal to/*for John.*
b. *The same thought occurred to/*for me.*

(12) a. *This doesn't matter to/for me.*
b. *This is important to/for John.*
c. *This is forbidden to/for us.*
d. *This seems to/%for me the best option.*

There is some dialectal variation in the data in (12). In particular, *seem*-type raising predicates only dialectally license *for*-PPs, and most speakers of standard British or American English find it marginal or unacceptable. Furthermore, the two prepositions are not always completely equivalent semantically, a point I come back to directly. What is immediately relevant is that such morphological variation only exists in the case of the PPs in (12), but not in the case of the PPs in (11). The existence of this contrast is expected, even if not fully explained, under the assumption that only the PPs in (11) are arguments. Being arguments, their morphology can be closed in the lexicon. The morphological encoding of adjuncts (12), however, is generally relatively free, subject only to the available morphological inventory of the given language.

In Hungarian, dative thematic adjuncts are either marked by dative case or by the inflecting postposition *számára* 'for' (14).[7] Dative arguments require dative case (13).

(13) a. *Kati tetsz-ik János-nak / *János számára.*
Kate appeal-3SG John-DAT John for.3SG
'Kate appeals to/*for John.'

b. *Ez be-jön János-nak / *János számára.*
this in-come John-DAT John for.3SG
'John likes this.' [lit. 'This comes in to/*for John.']

(14) a. *Úgy tűn-t János-nak / János számára, hogy ... .*
so seem-PAST John-DAT John for.3SG that
'It seemed to/%for John that ... .'

b. *Ez fontos János-nak / János számára.*
this important John-DAT John for.3SG
'This is important to/for John.'

c. *Ez sok-at jelent nek-em / számomra.*
this much-ACC means DAT-1SG for.1SG
'This means a lot to/for me.'

[7] As far as terminology is concerned, I abstract over this morphological variation and continue using the term *dative thematic adjunct* to refer to expressions that are in fact marked by the postposition *számára* 'for'.

Thus the contrast that we have observed in English reappears in Hungarian.[8]

The second property in which the two types of dative expressions in question differ concerns their semantics. Dative arguments of *piacere*-predicates and of *verbs of mental appearance* are necessarily interpreted as experiencers, that is, they refer to participants whose mental state is described by the predicate. In the case of dative thematic adjuncts, however, it is generally possible to have a non-psych construal, too. Consider the following two sentences, the first of which has an argument dative, and the second has a dative thematic adjunct.

(15) a. *János-nak tetsz-ik a meleg idő.*
John-DAT appeal-3SG the warm weather
'Warm weather appeals to John.'

b. *János-nak nem számít a meleg idő.*
John-DAT not matter the warm weather
'Warm weather does not matter to John.'

If (15a) is uttered, John must have favorable dispositions towards warm weather. It does not matter whether he is aware of it or not, he must like warm weather by (15a). (15b), on the other hand, can be used to describe a property of John, and it need not tell us anything about his dispositions (cf. Arad 1998 for similar data). (15b) is compatible with the assertion that John himself in fact believes that warm weather matters to him – but somebody else (his coach, for example) knows that this is not true and John can play in warm weather just as well as in cold weather.

That dative thematic adjuncts need not be interpreted as experiencers is evident in light of the fact that most predicates that license them allow them to be inanimate. *For* tends to be preferred to the preposition *to* in these environments, but *to* is also grammatical, cf. (16b).

(16) a. *Garlic is good for the vocal cords*.
b. *Oceans are important to the environment*.
c. *We don't know what these particles were, but it doesn't matter for the theory*.

In Hungarian, most speakers require the complement of the postposition *számára* 'for' to be animate and therefore dative case is generally the only option with inanimates.

(17) *A fokhagyma jó a hangszalag-ok-nak / *hangszalag-ok számára.*
the garlic good the vocal.cord-PL-DAT vocal.cord-PL for.3SG
'Garlic is good to/for the vocal cords.'

[8] The cross-linguistic distribution of dative case and *for*-type P-elements is somewhat more complex. As far as I am aware, it generally holds that dative arguments (e.g. the *piacere*-class) cannot be coded by anything else than dative case if dative is available in the language. There is some variation in whether dative case is grammatical on thematic adjuncts (especially with modals and evaluatives), or only an equivalent of *for* can be used. In Russian, the preposition *dlja* 'for' is licensed on thematic adjuncts as an alternative to dative case, whereas Romanian only allows dative case with some of these predicates and it is more common to use the preposition *pentru* 'for' instead. In Italian and Czech, dative case is almost never allowed on thematic adjuncts and a preposition must be used (*per* in Italian and *pro* in Czech, both meaning 'for').

Some of the predicates that license a dative thematic adjunct have a [+animate] selectional restriction on it. *Kellemes* 'pleasant', *kellemetlen* 'unpleasant' and *kényelmes* 'comfortable' belong to this cross-linguistically identifiable group. This selectional restriction notwithstanding, a non-psych reading is still available.[9] Consider this Hungarian example.

(18) *Ez a helyzet kellemetlen János-nak / János számára.*
this the situation unpleasant John-DAT John for.3SG
'This situation is unpleasant to/for John.'

(18) is ambiguous the same way as (15b). It has a psych-reading that expresses John's mental state, but it can also be uttered in a context in which the speaker knows that John is in fact enjoying the situation. There is a slight preference to use *számára* 'for' to render the non-psych reading, but the ambiguity is generally present both with dative case and the postposition.

The third property in which dative thematic adjuncts differ from dative arguments is their optionality. Indeed, this is the crucial property which, I argue, makes them what they are. An argument can be optional in the sense that it is semantically closed in the lexicon and is not represented in c-structure (cf. Bresnan 1982). Using generally accepted terminology, this argument remains implicit. In English, dative arguments always have to be expressed phonologically; observe the contrast between *appeal to* and *matter*. The latter is analyzed here as taking a dative thematic adjunct.

(19) a. *This does not appeal *(to me).*
b. *This does not matter (to me).*

In Hungarian, a dative argument can be omitted in appropriate discourse, but it necessarily has to be interpreted as a definite implicit argument, whose intended referent is always recoverable from the context. By default, the referent of an implicit dative argument is identified with the speaker. Consider the following example.

(20) *Manapság nem népszerű ez a könyv, de régen tetsz-ett.*
these.days not popular this the book but formerly appeal-PAST
(i) 'This book is not popular these days, but it used to appeal to me.'
(ii) *'This book is not popular these days, but it used to appeal to someone.'
(iii) *'This book is not popular these days, but it used to appeal to people.'

The only possible interpretation of this sentence is the one given under (i), and there is no discourse which would license the interpretations (ii) and (iii). It is obvious that the existence of a definite dative experiencer argument is always entailed, and Hungarian only differs from English in allowing this dative argument to remain implicit.

The question is then how to approach (19b) or (21). These two sentences contain predicates that I claim to license an optional dative thematic adjunct.

[9] The only exception to this is 'tűnik' *seem*. This predicate necessarily reflects the mental state of the referent of its optional dative thematic adjunct.

(21) *A fehér galamb jelent-i a béké-t.*
the white dove mean-3SG the peace-ACC
'The white dove means peace.'

(21) describes a property of the white dove, rather than a relation between it and an individual or a set of individuals (cf. Jackendoff to appear, Cuervo 2003). It is evident that there is no definite implicit dative involved, in contrast with (20). We could still assume, however, that a dative thematic adjunct can be closed existentially or universally, in which case (21) corresponds roughly to (22a) or (22b). I use English for ease of exposition.

(22) a. *The white dove means peace to someone.*
b. *The white dove means peace to everyone.*

The objective or "*perspective-free*" (Jackendoff to appear) interpretation of (21) is certainly compatible with (22a), and (22b) seems to be a good paraphrase of what (21) means. I want to argue nevertheless that (21) is not equivalent to the construction in (22b) in semantic terms. It is known that universal quantification tolerates exceptions (23a), but it does not tolerate massive exceptions (23b,c).

(23) a. *The white dove means peace to everyone, but John does not know this symbol.*
b. *#The white dove means peace to everyone,*
though most people do not know this symbol.
c. *#The white dove means peace to everyone, though no one knows this symbol.*

(23b), and especially (23c), are contradictory. However, if the dative is unexpressed in the matrix clause, such construction generally become acceptable to most speakers.

(24) a. *The white dove means peace, though most people do not know this symbol.*
b. *The white dove means peace, though no one really knows this symbol.*

This effect is even stronger if some domain restriction is placed on the predicate; compare the following two sentences.[10]

(25) a. *#This book is historically important for everyone,*
though not for anybody anymore.
b. *This book is historically important, but not for anybody anymore.*

It does not seem to be the case therefore that a construction of the type *The white dove means peace* can be reduced to *The white dove means peace to everyone*. In other words, a dative thematic adjunct can be genuinely absent not only in the syntactic sense, but also in the sense of not being entailed by the predicate. As such, it is not represented as an implicit argument, and the predicates *important* or *mean*, as well as the rest of the predicates in (9) can be thought to be truly monadic. A dative thematic adjunct is only added optionally.

What I am claiming is that *This book is important* does not entail that the book is important for anyone, though it is entailed that someone will find it important. Crucially,

[10] This has been brought to my attention by Scott Grimm at the LFG06 conference in Konstanz.

however, the two participants in question are of two different syntactic types, and they may in fact co-occur in the same clause. Consider the following examples.[11]

(26) a. *To me, this book is important for mankind*.

b. *Számomra, ez a könyv fontos az emberiség-nek*.
for.1SG this the book important the mankind-DAT
'For me, this book is important to mankind.'

The first adjunct (the *to*-PP in (26a) and the postpositional phrase in the Hungarian example (26b)) is a real, non-thematic adjunct, external to the VP and to the event. It identifies an individual who is taken to be the immediate anchor of the model in which the embedded proposition is interpreted. These initial PPs are like time, place, and manner adjuncts which function semantically as clause-level operators. As Bresnan (1982) points out, these frames or anchors are entailed for every clause, but this does not make them arguments. In the next section, I briefly discuss these event-external experiencers. Continuing my previous practice, I make the terminological convenience of talking uniformly about dative adjuncts irrespective of the actual encoding (dative case or P-marker).

2.3. *Non-thematic dative adjuncts*

An event-external experiencer dative can be added to any declarative sentence, as in the following English and Hungarian examples.

(27) a. *To her and her colleagues, this is simply a great opera*.
b. *For me, it is first and foremost an intellectual exercise*.
c. *To me, he had a great life in London - seeing friends, eating out, having time to read ... .*

(28) *Számomra / nek-em, a barátság az barátság*.
for.1SG DAT-1SG the friendship that friendship
'For/to me, friendship is friendship.'

These expressions are adjuncts, which is evidenced by the fact that their morphology is not fixed (cf. 2.2). I analyze them as non-thematic, regular adjuncts. In this subsection, I briefly discuss some properties by which these differ empirically from dative arguments and event-internal dative thematic adjuncts.

First of all, event-external dative adjuncts do not need to be licensed by specific classes of predicates. (27) and (28) show that they can be inserted quite freely. Second, they are ungrammatical or very marginal in a predicate-internal c-structure position. This is true even in Hungarian, a language which is known to be non-configurational (cf. É. Kiss 2002). Thus whereas a dative thematic adjunct can be inserted between its predicate and the subordinate clause (29a), a non-thematic dative adjunct cannot (29b). It typically occurs on the left edge of the clause (29c).

[11] This construction sounds best in both languages if a different morphological marker is chosen for the two experiencers.

(29) a. *Fontos nek-em, hogy itt-marad-j*.
important DAT-1SG that here-stay-SUBJUNCTIVE.2SG
'It is important for me that you should stay here.'

b. ??/*_Butaság nek-em, hogy itt-marad-sz_.
stupidity DAT-1SG that here-stay-2SG
'It is a stupidity to me that you stay here.'

c. *Nek-em butaság, hogy itt-marad-sz*.
DAT-1SG stupidity that here-stay-2SG
'To me, it is a stupidity that you stay here.'

Third, dative thematic adjuncts can host anaphors in Hungarian whereas non-thematic dative adjuncts cannot, or only very marginally.

(30) a. *Egymás-nak fontos-ak vagyunk*.
each.other-DAT important-PL are.1PL
'To each other, we are important.'

b. ??/*_Egymás-nak vicces-ek vagyunk_.
each.other-DAT funny-PL are.1PL
'To each other, we are funny.'

For a more detailed discussion of the syntactic differences between thematic and non-thematic datives, I refer the reader to Rákosi (2006).

Though they do not receive a thematic role, the event-external datives are also interpreted as experiencers in the sense of primarily reflecting upon their referent's dispositions, rather than on their knowledge state. Dative adjuncts minimally contrast with *according to*-type adjuncts in this respect. Imagine a context in which *John* knows that *Kate* is ugly.

(31) **John knows that Kate is ugly but**

(a) #*according to him, she is still beautiful*.

(b) *to him, she is still beautiful*.

In such a context, (31b) can be used felicitously, but (31a) cannot. *According to him* introduces a report on what John knows, whereas *to him* introduces a report on what John feels. *To him* relativizes the interpretation of the proposition *Kate is beautiful* to a model determined by John's disposition or feelings, and this model can in principle be compatible with a model that represents John's knowledge state and which contains the proposition *Kate is ugly*. This shows that non-thematic dative adjuncts have experiencer semantics. Indeed, this is the reason why they are discussed here together with thematic datives. However, being external to the event, they do not receive a thematic role, and their experiencerhood can be thought of as a semantic role, as this notion is understood by Asudeh & Toivonen (2005, 2006).

3. Datives in two thematic domains

3.1. *The Theta System of Reinhart (2000, 2002)*

I briefly introduce here the Theta System of Reinhart (2000, 2002), which is a lexicalist thematic theory that provides us with the tools that enable us to capture the behavior of dative thematic adjuncts, and to compare them with dative arguments. I focus on those aspects of the Theta System that are immediately relevant.

Reinhart's proposal shares some of its basic background assumptions with Dowty's (1991) proto-role analysis and with subsequent LFG-based proto-role analyses (such as Ackerman 1992, Zaenen 1993, and Alsina 1996). In particular, the traditional system of discrete thematic roles is rejected in the Theta System, and the exact interpretation of thematic content is thought to be possibly context-dependent. The Theta System differs from proto-role accounts as well as from the standard lexical mapping theory of LFG (cf. Bresnan & Kanerva 1989), inasmuch as thematic information is decomposed into two binary features: +/–*cause* and +/–*mentally involved*. A traditional agent argument, for example, is coded as [+c+m], since an agent-type participant is causally responsible for the event and he/she is also mentally involved. A patient is coded negatively as a participant that is not causally responsible and whose mental state is not necessarily relevant in the interpretation of the event: [–c–m]. The value of either attribute can be left unspecified in the lexicon, as in the case of the subject argument of *open*, which can be interpreted as an agentive or a non-agentive cause: [+c]. Traditional thematic role labels are not recognized formally in the Theta System and are not used to characterize the thematic content of arguments. Terms like *agent*, *patient*, and the like are at best names of semantic roles. In (32), I list the thematic roles of the Theta System and the semantic roles they correspond to.

(32) ***Feature decomposition of thematic roles in the Theta System***

THEMATIC ROLES	CORRESPONDING SEMANTIC ROLES
[+c+m]	agent
[+c–m]	instrument
[–c+m]	experiencer (*worry*-type object experiencers)
[–c–m]	theme, patient
[+c]	cause
[+m]	sentient, experiencer (*like*-type subject experiencers)
[–m]	subject matter (cf. Pesetsky 1995), locative source
[–c]	goal, benefactive, recipient, experiencer (*appeal to*)

In LMT, features are used to decompose argument functions. The features [+/–r(estrictive] and [+/–o(bjective)] mediate between the thematic content of an argument and the syntactic function the argument is mapped onto. These features are thus essential for the mapping. In the Theta System, the [+/–c] and the [+/–m] features are interpretable, and they are used to encode thematic information directly.

Reinhart's (2000, 2002) mapping proposal has two dimensions. One is a distinction between external and internal arguments. This is not directly relevant for us, and I refer the reader to the works cited for the details. The other dimension concerns the role of case in

the mapping.[12]

(33) ***Mapping of arguments in the Theta System: the case dimension***

Given an n-place verb, n>1

a. Mark the verb with the ACC feature if the entry includes both a [+]-role (i.e., [+c+m], [+m] or [+c]), and a fully specified cluster with the [–c] feature (i.e., [–c+m] or [–c–m]).

b. The [–] unary features (i.e., [–m] and [–c]) require inherent case (or an adposition).

Since the EPP is assumed, the argument which does not receive lexically specified case generally becomes a subject. In case of monadic entries, the single argument is always a subject. For our purposes, this can be adopted into an LFG-framework as follows.

(34) ***An LFG-theoretic reconsideration of (33)***

a. Assume the Subject Condition.

b. An argument specified as [–c+m] or [–c–m] is mapped onto OBJ in the presence of an argument specified as [+c+m], [+m] or [+c].

c. An argument specified as [–m] or [–c] is mapped onto OBL_{Θ}.

d. Other arguments, including the single argument of a monadic predicate, are mapped onto SUBJ.

(34) will be sufficient for our purposes.

We need two further ingredients for the analysis. First, uniqueness is recognized in the Theta System as a well-formedness condition on argument structures (cf. Marelj 2004). I make specific arguments in Rákosi (2006) that uniqueness can be interpreted as directly applying to thematic features, rather than to semantic roles. In other words, an argument structure which includes two [–c–m] roles at the same time is ungrammatical, even though one argument could be interpreted as a theme (i.e. unaffected entity that moves) and the other as a patient (affected entity), and thus they would not share the same semantic role. Second, Marelj (2004:67) proposes the following rule on unary thematic roles.

(35) ***The Principle of Full Interpretation*** (Marelj: 2004, 67)

For the purposes of interpretation, all thematic roles must be fully specified.

[+c] is a unary thematic role in the sense that the *mental state relevant* attribute is underspecified. The subject of the verb *open* has an underspecified subject of this sort (36a). What (35) says is that in a given context of use, the [+c] argument will be interpreted either as [+c**+m]**, as in (36b), or as [+c**–m**], as in (36c).

[12] Notice that this mapping proposal does not rely on a thematic hierarchy. The traditional thematic hierarchy cannot have a formal role in the Theta System since traditional thematic role labels are not used to specify thematic content.

(36) a. *open*: < [+c] [–c–m] >
b. *John*$_{[+c+\mathbf{m}]}$ *opened the window*$_{[-c-m]}$.
c. *The wind*$_{[+c\mathbf{-m}]}$ *opened the window*$_{[-c-m]}$.

This 'expansion' is governed by uniqueness, that is, it cannot result in two identical fully specified roles.

Finally, let me make a remark on how experiencers are encoded in the Theta System. As I have shown in (32), a traditional experiencer may be encoded as [–c+m], [+m] or [–c]. [–c+m] is the specification that the object of *worry*-type predicates receive. The subject of *like*-type subject experiencers is [+m]. Dative experiencers are coded as [–c]. In Rákosi (2006), I argue that this difference is not arbitrary, but there indeed exist thematically relevant differences between arguments that are otherwise all treated as experiencers. For now, it suffices to be aware that by (35) each of these three roles can be interpreted as [–c+m], and therefore there is a possible construal under which the thematic content of these three different types of experiencers is the same. In particular, the [–c] encoding of dative arguments helps us to capture the known conceptual and morphological relatedness between benefactives, recipients and dative experiencers. The generalization over these semantic roles is that they characterize participants that are not causally related to the event, but their mental state can be relevant.

I now turn to the analysis of the dative argument - dative thematic adjunct distinction that I have made empirical arguments for in section 2. The analysis is going to be executed as an extension of the Theta System.

3.2. *Dative experiencer arguments in the Theta System*

Predicates that have dative experiencer arguments (7) are represented in the form of the following type of dyadic lexical entry.

(37) *tetszik* 'appeal to < [–c–m] [–c] >'

The dative argument is coded as [–c], as discussed at the end of subsection 3.1. The subject argument of these predicates is coded as [–c–m] in the Theta System. The [–m] feature is self-explanatory, given that the mental state of a stimulus-type participant is not relevant in the event. The [–c] specification of the subject is in accordance with Pesetsky's claim that *piacere*-predicates do not have a cause(r) subject. For Pesetsky, this argument bears the role *target* (*of emotion*). A target-type participant is simply evaluated, but it does not affect the experiencer.[13]

The mapping of the entry in (37) is described in (38).

[13] As Pesetsky (1995) points out, *appeal to* minimally differs from *please* in this respect. *Please* does have a cause subject. One consequence is that only *please* can be passivized.

(i) a. **John was appealed to by the film*.
b. *John was pleased by the film*.

(38) *tetszik*$_V$ 'appeal to < [–c–m] [–c] >

	$OBL_{[-c]}$	[34c]
SUBJ		[34d, *Subject Condition*]

I now repeat (8a) as (39). (40) is the simplified f-structure of (39).

(39) *Te tetsz-el János-nak.*
you appeal-2SG John-DAT
'You appeal to John.'

(40)

SUBJ	PRED	'PRO'
	CASE	NOM
	PERS	2
	NUM	SG
PRED	'appeal to < (SUBJ) ($OBL_{[-c]}$) >'	
$OBL_{[-c]}$	PRED	'JOHN'
	CASE	DAT
	PERS	3
	NUM	SG

I have argued in 2.2. that dative experiencer arguments are morphologically closed; they can only be coded by dative case. Following Bresnan & Kaplan's (1982) analysis, I assume that *to* is a non-predicative case-marker that is used to code [–c]-arguments, including experiencers of *appeal to* predicates. I consider dative case in Hungarian to have the same function. It is a *semantic case* in the sense of Butt & King (2005).[14] The lexical entry for (this use of the) Hungarian dative case can be described as in (41), in the manner of Butt & King's analysis of Urdu/Hindi semantic case markers.

(41) *-nVk*: (↑CASE)=DAT
($OBL_{[-c]}$↑)

This requires Hungarian dative case to be associated with a [–c] oblique argument.

Finally, I have shown in 2.2. that dative experiencers of *piacere*-predicates and of verbs of mental appearance can only be interpreted as experiencers. This can be derived in the Theta System as follows (cf. Marelj 2004). By the Principle of Full Interpretation (35), every argument has to be fully specified for the purposes of interpretation. In the current case, the [–c] argument is underspecified, and two different interpretations can be

[14] Hungarian dative case satisfies the two defining characteristics of semantic case that Butt & King (2005) propose: (i) predictability via the formulation of generalizations across predicates and constructions, and (ii) a subjection to syntactic restrictions (such as only appearing on certain grammatical functions).

computed.

(42) a. *tetszik* 'appeal to < [–c–m] [–c**+m**] >'
b. **tetszik* 'appeal to < [–c–m] [–c**–m**] >'

(42b), however, is ruled out by the uniqueness constraint, since two [–c–m] roles co-occur in the same argument structure. Therefore, only (42a) is a well-formed interpretation, and it indeed requires the dative argument to be construed as an experiencer.

3.3. *Dative thematic adjuncts in the Theta System*

I argued in 2.2. that modal and evaluative predicates, as well as the verbs listed in (9) do not select for a dative argument. They only license a dative thematic adjunct. Let me first characterize the argument structure of these predicates. In the previous subsection 3.2, I argued that the subject of a verb that takes a dative experiencer argument is a non-cause *target*, or [–c–m] in terms of the Theta System. In contrast, the subject of a predicate that takes a dative thematic adjunct can often be interpreted as a potential cause. Consider the following English examples.

(43) a. *The situation was unpleasant for John*.
b. *The water was not enough for all the soldiers*.
c. *This matters to me a lot*.

The PPs refer to participants which are all affected in the relevant events. An affected participant generally occurs at one end of a causality chain which starts with a cause. Affectedness is not included among the thematic features in the Theta System, but being a cause is thought to be a thematically relevant property. I assume consequently that the subject of the predicates in (43) is a potential cause, or [–m] in terms of the Theta System, underspecified for the *cause*-feature. I assume furthermore that the predicates that license [–c] dative thematic adjuncts all have a [–m] subject argument. We can possibly derive the licensing condition of dative thematic adjuncts from this, on analogy with the licensing of non-argument instruments, which I take to be thematic adjuncts of the [+c–m]-type. It is well-known that an instrument phrase is licensed by any predicate that has an agent argument ([+c+m]). Similarly, a [–m] subject, which can potentially be interpreted as a cause, can license a non-cause participant [–c], which can be affected in the event denoted by the predicate.

The lexical entry for the representative predicate *fontos* 'important' is given in (44). For expository purposes, I represent the optional thematic adjunct in brackets outside the argument list and outside the semantic form of the predicate. This notation is not intended to mean that the dative is a non-thematic argument, nor do I want to suggest that it is in fact part of the argument list in any way.

(44) *fontos* 'important < [–m] >' ([–c])

(44) simply tells us that the semantic form of the predicate *fontos* 'important' is such that it optionally can license a thematic adjunct. But (44) is a monadic entry, and its single

argument will be mapped onto SUBJ by the Subject Condition.

We have seen that this adjunct can be coded by dative case as well as by the postposition *számára* 'for'. I assume that these two are not simply case-markers, but are predicative P-elements (cf. Bresnan 1982, footnote 12). This entails that dative case has another lexical entry in Hungarian, which crucially differs from the case-marker (41) in having a PRED feature. The two relevant lexical entries therefore minimally encode the following information.

(45) a. *-nVk*$_2$ 'to < (OBJ) >'
b. *számára* 'for < (OBJ) >'

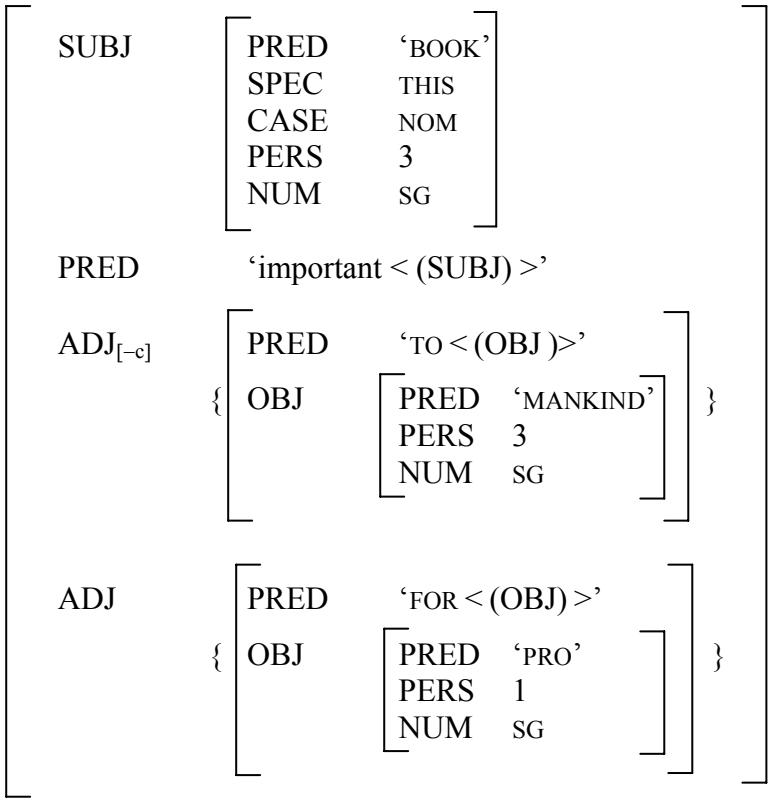
(45) includes only the bare minimum that is needed now. Most importantly, it should also be specified that the dative in (45a) is still a morphological case, whereas (45b) is a postposition.[15] What is important, however, is that these markers are predicative. Since the postposition *számára* 'for' only has a predicative use, it follows that it cannot mark arguments, since PRED features cannot be unified.

I now repeat (26b) as (46). Recall that this sentence has two adjuncts: the initial postpositional phrase is a non-thematic adjunct, and the predicate-internal dative DP is a thematic adjunct. I focus on the essentials in the f-structure (47). For expository purposes, I have placed the thematic adjunct into a distinct set. This need not be necessary: I assume that the thematic tag on an adjunct sufficiently identifies it as being of a category that has its own distinguishing syntactic properties.

(46) *Számomra, ez a könyv fontos az emberiség-nek.*
for.1SG this the book important the mankind-DAT
'For me, this book is important to mankind.'

[15] It is well-known though in the literature on Hungarian that case markers have developed historically from postpositions. There are many synchronic similarities between dative case and *számára* 'for'. Both take a bare (non-casemarked) complement, which I simply analyze below in (47) by assuming that the CASE feature is not defined on the complement. Besides, the complement of the postposition *számára* 'for' can only marginally be extracted out of the PP, which shows that the two make up a tight morphological unit.

(47)



Finally, I need to account for the fact that dative thematic adjuncts do not have to be interpreted as experiencers (cf. 2.2). To be able to do so, I assume a relativized understanding of the uniqueness constraint (Rákosi 2006). The basic idea is that thematic arguments and thematic adjuncts represent two distinct thematic domains.

(48) **Relativized thematic uniqueness**

a. The thematic specification of the arguments of a predicate is unique.
b. The thematic specification of the thematic adjuncts of a predicate is unique.
c. Uniqueness is relative to the given thematic domain of application.

By the Principle of Full Interpretation (35), the entry in (44) can be fully specified thematically in four different ways, assuming the thematic adjunct is present.

(49) a. *fontos* 'important < [+**c**–m] >' [–c+**m**]
b. *fontos* 'important < [+**c**–m] >' [–c–**m**]
c. *fontos* 'important < [–**c**–m] >' [–c+**m**]
d. *fontos* 'important < [–**c**–m] >' [–c–**m**]

(49b,d) represent the non-experiencer uses of the thematic adjunct. (49d), in particular, is well-formed even if the same thematic role occurs twice ([–c–m]). This is so because one is assigned to an argument, and the other is assigned to a thematic adjunct, and the uniqueness constraint has been relativized in (48) to apply over these two domains distributively.

4. Conclusions

In this paper, I have explored the possibility that so-called circumstantial PPs can be systematically analyzed as thematic adjuncts. I have focused on dative experiencers, and I hope to have shown that there is empirical motivation behind the assumption that not only arguments, but also certain adjuncts can receive a thematic role. One possible objection to this step may be that it makes our grammar less restrictive, as the inventory of basic syntactic categories is enriched. I believe the introduction of thematic adjuncts is motivated, at least for the following reasons.

First of all, notice that in the strict sense, I have not introduced another basic syntactic category. A thematic adjunct is a type of adjunct which differs from other adjuncts in receiving thematic specification. I have argued that the Theta System of Reinhart (2000, 2002) provides us with tools to express this thematic specification without actually expanding the inventory used in the thematic coding of arguments. Second, by the introduction of thematic adjuncts it becomes in principle possible to eliminate the notion of 'possible arguments' from the grammar. I have shown that this can be achieved with respect to dative experiencers: there is no optional dative experiencer argument. If a dative experiencer is optional, it is taken to be an adjunct in the current proposal. In Rákosi (2006), I argue that instruments, comitatives, and benefactives can be similarly analyzed. Third, thematic adjuncts can be distinguished systematically from event-external, non-thematic adjuncts. Typically, thematic adjuncts are licensed in the presence of a designated type of argument, as is true of instruments and the datives discussed. Event-external adjuncts are generally not selected by the predicate, therefore they cannot be licensed as thematic expressions. I have tried to present independent evidence for this partition of adjuncts. Notice that thematic datives have been argued to refer to possibly affected participants. Being affected is typically a property associated with thematic entities. Event-external dative adjuncts cannot refer to affected participants. Furthermore, I have shown in 3.3. that the thematic analysis of certain dative adjuncts gives the right predictions with respect to how they can be interpreted. Fourth, the current proposal combines what I believe to be the attractive aspects of the traditional LFG approach to circumstantial PPs with the insight of Asudeh & Toivonen (2005, 2006), inasmuch as circumstantial PPs are assigned a thematic role but they are still treated as adjuncts.

References

Ackerman, Farrell. 1992. Complex Predicates and Morpholexical Relatedness: Locative Alternation in Hungarian. In Ivan A. Sag & Anna Szabolcsi eds. *Lexical Matters*. Stanford, CA: CSLI Publications. 55-83.

Alsina, Alex. 1996. *The Role of Argument Structure in Grammar. Evidence from Romance*. Stanford: CSLI.

Arad, Maya. 1998. *VP-structure and the Syntax-Lexicon Interface*. PhD dissertation. University College of London.

Asudeh, Ash & Toivonen, Ida. 2005. Copy Raising and its Consequences for Perceptual Reports. Pre-publication draft to appear in Jane Grimshaw, Joan Maling, Christopher Manning, Jane Simpson & Anni Zaenen eds. *Architectures, Rules and Preferences*: *A Festschrift for Joan Bresnan*. Stanford CSLI. http://http-server.carleton.ca/~asudeh/

Asudeh, Ash & Toivonen, Ida. 2006, June 1. Copy Raising and Perception. Submitted draft for semanticsarchive.net. http://http-server.carleton.ca/~asudeh/

Belletti, Adriana & Rizzi, Luigi. 1988. Psych-verbs and θ-theory. *Natural Language and Linguistic Theory* 6. 291-352.

Bresnan, Joan. 1982. Poliadicity. In Joan Bresnan ed. *The Mental Representation of Grammar*. Cambridge, MA: The MIT Press. 149-172.

Bresnan, Joan & Kanerva, Jonni M. 1989. Locative Inversion in Chicheŵa: a Case Study of Factorization in Grammar. *Linguistic Inquiry* 20 (1). 1-50.

Bresnan, Joan & Kaplan, Ronald M. 1982. Grammatical Representation. In Joan Bresnan ed. *The Mental Representation of Grammar*. Cambridge, MA: The MIT Press. 173-281.

Butt, Miriam & King, Tracy H. 2005. The Status of Case. In Veneeta Dayal & Anoop Mahajan eds. *Clause Structure in South Asian Languages*. Berlin: Kluwer Academic Publishers.

Cinque, Guglielmo. 1999. *Adverbs and Functional Heads*. *A Cross-linguistic Perspective*. New York/Oxford: OUP.

Cinque, Guglielmo. 2006. Complement and Adverbial PPs: Implications for Clause Structure. In Guglielmo Cinque. *Restructuring and Functional Heads. The Cartography of Syntactic Structures*, *Volume 4*. Oxford: OUP. 145-166.

Cuervo, María C. 2003. *Datives at Large*. PhD dissertation. Cambridge, MA.: MIT.

Dalmi, Gréte. 2005. *The Role of Agreement in Non-finite Predication*. Amsterdam: John Benjamins.

É. Kiss, Katalin. 2002. *The Syntax of Hungarian*. Cambridge: Cambridge University Press.

Ernst, Thomas. 2002. *The Syntax of Adjuncts*. Cambridge: CUP.

Fillmore, Charles. J. 1994. Under the Circumstances (Place, Time, Manner, etc.). In Susanne Gahl, Andy Dolbey & Christopher Johnson eds. *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society. General session dedicated to the contributions of Charles J. Fillmore*. Berkeley Linguistics Society. 158-172.

Grimshaw, Jane. 1990. *Argument Structure*. Cambridge, Mass.: The MIT Press.

Grimshaw, Jane. 2005. Datives, Feet and Lexicons. In *Words and Structure*. Stanford: CSLI. 107-141.

Jackendoff, Ray. To appear. Experiencer Predicates and Theory of Mind. In *Language, Culture, Consciousness: Essays on Mental Structure*. MIT Press. http://people.brandeis.edu/~jackendo/

Komlósy, András. 1994. Complements and Adjuncts. In Ferenc Kiefer & Katalin É. Kiss eds. *The Syntactic Structure of Hungarian. Syntax and Semantics* Volume 27. San Diego: Academic Press. 91-178.

Levin, Beth & Rappaport Hovav, Malka. 1995. Unaccusativity at the Syntax - Lexical Semantics Interface. Cambridge, Mass.: The MIT Press.

Marelj, Marelj. 2004. *Middles and Argument Structure Across Languages*. PhD Dissertation. University of Utrecht.

Pesetsky, David. 1995. *Zero syntax. Expriencers and Cascades*. Cambridge, Mass.: The MIT Press.

Rákosi, György. 2006. Dative Experiencer Predicates in Hungarian. PhD dissertation. University of Utrecht.

Reinhart, Tanya. 2000. *The Theta System: Syntactic Realization of Verbal Concepts.* UIL-OTS Working Papers in Linguistics 00.0. University of Utrecht.

Reinhart, Tanya. 2002. The Theta System - an Overview. *Theoretical Linguistics* 28. 229-290.

Tóth, Ildikó. 2000. *Inflected Infinitives in Hungarian*. PhD dissertation. Tilburg University.

Zaenen, Anni. 1993. Unaccusativity in Dutch: Integrating Syntax and Lexical Semantics. In James Pustejovsky ed. *Semantics and the Lexicon*. Dordrecht: Kluwer. 129-161.

APPOSITION AS COORDINATION: EVIDENCE FROM AUSTRALIAN LANGUAGES

Louisa Sadler Rachel Nordlinger

University of Essex University of Melbourne

Proceedings of the LFG06 Conference

Universität Konstanz, Germany

Miriam Butt and Tracy Holloway King (Editors)

2006
CSLI Publications
http://www-csli.stanford.edu/

Abstract

Using data from a range of Australian languages, in this paper we argue for an analysis of various nominal appositional structures as syntactic coordinations (i.e. as hybrid f-structures) in LFG. We show that this provides a simple and straightforward account of the surface syntactic similarities among a range of juxtaposed construction types, while the differences between the constructions can be accounted for in the mapping to the semantics. We propose meaning constructors to capture the semantic differences between coordination and apposition.[1]

1 Introduction

Using data from a range of Australian languages, in this paper we argue for an analysis of various nominal appositional structures as syntactic coordinations (i.e. as hybrid f-structures) in LFG. We show that this provides a simple and straightforward account of the surface syntactic similarities among a range of juxtaposed construction types, while the differences between the constructions can be accounted for in the mapping to semantic structure. We propose meaning constructors to capture the semantic differences between coordination and apposition, adapting the standard treatment of the semantics of NP coordination to asyndetic coordination and providing a first proposal for a semantics for appositions of these types in LFG.

The present paper is organised as follows. In Section 2 we outline the LFG analysis of coordination by way of background. Section 3 introduces the use of simple nominal juxtaposition structures in Australian languages and the range of interpretations which they receive. In section 4 we provide an analysis of the syntax and semantics of these constructions which captures both their similarities and the differences between them. Section 5 then briefly discusses the occurrence of discontinuous juxtapositions and how they fit into our proposal. We conclude in section 6 with some remarks of a quite preliminary nature which situate our work within a wider perspective.

2 LFG Analysis of Coordination

A standard LFG analysis of coordination (Dalrymple and Kaplan, 2000; Dalrymple, 2001) assumes the coordination schema in (1)[2] mapping onto a hybrid f-structure containing both a set of f-structures, corresponding to the conjuncts, and a number of non-distributive features, for example the CONJ feature, representing the conjunction, and the resolved agreement features. The f-structure corresponding to the subject NP in the Spanish example (2) from Dalrymple and Kaplan (2000) is therefore as in (3):

(1) $\text{XP} \longrightarrow \underset{\downarrow \in \uparrow}{\text{XP}} \quad \text{CONJ} \quad \underset{\downarrow \in \uparrow}{\text{XP}}$

[1]We gratefully acknowledge the financial support of the British Academy in funding this work as part of the project *Coordination Strategies in Australian Aboriginal Languages* (SG-39545).

[2]Where X ranges over categories such as NP, VP, N, V etc.

(2) *Jose y yo hablamos.*
Jose and I speak.PRES.1PL
'Jose and I are speaking.'

(3)
$$\begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{PERS} & 1 \\ \text{NUM} & \text{PL} \end{bmatrix} \\ \text{CONJ} & \text{'AND'} \\ \left\{ \begin{bmatrix} \text{PRED} & \text{'JOSE'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 3 \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 1 \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix}$$

The distinction between distributive and non-distributive properties is introduced in Dalrymple and Kaplan (2000): if a *distributive* property (e.g. case marking) holds of a set it must hold of every member of the set (i.e. each member of the set must have the same value for these features); *nondistributive* properties (e.g. the CONJ feature), on the other hand, hold of the set itself (and therefore appear in the outer layer of the hybrid structure). These are defined in (4).

(4) For any *distributive* property P and set s, $P(s)$ iff $\forall f \in s.P(f)$.
For any *nondistributive* property P and set s, $P(s)$ iff P holds of s itself.
(Dalrymple and Kaplan, 2000, 779)

As shown in (3) one of the characteristic properties of coordination is the presence of a non-distributive index for the whole set, calculated from the conjuncts via principles of feature resolution (here, 1SG and 3SG are resolved to 1PL). Dalrymple and Kaplan (2000) propose a mechanism for syntactic feature resolution of the non-distributive (INDEX) PERS and GEND features involving closed sets as feature values and 'combining' values by set union. For example, if first person is represented as {S,H}, second person as {H} and third person as the empty set, {}, then the following holds:

(5) {S,H} (1ST) ∪ {H} (2ND) = {S,H} (1ST)
{S,H} (1ST) ∪ {} (3RD) = {S,H} (1ST)
{H} (2ND) ∪ {} (3RD) = {H} (2ND)
{} (3RD) ∪ {} (3RD) = {} (3RD)

Similarly, a two gender (M, F) system with resolution to the masculine works as follows (with MASC corresponding to the set {M} and FEM to the empty set):

(6) {M} (MASC) ∪ {M} (MASC) = {M} (MASC)
{M} (MASC) ∪ {} (FEM) = {M} (MASC)
{} (FEM) ∪ {} (FEM) = {} (FEM)

As shown in (7), the coordination schema for NP coordination in a language with syntactic feature resolution involves simple f-descriptions which ensure that the PERS and GEND features of each NP conjunct be a subset of the PERS and GEND features of the set.

(7) NP $\longrightarrow$

NP	CONJ	NP
$\downarrow \in \uparrow$		$\downarrow \in \uparrow$
$(\downarrow \text{PERS}) \subseteq (\uparrow \text{PERS})$		$(\downarrow \text{PERS}) \subseteq (\uparrow \text{PERS})$
$(\downarrow \text{GEND}) \subseteq (\uparrow \text{GEND})$		$(\downarrow \text{GEND}) \subseteq (\uparrow \text{GEND})$

Resolution of the NUM feature, on the other hand, is not purely syntactic, as shown by the following mimimally contrasting examples:

(8) The president and chief executive are attending the meeting in Beirut.
The president and chief executive is attending the meeting in Beirut.

Cases of boolean coordination, as in (8b), thus show NUM resolution to be semantically based — for a language such as English, when the *and* involved in an NP coordination is so-called 'group-forming' *and* (as in (8a) and the Spanish (3)) it can be associated with an equation specifying that (↑ INDEX NUM) = PL (Dalrymple, 2003).

3 Nominal Juxtaposition in Australian Languages

In many Australian languages NP coordination is achieved through simple juxtaposition. In the following Nyangumarta example, the coordinated subject 'the two kangaroos and one goanna' is encoded by the juxtaposition of the NPs 'two kangaroos' and 'one goanna', with no coordinator relating them. The fact that these are to be interpreted as a single coordinated NP, however, is made clear by the verbal morphology, which agrees with the resolved features of third person plural.

(9) *Pala-nga ngatu jarri-nya-pinti-ngi, mima-nikinyi-yi puluku, kujarra*
that-LOC stationary INCH-NM-ASS-LOC wait.for-IMPF-3PL.SUB 3DU.DAT two
kangkuru-jirri waraja yalapara.
kangaroo-DU one goanna.
'And there, on the finishing line, the two kangaroos and one goanna waited for those two.' (Sharp, 2004, 315, (9.61):Nyangumarta)

It seems reasonable to assume that such coordinate structures receive precisely the same syntactic treatment as (2) above, so that they differ only in the presence/absence of a coordinator. If this is correct, then the f-structure corresponding to the coordinated subject in (9) is that in (10), with resolved INDEX features but no CONJ feature in the outer f-structure.

(10)
$$\left[\begin{array}{l} \text{INDEX} \left[\begin{array}{ll}\text{PERS} & 3\\ \text{NUM} & \text{PL}\end{array}\right] \\ \left\{\begin{array}{l}\left[\begin{array}{ll}\text{PRED} & \text{`GOANNA'}\\ \text{INDEX} & \left[\begin{array}{ll}\text{NUM} & \text{SG}\\ \text{PERS} & 3\end{array}\right]\end{array}\right] \\ \left[\begin{array}{ll}\text{PRED} & \text{`KANGAROO'}\\ \text{INDEX} & \left[\begin{array}{ll}\text{NUM} & \text{DUAL}\\ \text{PERS} & 3\end{array}\right]\end{array}\right]\end{array}\right\}\end{array}\right]$$

While this seems to give a straightforward treatment of asyndetic NP coordination, Australian languages with such NP coordination structures frequently use juxtaposition in a range of other 'appositional-like' constructions as well, such as appositive modifier constructions, generic-specific constructions, part-whole constructions, among others.[3] The following exemplify a range of different interpretations associated with nominal juxtapositions: with the exception of the coordination in (11) all of these are frequently analysed as appositional constructions in Australian language descriptions (e.g. Blake 1979, 1983, 2001, Evans 1995, Heath 1978, 1984, etc.).[4] These constructions all have in common the fact that they involve the juxtaposition of NP elements in the same grammatical function, as evidenced by the fact that the nominals involved are all inflected for the same case feature. In (11) we see a straightforward nominal coordination; in (12) we see a generic-specific construction, with the generic noun *wanku-ya* juxtaposed to the specific noun *kulkiji-y*; (13) exemplifies a part-whole construction with the whole nominal ('bundle') juxtaposed to the part nominal ('fighting stick'); and (14) and (15) illustrate two variants of straightforward appositional constructions – a nominal-nominal appositional construction in (14) in which 'old man' is apposed to 'husband' in subject function,[5] and a nominal-pronominal appositional construction in (15) in which the coordinated NP 'those men and women' is apposed to the coreferential third person plural pronoun *bi-l-da*.

(11) *Niya* *kurrka-tha* *barruntha-ya* *wuran-ki* *nguku-y.*
3SG.NOM take-ACT yesterday-LOC food-MLOC water-MLOC
'Yesterday he took (with him) food and water.' (Evans, 1995, 250:Kayardild)

[3]Note that it is often hard to determine from the data available whether these constructions involve juxtaposed NPs or juxtaposed Ns: the presence of a demonstrative will sometimes make this clear, but in general, 'bare' Ns can have referential NP meanings and constitute a full NP on their own. The syntactic analysis we present applies equally well to either structural possibility and we simply use variables in our phrase structure rules to range over both category options (see §4).

[4]Obviously languages will differ in terms of the range of constructions that they encode with NP juxtaposition. Those exemplified here, however, are fairly typical.

[5]Note that the two apposed nominals come before the auxiliary *gin-amany* here, showing them to jointly belong to an NP constituent since the Wambaya auxiliary must always be the second constituent in the clause (Nordlinger 1998).

(12) *Dathin-a dangka-a niya wumburung-kuru raa-ja wanku-ya*
that-NOM man-NOM 3SG.NOM spear-PROP spear-ACT elasmobranch-MLOC
kulkiji-y.
shark-MLOC
'That man speared a shark with a spear.' (ibid, 244: Kayardild)

(13) *kawuka jardiyali*
bundle fighting.stick
'a bundle of fighting sticks' (ibid, 249: Kayardild)

(14) *Garidi-ni bungmanyi-ni gin-amany yanybi.*
husband.I-ERG old.man.I-ERG 3SG.M.A-P.TWD get
'(Her) old man husband came and got (her).' (Nordlinger, 1998, 133: Wambaya)

(15) *Dathin-a maku-wa bithiin-da bi-l-da warra-j.*
that-NOM woman-NOM man-NOM 3-PL-NOM go-ACT
'Those men and women are going.' (Evans, 1995, 249: Kayardild)

A further type of juxtaposed construction common to Australian languages is the inclusory construction (Singer, 2001, 2005) (also known in the literature as the 'plural pronoun construction' (Schwartz, 1988)), in which a plural pronoun referring to the superset is combined with a subset nominal. In many languages the inclusory construction involves simple juxtaposition of the two elements, as in the following from Kayardild:

(16) *Nga-rr-a kajakaja warra-ja thaa-th.*
1-DU-NOM daddy.NOM go-ACT return-ACT
'Daddy and I will go' (lit. 'We two, including daddy, will go') (Evans 1995:249)

Appositional structures, in which we loosely group the non-coordinated examples above, have received very little attention in the LFG literature and as a consequence the analysis of these constructions, and their potential structural relationship to NP coordination, raises a number of interesting issues. In particular, (i) how are the various juxtaposed constructions related syntactically in these languages?; (ii) how is coordination to be defined in these languages as distinct from other juxtaposed constructions?; (iii) how are all of these juxtaposed constructions to be analysed? It is to these questions that we turn in the remainder of this paper.

4 Analysis of syntactic juxtapositions

In very many cases there appear to be no clear *syntactic* grounds for distinguishing between coordinations (on the one hand) and other (mainly appositional) uses to which syntactic juxtapositions can be put. Case marking patterns and phrase structure constraints (where

these exist) are generally consistent across all such juxtaposed constructions, and indeed all are consistent with the general definitions of coordination in the literature, such as the following:

> An element in construction with a coordinate constituent must be syntactically construable with each conjunct[6] (Wasow)
>
> The term *coordination* refers to syntactic constructions in which two or more units of the same type are combined into a larger unit and still have the same semantic relations with other surrounding elements (Haspelmath, 2004, 34)
>
> A coordination is a construction consisting of two or more members which are equivalent as to grammatical function, and bound together at the same level of structural hierarchy by means of a linking device[7] (Dik, 1968, 25)

The fact that these nominal coordinations and appositions show no *syntactic* distinctions suggests an analysis that treats them as essentially a single type of syntactic construction that can be associated with a range of different semantics. More specifically, we propose an analysis in LFG in which juxtaposed constructions such as those exemplified in (11-16) above are treated as f-structure coordinations as in (10), that is, as involving hybrid f-structures as the value of a single grammatical function.[8] The various constructions may differ at f-structure, as we shall see, in terms of the agreement features of the set (i.e. whether they involve feature resolution or not), and then are further differentiated in the mapping to the semantic structure. Thus, we propose that all of these constructions are licensed by the basic phrase structure schema in (17), with different annotations depending on issues of feature resolution and semantics, as discussed in detail in the following sections. In this schema we use X as a metavariable ranging over the categories N, N' and NP. In other words, the basic schema allows juxtapositions of any of these categories, just as long as the juxtaposed elements are of the same categorial type (NP with NP, N with N, etc.).

(17) $\mathrm{X} \longrightarrow \underset{\downarrow \in \uparrow}{\mathrm{X}} \quad \underset{\downarrow \in \uparrow}{\mathrm{X}}$

4.1 Coordination vs. Apposition

On this view, the f-structure corresponding to the apposition in (14) is as in (18). Apart from the value of the non-distributive (INDEX) features of the set, this f-structure is structurally identical to that associated with the coordination in (9) ((10) repeated as (19)).

[6] Note that this assumes that a coordination is structurally a single constituent, and thus does not allow for discontinuous coordination (cf. §5).

[7] Note that this definition is in fact strictly inapplicable even to our regular coordination examples in that it requires the presence of an overt coordinator.

[8] Of course, an f-structure coordination analysis may not be appropriate for other types of nominal juxtapositions (e.g. possessive and other clearly modificational structures), nor for all types of 'appositional' constructions cross-linguistically. We are focussing here on the constructions exemplified above, particularly on appositive modifier constructions (called appositions in the Australianist literature).

(18) **Apposition:**

$$\begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \left\{ \begin{bmatrix} \text{PRED} & \text{`HUSBAND'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 3 \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{`OLD.MAN'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 3 \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix}$$

(19) **Coordination:**

$$\begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{PERS} & 3 \\ \text{NUM} & \text{PL} \end{bmatrix} \\ \left(\begin{bmatrix} \text{PRED} & \text{`GOANNA'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 3 \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{`KANGAROO'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{DUAL} \\ \text{PERS} & 3 \end{bmatrix} \end{bmatrix} \right) \end{bmatrix}$$

This analysis directly reflects the fact that there is no visible syntactic distinction within the nominal strings themselves between nominal coordination and nominal apposition. In fact, the nominal phrase in (14) is itself ambiguous between a coordinative and an appositional interpretation, disambiguated only by the verbal morphology. In Wambaya (14) the auxiliary form *gin-amany* '3SG.M.A-P.TWD' determines that the SUBJ is 3SG. If this example meant 'the old man and her husband (they)....' then the finite auxiliary would be encoded with 3DU. Crucially, the formal differences lie only in the agreement features of the set; there is no visible syntactic distinction within the nominal structure itself. Thus, as far as the syntax is concerned, our analysis needs to be able to account for the fact that the same nominal f-structure may sometimes involve feature resolution (i.e. in a coordination structure), and sometimes not (i.e. in an appositional structure).

4.2 Coordinate Meanings

As we have seen, nominal juxtapositions can have coordinate meanings, involving syntactic feature resolution and the construction of a coordinate semantics. For present purposes, we follow Dalrymple and Kaplan (2000) in our analysis of feature resolution but clearly the details of syntactic GEND resolution will differ considerably in a language like Wambaya which distinguishes four genders (e.g. MA, FEM, NEUT, VEG), and exhibit defaults and underspecification in gender agreement. The template for feature resolution in coordinate structures given in (20) simply introduces the annotations proposed by Dalrymple and Ka-

plan (2000) and discussed in §2 above.[9] This template is associated with each constituent in the phrase structure rule, as in (21).

(20) NP-CNJT: $(\downarrow \text{IND PERS}) \subseteq (\uparrow \text{IND PERS})$
$(\downarrow \text{IND GEND}) \subseteq (\uparrow \text{IND GEND})$

(21) $\text{X} \longrightarrow \underset{\substack{\downarrow \in \uparrow \\ @\text{NP-CNJT}}}{\text{X}} \quad \underset{\substack{\downarrow \in \uparrow \\ @\text{NP-CNJT}}}{\text{X}}$

As for the semantics of NP coordination, Dalrymple (2001) associates the semantic contribution **g-and** (group-forming *and*) in (22) with the coordinator. The semantics of **g-and** forms a plural individual from two individuals (in the glue, it consumes a meaning of type e and produces a resource which will consume a meaning of type e to produce a meaning of type e). For more than binary coordination, a further semantic contribution **g-and2**, involving the ! (of course) operator, can be used any number of times (including zero), each time adding an individual into the group.

(22) **g-and** λ X. λ Y. { X,Y } :
$(\uparrow \in)_{\sigma<e>} \multimap [(\uparrow \in)_{\sigma<e>} \multimap \uparrow_{\sigma<e>}]$

(23) **g-and2** λ X. λ Y. { X } $\cup$ Y :
$!(\uparrow \in)_{\sigma<e>} \multimap [\uparrow_{\sigma<e>} \multimap \uparrow_{\sigma<e>}]$

(24) *and* $(\uparrow \text{CONJ}) = \text{AND}$
[g-and]
[g-and2]

The situation in our case is a little more complicated, however as there is no coordinator in the structure to associate the semantics of **g-and** with.

Notice also that in languages (such as these) with three number distinctions (singular, dual and plural), it is not possible simply to associate the use of the group-forming semantics with NUM resolution to PL, because the syntactic NUM of a group containing just a pair is DU. For present purposes, which are largely illustrative, we restrict ourselves to binary coordination, and define the NUM resolution as in (25). This captures the generalisation that *either* the overall number is DU (i.e. when two singular nominals are coordinated) *or* (at least) one of the constituents is non-singular, in which case the overall number is PL.

(25) BINARY: $\{(\uparrow \in \text{INDEX NUM}) \neq \text{SG} \wedge (\uparrow \text{INDEX NUM}) = \text{PL}\}$
$\mid (\uparrow \text{INDEX NUM}) = \text{DUAL}$

[9] Templates are a simple and convenient means of naming a collection of f-descriptions. Because templates can call other templates, they can be organised to express linguistic generalisations succinctly. See Dalrymple et al. (2004).

To complete the interpretation of nominal juxtapositions as coordinative, we need to associate the template BINARY and **g-and** with the phrase structure rule in (21) (restricting attention to cases of binary coordination). Since there is no coordinator to associate them with, we arbitrarily associate them with one of the daughter constituents.

(26) X $\longrightarrow$ X (↓ ∈ ↑) @NP-CNJT @BINARY **g-and** ; X (↓ ∈ ↑) @NP-CNJT

Our analysis of the juxtapositions with coordinate semantics is thus analogous to the analysis of (non-juxtaposed) coordinate constructions in other languages (Dalrymple and Kaplan 2001, Dalrymple 2001). In the next section we see how this same general approach can also provide an analysis of appositional juxtapositions.

4.3 Appositional Meanings

In appositional juxtapositions the juxtaposed constituents are co-referential and there is no feature resolution at the level of the set: the features of the set are the same as the features of each of the members. Thus, in our terms, appositional constructions generally involve INDEX sharing between the set and the members of the set, as well as the construction of an appositional semantics.

In order to capture the sharing of INDEX features between the set members and the set itself, we define the appositional template in (27), which is associated with each of the daughter constituents in the appositional phrase structure rule, as in (28). This template ensures that the INDEX features of each daughter constituent are shared with the INDEX features of the set (i.e. a set containing two 3SG daughters will likewise have 3SG INDEX features).

(27) NP-APPOS*:* (↓ IND) = (↑ IND)

(28) X $\longrightarrow$ X (↓ ∈ ↑) @NP-APPOS ; X (↓ ∈ ↑) @NP-APPOS

In the interests of clarity, we assume here that all INDEX features in appositional constructions will be shared between the members and the set. This is potentially an oversimplification, since it may well be the case that there will be instances of appositions in which the f-structures may differ in one or more INDEX features despite being descriptions of the same real world entity. A circumstance where this might arise could be where apparent person mismatches are allowed in appositonal structures (e.g. in the English 'us linguists', 'you children'). A further tricky area concerns gender, where a complicating factor in the interpretation of appositional data is the fact that nouns have both INDEX GEND

and CONC GEND features, and these may not match. Well-known cases of 'mismatch' nouns include the Serbo-Croatian collective nouns of the second declension, such as *deca* 'children', which are analysed as FEM.SG CONCORD but NE.PL INDEX by Wechsler and Zlatić (2003). The potential for non-matching between CONCORD and INDEX in GEND complicates the interpretation of putative mismatches in appositional structures in the languages we are concerned with, because of course it may be the case that such examples involve nouns differing in CONCORD GEND but not in INDEX GEND. Other cases of gender mismatch in appositional constructions could possibly come from generic-specific constructions in which hyponyms and hypernyms clearly belong to different gender classes (e.g. VEgetable and NEuter), but we leave investigation of whether this occurs to further research. Should plausible examples emerge, these constructions could be captured by modifying the above analysis in a number of ways. One possibility would be to have only one daughter in the appositional phrase structure rule contribute INDEX features to the set (i.e. be associated with the NP-APPOS template above), with the INDEX features of the other daughter only partially shared, or not shared at all.

Turning now to the semantics of appositional constructions, as a first approximation we take the semantics of appositional juxtapositions to be basically intersective (applying to property-denoting nominal (rather than NP) meanings). One possibility is something comparable to boolean *and* (as in the joint reading of *five linguists and philosophers*), taking two sets of properties and intersecting them (see Dalrymple (2004)):

(29) **b-and** λ X. λ Y. X $\sqcap$ Y

An alternative, which is the one we will follow here, is to model the semantics of apposition on the semantics of nominal modification, as follows:

(30) **appos** λ Q. λ P. λ X. Q(X) $\wedge$ P(X):
[((↑ ∈)$_\sigma$ VAR) $\multimap$ ((↑ ∈) $_\sigma$ RESTR)] $\multimap$
[[((↑ ∈)$_\sigma$ VAR) $\multimap$ ((↑ ∈) $_\sigma$ RESTR)]
$\multimap$ [(↑ $_\sigma$ VAR) $\multimap$ (↑ $_\sigma$ RESTR)]]

On the meaning side, this is a function which applies to two nominal ($< e, t >$) meanings and produces an abstraction over a logical conjunction of predications holding of this individual (so it takes two nominal meanings and produces a nominal meaning, where nominal meanings are of type $< e, t >$). On the glue side the meaning constructor consumes one nominal contribution and then the other nominal contribution to produce the meaning of the NP as a whole.

We can therefore complete our analysis of appositional juxtapositions by arbitrarily associating the **appos** semantics with some daughter in the appositional phrase structure rule:

(31) X $\longrightarrow$ X X
↓ ∈ ↑ ↓ ∈ ↑
@NP-APPOS @NP-APPOS
appos

In order to see how this works, consider the nominal apposition in the Wambaya example (14). The semantics associated with each of the nominals in this construction is given in (32) and (33).

(32) *garidi-ni (*husband.I-ERG*)* λ X. husband(X): ($\uparrow_\sigma$ VAR) $\multimap$ ($\uparrow_\sigma$ RESTR)

(33) *bungmanyi-ni (*old.man.I-ERG*)* λ X. old.man(X): ($\uparrow_\sigma$ VAR) $\multimap$ ($\uparrow_\sigma$ RESTR)

(30) consumes (32) and (33), which results in the following nominal meaning:

(34) *garidi-ni bungmanyi-ni* λ X. old.man(X) $\wedge$ husband(X):
($\uparrow_\sigma$ VAR) $\multimap$ ($\uparrow_\sigma$ RESTR)

Note that in these languages, a bare nominal such as (32) or (33) (or indeed (34)) may be interpereted predicatively, but may also be given a range of NP meanings in context (e.g. 'the boy', 'a boy', 'boys') - pronouns and demonstratives may occur in "determinizing" function but are by no means obligatory in the production of full (referential) NP meanings. In these cases, where there are no demonstratives or pronouns, we take it that additional meaning constructors (not associated with lexical material) must be available to lift nominals into the appropriate range of NP meanings.[10]

To summarise, we can account for the use of syntactic juxtaposition to encode both coordinate and appositional constructions by making two alternative sets of annotations available for the "coordinate" NP rule, (26) and (31), as follows:

- Annotate each dtr @NP-CNJT and some dtr @BINARY and **g-and** ; **OR**
- Annotate each dtr @NP-APPOS and some dtr **appos**

4.4 Other juxtaposed constructions

This analysis also provides a straightforward account of the other juxtaposed constructions discussed in §2, namely generic-specific, part-whole and inclusory constructions. Generic-specific and part-whole constructions are simply appositional-like structures, licensed by (31). This treatment is consistent with Australianist descriptions that treat such constructions as consisting of apposed nominals (e.g. Blake 1983, Evans 1995, Heath 1978, etc.).

The f-structure corresponding to the juxtaposed (generic-specific) construction in (35) is given in (36).[11] Standard nominal lexical entries along the lines of (32) for *wanku-ya* (elasmobranch-MLOC) and *kulkiji-y* (shark-MLOC) combine with the appositional meaning constructor to give (37):

[10]Our account of the semantics of apposition *per se*, on the other hand, must be extended to deal with examples in which it is clear that full NPs (e.g. of type e) occur in apposition.

[11]Each member of the set in (36) is marked for modal case (which contributes information to the sentential f-structure).

(35) *Dathin-a dangka-a niya wumburung-kuru raa-ja wanku-ya*
that-NOM man-NOM 3SG.NOM spear-PROP spear-ACT elasmobranch-MLOC
kulkiji-y.
shark-MLOC
'That man speared a shark with a spear.' (Evans, 1995, 244: Kayardild)

(36)
$$\left[\begin{array}{l} \text{INDEX} \quad \left[\begin{array}{ll}\text{PERS} & 3\\ \text{NUM} & \text{SG}\end{array}\right] \\ \left\{\begin{array}{l}\left[\begin{array}{ll}\text{PRED} & \text{'ELASMOBRANCH'}\\ \text{INDEX} & \left[\begin{array}{ll}\text{NUM} & \text{SG}\\ \text{PERS} & 3\end{array}\right]\end{array}\right] \\ \left[\begin{array}{ll}\text{PRED} & \text{'SHARK'}\\ \text{INDEX} & \left[\begin{array}{ll}\text{NUM} & \text{SG}\\ \text{PERS} & 3\end{array}\right]\end{array}\right]\end{array}\right\}\end{array}\right]$$

(37) *wanku-ya kulkiji-y* (elasmobranch-MLOC shark-MLOC)
λ X. elasmobranch-fish(X) $\wedge$ shark(X): ($\uparrow_\sigma$ VAR) $\multimap$ ($\uparrow_\sigma$ RESTR)

Inclusory constructions are a particularly interesting case as the features of the set overall are identical to the features of one member of the set, but not the other. Consider the following example:

(38) *Nga-rr-a kajakaja warra-ja thaa-th.*
1-DU-NOM daddyNOM go-ACT return-ACT
'Daddy and I will go' (lit. 'We two, including daddy, will go') (Evans, 1995, 249: Kayardild)

In these constructions a pronominal referring to the superset (here 'we two') is juxtaposed with a nominal representing just one member of the set (here 'daddy') (see Singer 2001 for discussion). The INDEX features of the whole are those corresponding to the INDEX features of the pronominal, in which the features of the single member must be included. Thus, inclusory constructions are a composite of the coordination and appositional schemas presented in (26) and (31) above. The constituent corresponding to the superset pronominal carries the appositional template (specifying that its INDEX features are identical to the INDEX features of the whole) and the constituent corresponding to the single member carries the coordination template (specifying that its INDEX features must be a subset of the INDEX features of the whole).

(39) NP $\longrightarrow$ NP , NP
$\downarrow \in \uparrow$ $\downarrow \in \uparrow$
@NP-APPOS @NP-CNJT

(40) **inclusory**:

$$\begin{bmatrix} \text{INDEX} & \begin{bmatrix} \text{PERS} & 1 \\ \text{NUM} & \text{DUAL} \end{bmatrix} \\ \left\{ \begin{bmatrix} \text{PRED} & \text{'DADDY'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{PERS} & 3 \end{bmatrix} \end{bmatrix} \right. \\ \left. \begin{bmatrix} \text{PRED} & \text{'PRO'} \\ \text{INDEX} & \begin{bmatrix} \text{NUM} & \text{DUAL} \\ \text{PERS} & 1 \end{bmatrix} \end{bmatrix} \right\} \end{bmatrix}$$

The semantics of the inclusory is that one member denotes a group and the other member of the set contributes a further restriction over the group by providing a specification about a member of the group.

4.5 Summary of analysis

The range of juxtaposed NP constructions in Australian languages can be accounted for relatively simply with an account in which nominal-nominal sequences have the same essential f-structure, but correspond to 3 different feature resolution patterns, as in (39)-(41), and map onto a range of different semantics correlated with these three different patterns.

(41) **coordination – X $\supseteq$ Y, Z**

$$\begin{bmatrix} \text{INDEX} & [\text{X}] \\ \left\{ \begin{bmatrix} \text{INDEX} & [\text{Y}] \end{bmatrix} \right. \\ \left. \begin{bmatrix} \text{INDEX} & [\text{Z}] \end{bmatrix} \right\} \end{bmatrix}$$

(42) **apposition – X = Y, Z**

$$\begin{bmatrix} \text{INDEX} & [\text{X}] \\ \left\{ \begin{bmatrix} \text{INDEX} & [\text{Y}] \end{bmatrix} \right. \\ \left. \begin{bmatrix} \text{INDEX} & [\text{Z}] \end{bmatrix} \right\} \end{bmatrix}$$

(43) **inclusory – X = Y $\supseteq$ Z**

$$\begin{bmatrix} \text{INDEX} & [\text{X}] \\ \left\{ \begin{bmatrix} \text{INDEX} & [\text{Y}] \end{bmatrix} \right. \\ \left. \begin{bmatrix} \text{INDEX} & [\text{Z}] \end{bmatrix} \right\} \end{bmatrix}$$

This approach exploits the flexible architecture of LFG to account for the fact that these constructions are structurally similar – all consisting of juxtaposed nominals in the c-structure, and hybrid structures in the f-structure – yet semantically distinct. This seems to capture the intuition that appositions are closely related to coordinations, while still permitting us to capture the (mainly semantic) difference between coordination and apposition.

5 Discontinuity

Of course, these being Australian languages, all of the structures can also be discontinuous. The following examples illustrate discontinuous coordination constructions (44), generic-specific constructions (45), and inclusory constructions (46):

(44) *Ngul ngay kirk kempthe kal-m thul=yuk*
then 1SG(ERG) spear(ACC) apart carry-P.IPFV woomera(ACC)
'I used to carry spears and woomeras separately' (Kuuk Thaayorre, Gaby 2006)

(45) *Ngayika ati-ntji ari-li thuwarr-ku.*
I meat-DAT eat-APASS snake-DAT
'I'm eating snake.' (Blake, 2001, 419, ex 8: Kalkatungu)

(46) *Wey, ngali yancm ngan waanharr iipal*
hey 1DU:EXCLNOM go:P.IPFV relative e.brother from.there
'Hey, my brother and I have come here' (Kuuk Thaayorre, Foote 1977, cited in Gaby 2006)

While we do not have the space for detailed discussion here, the occurrence of discontinuous coordinations and appositions is not problematic. Our analysis will extend straightforwardly to these cases on the (rather standard) assumption that each daughter constituent in the phrase structure rule is optional, thereby allowing for each one to occur alone in an NP in the c-structure, and be unified into a hybrid structure at f-structure.[12]

6 Conclusion and broader implications

The flexible architecture of LFG provides a unified syntactic account of a range of juxtaposed nominal constructions common to Australian languages, while still capturing their semantic differences. In this paper we have shown how the use of hybrid f-structures can be extended beyond true (semantically) coordinated constructions to generic-specific, part-whole and other types of appositional constructions also, making a distinction between syntactic coordination (hybrid structures) and semantic coordination (corresponding to feature resolution and coordinate semantics). This approach has a number of broader implications:

[12]This will, of course, raise some technical issues such as ensuring that the relevant GF is a set, however such issues seem resolvable, and so we do not consider them to be an impediment to the analysis in principle.

(i) **syntactic vs. semantic coordination:**
This distinction between syntactic and semantic coordination allows for constructions that are both syntactically and semantically coordinated (i.e. true coordinations), semantically coordinated without being a coordinated structure in the syntax (i.e. Nyangumarta compounds (47)); and syntactically coordinated without being instances of semantic coordination (i.e. the appositional-like structures discussed above).[13]

(47) *Pipi-japartu-lu partany kalku-rnikinyi pulu.*
mother-father-ERG child keep-IMPF 3DU.SUB
The mother and the father (the parents) looked after the child (Nyangumarta, Sharp 2004: 312 (9.50))

(ii) **boolean coordination:**
So called boolean coordination, such as the English *my friend and colleague*, *the president and commander-in-chief*, is no longer an outlier construction, but can now be seen in the context of a wider set of data. On our view, boolean coordination can be considered to be essentially similar to the appositional juxtapositions we discuss, the only difference being the presence of an overt coordinator. It is thus syntactically coordinated (having a hybrid f-structure), but semantically appositional (having no feature resolution and appositional semantics).

(iii) **application beyond Australian languages;**
One of the implications of our analysis of Australian nominal-nominal constructions is that appositions are syntactically the same as coordinations – the only difference being that there is resolution of features in the f-structure with the latter but not the former. Similar suggestions have been made in the literature, outside of the Australianist and LFG contexts (Quirk et al., 1985; Koster, 2000; de Vries, 2006; Van Eynde, 2005).

While it remains for further research to determine the extent to which our analysis can be applied to languages outside of the Australian context, it provides a way to capture this association between apposition and coordination in LFG terms.

References

Blake, Barry. 1979. *A Kalkatungu Grammar*. Canberra: Pacific Linguistics.

Blake, Barry J. 1983. Structure and word order in Kalkatungu: The Anatomy of a Flat Language. *Australian Journal of Linguistics* 3(2):143–175.

Blake, Barry J. 2001. Forty years on: Ken Hale and Australian languages. In Jane Simpson et. al., ed., *The Noun Phrase in Australian Languages*, pages 415–425. Canberra: Pacific Linguistics.

[13] cf. Culicover and Jackendoff (1997, 2005) and Yuasa and Sadock (2002) who also discuss mismatches between 'syntactic' and 'semantic' coordination, although, in LFG terms their data is relevant to mismatches between c-structure and f-structure, rather than between f-structure and semantic structure, which is our concern here.

Culicover, Peter and Ray Jackendoff. 1997. Semantic subordination despite syntactic coordination. *Linguistic Inquiry* 28:195–218.

Culicover, Peter and Ray Jackendoff. 2005. *Simpler Syntax*. Oxford: Blackwells.

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

Dalrymple, Mary. 2003. Noun Coordination: Syntax and Semantics. Talk given at the University of Essex.

Dalrymple, Mary. 2004. Noun coordination: Syntax and semantics. talk given at the University of Canterbury, New Zealand.

Dalrymple, Mary and Ron Kaplan. 2000. Feature indeterminacy and feature resolution in description-based syntax. *Language* 76(4):759–798.

Dalrymple, Mary, Ron Kaplan, and Tracy Holloway King. 2004. Linguistic generalizations over descriptions. In M. Butt and T. H. King, eds., *Proceedings of the LFG04 Conference*. Stanford, CA: CSLI Publications: http://www-csli.stanford.edu/publications.

de Vries, Mark. 2006. The syntax of appositive relativization: on specifying coordination, free relatives and promotion. *Linguistic Inquiry* 37:229–270.

Dik, Simon. 1968. *Coordination: its Implications for the Theory of General Linguistics*. Amsterdam: North Holland Publishing.

Evans, Nicholas. 1995. *A Grammar of Kayardild: With Historical-Comparative Notes on Tangkic*. Berlin: Mouton de Gruyter.

Gaby, Alice. 2006. *A Grammar of Kuuk Thaayorre*. Ph.D. thesis, University of Melbourne, Melbourne, Australia.

Haspelmath, Martin. 2004. Coordinating constructions: An overview. In M. Haspelmath, ed., *Coordinating Constructions*, pages 3–40. Amsterdam: John Benjamins.

Heath, Jeffrey. 1978. *Ngandi grammar, texts and dictionary*. Canberra: Pacific Linguistics.

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

Koster, Jan. 2000. Extraposition as parallel construal. Unpublished paper, University of Groningen.

Nordlinger, Rachel. 1998. *A Grammar of Wambaya, Northern Territory (Australia)*. Canberra: Pacific Linguistics.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Schwartz, Linda. 1988. Asymmetric Feature Distribution in Pronominal 'Coordinations'. In M. Barlow and C. A. Ferguson, eds., *Agreement in Natural Language*, pages 237–50. Stanford: CSLI Publications.

Sharp, Janet. 2004. *Nyangumarta: A Language of the Pilbara Region of Western Australia*. Canberra: Pacific Linguistics.

Singer, Ruth. 2001. The Inclusory Construction in Australian Languages. Unpublished Honours thesis, University of Melbourne.

Singer, Ruth. 2005. Comparing constructions across languages: a case study of the relationship between the inclusory construction and some related nominal constructions. Unpublished talk, ALT 2005.

Van Eynde, Frank. 2005. A head-driven treatment of asymmetric coordination and apposition. In S. Müller, ed., *Proceedings of HPSG 2005*, pages 396–409. CSLI Publications: http://www-csli.stanford.edu/publications, Stanford, CA.

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

Yuasa, Etsuyo and Jerry Sadock. 2002. Pseudo-subordination: a mismatch between syntax and semantics. *Journal of Linguistics* 38:87–111.

USING SUBSUMPTION RATHER THAN EQUALITY IN FUNCTIONAL CONTROL

Peter Sells

Stanford University

Proceedings of the LFG06 Conference

Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006

CSLI Publications

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

Abstract

In this paper I consider the typology of forward and backward control and raising structures, and argue that structure-sharing based on the relation of subsumption rather than equality provides a stronger linguistic basis for the typology (cf. Zaenen and Kaplan (2002, 2003)). I also consider the points of contact and divergence between structure-sharing and the 'copy theory of movement' that is prevalent in current minimalist syntax.

1. Functional Control in LFG

Non-derivational syntactic theories such as LFG or HPSG traditionally analyze raising and control in terms of structure-sharing between (in the first instance) the subject of a predicate and the subject of a complement of that predicate, stated as properties of a lexical form. The same grammatical information simultaneously acts as the subject of the raising or control verb, and the subject within the infinitival complement.*

On the basis of this structure-sharing, such approaches would seem like prime candidates for extensions to insightful accounts of backward control and raising, phenomena that Polinsky and Potsdam (2002a, 2002b, 2006) (P&P) have brought squarely into theoretical discussions of control and raising in general. The structures corresponding to 'forward' and 'backward' are shown in (1), where Δ_i marks an empty subject position.

(1) Forward and Backward Control and Raising:

a. Kim$_i$ seems/hopes [Δ_i to be singing]. (forward)

b. Δ_i seems/hopes [Kim$_i$ to be singing]. (backward)

In the LFG analysis, e.g., Bresnan (1982), the fact that control and raising are 'forward' in English is because the structure-sharing lexical forms select for a VP complement, which has no place for a 'downstairs' subject position. If English had a control predicate which selected for an S as its complement, English could allow backward control. In other words, while the fundamental control or raising properties of predicates might be essentially universal, the syntactic manifestation of the shared argument is a more parochial fact about the phrase structure category of the complement. Whether the phenomena are forward or backward is only determined by constraints on phrase structure configuration. In this paper I will show that conditions on a construction being forward or backward should be accounted for in terms of lexical entries, involving LFG's f-structure information. Specifically, I argue that it is necessary to express structure-sharing not via equality, but via the relation of subsumption, introduced below.

LFG traditionally analyzes subject raising and control as structure-sharing between a subject and the subject of a complement Bresnan (1982). The two positions are set equal: in unification terms, they share all properties; exactly the same grammatical information flows to each shared position (see (2)).

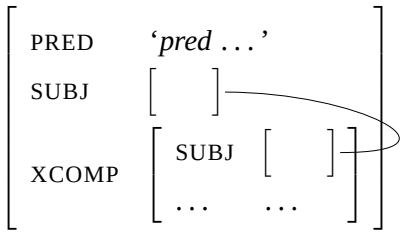
(2) Equality: SUBJ = XCOMP SUBJ (information flows between both positions)

Control and raising predicates in LFG are both subject to (2), and only differ in that the matrix subject is thematic in the case of control and non-thematic in the case of raising. The difference is represented in the lexical forms as shown in (3): thematic arguments appear within $\langle\ \ \rangle$, non-thematic arguments appear outside. The f-structure in (3) is the same for both control and raising predicates.

*I am grateful to Masha Polinsky and Eric Potsdam for help with examples and access to materials, and to the audience at LFG-06 in Konstanz for useful comments, in particular Ash Asudeh and Jonas Kuhn. Eric also provided a useful commentary on the original version of section 4.2. Parts of this material was presented at the Daegu Catholic University, Daegu, Korea, in July 2006, and I am grateful to Stan Dubinsky for comments on that occasion, some of which are incorporated here.

(3) a. Control Predicate – thematic subject:
(↑ PRED) = '*c-pred* ⟨(↑ SUBJ), (↑ XCOMP)⟩'
(↑ SUBJ) = (↑ XCOMP SUBJ)

b. Raising Predicate – non-thematic subject:
(↑ PRED) = '*r-pred* ⟨(↑ XCOMP)⟩(↑ SUBJ)'
(↑ SUBJ) = (↑ XCOMP SUBJ)

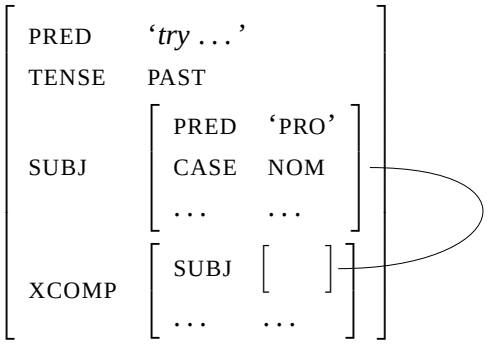


(4) gives the standard analysis of English *try*:

(4) a. She tried to leave.

try takes a VP complement in c-structure which corresponds to an XCOMP in f-structure.

b.

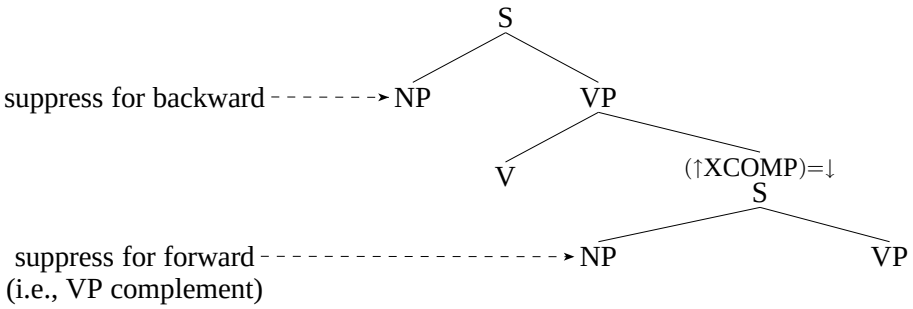


same information in both places

The LFG analysis effectively foreshadows the recent Minimalist-style analyses in which movement leaves a copy (an unpronounced copy – see Chomsky (1995)), in which both control and raising are analyzed via movement (e.g., Hornstein (1999), Polinsky and Potsdam (2002a)). There may be technical differences, depending on whether movement creates literal copies (giving type- but not token-identity, see e.g., Asudeh (2005)), or whether the same item is continuously 're-merged', as in Fox and Pesetsky (2005); see also section 4.2.

As mentioned above, under the standard equality account, the fact that control and raising are 'forward' in a language like English is a fact about constituent structure. Structure-sharing lexical forms select a VP complement, which cannot host a downstairs subject; if English had a control predicate selecting an S complement, it could allow backward control. The standard approach has the property that while the fundamental control/raising properties of predicates might be essentially universal, the manifestation of the shared argument is a more parochial fact about phrase structure(s).

(5)



Whether the phenomena are forward or backward is determined only by c-structure. In English, the XCOMP complement is a VP, not an S, and hence control and raising are forward. If the matrix subject position could be suppressed, the constructions could in principle be backward. However, it is not clear that the theory can restrict control to only backward control. In the best case, that would be a language in which the matrix subject position in (5) could/should be empty but the embedded subject position must be filled, which seems

counterintuitive. And a c-structure solution would seem to face real difficulties in a language where some predicates are 'forward' and some are 'backward'; but such languages exist.

The paper is organized as follows. In section 2, I present the empirical observations that lead to the typology of control and raising constructions. In section 3, I review the proposals of Zaenen and Kaplan (2002, 2003) to incorporate subsumption into the grammar, and extend them to the data at hand. In section 4, I consider alternatives in terms of the copy theory of minimalist syntax, and a different LFG analysis which does not use subsumption.

2. Control and Raising, Forward and Backward

In a series of papers, Polinsky and Potsdam have demonstrated the existence of backward control and raising, in a variety of languages. I briefly survey the basic data here, illustrating with examples from Tsez, Malagasy and Circassian. All of the data in this section is taken from Polinsky and Potsdam's work.

2.1. **Tsez** (Forward Raising, Backward Control)

Polinsky and Potsdam (2002a) argue that the Tsez verbalizer *-oqa* ('begin') is ambiguous between a control and a raising use. In addition, as a control predicate it requires backward control, while as a raising predicate it is forward. Tsez is a verb-final language.

(6) *-oqa* ('begin') is forward if raising or backward if control:

a. kid [ziya b-išr-a] y-oq-si (forward raising)
girl.II.ABS [cow.III.ABS III-feed-INF] II-begin-PAST.EVID
'The girl began to feed the cow.'

b. [kid-bā ziya b-išr-a] y-oq-si (backward control)
[girl.II-ERG cow.III.ABS III-feed-INF] II-begin-PAST.EVID
'The girl began to feed the cow.'

The syntactic analyses are those shown in (7):

(7) a. kid_i [t_i ziya b-išr-a] y-oq-si (forward raising)
girl.II.ABS [cow.III.ABS III-feed-INF] II-begin-PAST.EVID
'The girl began to feed the cow.'

b. Δ_i [kid-bā_i ziya b-išr-a] y-oq-si (backward control)
[girl.II-ERG cow.III.ABS III-feed-INF] II-begin-PAST.EVID
'The girl began to feed the cow.'

The facts in (7)a are relatively straightforward: the raised argument passes the usual tests for being non-thematic with respect to the matrix predicate, and the verb agrees in noun class with it. Note that in the embedded clause the verb agrees in noun class with the absolutive argument, the typical agreement pattern.

The facts in (7)b are more unusual – the matrix verb apparently agrees with the embedded clause ergative subject. This would be the only instance of agreement with an ergative. Polinsky and Potsdam (2002a) argue that Δ_i in (7)b represents the thematic subject position of the control verb, and that the verb agrees with this position, maintaining the generalization that agreement is with an absolutive). So this is backward control – the matrix and embedded subject positions are shared, but the overt argument is in the lower position.

2.2. **Malagasy** (Forward and Backward Control)

Polinsky and Potsdam (2002b) present evidence for forward and backward control in Malagasy, a verb-initial language. Forward control obtains with a verb like *try*, which shows the two syntactic patterns in (8), with the analysis proposed by P&P indicated by the bracketing and with Δ_i indicating the empty position.

(8) a. m-an-andrana [m-i-tondra ny fiara Δ_i] Rabe$_i$ (forward control)
PRES-ACT-try [PRES-ACT-drive the car] Rabe
'Rabe is trying to drive the car.'

b. m-an-andrana Rabe$_i$ [m-i-tondra ny fiara Δ_i]
PRES-ACT-try Rabe [PRES-ACT-drive the car]
'Rabe is trying to drive the car.'

The order of constituents here varies, as the two matrix clause arguments of *try* may appear in either order. Other verbs, such as *begin*, only appear in the pattern of (8)a, which P&P analyze as due to backward control.

(9) a. m-an-omboka m-i-tondra ny fiara Rabe (backward control)
PRES-ACT-begin PRES-ACT-drive the car Rabe
'Rabe is beginning to drive the car.'

b. *m-an-omboka Rabe [m-i-tondra ny fiara]
PRES-ACT-begin Rabe [PRES-ACT-drive the car]
'Rabe is beginning to drive the car.'

To account for this difference, Polinsky and Potsdam (2002b) argue that the correct analysis of (9)a is as backward control, with the analysis in (10):

(10) m-an-omboka [m-i-tondra ny fiara Rabe$_i$] Δ_i
PRES-ACT-begin [PRES-ACT-drive the car Rabe]
'Rabe is beginning to drive the car.'

As *Rabe* is not a constituent in the matrix clause, and as there is only one overt matrix clause argument, the bracketed phrase, no reordering at the matrix clause level is possible (at least, visible).

2.3. **Circassian** (Forward and Backward Raising)

Backward raising is illustrated in the Circassian data in (11) (from Polinsky and Potsdam (2006) and Polinsky (p.c.)); here the verb 'begin' only has raising uses:

(11) a. ŝalexe-r [pjəsme-r-q'əʁ zeč'e-m-jə atxənew] Ø-fjež'aʁe-x
boys-ABS letter-ABS-EMPH all-ERG-CONJ write-INF 3ABS-began-3ABS.PL
'The boys began to write the stupid letter all.' (forward raising)

b. [ŝalexe-m pjəsme-r-q'əʁ atxə-new] zeč'e-r-jə Ø-fjež'aʁe-x
boys-ERG letter-ABS-EMPH write-INF all-ABS-CONJ 3ABS-began-3ABS.PL
'The boys began to write the stupid letter all.' (backward raising)

(12) a. ṣ̂alexe$_i$-r [Δ_i pjəsme-r-q'əʁ zeč'e-m-jə atxənew] ∅-fjež'aʁe-x
boys$_i$-ABS [Δ_i letter-ABS-EMPH all-ERG-CONJ write-INF] 3ABS-began-3ABS.PL
'The boys began to write the stupid letter all.'

b. Δ_i [ṣ̂alexe$_i$-m pjəsme-r-q'əʁ atxə-new] zeč'e-r-jə ∅-fjež'aʁe-x
Δ_i [boys$_i$-ERG letter-ABS-EMPH write-INF] all-ABS-CONJ 3ABS-began-3ABS.PL
'The boys began to write the stupid letter all.'

The evidence that P&P present for true raising includes a variety of diagnostics which I do not review here; here we see their evidence involving a floated quantifier. Absolutive is the appropriate case for the subject of 'begin' while ergative is the appropriate case for the subject of 'write'. In (12)a, the lower subject position would be ergative (determined by 'write'), which floats an ergative quantifier in the lower clause, even though the matrix raised subject is absolutive, as dictated by 'begin'. In (12)b, we have the opposite situation: the lower ergative subject floats an absolutive quantifier in the matrix clause. This shows that although the overt NP is phonologically overt in the lower clause, it still has raised into the higher clause. P&P argue that the only viable Minimalist analysis of the full range of facts involves treating both construction types as movement, with different strategies of chain reduction – spell-out of either the head (forward) or tail (backward) of the chain.

The fact that the chain actually has two differing cases is a problem for all approaches which assume that what is shared is a feature structure larger than an INDEX (see e.g., the discussion in Potsdam (2006)).[1] I address these issues below in section 4.3. Zaenen and Kaplan (2002) note examples in German where case cannot be shared between two positions, and propose to restrict equality or subsumption by the Restriction Operator of Kaplan and Wedekind (1993). Of course, case is only a problem to the extent that there is a CASE attribute represented in f-structure: if case is constructive, as proposed by Nordlinger (1998), it is only concerned with GF information in f-structure. Ergative case could have the entry shown in (13):

(13) Ergative case:

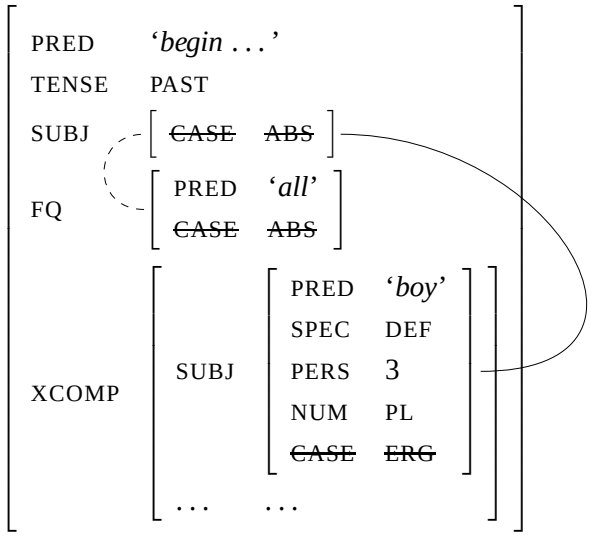
((SUBJ ↑) SUBJ) = ↓
((SUBJ ↑) OBJ)

It is in fact an open question as to whether a CASE feature is necessary in f-structure in LFG (see Spencer and Otoguro (2005) for relevant discussion).

In terms of an equality-based LFG analysis, we could assign (11)b the f-structure in (14), where the grammatical features of the embedded SUBJ are shared up to the matrix SUBJ, which can be associated with a Floated Quantifier at that level. To gloss over the problem of case for now, the strikethroughs in (14) indicate that the apparent case conflict is not considered a problem.

[1] Index-sharing for the analysis of control is standard in HPSG, e.g., Pollard and Sag (1994), which therefore has no problem with case.

(14) F-structure of (11)b, ignoring the case conflict:



As noted above, the fact that predicates can be forward or backward seems to be naturally analyzed within the LFG account of functional control based on equality – the relevant features of the subject are present in both matrix and embedded subject structures. For a languages like Circassian, we simply propose a solution which allows the matrix subject position in the c-structure to be absent.

However, Tsez is problematic under this view. In Tsez, the predicate 'begin' is forward if it is raising, and backward if it is control, so there cannot be any general requirement in the c-structure of the language one way or the other as to which subject positions are obligatorily filled or absent. The equality-based account will simply allow either possibility for either type of verb, incorrectly.

It is clear that the restrictions on forward or backward functional control need to be relativized to particular verb forms – they have to be encoded in the lexical entries of verbs. This means that the restrictions have to be stated at the functional level in LFG. In the following section, I introduce the mechanism to accomplish this.

3. Functional Control Based on Subsumption

These problems all find a simple solution if structure-sharing is asymmetric, as I will show below. The asymmetry comes from the use of the relation of subsumption rather than equality in the statement of structure-sharing.

3.1. Subsumption

Zaenen and Kaplan (2002, 2003) proposed to analyse some cases of structure-sharing in terms of the relation of subsumption, rather than equality. They anticipated the need to express restrictions on information flow in the lexical entries of verbs, and what I present here is an extension of their proposals. For subsumption, f_1 subsumes f_2 if the information associated with f_1 is a subset of that associated with f_2 – in other words, f_1 is more general than f_2. An example from Zaenen and Kaplan (2002) is shown in (15):

(15) Subsumption – the left subsumes (is more general than) the right

$$\begin{bmatrix} A & \begin{bmatrix} C & + \end{bmatrix} \end{bmatrix} \sqsubseteq \begin{bmatrix} A & \begin{bmatrix} C & + \\ D & - \end{bmatrix} \\ B & E \end{bmatrix}$$

(Equality is mutual subsumption.)

In many languages, the agreement information on a verb subsumes the information on the agreed-with subject; for example, the verb may inflect for person and number, while the subject may be coded for person, number and gender. Shieber (1992) discusses an application of subsumption to coordinations in the complement of English *be*. Blevins (2006) provides a fuller discussion of the linguistic relevance of subsumption, and motivations for its necessity in some analyses, though his particular approach to raising predicates differs from what I present below.

Zaenen and Kaplan apply their proposals to Partial VP Fronting in German and Stylistic Inversion in French. Their analyses entail for each language that some functional control relations are stated in terms of equality and some in terms of subsumption. Here I will focus on how a stronger theory of raising and control emerges if the only relation available is subsumption. To see how subsumption is relevant, let us consider the German raising and control examples in (16). All the German examples presented here are main V2 clauses, with an initial 'topic' preceding the finite verb.

(16) a. [Ein Aussenseiter] schien hier eigentlich nie [zu gewinnen]. (raising)
[An outsider] seemed here actually never [to win]
'An outsider never actually seemed to win here.'

b. [Ein Aussenseiter] versuchte hier noch nie [zu gewinnen]. (control)
[An outsider] tried here still never [to win]
'An outsider never tried to win here.'

These are straightforward examples in which the initial subject (bracketed) is also the subject of the embedded XCOMP (bracketed).

German exhibits a construction which looks like backward raising, in the famous case of Partial Fronting of a VP including a subject, over a raising verb, in (17)a, which contrasts with the control verb in (17)b:

(17) Fronted VP containing subject:

a. [[Ein Aussenseiter] zu gewinnen] schien hier eigentlich nie. (raising)
[[An outsider] to win] seemed here actually never
'An outsider never actually seemed to win here.'

b. *[[Ein Aussenseiter] zu gewinnen] versuchte hier noch nie. (control)
[[An outsider] to win] tried here still never
'An outsider never tried to win here.'

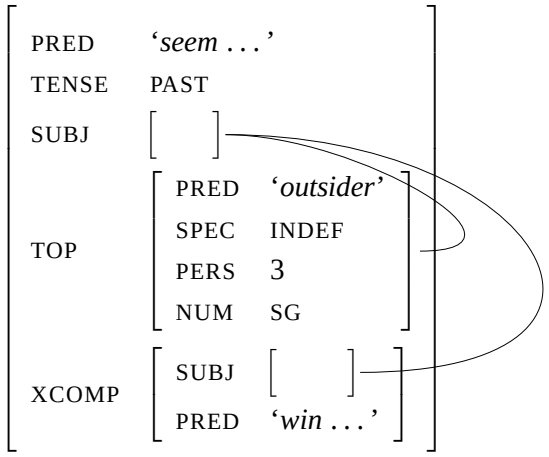
The initial topic is the XCOMP selected by the matrix predicate, and the effective grammatical subject of the clause seems to be the inner bracketed part of the topic *ein Aussenseiter*: the matrix verb shows agreement in present tense (examples in (18) from Haider (2002)), and this subject has nominative case.

(18) a. [Ein Wunder ereignet] hat sich hier noch nie.
[A(NOM) miracle occurred] have.3SG REFL here never ever
'A miracle has never ever occurred here.'

b. [Wunder ereignet] haben sich hier noch nie.
[miracles(NOM) occurred] have.3PL REFL here never ever
'Miracles have never ever occurred here.'

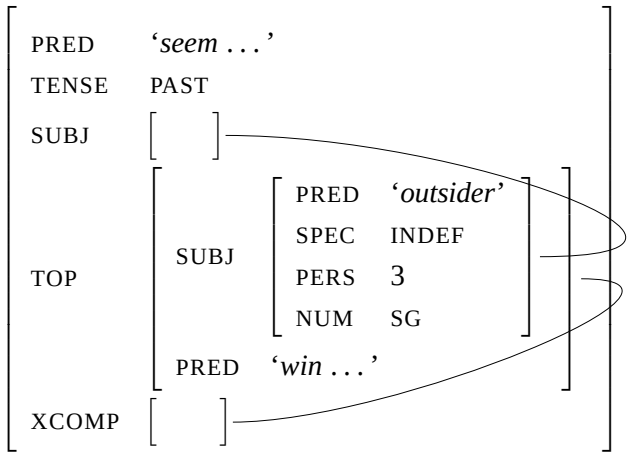
The regular 'forward raising' example with *ein Aussenseiter* alone in the initial position in (16)a has the f-structure in (19), and (17)a has the f-structure in (20).

(19)



The fronted constituent is TOPIC and SUBJ, and 'seem' dictates that SUBJ = XCOMP SUBJ.

(20)



The fronted constituent is TOPIC and XCOMP, and 'seem' dictates that SUBJ = XCOMP SUBJ.

Apart from what is topicalized (SUBJ or XCOMP), the f-structures are identical, and (20) looks like a case of backward raising, allowed as a possibility by the equality-based structure sharing equation SUBJ = XCOMP SUBJ. The non-derivational proposals of Hudson (1997) (Dependency Grammar) and Meurers and De Kuthy (2001) (HPSG) for the raising verbs are rather similar to that of Zaenen and Kaplan, in that they allow the embedded subject to act as the subject in matrix clause. I will show below that the assumption of full structure-sharing for raising verbs is incorrect.

However, the problem for the LFG analysis now is how to account for the forward-only restriction on *versuchen*. For the data in (17), Zaenen and Kaplan (2002, 2003) proposed to introduce subsumption, where information only flows one-way, only from the general position to the specific position. With SUBJ and XCOMP SUBJ, there are two options. (21)a defines a 'forward' predicate, (21)b a 'backward' predicate.

(21) Subsumption

a. SUBJ $\sqsubseteq$ XCOMP SUBJ (Information only flows down from SUBJ; whatever information the subject has in the matrix it has in the embedded constituent, but not vice versa)

b. XCOMP SUBJ $\sqsubseteq$ SUBJ (Information only flows up to SUBJ; whatever information the subject has in the embedded constituent it has in the matrix, but not vice versa)

For the data in (17), Zaenen and Kaplan propose to analyze *scheinen* with equality as in (2), but *versuchen* with subsumption as in (21)a. This has the effect of allowing the subject of the raising verb to be either in matrix or embedded subject position, as in the f-structures above, while the control verb must have its matrix subject position filled, as I show below in (23). In other words, German allows forward or backward raising, but only forward control. The f-structure for (17)b, which is ungrammatical, is shown in (23). The subsumption relation is shown by the directional arrow on the curved line in the f-structure representation. I assume that the finite verb determines agreement and case features of its subject, but nothing more.

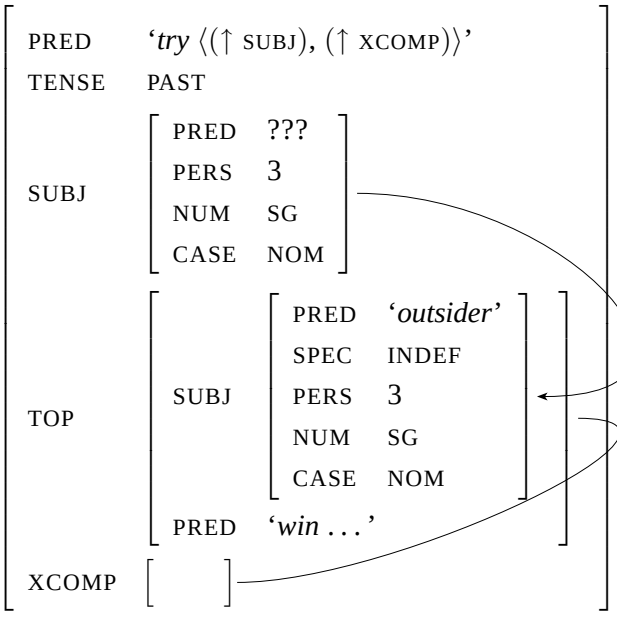
(22) German – Zaenen and Kaplan (2002):

a. Raising: SUBJ = XCOMP SUBJ
subject can be upstairs or downstairs
(as above in (19) and (20))

b. Control: SUBJ $\sqsubseteq$ XCOMP SUBJ
subject can only be upstairs
(see (23))

(17) b. *[[Ein Aussenseiter] zu gewinnen] versuchte hier noch nie.
[[An outsider] to win] tried here still never
'An outsider never tried to win here.'

(23)



This is ill-formed, as the matrix nucleus is INCOMPLETE: there is no PRED information about the thematic SUBJ of *try* (following (22)b).

The position which other constraints (in particular, COMPLETENESS) require to be filled must be the one to the left of the $\sqsubseteq$ symbol. (24) presents a fuller summary of the consequences of the subsumption approach:

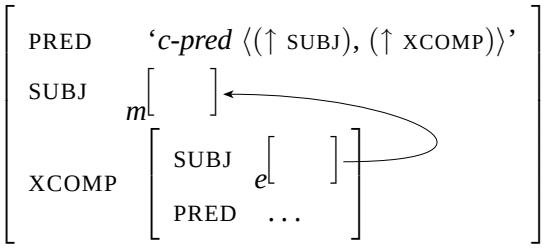
(24) Typology of Control and Raising: The position which other constraints (in particular, COMPLETENESS) require to be filled must be the one to the left of the $\sqsubseteq$ symbol.

a. SUBJ $\sqsubseteq$ XCOMP SUBJ: Control: forward only; Raising: forward, or backward
A control predicate has a thematic subject, so it requires a SUBJ with a PRED. As information only flows down to XCOMP SUBJ, it is SUBJ which must be expressed, or else it will have no information (see (23)). In this way, (24)a predicts forward control: SUBJ must be expressed, and XCOMP SUBJ cannot. A raising predicate does not require a thematic subject, and (24)a does not constrain its position.

b. XCOMP SUBJ $\sqsubseteq$ SUBJ: Control: backward only; Raising: backward only.
As information only flows upward, unless the XCOMP SUBJ is expressed overtly, the XCOMP will be incomplete (no information about the PRED of its SUBJ). So the XCOMP SUBJ must be expressed, and its information flows up to the matrix SUBJ, meaning that that position cannot be expressed overtly (by LFG's principle of Uniqueness). (24)b predicts that XCOMP SUBJ is expressed and that SUBJ is not.

What is important here is that subsumption directly determines the properties of $\frac{3}{4}$ of the space of the phenomena (see (24)); equality is merely compatible with the entire space. The specific prediction for backward control is shown in (25).

(25) Backward Control: (24)b
The matrix subject position contributes to *m*, the embedded subject position contributes to *e*. If only the matrix subject position is filled, *e* does not get a PRED value (at least), and the XCOMP is INCOMPLETE.



3.2. Thematic Positions and Apparent Backward Constructions

Viewed in the typology developed by P&P, German raising turns out to show only consistent 'forward' properties, and applying equality as in (2) to it makes incorrect predictions. To see this, it is necessary to consider what P&P call 'false backward raising', illustrated by the Greek data in (26), from Polinsky and Potsdam (2005):

(26) a. i dhaskoli stamatisan/*stamatise [na malonun tus mathites]
the teacher.PL stop.3PL/*stop.3SG [COMP scold.3PL the students]
'The teachers stopped scolding the students.' (forward raising)

b. stamatisan/*stamatise [na malonun i dhaskoli tus mathites]
stop.3PL/*stop.3SG [COMP scold.3PL the teacher.PL the students]
'The teachers stopped scolding the students.' (false backward raising)

While the matrix predicate in (26)b agrees with the embedded subject 'teachers', Polinsky and Potsdam (2006) show that there is no evidence of full syntactic presence of that subject in matrix clause, in contrast to the evidence from Tsez, Malagasy and Circassian presented in section 2. P&P call this 'false backward raising', where the only evidence for the subject being in the matrix clause is agreement and case, properties which are determined by matrix predicate. P&P analyze this in Minimalist terms as long-distance Agree from the matrix Tense to the embedded subject, where this operation of Agree is also enough to satisfy the matrix EPP.

Returning to German, the crucial example (17)a also shows no clear evidence of a syntactic subject in the matrix clause – it is well-known that a subject inside the fronted VP only shows embedded and/or narrow scope behavior (see e.g., Netter (1991), Meurers and De Kuthy (2001)).

(17) a. [[Ein Aussenseiter] zu gewinnen] schien hier eigentlich nie.
[[An outsider] to win] seemed here actually never
'An outsider never actually seemed to win here.'

Under VP fronting, the embedded subject cannot be the antecedent of an anaphor (see (27)), nor can it float a quantifier (see (28)), in the matrix clause:

(27) *[Ein Aussenseiter$_i$ zu gewinnen] scheint [seiner$_i$ Mutter] hier nie.
an outsider(NOM) to win seem.3SG [his(DAT) mother] here never
'No outsider seems to his mother to win here ever.'

(28) a. [Ein Fehler unterlaufen] ist meinem Lehrer noch nie.
a mistake(NOM) happened be.3SG my(DAT) teacher still never
'So far my teacher has never made a mistake.'

b. [Fehler unterlaufen] sind meinem Lehrer nicht viele.
mistakes(NOM) happened be.3PL my(DAT) teacher not many
'My teacher has not made many mistakes.'

c. ?*[Fehler unterlaufen] sind viele meinem Lehrer nicht.

d. ?*[Fehler unterlaufen] sind meinem Lehrer viele nicht.

While (28)a may look like an example of a floated quantifier, the fact that other positions of *viele* are ungrammatical suggests otherwise. It seems that either *viele* or *nicht viele* ('not many') is extraposed in the first example, and hence is forced into a clause-final position. The examples do not involve true quantifier float from the matrix subject. In other words, all of the properties of the embedded subject are not shared up to the matrix subject, in contrast to what the equality-based account of German raising ((22)a) predicts.

This leads to the conclusion that (17)a is an example of false backward raising. In other words, German does not have equality in raising (does not have (2)), but only has forward subsumption as in (21)a. Let us see how (17)a is well-formed. Berman (2003) proposed an LFG analysis of German clause structure in which the 'subject condition' (like the 'EPP' in Minimalist syntax) can satisfied by agreement features alone. This is possible just in case the matrix predicate has a non-thematic subject, as with a raising predicate, or some other kind of impersonal predicate. Berman (2003, 58) restricts this possibility to 3rd singular subjects, via the following constraint on f-structures:

(29) All f-structures must have a PRED-feature, unless they are specified for third person singular.
$(\forall f)[\neg(f\ \textsc{pred}) \Rightarrow [(f\ \textsc{num}) = \textsc{sg}, (f\ \textsc{pers}) = 3]]$

This restriction is to explain why the impersonal clause in (30)a in German is well-formed, but not that in (30)b:

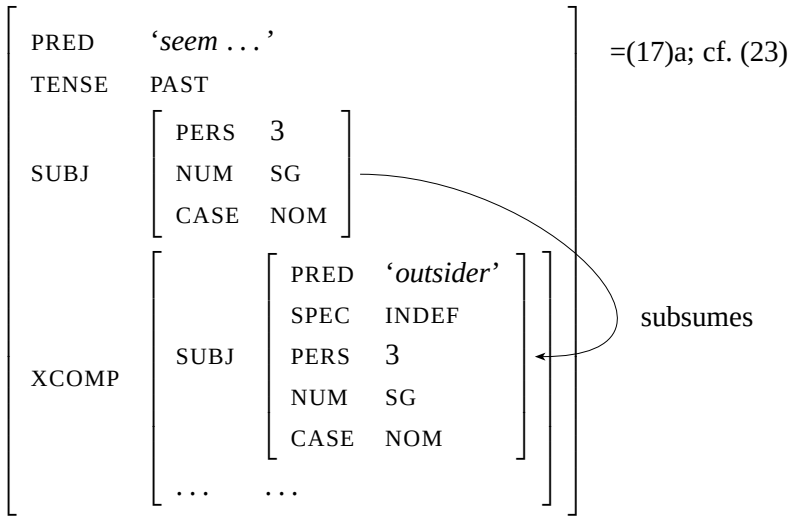
(30) a. ... weil getanzt wird. (3sg subject)
because dance.PASS be.PRES
'... because there is dancing'

b. * . . . weil getanzt werden. (1pl/3pl subject)
because dance.PASS be.PRES

However, to cover the false backward raising structures discussed in this paper, we need to allow for PRED-less f-structures that are 3rd plural as well as 3rd singular, as in (28)a/b. I propose that the f-structure of (30)b is not well-formed because of the fact that the singular morphology does not disambiguate 1st and 3rd persons. There is no evidence in (30)b as to whether the subject's person is 1 or 3. However, in the examples in (28), the f-structure for each is unambiguously 3rd person, due to the presence of the noun *Fehler*. Hence, I propose that the condition in (29) be revised to apply to all unambiguously-marked 3rd person f-structures, with no restriction on number.

The assumption that no PRED is necessary in the matrix clause allows us to analyze (17)a via subsumption (forward raising); the example is grammatical as the agreement features of the matrix subject are shared down to the embedded subject, with the subject further specified by its overt form in the embedded constituent, consistent with (21)a.

(31)



Such an f-structure is not possible for 'try' (see (23)), as any predicate with a thematic subject requires a PRED value for that subject.

False backward raising is possible in a language which has only forward raising but which, like German, allows the Subject Condition to be satisfied by only agreement features.

4. Further Thoughts

4.1. Subsumption

Blevins (2006) writes: "A standard model of LFG can, for example, be modified to run in 'subsumption mode' by replacing all identity relations between attributes by appropriate subsumption relations". However, my argument here is that moving to subsumption does in fact have important analytic consequences. Equality as in (2) is too permissive, and might even be removed from the options in Universal Grammar, keeping (21) instead, with languages or predicates being classified as forward (like German; (21)a), backward ((21)b), or both. Fang (2006) presents an account of Chinese VP structure which crucially relies on the notion of subsumption, and for which equality would not work, as there is an asymmetry in the structures so related.

4.2. Equality, Subsumption, and the Copy Theory of Movement

Does equality or subsumption correspond to the copy theory of movement, especially the one with 're-merge', as in Zhang (2004), Fox and Pesetsky (2005) and Hornstein et al. (2006)? On the face of it, the description of 'remerge' is that the same information is simultaneously present in several places at once, so this looks like equality.

Subsumption can be backward (upward) or forward (downward), but movement is only upwards. In the most recent versions of Minimalism, feature valuation happens from a probe, downwards via Agree to a goal in a lower position. Feature valuation itself adds information, but then some valued features are checked off and disappear. If features are "checked off" as movement operates upwards, then the higher copy actually is informationally less specific than the lower copy, though it may be that the correct notion here is one of being 'informationally most complete': an element having all of its relevant features checked (in the sense of being licensed), without them necessarily having disappeared from the structure altogether.

Potsdam (2006) presents the most complete discussion of the consequences for Spell-Out of backward control and raising. He follows the Chain Reduction Principles of Nunes (2004): only one copy can be pronounced, and the pronounced copy is the one with the fewest unchecked features. With the backward structures considered here, the case feature of the relevant argument can be checked in the lower clause, and can be checked again in the higher clause. As a consequence, the case feature is as equally checked in the lower or the higher clause, and so in principle Spell-Out could apply optionally to either copy (cf. (11)b/(14)). Potsdam proposes that the case feature can be assigned a value multiple times, with each successive valuation overwriting the previous one.

Without further elaboration, such a system predicts optional forward or backward control or raising, but it cannot force only backward structures. Forward-only structures are straightforward: the argument is assumed not to be able to get case in the lower clause, as in English. Moreover, it does not seem to tie the Spell-Out of case to the position of the argument – e.g., Ergative in the lower clause and Absolutive in the higher clause for (11)b/(14). As the chain is formed, the argument picks up Ergative in the lower clause and Absolutive in the higher clause. If Absolutive overwrites Ergative, then strictly speaking, the chain only has Absolutive for a case value, and it is not clear how Ergative could ever Spell-Out. If the chain has both values, it somehow has to 'know' that Ergative Spells-Out on the downstairs copy and Absolutive on the upstairs copy.

Now, the obvious solution to this dilemma is to have cyclic Spell-Out: the Ergative Spells-Out in the lower clause before the chain into the upper clause is even formed. And if Spell-Out does not take place, the chain is formed by movement into the upper clause, the case value gets rewritten as Absolutive and that is what Spells-Out. However, this now divorces Spell-Out from the predicate that governs it: an argument would be Spelled-Out overtly in a lower clause without having any access to the predicate/construction type of the upper clause, which is precisely the locus of whether the construction should be forward or backward. As far as I can see, it is the properties of the upper clause (e.g., properties of the governing predicate) which determine whether a construction is forward or backward. The Spell-Out mechanisms of the Minimalist Program currently do not seem to have any means for encoding this, as Spell-Out is not relativized to properties of the complements of verbal heads.

Eric Potsdam (p.c.) presents two interesting avenues for lines of development of these ideas. One is that, if two copies in a chain are equally informationally specific, but independent principles of PF only allow one to Spell-Out, then there should be two outputs – an upstairs Spell-Out, and a downstairs Spell-Out. However, there could be other and independent principles which further restrict the options: for example, some other property of the language that disfavors Spell-Out in the matrix clause would therefore bias towards a backward construction with Spell-Out in the embedded clause. In this form, such an approach would not be suitable for a language in which specific predicates are forward and others are backward, for it presumes language-wide conditions interacting with Spell-Out.

The other idea would be to let predicates determine the category of their overall complement, in a way that interacts with phases or domains of Spell-Out. For example, a forward control predicate would take a complement αP such that the subject within it is not Spelled-Out (αP would not be a Spell-Out domain, while a vP within αP may be such a domain). A backward control predicate would take a complement βP bigger than αP, such all the elements within βP need to Spell Out. This then relativizes Spell-Out to properties of the governing predicate, through its complement selection.

Putting these observations back in the context of LFG, the position of the overt argument is determined at f-structure, as I have stressed in this paper – the f-structure information in the selecting head controls the c-structure appearance of arguments. In constrast, I think it would be odd in any theory to directly relativize phrase structure configuration to a particular head: for example, a language in which one verb requires a preceding NP but the next verb requires a following NP. It seems to me that a need for such a description would be highly unexpected, and it would amount to direct access from heads to c-structure positions of arguments. It is in fact much more natural that the conditions on forward and backward structures come from f-structure, rather than c-structure.

Multiple case is only an issue if there is a case feature with a value; this is not a necessary part of the LFG analysis, for case could be given a solely constructive role (e.g., as in (13)). In fact, there could be a nice prediction following on from the ideas developed by Spencer and Otoguro (2005): it would be that backward constructions are possible in languages where case has only a GF-constructive role, but not possible in languages where a CASE attribute is present inside the GF, with a variety of values (say, for the purposes of case agreement).

4.3. 'Backward Subsumption' or Backward Obligatory Anaphoric Control?

The alternative to subsumption is to treat control and raising as always involving a coindexed PRED 'PRO' in f-structure, to prevent overt expression of the argument (cf. the analysis of control in HPSG, e.g., Sag and Pollard (1991)). All we have to do is specify obligatory anaphoric control, sharing of INDEX as in line (i) in (32), and then make sure that a PRED 'PRO' is somewhere in f-structure, in shown in lines (ii) or (iii):

(32) Control Predicate – thematic subject:
($\uparrow$ PRED) = '*c-pred* $\langle$($\uparrow$ SUBJ), ($\uparrow$ XCOMP)$\rangle$'

Raising Predicate – non-thematic subject:
($\uparrow$ PRED) = '*r-pred* $\langle$($\uparrow$ XCOMP)$\rangle$($\uparrow$ SUBJ)'

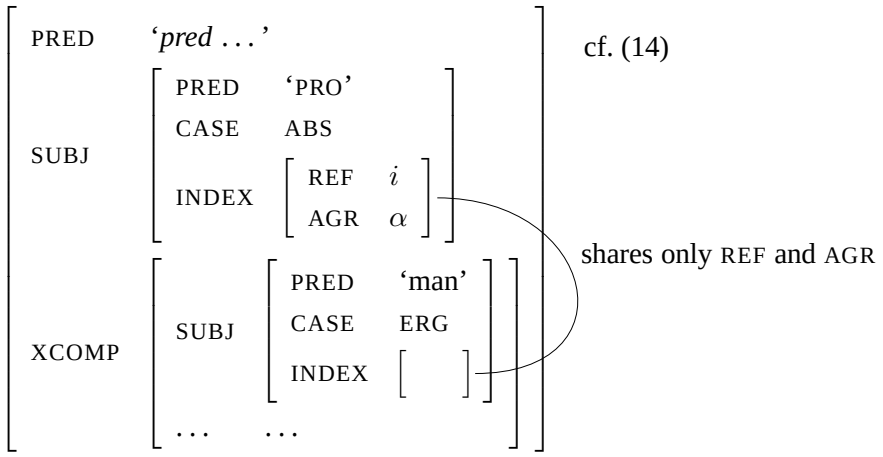
(i) ($\uparrow$ SUBJ INDEX) = ($\uparrow$ XCOMP SUBJ INDEX) (obligatory anaphoric control)
(ii) { ($\uparrow$ XCOMP SUBJ PRED) = 'PRO' (forward; lower position unavailable)
(iii) | ($\uparrow$ SUBJ PRED) = 'PRO' } (backward; higher position unavailable)

This is perfectly viable solution, subject to two provisions: raising is effectively treated as always being 'Copy Raising' (see e.g., Asudeh (2002, 2004)), and the backward cases have to be treated as in line (iii) of (32), with a higher pronoun.[2]

Now, further assume that INDEX has two parts, REF and AGR (cf. Bresnan (2001)), as in (33). This 'PRO' analysis explains which position is empty (by positing a 'PRO' there), and has no problem with different cases upstairs and downstairs, for case is not part of what is shared, only REF and AGR are:

[2]Polinsky and Potsdam (2006) and Potsdam (2006) have argued that this kind of analysis is not appropriate for some instances of backward control and raising.

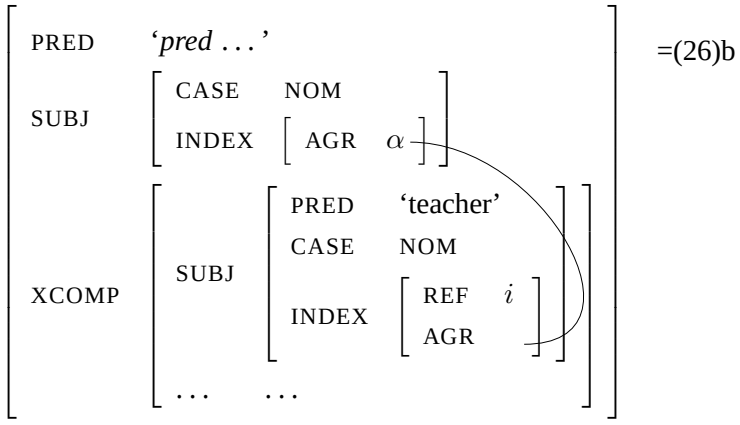
(33)



However, this analysis creates apparent violations of Principle C of the Binding Theory, and requires expletive subjects of raising predicates to be PRED ‘PRO’, rather than to be lack a PRED value altogether (because what is shared is only INDEX, in a structure that also is specified for case). In addition, it does not seem to have the right properties to account for false backward raising, which would have the account sketched in (34); in this situation, all that is shared is the AGR part of INDEX:

(34) False backward raising:

(↑ SUBJ INDEX AGR) = (↑ XCOMP SUBJ INDEX AGR)
(cf. ‘long-distance agreement’ in Minimalist accounts)



To see what the consequences of this kind of analysis are, we have to look at the overall space of control and raising types. (35) shows this for a subsumption-only account and (36) shows it for an equality-only account:

(35) Subsumption-only mode

a. Forward subsumption for forward control and raising; which also allows false backward raising in a language in which AGR alone can satisfy the Subject Condition; or

b. Backward subsumption for backward control and raising.

(36) Equality-only mode

a. Equality of INDEX plus downstairs PRED ‘PRO’ for forward control and raising; or

b. Equality of INDEX plus upstairs PRED 'PRO' for backward control and raising; or

c. Equality of AGR for false backward raising in a language in which AGR alone can satisfy the Subject Condition.

The crucial difference is that false backward raising cannot be a special case of one of the other types, as the informational unit that is shared is different (INDEX or AGR). So on this approach, false backward raising does not fall out as a special property of forward raising (why are (36)a and (36)c related?), though it does with subsumption (both are part of (35)a).

Now we know that German cannot have traditional equality across the board, due to the ungrammaticality of *(17)b. We also know that German does not have equality of INDEX: the expletive subject in false backward raising has no referential index, on Berman's analysis. Hence German requires the option in (36)c, equality of AGR; and this must only be used for cases of false backward raising.

References

Asudeh, Ash. 2002. Richard III. In Mary Andronis, Erin Debenport, Anne Pycha, and Keiko Yoshimura (eds.), *Papers from the 38th Regional Meeting*. Chicago, Chicago Linguistics Society, 31–46.

Asudeh, Ash. 2004. *Resumption as Resource Management*. Doctoral dissertation, Stanford University.

Asudeh, Ash. 2005. Control and semantic resource sensitivity. *Journal of Linguistics* 41, 1–47.

Berman, Judith. 2003. *Topics in the Clausal Syntax of German*. Stanford, CSLI Publications.

Blevins, James P. 2006. Feature-based grammar. To appear in R. Borsley and K. Börjars (eds.) *Non-transformational syntax*, Blackwell Publishing.

Bresnan, Joan. 1982. Control and complementation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*. Cambridge, Mass., MIT Press, 282–390.

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

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MIT Press.

Fang, Ji. 2006. *The Verb Copy Construction and the Post-Verbal Constraint In Chinese*. Doctoral dissertation, Stanford University.

Fox, Danny, and David Pesetsky. 2005. Cyclic linearization of syntactic structure. *Theoretical Linguistics* 31, 1–45.

Haider, Hubert. 2002. Mittelfeld phenomena (case #64). In Henk van Riemsdijk and Martin Everaert (eds.), *The Syntax Companion*. Oxford, Blackwell Publishing.

Hornstein, Norbert. 1999. Movement and control. *Linguistic Inquiry* 30, 69–96.

Hornstein, Norbert, Jairo Nunes, and Kleathes K. Grohmann. 2006. *Understanding Minimalism*. Cambridge, Cambridge University Press.

Hudson, Richard. 1997. German partial VP fronting. Ms. University College, London.

Kaplan, Ronald M., and Jürgen Wedekind. 1993. Restriction and correspondence-based translation. In *Proceedings of the 6th conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA. Association for Computational Linguistics, 193–202.

Meurers, Walt Detmar, and Kordula De Kuthy. 2001. Case assignment in partially fronted constituents. In Christian Rohrer, Antje Roßdeutscher, and Hans Kamp (eds.), *Linguistic Form and its Computation*. Stanford, CSLI Publications, 29–63.

Netter, Klaus. 1991. Clause union and verb raising phenomena in German. Research Report 91-21, Saarbrücken, DFKI.

Nordlinger, Rachel. 1998. *Constructive Case: Evidence from Australian Languages*. Stanford, Dissertations in Linguistics, CSLI Publications.

Nunes, Jairo. 2004. *Linearization of Chains and Sideward Movement*. Cambridge, MIT Press.

Polinsky, Maria, and Eric Potsdam. 2002a. Backward control. *Linguistic Inquiry* 33, 245–282.

Polinsky, Maria, and Eric Potsdam. 2002b. Backward control: Evidence from Malagasy. In Andrea Rackowski and Norvin Richards (eds.), *Proceedings of AFLA VIII*. (MIT Working Papers in Linguistics Vol. 44), Dept. of Linguistics, MIT, 257–272.

Polinsky, Maria, and Eric Potsdam. 2005. Backward raising: Theoretical and empirical options. Paper presented at the workshop on 'New Horizons in the Grammar of Raising and Control', LSA Summer Linguistic Institute Workshop, Harvard University, July 2005.

Polinsky, Maria, and Eric Potsdam. 2006. Expanding the scope of control and raising. To appear in *Syntax* 9.

Pollard, Carl, and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, University of Chicago Press and Stanford, CSLI Publications.

Potsdam, Eric. 2006. Backward object control in Malagasy and principles of chain reduction. Ms. University of Florida.

Sag, Ivan, and Carl Pollard. 1991. An integrated theory of complement control. *Language* 67, 63–113.

Shieber, Stuart M. 1992. *Constraint-Based Grammar Formalisms: Parsing and Type Inference for Natural and Computer Languages*. Cambridge, Mass., MIT Press.

Spencer, Andrew, and Ryo Otoguro. 2005. Limits to case – A critical survey of the notion. In Mengistu Amberber and Helen de Hoop (eds.), *Competition and Variation in Natural Languages: The Case for Case*. Amsterdam, Elsevier, 119–145.

Zaenen, Annie, and Ronald M. Kaplan. 2002. Subsumption and equality: German partial fronting in LFG. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG02 Conference*. Stanford, CSLI Publications, 408–426. (At `http://cslipublications.stanford.edu/LFG/7/lfg02.html`).

Zaenen, Annie, and Ronald M. Kaplan. 2003. Subject inversion in French: Equality and inequality in LFG. In Claire Beyssade, Olivier Bonami, Patricia Cabredo Hofherr, and Francis Corblin (eds.), *Empirical Issues In Syntax and Semantics 4*. Paris, Presses de l'Université Paris-Sorbonne, 205–226.

Zhang, Niina. 2004. Move is remerge. *Language and Linguistics* 5, 189–208.

NORWEGIAN *WHEN*-CLAUSES

Nola M. Stephens
Stanford University

Proceedings of the LFG06 Conference

Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)

2006

CSLI Publications

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

Abstract

Norwegian has two connectives meaning *when*: *da* and *når*. *Da*- and *når*-clauses have been treated as relative clauses that differ essentially in aspectual features (cf. Faarlund *et al*. 1997). I show that this view fails to fully account for the data, and I argue instead that *da* and *når* are crucially different in their lexical-syntactic properties. Whereas *når* always introduces relative clauses (bound or free), *da* can only introduce relative clauses that have a lexical head. *Da*-clauses without a lexical head are subordinate clauses adjoined directly to the matrix clause. Corpus data reveal that the actual aspectual properties of *da*- and *når*-clauses correlate with the status of the clause as relative or non-relative.

1. Introduction

Norwegian, like German, Dutch and Danish, exhibits two temporal connectives that correspond to the English *when* (see Vikner 2004). Though their syntactic properties have been largely ignored, these two connectives, *da* and *når*, constitute classic ingredients of Norwegian grammars. In general, both *da*- and *når*-clauses are regarded as relative clauses that differ crucially in aspectual features (see Faarlund, Lie & Vannebo 1997). According to normative tradition, *da* introduces clauses denoting past episodic events, while *når* is used with present, future, and habitual past events. The examples in (1) and (2) illustrate clauses that conform to this tradition.

(1) Det skjer [når folket vil]
That happens [NÅR people-DEF will]
'That happens when the people want.'

(2) Jeg fartet mye rundt (…) [da jeg var korrespondent]
I traveled much around [DA I was correspondent]
'I traveled around a lot (…) when I was a correspondent.'

As shown in (3)-(4), *da*- and *når*-clauses also appear with overt lexical heads.

(3) Hver dag [når vi kom på jobb] (…)
Every day [NÅR we came to job]
'Every day when we came to work, (…).'

(4) I en tid [da boksalget synker] (…)
In a time [DA book-sale-DEF sink]
'In a time when the book's sales are sinking (…).'

Within the framework of LFG, I consider each type of temporal clause above and show that the traditional analysis fails to adequately account for the data. I propose that the essential differences between *da*- and *når*-clauses lie in their lexical-syntactic properties, not in their aspectual features. Specifically, the connective *når* always introduces a relative clause. *Når*-clauses that modify a lexical head as in (3) are bound relatives, while those without lexical heads (e.g. (1)) are free relatives. In contrast, although *da* is also used for bound relatives (see (4)), *da* cannot introduce free relatives. Unlike the *når*-clause in (1), the *da*-clause in (2) is adjoined directly to the matrix clause. Section 2 below addresses *når*-clauses and argues that, in sentences like (1) and (3), *når* consistently appears in the specifier position of a relative CP and has one lexical entry. Section 3 turns to the more controversial status of *da*, arguing that *da* is positioned in C and requires two lexical entries, one for relative clauses and one for non-relative clauses. Section 4 then shows that the actual aspectual properties of these clauses as revealed by corpus data support the purposed analysis.

2. *Når*

2.1 Properties of når

Når is generally regarded as an *hv*-word (*wh*-word). Like other *hv*-words, *når* can introduce direct and indirect questions as shown in (5) and (6), respectively.

(5) Når skal vi politikere bli voksne?
NÅR shall we politicians become grown
'When will we politicians grow up?'

(6) Jeg spurte ham om når hun kom tilbake.
I asked him about NÅR she comes back
'I asked him when she'll come back.'

A typical feature of *hv*-words is that they also introduce both bound and free relative clauses. Significantly, temporal *når*-clauses that lack an overt lexical head (e.g. (1)) exhibit behavior indicative of free relatives. For example, like free relatives and *hv*-words in general, they allow expansion (see Bresnan & Grimshaw 1978). One manifestation of this property in Norwegian is the addition of *som helst* (lit. 'as rather'):

(7) [Når som helst jeg har søkt om råd] har han gitt meg veiledning[1]
[NÅR SOM HELST I have sought about advice] has he given me direction
'Whenever I have looked for advice, he has given me direction.'

Moreover, as discussed in Faarlund *et al.* (1997), *når*-clauses as in (1) accommodate certain ambiguities typical of free relatives. They have a *general* or a *specific* interpretation, dependent upon whether they refer to something (here: a time) that is left undetermined (see (8)) or something that is specified in the utterance (see (9)).

(8) Kom [når du har lyst] (Faarlund *et al.* 1997, p. 1053) (*general interpretation*)
come [NÅR you have desire]
'Come when you want.'

(9) Hunden kom [når ho ropte] (*ibid*, p. 1054) (*specific interpretation*)
dogs-DEF came [NÅR she called]
'The dogs came when she called.'

Lastly, *når*-clauses without lexical heads behave like relatives in that they allow long-distance dependencies (cf. (10)-(11), due to Helge Lødrup, p.c.).

(10) Napoleon var faktisk på Korsika på den tiden [når du påstår at
N. was actually on Corsica at the time [NÅR you claim that
han ledet hæren i Italia_]
he lead army-DEF in Italy]
'N. was actually in Corsica at the time when you claim that he led the army into Italy.'

[1] http://www.yogasenteret.no/Artikler/artikkel.php?article_id=48

(11) Napoleon var faktisk på Korsika [når du påstår at han ledet hæren i Italia_]
'N. was actually in Corsica when you claim that he led the army into Italy.'

In light of these observations, both types of *når*-clauses should be given a relative clause structure.

Given the status of *når* as an *hv*-word, I adopt the well-precedented approach of assigning it to SpecCP (see, e.g., Dalrymple 2001). This analysis is supported by the fact that earlier stages of Norwegian (see (12)) and certain modern dialects (see (13)) allow *når* to immediately precede a complementizer.

(12) [Naar som helst **at** for[ne] Jacob hafde (…)]
[NÅR SOM HELST AT aforementioned Jacob had
'Whenever the aforementioned Jacob had (…).' (Absalon Pederssøn Beyer diary 1563[2])

(13) Selv [når **at** det stormer som verst] må du (…)[3]
Self [NÅR that it storms like worst] must you
'Even when it storms the worst, you must (…).'

Since *at* appears in C, an analysis of *når* as the head of CP would fail to explain the distribution in (12)-(13). Moreover, examples such as (7) also suggest that *når* should not be analyzed as a complementizer since *når* can be expanded to an element larger than that typically associated with C. In the ensuing discussion, I follow the work of Groos & van Riemsdijk (1981), Grosu (2003) and others and regard the *wh*-phrase to be in SpecCP in both bound and free relative clauses.

2.2 Analysis of når

In general, Norwegian is a head-initial V2 language.[4] Nevertheless, because V2 effects correlate with assertive force (Andersson 1975, Wechsler 1991, Sells 2001), none of the *when*-clauses relevant to this discussion will exhibit V2 syntax. In other words, since these are adjunct clauses, they, like most embedded clauses in Norwegian, are not assertions, and their finite verbs will necessarily appear in V regardless of which elements in the CP are filled.

Given the discussion so far, I propose the partial phrase structure rules for Norwegian in (14) and the lexical entry for *når* in (15).[5] Note that this lexical entry only applies to the use of *når* in relative clauses (hence the specification STMT-TYPE = REL) and does not account for interrogative uses of *når*.

(14) CP → CP XP
↑=↓ (↑ADJ)=↓
CP → XP C'
(↑DF) =↓ ↑=↓
C' → C IP
↑=↓ ↑=↓
NP → N' CP
↑=↓ (↑ADJ)
(↑INDEX)=(↓TOPIC INDEX)
(↓TOPIC)=(↓GF* GF)

[2] Thanks to H. Lødrup for this example, which is available at <http://www.dokpro.uio.no/cgi-bin/litteratur/oratxtprod.cgi?tabell=beyer&id=dagbok008&frames=Nei&offset=25896&lengde=11#sted>.
[3] http://www2.bi.no/biforum/bi298/07_2_98.htm
[4] For simplicity, I treat all V2 clauses as CPs with the finite verb fixed in C (though see Sells 2001).
[5] Though the notation is abbreviated throughout, I assume all adjuncts to be sets.

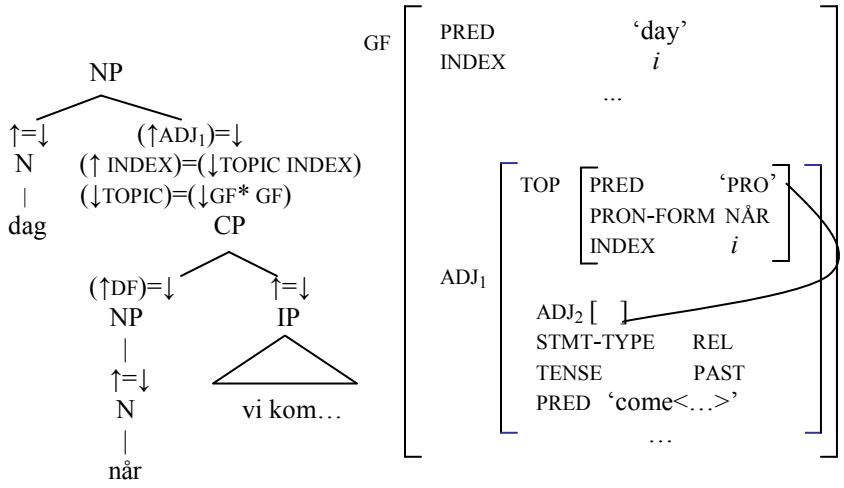
(15) *når* N

(ADJ ↑)	((TOPIC ↑) STMT-TYPE) = REL
(TOPIC ↑)	(TOPIC ↑) TENSE
(↑ PRED) = 'PRO'	(((ADJ TOPIC ↑) PRED) = 'PRO')
(↑ PRON-FORM) = NÅR	CAT((ADJ TOPIC ↑), N)
(↑ INDEX) = *i*	

Since adjuncts are represented by a variety of different phrase types, the lexical category of *når* is subject to debate. Even so, NPs are among the possible phrase types for adjuncts, and I treat *når* as belonging to category N, paralleling analyses for other *wh*-words (see Bresnan & Grimshaw 1978).

Applying (14) and (15) to bound relatives yields the c- and f-structures in (16).

(16) *Hver dag når vi kom* (…) (cf. (3))

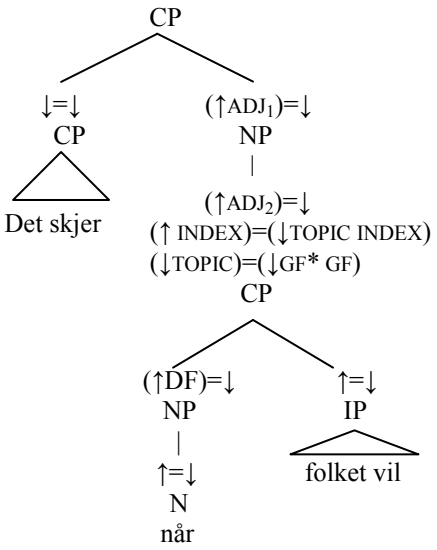


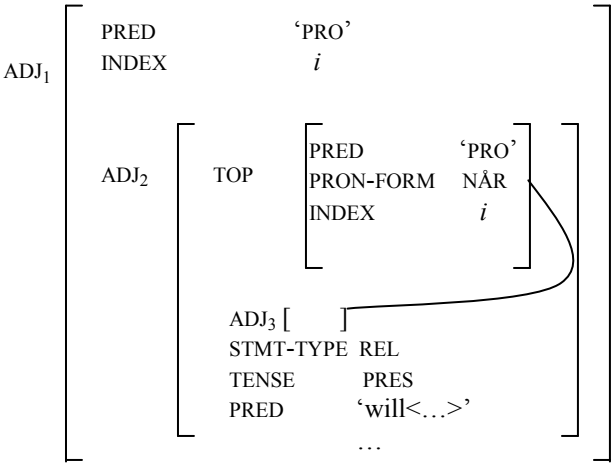
Taking this structure to be representative of relative clauses introduced by *når*, note that the f-structure of the CP is embedded in a larger f-structure and contains a TOPIC that shares an index with an element of the larger f-structure. This index sharing marks the semantic relationship between the head of the relative clause (*dag* in (16)) and *når* in SpecCP and is ensured by the functional annotation on CP and by the lexical specification of *når* as a TOPIC (see, e.g., Falk 2001 concerning the choice of TOPIC for this position). *Når*'s lexical entry also dictates that its f-structure is not only a TOPIC but also an ADJUNCT, and the functional uncertainty equation on CP guarantees that the TOPIC will structure share this ADJUNCT. This linking of the TOPIC to a GF appropriately integrates the TOPIC into the f-structure in fulfillment of the extended coherence condition (see Bresnan 2001). Moreover, it is this functional uncertainty equation that is responsible for licensing the unbounded dependencies. In essence, this equation ensures that the ADJUNCT representing the gap within the relative clause is linked to the TOPIC which is, in turn, co-indexed with the head of the relative clause.

Another notable property of this analysis is that the CP itself is headless. Because C is a functional category rather than a lexical one, the principle of endocentricity does not rule out the absence of C in the structure. Furthermore, information such as tense and clause-type that would be provided by C is supplied instead by *når*. Consequently, positing a C head would not contribute to the f-structure information and would, therefore, be ruled out by economy of expression (Bresnan 2001, Falk 2001). Similarly, the embedded IP lacks an I head. In this case, the information that would be provided by I is given in V, rendering I superfluous.

The above arguments apply equally to free relative clauses. Consider the c- and f-structure in (17).

(17) *Det skjer når folket vil* (cf. (1))





The key difference between the structures in (16) and (17) is that only the latter takes the PRED value of the N head of the relative clause from the lexical entry of *når*. Here, the operator CAT given in the lexical entry of *når* non-trivially dictates that an element of category N is available in the c-structure, thereby preserving the integrity of the phrase structure rules in (14) (see Asudeh 2002 where CAT accomplishes non-branching CP-over-IP). These phrase structure rules and the arguments presented here will also be relevant in the next section where *da*-clauses are addressed.

3. *Da*

Some uses of *da* and *når* differ transparently. For example, of the two, only *når* functions as an interrogative word, and only *da* can be used alone as a temporal adverb (see (18)-(19), respectively).

(18) a. Når skal vi gå?
when shall we go
b. *Da skal vi gå?

(19) a. Da var det lettere.
then was it easier
b. *Når var det lettere.

Nevertheless, because both *da* and *når* introduce temporal clauses, the extent to which they differ from each other as temporal connectives warrants clarification. In the following, I contrast the properties of *da* and *når* and argue that *da* is a complementizer that introduces relative clauses when a lexical head is present and non-relative clauses when a lexical head is absent.

3.1 Properties of da

As illustrated in section 1, and exemplified again in (20) and (21), two types of *da*-clauses are relevant to the present discussion.

(20) Forth var sammen med en kollega [da ugjerningen fant sted]
F. was together with a colleague [DA misdeed-DEF found place]
'Forth was with a colleague when the misdeed took place.'

(21) 18. mai 1993 var dagen [da danskene (…) sa ja til EU]
May was day-DEF [DA Danish-DEF.PL said yes to EU]
'May 18, 1993 was the day when the Danish said yes to the EU.'

Faarlund *et al.* (1997) recommend an analysis that considers *da* in both of the clause types above to be the head of a relative clause. Support for this view is purportedly provided by examples like (22) and (23) where *da* immediately precedes a *når*-clause. As (24) shows, *når* cannot precede *da* in this manner.

(22) Da, [når vi dveler ved dem], får vi anledning til[6] (…)
DA [NÅR we dwell upon that], receive we chance to
'When we dwell upon that, we get a chance to (…).'

(23) Da [når ho kom heim], var alt i orden (Faarlund, *et al.* 1997, p. 10)
DA [NÅR she came home], was everything in order
'When she came home, everything was in order.'

(24) *Når, da vi dveler ved dem, får vi anledning til (…)

Faarlund *et al.* suggest that *da* in all of its uses as a temporal connective is an adverb that heads a relative clause. The data in (25)-(26) prove this analysis to be problematic, however. Unlike *når*-clauses, *da*-clauses allow unbounded dependencies only in bound relatives (cf. (25) vs. (26)) (Helge Lødrup, p.c.).

(25) Napoleon var faktisk på Korsika på den tiden [da du påstår at
N. was actually on Corsica at the time [DA you claim that
han ledet hæren i Italia_]
he lead army-DEF in Italy]
'N. was actually in Corsica at the time when you claim that he led the army into Italy.'

(26) *Napoleon var faktisk på Korsika [da du påstår at han ledet hæren i Italia_]
'N. was actually in Corsica when you claim that he led the army into Italy.'

If *da* were the head of the relative clause, one would expect sentences like (26) to be perfectly grammatical (cf. (10)-(11)). Hence, while I agree with Faarlund *et al.* that examples like (22) and (23) involve adverbial uses of *da*, I maintain that a different analysis is needed for subordinate clauses introduced by *da*. Specifically, subordinate *da*-clauses with lexical heads are relative clauses, while those without modify the matrix clause as a whole and are, therefore, non-relative clauses. Accordingly, I treat the former as adjunct clauses that adjoin to an N and the latter as adjunct clauses that adjoin directly to CP.

At this point, it remains for us to determine whether *da* is best understood as a complementizer or an *hv*-like element in SpecCP. Unlike *når*, *da* does not provide independent motivation for analysis as an *hv*-word. For example, it cannot be employed in question formation as illustrated in (18b). Furthermore,

[6]http://66.102.7.104/search?q=cache:gH5kiSe_OjEJ:www.ethikon.no/files/documents/altoppslukende_fokus.doc+%22da+n%C3%A5r+vi%22&hl=no

though *når* can be followed by the complementizer *at* in some cases, this option is not open to *da* (27), nor does *da* allow expansion (28):

(27) *Jeg fartet mye rundt i USA [da at jeg var korrespondent] (cf. (12)-(13))
(28) *Jeg fartet mye rundt i USA [da som helst jeg var korrespondent] (cf. (7))

Finally, since headed relative clauses with *da* allow unbounded dependencies while the other *da*-clauses do not, a free-relative-clause analysis of the latter is untenable. Thus, I treat *da* as a different type of complementizer in relative clauses and in non-relative clauses.

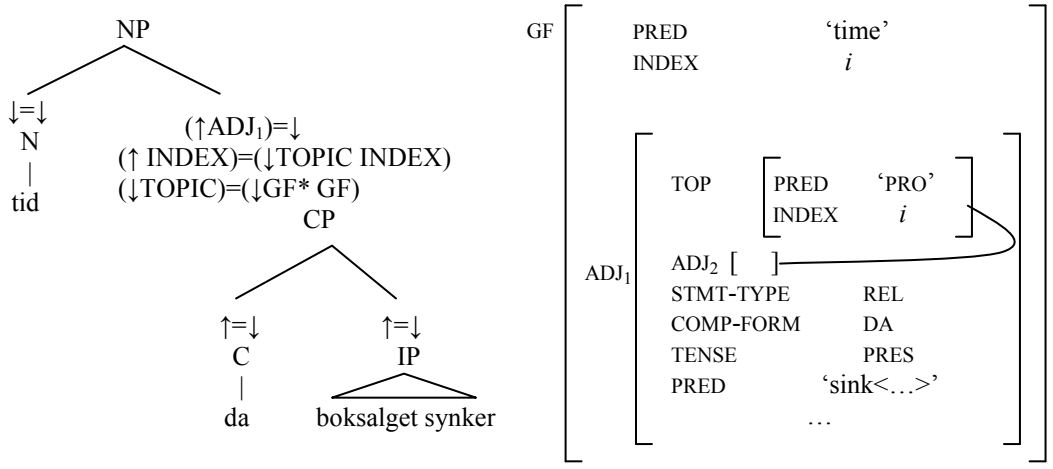
3.2 Analysis of da

Keeping in mind the phrase structure rules and analysis presented above, I propose the following lexical entry for *da* as it appears in relative clauses:

(29) *da*-REL C (↑ STMT-TYPE) = REL
(↑ COMP-FORM) = DA
(↑ TENSE)
(↑ ADJ)
(↑ TOPIC)
(↑ TOPIC PRED) = 'PRO'

This entry will interact with the proposed phrase structure rules to derive the c- and f-structures in (30). Note that (30) closely parallels the structures for *når* clauses in section 2.2. The main difference is that the information about the clause's TOPIC, ADJUNCT, tense and clause-type is housed in C, rather than SpecCP. The indexing between the head and the TOPIC of the relative clause is preserved by the annotation on the CP and the specification in *da*'s lexical entry that there is a TOPIC in the f-structure with the PRED value of 'PRO'. Unlike the case with *når*, this PRED value is not optional. As a result, the analysis makes the correct prediction that no overt element can appear in SpecCP. If there were an element competing for this position and thereby supplying an additional PRED value, the f-structure would violate the uniqueness principle since PRED values never unify. As with the *når*-clauses the functional uncertainty equation annotated on CP licenses the relative clause's unbounded dependencies. In turn, *da*'s lexical entry provides a GF, the ADJUNCT, for the TOPIC to associate with in fulfillment of the coherence principle.

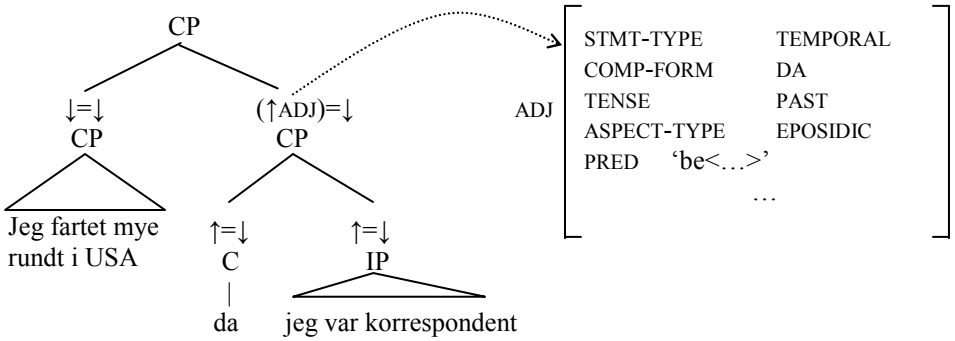
(30) *I en tid da boksalget synker* (…) (cf. (4))



The final structure to be presented here is that of *da* in non-relative clauses. As noted above, this type of *da*-clause does not contain a gap and should be regarded as a sister to the main clause. As such, the functional uncertainty equation employed above should not be involved in non-relative *da*-clauses. In the absence of any motivation for a TOPIC position, the *da* in non-relative clauses will specify that its f-structure does not contain a TOPIC. This ensures that non-relative *da*-clauses cannot appear in the position of relative clauses, as this would render the annotations on the relative CP that pertain to the TOPIC unsatisfiable. For *da* in non-relative clauses, I propose the lexical entry in (31) and the c- and f-structures in (32).

(31) *da* C (↑ STMT-TYPE) = TEMPORAL
(↑ COMP-FORM) = DA
(↑ TENSE)
¬(↑TOPIC)

(32) *Jeg fartet mye rundt i USA da jeg var korrespondent* (cf. (2))



The TEMPORAL clause-type specification in *da*'s lexical entry marks that the f-structure in question is a (non-relative) temporal ADJUNCT, and the f-structure is complete and coherent without the structure sharing assigned to the other clause types. In essence, unlike the *da* that introduces relative clauses, the *da* illustrated in (32) introduces clauses that adjoin to IP and needs a different lexical entry to account for the lack of unbounded dependencies and the absence of a TOPIC.

4. Aspectual Properties

Data compiled from the *Oslo Corpus of Tagged Norwegian Texts* (the *bokmål* section, 18.5 million words) reveal that the actual aspectual properties of Norwegian *when*-clauses support the distinctions drawn above. In particular, examples were found of all aspectual types for relative *when*-clauses (all *når*-clauses and a subset of the *da*-clauses) but not for non-relative *when*-clauses (the remaining *da*-clauses). Whereas the normative tradition holds that *når* appears in all but past episodic clauses, the example given in (33) illustrates that *når*-clauses can also be used for past episodic events.

(33) Heidi Tjugum var der hun skulle [når danskene skjøt] (*past episodic*)
H. T. was where she should [NÅR Dane-DEF.PL shot]
'Heidi Tjugum was where she should have been when the Danes shot.'

Similarly, the clauses in (34)-(36) show that relative uses of *da* are not limited to past episodic events.

(34) I de få perioder [da partiet har satt (…) parti fremst], har (…) (*past habitual*)
In the few periods [DA party-DEF has set party foremost], has
'During the few periods when the party has put party first, has (…)'

(35) Slik kan det gå i disse tider [da røyking er en kardinalsynd] (*present habitual*)
So can it go in these times [DA smoking is a cardinal-sin]
'So can it go in these times when smoking is a cardinal sin.'

(36) Lovprisningen fortsetter til denne dag [da hans etterkommere (…) sitter
Praise continues until the day [DA his descendant-DEF sit
på Thailands krone [sic., Intended: trone]] (*future*)
on Thailand's throne]
'Praise will continue until the day when his descendants sit on Thailand's throne.'

Despite the flexibility of *da* used in relative clauses, *da* in non-relative clauses appears to conform to the normative tradition which holds that *da*-clauses only describe past episodic events (see Vikner 2004 who makes the same observation for Danish *da*-clauses). This generalization evidently holds regardless of tense:

(37) Berit (…) *fikk* sin første bunad [da hun var fire år gammel] (*simple past*)
B. receive-PRET her first bunad [DA she was five years old]
'Berit (…) got her first national costume when she was five years old.'

(38) [Da det verste spetakkelet *hadde git* seg] hørte vi (…) (*perfect*)
[DA the worst noise had given-PAST.PART self] heard we
'When the worst noise was over, we heard (…).'

(39) Det er stille i bygningen [da Jeremy *låser* seg inn] (*historic present*)
It is quiet in town-DEF [DA Jeremy lock-PRES self in]
'It's quiet in the town when Jeremy locks himself in.'

In essence, for some speakers (though not all) the aspectual features of temporal *when*-clauses are impendent of whether *da* and *når* is used and are contingent instead upon the status of the clause as relative or non-relative.

5. Conclusion

The aspectual distinctions championed by normative grammarians fail to appropriately capture the differences between temporal *da*- and *når*-clauses. The connective *når* is best analyzed as a topic element that introduces either a bound or a free relative clause, appropriately accommodating unbounded dependencies in either case. Temporal *da*-clauses, on the other hand, are not all relative clauses. Instead, only *da*-clauses with lexical heads should be assigned a relative clause structure. Thus, even though *da* always appears in C, it requires two lexical entries: one for relative clauses that licenses unbounded dependencies and one for clauses that are sister to the matrix clause that prohibits unbounded dependencies. As shown in section 4, the aspectual features of these clauses are consonant with the proposed divisions. The only *when*-clauses in the corpus that exhibited clear aspectual restrictions were the non-relative *da*-clauses. The relative *da*-clauses patterned like *når*-clauses in permitting all aspectual types. Ultimately, I have shown that temporal *da*- and *når*-clauses must be evaluated on the basis of their syntactic structure, not merely on their lexical form and aspectual features.

Acknowledgements: I am especially grateful to Peter Sells and Helge Lødrup for their generous help and guidance. Thanks also to Joan Bresnan, Bruno Estigarribia**,** Florian Jaeger, and Ivan Sag for valuable comments and to Helge Lødrup, Janne Bondi Johannessen and Jakob Snilsberg for help with the data. Unless otherwise specified, all examples presented here were taken from the *Oslo Corpus of Tagged Norwegian Texts* (http://www.tekstlab.uio.no/norsk/bokmaal/english.html).

References

Andersson, Lars-Gunnar. 1975. *Form and Function of Subordinate Clauses*. Gothenburg University, Dept. of Linguistics.

Asudeh, Ash. 2002. "The syntax of preverbal particles and adjunction in Irish," in M. Butt and T. H. King (eds.), *Proceedings of the LFG02 Conference*. Stanford, California, CSLI Publications.

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

Bresnan, Joan and Jane Grimshaw. 1978. The syntax of free relatives in English. *Linguistic Inquiry* 9, 331-391.

Dalrymple, Mary. 2001. *Lexical-Functional Grammar: Syntax and Semantics* 34. London, Academic Press.

Faarlund, Jan, Svein Lie and Kjell Vannebo. 1997. *Norsk Referanse-Grammatikk*. Oslo: Universitetsforlaged Oslo.

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

Groos, Anneke and Henk van Riemsdijk. 1981. Matching effects in free relatives: A parameter of Core Grammar. In A. Belletti, L. Brandi, and L. Rizzi (eds.), *Theory of Markedness in Generative Grammar*. Scuola Normale Superiore, Pisa, 171-216.

Grosu, Alexander. 2003. A unified theory of 'standard' and 'transparent' free relatives. *Natural Language and Linguistic Theory* 21, 247-331.

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

Vikner, Carl. 2004. The semantics of Scandinavian 'when'-clauses. *Nordic Journal of Linguistics* 27, 133-167.

Wechsler, Stephen. 1991. Verb second and illocutionary force. In K. Leffel and D. Bouchard (eds.). *Views on Phrase Structure*. Dordrecht, Kluwer, 177-191.

ESTONIAN TRANSITIVE VERBS AND OBJECT CASE

Anne Tamm
University of Florence

Proceedings of the LFG06 Conference
Universität Konstanz
Miriam Butt and Tracy Holloway King (Editors)

2006
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

This article discusses the nature of Estonian aspect and case, proposing an analysis of Estonian verbal aspect, aspectual case, and clausal aspect. The focus is on the interaction of transitive telic verbs (*write, win*) and aspectual case at the level of the functional structure. The main discussion concerns the relationships between aspect and the object case alternation. The data set comprises Estonian transitive verbs with variable and invariant aspect and shows that clausal aspect ultimately depends on the object case. The objects of Estonian transitive verbs in active affirmative indicative clauses are marked with the partitive or the total case; the latter is also known as the accusative and the morphological genitive or nominative. The article presents a unification-based approach in LFG: the aspectual features of verbs and case are unified in the functional structure. The lexical entries for transitive verbs are provided with valued or unvalued aspectual features in the lexicon. If the verb fully determines sentential aspect, then the aspectual feature is valued in the functional specifications of the lexical entry of the verb; this is realized in the form of defining equations. If the aspect of the verb is variable, the entry's functional specifications have the form of existential constraints. As sentential aspect is fully determined by the total case, the functional specifications of the lexical entry of the total case are in the form of defining equations. The general well-formedness conditions on functional structures secure the sensitivity of aspectual case to verb classification.

1. Introduction*

Understanding or representing the Estonian aspect in any functional or formal grammar is complicated, since the application of previously used terminology does not apply smoothly to the phenomena. The main problem is that aspectual terminology was developed to explain and represent verbal aspectual phenomena (e.g. Dahl 1985, Comrie 1976, the Slavic, or Hungarian tradition, Kiefer s.a.) or compositional and quantification-related phenomena (Krifka 1992, Tenny1994, Verkuyl 1993). The article discusses data where neither verbal aspectual nor compositional approaches lend themselves easily to a straightforward account of the Estonian aspectual object case alternation. Verbal aspect does not determine the aspect of the clause as in the Slavic languages, nevertheless, transitive clauses must be specified for aspect as Slavic clauses are; the quantification of the inner argument does not have the aspectual effect familiar from English or Dutch; however, Estonian clausal aspect is not entirely determined by the verb. Rather, the alternation of the partitive (1) and total (accusative, morphologically genitive or nominative) (2) object cases corresponds to aspectual oppositions.

(1) *Mari kirjutas raamatut.*
M.nom write.3.sg.past book.part
'Mari was writing a/the book.'
(2) *Mari kirjutas raamatu.*
M.nom write.3.sg.past book.tot
'Mari wrote a book.'

Typologcal works do not consider Estonian as a language with fully grammaticalized aspect (Metslang (2001:444), Metslang and Tommola (1995: 300-301)). Grammatical aspect in Estonian has not developed into a consistent grammatical category, but it emerges in the object case alternation as illustrated above. The article proposes a way to understand and represent the object case alternation for some classes of transitive verbs: the telic verb classes of the types *write* (an accomplishment verb) and *win* (an achievement verb). As opposed to the type of telic verbs such as *write*, telic achievement verbs of the *win* type do not have aspectual case alternation (3). Although achievement verbs are traditionally considered telic, the partitive object case marking reveals that the telicity type at hand could be different from the typical cases of telicity, being perhaps reduced.

(3) *Itaalia võitis Saksamaad jalgpallis 2:0.*
Italy won Germany.part in football 2:0
'Italy won Germany in football with the result 2:0.'

* I acknowledge the support of OTKA K 60595, and I thank the participants of the LFG06 Conference for many valuable discussions. I am grateful for the comments of the reviewer, which have encouraged me to address many unclear issues; the remaining errors are mine.

The verb classification tries to accommodate the systematic compatibility of verb classes with aspectual object case marking patterns. The questions are presented as follows: Section 2 discusses the aspectual object cases and their place in the Estonian case system; Section 3 shows that the object case alternation does not reflect oppositions in specificity, definiteness or quantification. Section 4 opens the discussion of the aspectual hypotheses with a review of the phenomena from the outer aspect (perfectivity-imperfectivity) perspective. While the correlations are not exact, the alternation in aspectual case marking is in core cases part of other general two-way case marking strategies employed in Estonian grammar. Section 5 addresses the inner aspect relatedness of the object case alternation. Section 6 views the relatedness of objects and aspect in comparison to other lexicalist approaches and points out that the Estonian object case phenomena cannot be accounted for by, for instance, thematic role based approaches. Section 7 words a proposal to improve some problems of earlier approaches and sketches a possible analysis in LFG, and Section 8 is a conclusion.

2. Aspectual object cases and their place in the case system

Aspectual meanings have developed as part of a grammatical marking system where nominal case marking is a general strategy of encoding grammatical, lexical, semantic, and pragmatic meanings. More specifically, in Estonian, grammatical functions are distinguished by case marking; also mood categories, such as imperative, influence the case marking of the core grammatical functions; thematic roles determine the case marking of some arguments; the semantic properties of the NP determine the case marking; and pragmatic information determines partly the case marking of plural count and mass singular subject NPs. The case system of Estonian comprises 14 cases (Erelt et al 1997) (4).

(4)

Nominative	book	raamat
Genitive	of a book	raamatu
Partitive	(of) a book	raamatut
Illative	into the book	raamatusse
Inessive	in a book	raamatus
Elative	from (inside) a book	raamatust
Allative	onto a book	raamatule
Adessive	on a book	raamatul
Ablative	from the book	raamatult
Translative	in(to), as a book	raamatuks
Terminative	until a book	raamatuni
Essive as a book	as a book	raamatuna
Abessive	without a book	raamatuta
Comitative	with a book	raamatuga

Case is a strategy of marking nominal entities, including participles and infinitives. Case marking is widely applied historically and in the present-day Estonian as a strategy of differentiating oppositions in several, mainly mood, categories. The verb form and the argument NP case encoding differ from the verb's and its argument's morphological form in active affirmative indicative sentences. For instance, total objects in imperative and indicative sentences are differentiated by nominative and total case marking respectively. The case alternations appear between other cases than in active affirmative indicative sentences; in the impersonal category, the alternation is between nominative and partitive, whereas in the imperative, the object is either nominative or partitive. In the evidential mood, the personal present participle is marked with partitive, and in negation, the object and subject may be partitive (see Nemvalts (1996), Tamm (2004a) and to appear for further details). Explaining case phenomena, reference to selection strategies based on case or grammatical function hierarchies is a widespread practice in works dealing with Finnic (Maling 1993, Ackerman and Moore 2001). For instance, there is a tendency of the nominative case marking of the highest GF of the sentence (Maling 1993). In sum, case is a much-employed grammatical device in Estonian as in the Finno-Ugric languages in general.

One of the case alternations that are related to sentential semantics and function is the aspectual object case marking. Differently from the case opposition phenomena in the mood categories, the range of possibilities depends on the aspectual lexical semantics of the verbs. Object NPs can be marked either with the morphological *partitive* or with the *total case* (also referred to as the accusative case). The latter is morphologically realized as the genitive, if singular, or the nominative, if plural, a numeral, a certain quantizer and also in some infinitival constructions, imperatives, and impersonal sentences. A note on current terminological debates is in order. Firstly, there are proposals to change the traditional term "total object" to "accusative object" (Pusztay 1994, Hiietam 2003). "Total object" (*totaalobjekt, täissihitis*) is the most frequently used term in Estonian linguistics for the NP in object function that is case-marked with the morphological genitive or nominative. In international sources, the total case is frequently referred to as "accusative", since it is one of the object cases, and establishing an analogy with Finnish, where personal pronouns have a separate morphological accusative. However, the total-accusative case cannot be considered as "the" object case since many objects are marked with other cases as mentioned above. In addition, the total case marks measure adverbials on partly similar semantic grounds with object marking. In semantics and functional linguistics oriented writings, "total" is a term that metaphorically conveys the "totally" bounded or finished nature of the transitive clause.

Secondly, the term "partitive" covers a variety of concepts in linguistics. Partitive is used as the traditional name for a morphological case, also, as the name of the inherent Case in GB theory associated with indefiniteness, and as a semantic notion associated with partial interpretation. Relating the object case alternation primarily to aspect and not to configurational positions, this article defines it as semantic alternation and the cases involved as aspectual semantic cases. Semantic case is here understood as in Butt and King (2005) (cf. also Butt 2006), that is, a type of case about which regular semantic generalizations can be made and that has the following characteristics: it is predictable via the formulation of generalizations across predicates and constructions (here, aspectual generalizations) and subject to syntactic restrictions, such as restrictions on grammatical functions of the NPs where the case can appear (appearing, in this case, only on measure adverbials, subjects, or objects). This article regards the Estonian partitive as a morphological case that changes the semantic interpretation of a predicate so that it differs from the semantic interpretation of the predicate in a sentence with the total (accusative) object.

As opposed to the term "total", the term "partial", used in Estonian grammars, does not transparently cover the semantic content of the partitive case. The partitive marked NP's denotation generally cannot be understood as "part of" the denotation of the object NP's referent. Events described in clauses with "partial" objects do not necessarily reflect any "partial" progress of the event either (see examples (5) and (9)). Frequently, the partitive marked object noun phrase has no referent. Therefore, nothing related to parts can serve as a cover term for the "partial" object phenomena, and the morphological form related term "partitive" is preferred to the semantically (wrongly) suggestive "partial".

Not only objects but also subjects and measure adverbials have "split" case-marking. For subjects, the alternation is between the morphological partitive versus nominative; the measure adverbials have a three-way split into the total case and the morphological partitive versus nominative. The general pattern of aspectually relevant object, adverbial, and subject case-marking in Estonian is presented in Table 1. After this overview, the article confines the discussion to objects of affirmative indicative active sentences that are singular count nouns.

Table 1. Subject, object, and durative adverbial case-marking in Estonian

Subject cases	Object cases	Adverbial cases
Nominative *Partitive* (alternation in plural count and singular mass nouns only (in the "existential sentences") and in existence negation)	*Genitive* (total) (singular) *Nominative* (total) (plural, numerals, etc) *Partitive* (singular and plural, in most negative sentences)	*Genitive* (total) (singular) *Nominative* (total) (plural, singular for numerals and some quantizers, etc) *Partitive* (plural, and singular of mass NPs and in some data on negation)

The Estonian aspectual case marking is a result of combining a grammar system imbued with case-marking practices of oppositions with an aspectual verb classification that is fairly comparable to better-studied languages. The following section, however, considers first the arguments for not choosing for an NP related account of the object case alternation.

3. Case alternation and NP relatedness

This section shows that the object case alternation of singular count noun objects of affirmative indicative active sentences reflects oppositions that are not primarily oppositions of specificity, definiteness or quantification.

In terms of a possible "definiteness hypothesis", the total case marks definite NPs, and the partitive marks indefinite NPs. The strongest arguments for the accusative-definiteness link in support of this hypothesis can be found in Hiietam (2003). Other Estonian works that discuss the NP properties of the object case (Rajandi and Metslang 1979) and earlier relevant works on the NP-related case in Finnish, such as de Hoop (1993) or Belletti (1988), include instances of plural and mass partitive NPs, which are rather an exception than a rule (cf. Kiparsky 2001, 1998).

The problem is that the "definiteness hypothesis" covers some frequent but not all instances of the total case phenomenon and does not cover adequately the partitive case either. The so-called "partitive verbs" in Estonian grammars and previous works on verbs and object case (Erelt et al 1993, Kont 1963, Tauli 1968, 1983, Metslang 2001, Klaas 1996, 1999) are not verbs that have indefinite objects (5).

(5) *Jaan usub/usaldab/näeb/laseb presidenti.*
J.nom believe/trust/see/shoot.3.sg president.part
'John believes/trusts/sees/shoots at the president.'

The object NP ("president") is marked with the partitive, and the partitive marked object, "president," is not indefinite. Further, there is a large class of verbs, creation verbs, in sentences with typically indefinite object NP referents that nevertheless occur with total case-marking. Some verbs regularly allow for variation in their aspectual behavior (see the tests in Section 5), demonstrated in the following examples with the verb *kirjutama* 'write' (1) and (2), repeated here as (6) and (7); however, relating the alternation and the total case marking to definiteness is, considering the creation verb, which typically brings new referents to discourse, not plausible.

(6) *Mari kirjutas raamatut.*
M.nom write.3.sg.past book.part
'Mari was writing a/the book.'
(7) *Mari kirjutas raamatu.*
M.nom write.3.sg.past book.tot
'Mari wrote a book.'

Moreover, there are no verbs that would give rise to regular minimal pairs on the basis of opposite object case alternation that would confirm the definiteness hypothesis. In some cases, the opposite can be true. For instance, the object case alternation with the verb *leidma* 'find' in (8) and (9) provides negative evidence for the "definiteness hypothesis". The data bear resemblance to the Finnish and Scottish Gaelic data (Kiparsky 1998, Ramchand 1997). In a sentence with the total object, such as *leidsin võtme* 'I found a key' the total object NP "key" is new in the discourse, in the sentence *leidsin võtit korduvalt* 'I found the key several times' the partitive marked NP with "key" is definite (or specific).

(8) *Leidsin võtme.*
Find.past.1.sg key.tot
'I found a/some key.'
(9) *Leidsin võtit korduvalt.*
Find.past.1.sg key.part repeatedly
'I found the key several times.'

Therefore, the case alternation cannot be related to the alternation of indefinite-definite features of the respective NPs; the definiteness or indefiniteness and case cannot be related via the discourse requirements of the verbs either. The definiteness hypothesis, which assumes a link between the feature of definiteness and the total object case or a link between indefiniteness and the partitive object case, does not find sufficient support from the data.

Neither can the quantification of the object be associated with the case alternation. Since the quantification of the object in the sentences (6) and (7) or (8) and (9) remains constant while the case alternates, also the quantification of the NP and case are not related.

4. Introducing the aspectual hypotheses and outer aspect

This section opens the discussion of the aspectual hypotheses with a review of the phenomena from the outer aspect (perfectivity versus imperfectivity) perspective and shows that while the correlations are not exact, the alternation in aspectual case marking is in core cases part of other general two-way case marking strategies employed in Estonian grammar.

The existence of clearly aspectual (event structural) verb classes and their typical occurrence with either total or partitive case-marking suggests that even if there is a tendency of total-definite and partitive-indefinite correlation, an aspect-related hypothesis plausibly covers more data than the definiteness hypothesis and related NP based hypotheses. An aspectual hypothesis is also the hypothesis that has found more followers in discussions of Finnish aspectual case alternation (e.g., Vainikka and Maling 1996, Nelson 1998, 2003, Kiparsky 1998, 2001). Frequent discussion around relating Estonian aspectual phenomena to the Russian aspectual terminology (Metslang and Tommola 1995, Rätsep 1957, Pihlak 1982) shows that Estonian aspectual phenomena are at least in some respects comparable to the Slavic ones. Also, for instance, in an early generative Estonian grammar by Harms (1962:131) discusses under "'Aspectual' Partitive Object" examples that are clearly similar to the Russian secondary imperfective, such as *ma võtan raamatut ära* 'I am taking the book away' with the completive particle *ära* (approximately meaning 'up, away, done, completed') and the partitive object.

Aspect is a phenomenon that is discussed at different levels of description and phenomena (Verkuyl 1993, Smith 1991). This section discusses the "outer aspect", "perfectivity", or grammatical aspectual hypothesis. The hypothesis may be worded as follows: case alternation and the viewpoint aspect are related. More specifically, the partitive case marking corresponds to the imperfective viewpoint aspect and the total case marking reflects the perfective viewpoint aspect. The following examples demonstrate a test that indicates whether the aspect is inside or outside the described event by checking the possibility of temporal overlap or sequencing. If two sentences receive an interpretation of temporal sequencing, the first sentence is perfective. According to this test of perfectivity, the sentence with the total object (10) is a perfective sentence, since starting work at the university is interpreted by having started temporally after the event of writing her book. Inserting "…and then …" between the sentences is possible and felicitous in case of the perfective sentence.

(10) *Mari* *kirjutas* *raamatu. Ta läks ülikooli tööle.*
M.nom write.3.sg.past book.tot she went to work at the university
'Mari wrote a book. She went to work at the university.'

On the contrary, sentences that are imperfective do not allow the temporal sequencing. According to this test of perfectivity, the sentence with the total object (11) is an imperfective sentence, since starting work at the university is not interpreted by having started temporally after the event of writing her book; instead, the events have a temporal overlap. Inserting "…and then …" between the sentences is anomalous.

(11) *Mari* *kirjutas* *raamatut. Ta läks ülikooli tööle.*
M.nom write.3.sg.past book.part She went to work at the university
'Mari was writing a/the book. She went to work at the university.'

Evidence from (10) and (11) suggests that perfectivity determines the total case marking of the object NPs and imperfectivity determines the partitive marking. However, there are instances that can be considered as counterevidence: partitive plural NPs (not discussed here), several event verbs, such as some psych-verbs, some inchoative verbs and degree achievement verbs. The temporal sequencing test shows that, despite the partitive case marking of the objects, is sentences with several event verbs have perfective aspect, such as the verbs *alustama, algama* 'start, begin', *solvama* 'offend', *võitma* 'win', *rikkuma* 'ruin', *ehmatama* 'frighten' (12).

(12) *Mari* *ehmatas* *Jürit/#Jüri.* *Kuuldus karje.*
M.nom frighten.3.sg.past George.part/#gen was-heard a scream
'Mari frightened George.' (OK: And then a scream could be heard.)

The test shows that perfectivity appears with degree achievement verbs, with partitive object marking (13).

(13) *Firma* *laiendas* *(kahe tunniga)* *teed.* *Algas töö.*
Firm.nom widen.3.sg.past in two hours road.part The work started.
'The firm widened the road (in two hours).' (in the sense of somewhat, to some extent)
(OK: And then, the work started.

Even if the effect of the test is weaker than with total objects, in conclusion, evidence from partitive objects in perfective sentences suggests that the object case alternation cannot be related to the perfectivity-imperfectivity features despite strong correlations. Therefore, another sentential aspectual hypothesis, the "resultativity hypothesis", is reviewed next. The resultativity hypothesis may be worded as follows: Case marking of objects reflects whether or not the sentence describes a result; partitive case marking corresponds to the irresultative aspectual interpretation and the total case marking corresponds to the resultative aspectual interpretation. A suitable test contains reference to a result state that does not change. However, the partitive case marking does not correspond to the irresultative aspectual interpretation as sentence with a partitive object (14) describes a clear result or outcome, specified in the sentence (as the result of the game, 2:0).

(14) *Itaalia* *võitis* *Saksamaad* *jalgpallis 2:0.*
Italy won Germany.part in football 2:0
'Italy won Germany in football with the result 2:0.'

In sum, case alternation cannot be related to the resultativity features and the notion of the result state. Therefore, the third possible aspect-related hypothesis, the "boundedness hypothesis", is reviewed next. The boundedness hypothesis may be worded as follows: Case marking of objects reflects whether or not the sentence contains linguistic means to refer to a boundary, boundedness in a wider sense, either aspectual or NP-related; partitive case marking corresponds to the non-bounded interpretation and the total case marking corresponds to the bounded interpretation. However, phrases that bound the situation do not appear only in sentences with total objects. For instance, sentence (15) has a partitive object and a terminative phrase that serves as the bounder of the situation.

(15) *Mari* *saatis* *lauljat* *ukseni.*
M.nom accompany.3.sg.past singer.part door.termin
'Mari accompanied the singer until the door, Mari saw the singer to the door.'
(16) *Mari* *solvas* *Toomast* *südamepõhjani.*
M.nom insult.3.sg.past Thomas.part bottom- of-the-heart.terminative
'Mary insulted Thomas deeply (to the bottom of his heart).'

Since partitive objects may be in sentences with situation bounders, case alternation cannot be related to a simple idea of boundedness of the situation. The outer aspect hypothesis must combine with a more lexicon related hypothesis in order to explain the data. Therefore, the following section 5 narrows the aspectual domain to inner aspect and returns to issues of outer aspect in the section that follows Section 5.

5. Inner aspect and case

This section addresses the inner aspect relatedness of the object case alternation. As previously mentioned, in the Estonian examples with case alternation, the quantification of the object NP does not contribute a significant feature in the aspectual composition of the sentence, as it is the case with English. The results of the aspectual tests with durative and time frame adverbials demonstrate that the aspectual effect of the Estonian partitive on the singular quantized NP object is comparable to the aspectual effect of the English plural (19). This rather points to the relation between the total case and the terms delimitedness (Tenny 1994), telicity (Krifka 1992), plus terminativity (Verkuyl 1993) or boundedness (Kiparsky 1998) and the relation between total case and the terms non-delimitedness, atelicity, minus terminativity or non-boundedness.

(17) Mary wrote the book in a year.

(18) *Mari* *kirjutas* *raamatu* *ühe* *aastaga.*
M.nom write.3.sg.past book.tot one.gen year.comitative
'Mari wrote a/the book in a year.'

(19) Mary wrote books for years.

(20) *Mari* *kirjutas* *raamatut* *terve* *aasta.*
M.nom write.3.sg.past book.part whole.tot year.tot
'Mari was writing a/the book for whole a year.'

While sentence (20) with a partitive object is atelic, sentence (18) with the total object is telic according to the results of standard telicity tests. Therefore, the "telicity hypothesis" is checked next. The inner aspectual "telicity" hypotheses may be divided in two in this section, the endpoint-related and the quantization-related ones. First the endpoint-related hypothesis is reviewed: predicates with an endpoint have a total object; predicates without an endpoint have a partitive object. However, many of the examples in section 4 show that the presence of a vague notion of an endpoint, or a boundary, which can be associated with the properties of the predicates, complements, measure phrases or adjuncts does not correlate with the total object case marking. Therefore, the quantization-related telicity hypothesis is addressed; it can be divided in two: the plus-principle (following the line of thought in Verkuyl 1993, Kiparsky 1998) and incremental-theme (following Krifka 1992) ones.

The plus-principle hypothesis might be as follows: object case alternation depends on the value of the T feature; the +T predicates have total objects, -T predicates have partitive objects. Several accounts of aspect view a sentence's aspectual properties as being determined by more components in a sentence than the verb alone: for instance, direct objects and their corresponding NP's quantificational character. A closer look reveals some challenging contrasts between the Germanic and Finnic languages. Verkuyl's two main principles of modeling aspect, the so-called Plus Principle and the assumption of aspectual phenomena at two syntactic levels are not directly helpful for modeling Estonian case and aspect matters. First, the aspect of a verb-argument complex cannot be composed on the basis of the verb's (temporal) feature and the (atemporal) quantificational properties of the argument as envisaged by Verkuyl. The examples (1) and (2) above show that the expected variation in the aspectual value of the sentence is not paralleled by the difference in the object NP properties: the quantification of the object NP *raamatu(t)* 'book.tot/part' remains constant. Instead, it is the partitive and total case-marking that correlates with the aspectual oppositions in these examples. The issue of composition is more complicated since the prediction of most theories is that sentences with bare plural nouns are not quantized, they are unbounded. This prediction is not borne out, since sentence (21) can have a quantized, bounded, interpretation regardless of the bare plural (partitive-marked) object.

(21) *Mari* *kirjutas* *raamatuid.*
M.nom write.3.sg.past book.part
'Mari did some book-writing.'

Therefore, the reason for why the omission of an object leaves an aspectually underspecified sentence (e.g., *Kirjuta!* 'Write!'), which may be telic or atelic, is not the lack of the information about the quantization of the object, but the lack of evidence about the exact aspect, which would be obtained from object case.

The second telicity hypothesis is discussed next. It may be worded as follows: sentences denoting quantized events have total objects; sentences denoting cumulative events have partitive objects. The quantization telicity hypothesis is divided in two in literature. The more semantics based approaches formalize the idea that the quantized nature of the predicate is determined by the quantized nature of the incremental theme (Krifka 1992) and that leads to a possible formulation of the hypothesis: the total objects are objects of predicates that have quantized incremental themes. Kiparsky (1998) has pointed out that not only verbs with incremental themes have accusative objects in Finnish. To a lesser extent, the claim is true for Estonian. Taking a more general idea of a relation between event quantization and total case marking as the basis for the hypothesis, another problem occurs. Namely, Depraetere (1995) distinguishes two types of aspectual oppositions, those of boundedness and telicity, which are not distinguished in works following Krifka (1992). In sentence (22), the event can be classified as telic and quantized via the quantized path, one kilometre; however, the object of the sentence is partitive.

(22)	*Takso*	*sõidutas*	*Peetrit*	*ühe*	*kilometeetri.*
	Taxi.nom	drive.3.sg.past	Peeter.part	one.tot	kilometer.tot
	'The taxi drove Peeter for one kilometer.'				

The total case appears on the adverbial, or it appears on the adverbial and the object with a change in the aspectual meaning (see Tamm 2006 for a discussion). The data shows that the telicity or quantization of the predicate does not correlate with the total object case marking, since sentence (22) is telic, but the object is partitive.

In the account of Kiparsky (1998), the role of the quantification of the objects is diminished and the verb classification is more fine-grained. Kiparsky establishes a direct link between the VP-boundedness (the term is based on non-homogeneity) and object case alternation of singular count NP objects. While the interaction between verbal aspect and clausal aspect cannot be related to the quantificational properties of the object NP, the event quantification itself and the case are be related. However, the crucial differences between Estonian and Finnish, mainly in verb classification concerning the case marking of stative verbs, aspectual particles (Tamm 2004b), and predicate complexes, suggest that the account of Estonian data must be formulated differently (see Tamm 2004a). For instance, Estonian is different from Finnish, as it does not allow telic (accusative) in progressive constructions with telic verbs (23) while Finnish does (Sulkala 1996).

(23)	**Olen*	*pileti*	*ostmas.*
	be.1.sg	ticket.tot	buy.mas-infinitive
	Meaning 'I am buying a ticket.'		

Moreover, Estonian aspectual bounding (perfective) particle (Tamm 2004b) appears with the total object and the verb (24) (the data is from Metslang 2001). This particle diverges from the particles that have a strong argument structural link as, for instance, English particles, and have a different semantics and syntax. In this case, the particle combines with an atelic verb that appears only with partitive case (25).

(24)	*Ta*	*suudles*	*tüdruku*	***ära.***
	s/he	kiss.3sgpst	girl.tot	particle
	'S/he did the kissing of a girl.'			
(25)	*Ta*	*suudles*	*tüdrukut/*tüdruku.*	
	s/he	kiss.3sgpst	girl.part/*tot	
	'S/he kissed the girl.'			

The data with particles suggest that an account of the aspectual composition of Estonian aspectual system and object case marking needs to accommodate additional elements compared to Finnish (Kiparsky 1998). Aspectual verb classes do exist in Estonian according to tests that do not involve any case alternation (Tamm 2003a), while the correspondence of the telicity of the predicate and the total object case is not absolute.

6. Relatedness of objects and aspect in other lexicalist approaches

This section views the relatedness of objects and aspect in comparison to other lexicalist approaches and points out that the Estonian object case phenomena cannot be accounted for by, for instance, thematic role based approaches. The more syntactic approaches may allow the following formulation of their hypothesis on the Estonian object case: total object is an affected object, and the total case marks the event measurer, which must be an internal argument. Tenny (1994) claims that universal principles of mapping between the lexicon and syntactic argument structure are governed by aspectual properties. More specifically, Tenny posits a link between the presence of a direct object (direct internal argument) and the expression of certain aspectual properties such as "delimitedness" or "measuring out of events". At first sight, this claim seems to be confirmed by Estonian data: objects and aspect are clearly related. However, a closer look reveals that Tenny's widely accepted aspectual interface hypothesis is too strong. Also, many of her formulations about the relationships between direct internal arguments (in LFG, objects) "delimitedness", "measuring out", and "internal change or movement" are not clear in view of the Estonian phenomena. Firstly, there are examples without any direct internal argument that, contrary to expectations, are compatible with Tenny's criteria for delimitedness and measuring out (*tutvuma* 'get acquainted'). Secondly, the relations between delimitedness, object case, verbs, and particles present a wider array of data than Tenny's theory can capture. For instance, there are sentences with verbs with an experiencer, and an agent or theme argument. The theme, not the experiencer argument is realized as the (total) object, while the experiencer undergoes an internal change and should, therefore, provide the measure for the event. A couple of examples are *andestama* 'forgive' and *unustama* 'forget' (26).

(26) *Mari* *unustas* *oma* *sõbra.*
M.nom forget.3sg.past his/her friend.tot
'Mari forgot her friend.'

The total (accusative) objects may be non-measuring arguments that do not delimit the situation, as in (27) and (28) with verbs such as *andma* 'give' or *lükkama* 'push'.

(27) *Andsin* *Marile* *raamatu.*
Give.1.sg.past to Mari. book.tot
'I gave a book to Mary.'

(28) *Mari* *lükkas* *käru* *poodi.*
M.nom push.3.sg.past. cart.tot to the store
'Mary pushed a/the cart to the store.'

These examples are problematic for Tenny's account of Finnish, where the distribution of accusative and partitive case should reflect the presence and absence of aspectual roles of the NP and delimitedness (see Tamm 2003b for more data on this issue). The aspectual nature of the sentences above cannot be dependent on the presence of the measure role of the argument but rather on the aspectual nature of the verb. This is the insight that is captured by the modified lexicalist thematic role based proposal, developed in Ackerman and Moore (2001). The aspectual role of "boundedness" is not linked to an argument but is part of predicate entailments (thematic ("proto"-) role entailments) that are involved only in aspectual object case encoding. However, it is problematic to account for Estonian total case in terms of case assignment based on predicate properties and thematic roles, since roles as such can be assigned to argument NPs by verbs. The data show that aspectual case-marking concerns both arguments and adjuncts (adverbials) (22) and also depends on the presence of the progressive construction (23) or an aspectual particle in the sentence (see examples (24) – (25)) and

not on the proto-role grid of the verb (28). The object case is on the one hand dependent on the verbs class and on the other hand independent of them, triggering a type shift on the verbs.

In sum, on the basis of Sections 2 to 6, the partitive and total cases are not primarily for marking oppositions of NP properties: quantification or definiteness. The case alternation reflects aspectual oppositions. Clausal aspect is largely but not entirely determined by the aspectual nature of the verbs; the same can be claimed about the object case marking. Estonian has clear aspectual verb classes that correlate with (a) the typical object case that occurs with these verbs (see for the data in Tamm 2004a) and (b) the aspectual interpretation (e.g. iterative (9) or not (1)) that the verbs have with partititve objects. Most verbs can occur in aspectually opposite sentences, but the conditions of the aspect-based assignment of the alternative object cases clearly vary according to verb classification. Instead of proposing principles for verb classes and establishing their typical object case, and instead of departing from object cases and establishing their link with aspect, those elements or factors are studied in their interaction. Differently from earlier accounts, the interaction is not formulated in terms of thematic or aspectual roles but in terms of features that reflect better the overall case marking strategies of Estonian and, therefore, are more independent of the exact lexical properties of the verbs. Historically, the case alternation stems from the semantics of the partitive NP in both Finnish and Estonian. The features do not capture aspectual composition based on object NP quantificational features and verbal temporal features. Instead, differently from the predominantly VP-aspectual Finnish aspectual case, the Estonian object case is better seen as if either completing or changing verbal aspect, thus mixing in its function the inner and outer aspectual levels. The following section seeks a representation for the compatibility of verbs and case, for the aspectual verb classes, and the fact that the case determines ultimately the aspectual nature of the sentence.

7. Proposal

7.1. Goals and insights

This section discusses the levels of description of the Estonian aspectual phenomena and then proposes a way to formulate the description in LFG.

Two of the main levels of representing the Estonian grammatical aspectual phenomena are morphology, since object case is involved, and semantics, since aspectual interpretations of sentences and lexical aspect are involved. In contrast to previous lexicalist accounts, this paper proposes a solution where the syntactic level of the functional structure and functional features are part of the analysis. A comparison can be made with the morpho-semantic interface as envisaged in Ackerman and Moore (2001), who crucially involve an aspectual proto-role, associated with lexical items. The presence of an entailment of this proto-role determines the morpho-semantic selection of the accusative (total) case for the object NP. However, relating the morphological case of objects and the notion of semantic boundedness that is based on the definition of telicity as in Krifka (1992) fails in telic sentences with accusative measure adverbials and telic verbs with partitive object case (*win, frighten*). The objects in those sentences are predicted to select accusative (total), since they are telic; however, their object may be partitive. An interface mismatch that needs to be adjusted appears, since a predicate may be semantically telic with or without having total object case marking in Estonian. Therefore, the following sections attempt to modify the link between the predicate aspect and object case. The options for adjusting the problem are in the semantics of the predicates, the ways of composition and mapping, the representation for the case itself, or syntax—or combined.

This approach has opted for a combined solution. More specifically, the analysis involves feature unification at the syntactic level of functional structure, where verbs and the case morphemes contribute aspectual features to syntax. There is also independent motivation, discussed above in Section 2, to use functional features and functional structure. Since case-marking alternation is a general strategy in Estonian, signaling oppositions in, for instance, voice or mood, the representation of the relations between predicates and case may plausibly be part of the functional structure.

In this analysis of transitive verbs, therefore, the interface with aspectual semantics is drawn between the functional structure and the semantic structure as in standard LFG (Glasbey (2001), Butt, Dalrymple and Frank (1997)). An important part of the analysis is the case morphology, which contributes features to the functional descriptions of lexical items. The choice to constrain sentential aspect simultaneously from case and verbs is based on the intuition about the current state of art in the

grammaticalization and lexicalization of aspectual meanings in Estonian object case and verbs. The intuition concerns evaluations about whether an object type occurs with a verb naturally or feels as coercion. The choice to account for aspectual composition in the f-structure syntax and not in the lexicon, which would mean that verbs contribute fully specified features, is based on those intuitions and considerations. Independent evidence from the work of Nordlinger and Sadler (2004) shows that encoding TAM on dependents instead of heads is a wider spread phenomenon.

7.2. The possibilities of the Lexical Functional Grammar framework

Importantly for this account, the LFG framework allows locating pieces of aspectual information and information about grammatical relations in many (discontinuous) constituents that may appear in several configurations in surface constituent structure syntax. Simultaneously, it allows locating them at one place at the other syntactic level, the functional structure. This effect is achieved by means of constraints that pertain to relations between the levels of representation. The account relies on parts of several previous analyses and methods, basically Tamm (2004). I apply the analyses of Constructive case in LFG as in Nordlinger (1997) and Nordlinger and Sadler (2004), in King (1995) on Russian, in Butt and King (2005) on semantic case, and in Lee (1999) on Korean. The approach of Toivonen (2001) to the interaction between the Swedish aspectual particles and verbs is adopted here for modeling the interaction between verbs and case, which is the basis for possible later elaboration of the interface with semantics. The basic advantage of the LFG framework is that it allows locating pieces of aspectual information and information about grammatical relations in many (discontinuous) constituents that may appear in several configurations in surface constituent structure syntax and locating them at one place at the functional structure.

7.2. Boundability and boundedness

This section concentrates on the terminology and on how the aspectual information from lexical entries specifies structures of syntactic representation. The main observation that this paper wishes to capture is that lexical entries provide partial but basic information about clausal aspect at the f-structural level of syntactic description. I discuss telic verbs. The data shows that regardless of their inner structure, telic verbs are only potentially telic, appearing in sentences either as atelic or telic, depending on the object case. The same problem appears with relating perfectivity and case. The proposal is that the terms telicity and perfectivity be dropped since confusion that may rise from the intuition that there is an "inner", event structural "telicity" or "perfectivity-punctuality" and an "outer", a grammatical aspectual "telicity" or "perfectivity". Considering also the states where total objects relate to the maximal coverage or containment of bounded space (see Tamm 2004a), the term boundedness is used. Also, as the borderline between two levels of aspect is unclear and case seems to be a transition phenomenon from inner aspect marking to outer aspect marking. As many earlier Estonian accounts suggest treating Estonian aspect in terms of boundedness and verbal boundability, based on the intuition that transitive verbs are either boundable or not, I propose the terms bounded or boundable for describing the two types of telic verbs (*frighten/win* versus *write*) and the terms bounded or non-bounded for sentences.

The proposal is to represent the information about the aspectual boundedness also in the functional descriptions of the verb entry. Exactly as a Slavic verb must be specified for aspect, an Estonian transitive verb's object must encode the aspectual value in the sentence. Object case alternating telic verbs such as *kirjutama* 'write' do not specify their aspect themselves but specify only that the sentence where the verb occurs must have aspect. The representation of the presence of aspect is comparable to the presence of an object. Transitive verb entries contain a lexical constraint about an object but do not specify the exact content of the object. In a well-formed sentence, a transitive verb does not occur without an object, that is, without a value of the OBJ attribute; by the same token, a boundable transitive verb does not occur without aspect and aspectual case marking on its object.

Objects and the type of aspect that relates to the Estonian total case cannot be related more tightly, since the study on Tenny's aspectual interface hypothesis showed that boundability and transitivity are independent lexical requirements of a given verb. However, on the one hand, the existence of

boundability is still dependent on transitivity, since grammatical aspect emerges in Estonian clearly and unambiguously only with transitive verbs. On the other hand, as discussed above, boundability and transitivity are similar in constraining the conditions of the well-formedness of a sentence. In order to capture this parallel, the valueless boundedness feature is formalized as an existential constraint (B) exactly as the attribute (OBJ). The presence of the existential constraint in the functional specifications associated with the lexical entry means that the attribute must obtain a value in order to form a well-formed functional structure.

In case of the transitive telic verbs such as *write* verbs, the functional description consists of the boundedness attribute (B) that is either valueless (the correspondent of the possibility of being bounded). In case of the transitive telic verbs such as the *win* verbs, that is, the partitive-object telic verbs that are telic in their own right and cannot have total objects, the functional description consists of the boundedness attribute (B) that has the value (MIN) meaning minimally bounded. In this way, the lexical entries encode lexical boundability or lexical boundedness, respectively.

The characterization of telic verbs is follows in (29).

(29)
Boundable verbs: *kirjutama* 'write'
Bounded verbs: *võitma* 'win'

In my classification, if a verb is telic and lexically bounded, then its boundedness feature is specified. Indications about the boundedness of the verb belong to the functional specifications in the verb entries and in the respective terminal node of the constituent-structure.

(30) *võitma* , *V* : (↑PRED)= 'WIN <(↑SUBJ), (↑OBJ)>'
(↑B)=MIN

(31)

$$\begin{bmatrix} \text{PRED} & \text{'WIN} \quad \text{<SUBJ, OBJ>'} \\ \text{B} & \text{MIN} \end{bmatrix}$$

These specifications have the form of defining equations as in the verb entry of *võitma* 'win' (30). In this case, boundedness is specified in the lexical entry of the verb and clausal aspect is determined by the verb. As a result of the mapping from constituent structure to functional structure, the f-structure is constrained to contain the specified boundedness feature, that is, an attribute with a "fixed" value (31). Having a fully specified feature (a defining equation) as part of its lexical entry, such as (↑B)=MIN, means for the verb that its boundedness is lexicalized, that it is an inherently perfective or telic, an inherently bounded verb. Since clausal aspect is modeled in terms of the unification of boundedness features in the f-structure, the failure in unification explains the restrictions on case marking patterns in the model where case contributes different values. This means that these verbs are not boundable by further elements and the range of aspectual case marking possibilities is restricted.

If verbs are boundable, their boundedness feature is valueless. They can be bounded, and the range of case marking possibilities is open. Indications about the boundability of the verb also belong to the functional specifications in the verb entry and are present at the terminal verb node of the c-structure. These specifications have the form of existential constraints in LFG as in (32).

(32) **kirjutama**, *V* : (↑PRED)= 'WRITE <(↑SUBJ), (↑OBJ)>'
(↑B)

In this case, clausal boundedness is not determined by the verb (by the lexical entry of the verb) but only as the result of the unification of features in the clausal f-structure (33). As a result of the

mapping from constituent structure to f-structure, the f-structure is constrained to contain only the attribute part of the boundedness feature, that is, an attribute without any value.

(33)
$$\begin{bmatrix} \text{PRED 'WRITE <SUBJ, OBJ>'} \\ \text{B} \end{bmatrix}$$

Having an existential constraint (↑B) means that the attribute B must be present in the f-structure feature matrix that corresponds to the verb in c-structure. As clausal aspect is modeled in terms of the unification of boundedness features in the functional structure, the possibility of the unification with features with different values explains the wider range of case marking patterns. In my model, the "underspecified" features become fully specified by the features of case-marked objects.

The next question is: given the incomplete f-structure, how will the values be obtained? Before discussing the verbs' contribution to the sentence and the interaction with case-marked objects, I present the features associated with the three types of case markers.

7.3. Inside-out constraints for features associated with case-marked objects

Boundedness is also the term for the aspectual features in the f-structure feature matrix, where the B attribute can have the value of MINimal (in minimally bounded sentences) or MAXimal (in maximally bounded sentences). A maximally bounded sentence denotes an event with clear boundaries and that cannot be continued. A minimally bounded sentence denotes an event that either has existing but unspecified clear boundaries or can be continued.

The total case is the case that encodes the maximal boundedness in sentences; it appears in sentences that denote an event with clear boundaries and that cannot be continued. The lexical entry of the total case contains a defining equation, an inside-out constraint for the maximal boundedness feature, (B↑)=MAX. The entry for the total case is presented in (34). A total case-marked nominal specifies the f-structure information in (35).

(34)

TOT: (↑CASE) = TOT
((OBJ ↑) B) = MAX

(35)
$$\begin{bmatrix} \text{B} & \text{MAX} \\ \text{OBJ} & \begin{bmatrix} \text{CASE} & \text{TOT} \end{bmatrix} \end{bmatrix}$$

The indication (OBJ ↑) secures that the higher f-structure contains an object to which the immediate f-structure containing the case-marked nominal belongs. The association between the nominal and its grammatical function is established by virtue of the case marker attached to it (cf. Nordlinger and Sadler 2004). I leave the semantic constraints that constrain the mapping between the f-structure and c-structure aside.

Partitive is the default case; it encodes only the constraint that the sentence is not maximally bounded (36). A constraint equation captures this constraint on the f-structures.

(36) PART1: (↑CASE) = PART
((OBJ ↑) B) ≠ MAX

Partitive object NPs specify the information in the f-structure feature matrix as in (37). If the f-structure matrix contained a B attribute with a MAX value, the structure would be ill-formed.

(37)
$$\begin{bmatrix} \text{OBJ} & \begin{bmatrix} \text{CASE} & \text{PART} \end{bmatrix} \end{bmatrix}$$

The general well-formedness conditions of LFG secure the sensitivity of aspectual case to verb classification and vice versa. The sentence is ill-formed as a result of a feature clash between the features specified by the total case, (B↑)=MAX, and the verb *võitma* 'win', (↑B)=MIN. Partitive marked objects and the bounded verb form well-formed minimally bounded sentences, since the verb entry constrains the f-structures to have a "minimally bounded" feature, and the features are unifiable, and the entry for partitive fixes that the structure should not contain a "maximally bounded" feature, which it does not. The two types of bounded sentences formed by the verb *kirjutama* 'write', which has an entry with an existential constraint, are also explained: the "minimal" and "maximal" values of the attribute are provided by the case-marked objects partitive plural and total, respectively.

(38) ***kirjutama*** 'write' … (↑B)
võitma 'win'… (↑B) = MIN

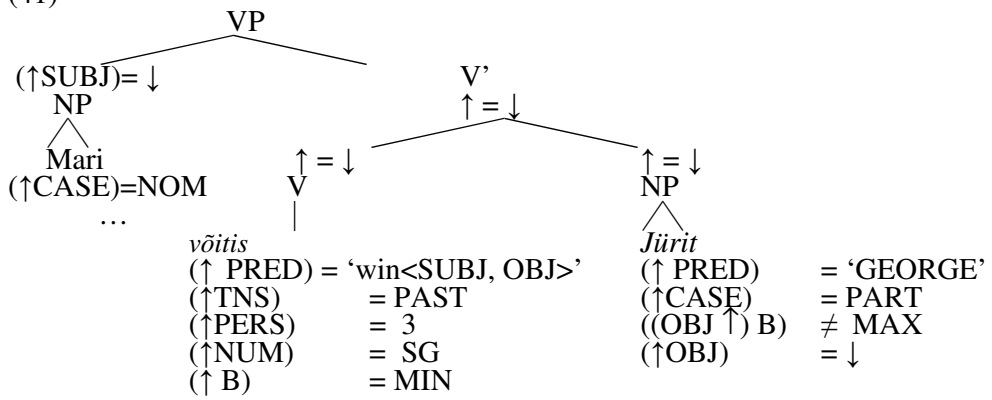
The following example (39) is analyzed below. This is an example of a bounded verb, and a minimally bounded sentence. The lexical entries for the verb and the object are represented as in (40) and the constituent structure of (39) is presented in (41).

(39) *Mari* *võitis* *Jürit.*
M.nom win.3.sg.past George.part
'Mary won George.'

(40)

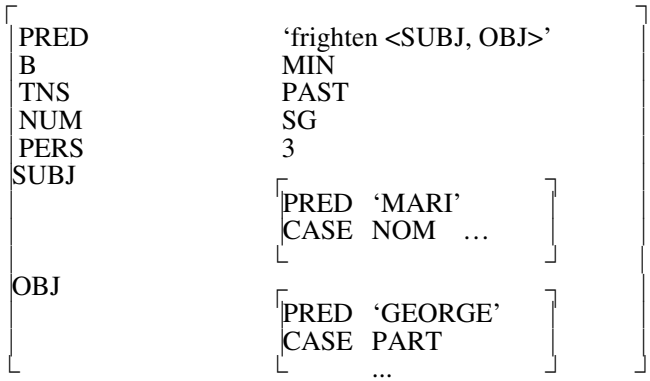
võitis	V	(↑ PRED)	= 'win <(↑ SUBJ), (↑ OBJ)>'
		(↑TNS)	= PAST
		(↑PERS)	= 3
		(↑NUM)	= SG
		(↑ B)	= MIN
Jürit	N	(↑ PRED)	= 'GEORGE'
		(↑CASE)	= PART
		((OBJ ↑) B)	≠ MAX

(41)



The corresponding functional structure of (39), *Mari võitis Jürit*, containing the relevant information, is unified without any violation of well-formedness conditions (42).

(42)



Also, boundable verbs in maximally bounded sentence (43) are unified without any violation of well-formedness conditions.

(43) *Mari* *kirjutas* *raamatu.*
M.nom write.3.sg.past book.tot
‘Mari wrote a book.’

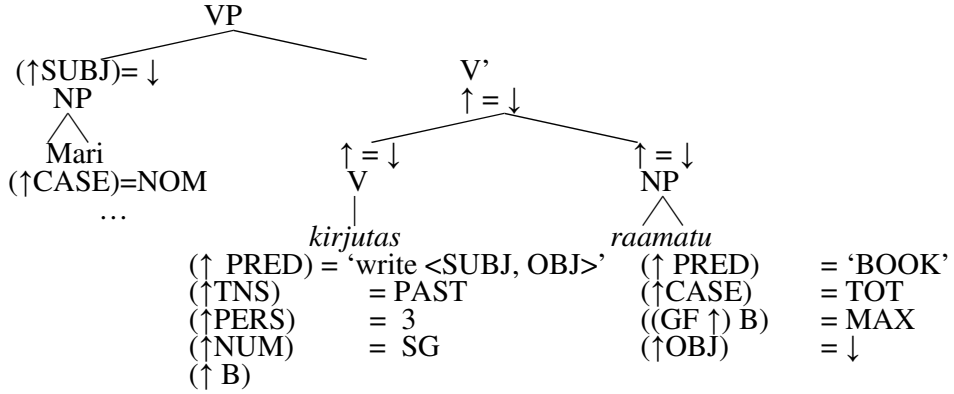
The lexical entries for the verb and the object of (43) follow in (44).

(44)

kirjutas V	(↑ PRED)	= ‘write <(↑ SUBJ), (↑ OBJ)>’
	(↑TNS)	= PAST
	(↑PERS)	= 3
	(↑NUM)	= SG
	(↑ B)	
raamatu N	(↑ PRED)	= ‘BOOK’
	(↑CASE)	= TOT
	((OBJ ↑) B)	= MAX

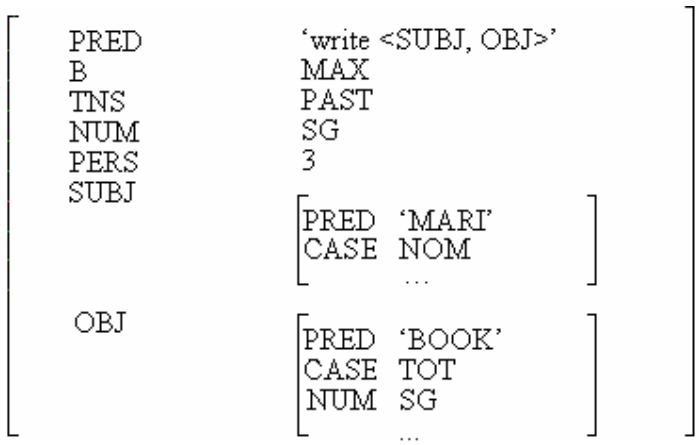
The constituent structure of *Mari kirjutas raamatu* (43) follows in (45).

(45)



The corresponding functional structure of (43) does not violate well-formedness conditions (46).

(46)



8. Conclusion

The Estonian aspectual phenomena, especially the relation between aspect and the object case alternation, bear resemblance to the better-known Finnish aspectual phenomena (e.g., Kiparsky 1998, 2005) and Scottish Gaelic (Ramchand 1997). The account presented here addresses new Estonian data about verbs and case alternation and proposes a way to deal with the syntactic side of the phenomenon in LFG. This contribution adds to the puzzles of the Finnic partitive case, mainly as presented in earlier sources such as Tamm (2004a) and Kiparsky (2005). The latter source on Finnish provides considerably more data and semantic insights than Kiparsky (1998), the previous, VP-compositional account of the Finnish partitive. As my account uses Lexical Functional Grammar, the composition is not within the VP, but at the syntactic level of functional structure and is modeled in terms of feature unification. Kiparsky (2005), in its handout form, does not aim at giving either a complete semantic or syntactic account. My shortcut choice to account for aspectual composition in the f-structure syntax and not in the semantics is motivated by the difficulties in formalizing identical output from several inputs (cf. the same difficulties with the Finnish data and explanation in Kiparsky 2005) and also by the general intuition that multiple and partly overlapping functional constraints reflect the grammaticalization of the Estonian aspect more flexibly, witnessed by the volatility in the judgements on grammaticality and interpretations by native speakers. It remains to be clarified what are the costs and benefits of placing a part of the explanation to syntax and how to articulate the semantic account of the phenomenon.

Alongside with providing Estonian data that allow drawing parallels between Finnish and Estonian, this article concentrates on one of the several differences of the Estonian aspectual system that lead to a more nuanced account of the data, namely, the Estonian transitive telic verb classes that pattern with partitive object case marking (e.g., *win, frighten*). Taking into account that some atelic accusative-object verb classes in Kiparsky (1998), (2005) have partitive objects in Estonian, it can be concluded that the partitive-object verb classes are more numerous in Estonian. This article proposes an aspectual verb classification for Estonian transitive "telic" verbs: they are either lexically (minimally) bounded (*win, frighten*) or lexically boundable (*write, read, find, give, push to the store*). This classification is the basis for the observed systematic compatibility of verb classes with certain clausal aspectual object case marking patterns. Clausal aspect is understood in terms of boundedness. A clause or a sentence is maximally bounded if it describes an event with clear boundaries that cannot be continued. Clausal boundedness is encoded in the form of features at the syntactic level of functional structures. This article studies those aspect-related attributes and values that transitive telic verbs contribute to the f-structure. The lexical entries for transitive verbs are provided with valued or unvalued boundedness features in the proposed LFG lexicon. If a verb is classified as lexically minimally bounded, its functional specifications contain a valued boundedness feature. This restricts the range of aspectual case marking possibilities. If verbs are boundable, their boundedness feature is

unvalued. Since clausal aspect is modeled in terms of the unification of boundedness features in the f-structure, the possibility of the unification of features with different values explains the wider range of case marking patterns. The features become fully specified in the process of the unification with the features of case-marked objects. Verbs fall into aspectual classes, distinguished from each other according to the pattern of the attributes and values in the functional specifications of the verbs' lexical entries. This verb classification of two verb classes is the basis for accounting for the interaction between Estonian aspect, verbs, and case; however, many other verb classes and aspectual phenomena of Estonian are not addressed here.

Referenes

Ackerman, Farrell and John Moore. 2001. *Proto-Properties and grammatical encoding: a correspondence theory of argument selection*. CSLI Publications, Stanford.

Belletti, Adriana. 1988. 'The Case of Unaccusatives.' *Linguistic Inquiry* 19: 1-34.

Butt, Miriam. 2006. *Theories of Case.* Cambridge University Press, Cambridge.

Butt, Miriam and Tracy Holloway King. 2005. 'The Status of Case.' In Veneeta Dayal and Anoop Mahajan (eds.) *Clause Structure in South Asian Languages.* Springer Verlag, Berlin.

Butt, Miriam, Mary Dalrymple and Anette Frank. 1997. 'An Architecture for Linking Theory in LFG.' In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG 97 Conference.* http://www-csli.stanford.edu/publications/

Comrie, Bernard. 1976. *Aspect. An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge University Press. Cambridge, London, New York, Melbourne.

Dahl, Östen. 1985. *Tense and Aspect systems.* Blackwell, Oxford.

Depraetere, Ilse. 1995. 'On the necessity of distinguishing between (un)boundedness and (a)telicity', *Linguistics and Philosophy* 18.1: 1 – 19.

Erelt, Mati, Tiiu Erelt, and Kristiina Ross. 1997. *Eesti Keele Käsiraamat*. [The Handbook of Estonian] . Eesti Keele Sihtasutus, Tallinn.

Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael and Silvi Vare. 1993. *Eesti Keele Grammatika II. Süntaks. Lisa: Kiri.* [The Grammar of the Estonian Language II. Syntax]. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn.

Erelt, Mati. 1985. 'MA-, -MAS- ja –MAST-infinitiivist eesti keeles.' [On -MA, -MAS and –MAST infinitive in Estonian] In: *Ars Grammatica*. Academy of Sciences of the Estonian SSR Institute of Language and Literature. Valgus, Tallinn: 4 – 22.

Glasbey, Sheila. 2001. 'Tense, Aspect and the temporal structure of discourse: towards an LFG account.' In: Miriam Butt and Tracy Holloway King. *Proceedings of the LFG01 Conferece*. CSLI Publications, Stanford.

Harms, Robert T. 1962. *Estonian Grammar*. Indiana University Publications. Uralic and Altaic Series. Vol.12. Indiana University/Mouton & Co., Bloomington/the Hague.

Hiietam, Katrin. 2003. *Definiteness and Grammatical Relations in Estonian*. PhD dissertation, Manchester.

de Hoop, Helen. 1996. *Case Configuration and NP Interpretation.* Garland, New York.

Kiparsky, Paul. 2005. *Absolutely a Matter of Degree: The Semantics of Structural Case in Finnish.* Handout CLS, April 2005.

Kiefer, Ferenc. s.d. *Jelentéselmélet.* [Semantic theory]. Corvina, Budapest.

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

Kiparsky, Paul. 2001. 'Structural Case in Finnish.' *Lingua* 111:315 – 376.

Kiparsky, Paul. 1998. 'Partitive Case and Aspect.' In: Miriam Butt and Willem Geuder (eds.) *The Projection of Arguments.* Stanford, CSLI Publications: 265 – 307.

Klaas, Birute. 1996. 'Similarities in case marking in Estonian and Lithuanian'. In: (ed.) Erelt, Mati (ed*.). Estonian: Typological Studies* I. Publications of the Department of Estonian of the University of Tartu 4:35 – 67.

Klaas, Birute. 1999. 'Dependence of the object case on the semantics of the verb in Estonian, Finnish and Lithuanian.' In: Erelt, Mati (ed.) *Estonian. Typological studies* III. Publications of the Department of Estonian of the University of Tartu 11:47 – 83.

Kont, Karl. 1963. 'Käändsõnaline objekt läänemeresoome keeltes.' [The declined Object in Baltic Finnic languages] ENSV *Teaduste Akadeemia Keele ja Kirjanduse Instituudi uurimused* IX, Tallinn.

Krifka, Manfred. 1992. 'Thematic Relations as Links Between Nominal Reference and Temporal Constitution.' In: Ivan Sag and Anna Szabolcsi (eds.), *Lexical Matters*. CSLI Publications, Stanford: 29 – 53.

Larsson, Lars-Gunnar. 1983. 'Studien zum Partitivgebrauch in den ostseefinnischen Sprachen.' *Acta Universitatis Upsaliensis. Studia Uralica et Altaica Upsaliensia* 15. Uppsala.

Lee, Hanjung. 1999. 'The Domain of Grammatical Case in Lexical-Functional Grammar' In: Miriam Butt and Tracy Holloway King (eds.) *Proceedings of the LFG99 Conference*. CSLI Publications, Stanford University. http://www-csli.stanford.edu/publications/

Maling, Joan. 1993. 'Of nominative and accusative: the hierarchical assignment of grammatical case in Finnish.' In Holmberg, Anders and Urpo Nikanne (eds.), *Case and other functional categories in Finnish syntax.* Mouton de Gruyter, Berlin: 49-74.

Metslang, Helle. 2001. 'On the Developments of the Estonian Aspect: the Verbal Particle ära.' In: Östen Dahl and Maria Koptjevskaja-Tamm (eds.) *The Circum-Baltic Languages: Their Typology and Contacts. Studies in Language Companion Series* 55. Benjamins, Amsterdam: 443 – 479.

Metslang, Helle. 1994. 'Temporal Relations in the predicate and the Grammatical System of Estonian and Finnish.' Oulun *Yliopiston Suomen ja saamen kielen laitoksen tutkimusraportteja* 39. Oulu.

Metslang, Helle and Hannu Tommola. 1995. 'Zum tempussystem des Estnischen.' [On the tense system of Estonian.] In Rolf Thieroff, *Tense Systems in European Languages II (Linguistische Arbeiten,* 338). Niemeyer, Tübingen: 299 – 326.

Nelson, Diane. 2003. 'Case and adverbials in Inari Saami and Finnish.' In: Anne Dahl and Peter Svenonius (eds.) *Proceedings of the 19th Scandinavian Conference of Linguistics. Nordlyd* 31.4:708 – 722.

Nelson, Diane. 1998. *Grammatical Case Assignment in Finnish*. Garland, New York.

Nemvalts, Peep. 1996. 'Case Marking on Subject Phrases in Modern Standard Estonian.' *Acta Universitatis Upsaliensis. Studia Uralica Upsaliensia* 25.

Nordlinger, Rachel and Louisa Sadler. 2004. 'Tense Beyond the Verb: Encoding Clausal Tense/Aspect/Mood on Nominal Dependents.' *Natural Language and Linguistic Theory* 22. 597–641.

Nordlinger, Rachel. 1998. 'Constructive Case: Evidence from Australian Languages.' *Proceedings of the LFG97 Conference.* CSLI Publications, Stanford.

Pereltsvaig, Asja. 2001. 'On accusative adverbials in Russian and Finnish, In: A. Alexiadou and Peter Svenonius (eds.). *Adverbs and Adjunction*, Linguistics in Potsdam 6: 155 – 176.

Ramchand, Gillian Catriona. 1997. *Aspect and Predication: The Semantics of Argument Structure*. OUP.

Pihlak, Ants. 1982. 'Vene aspektikategooria ja eesti ajakategooria suhtest.' [On the relation between the Russian category of aspect and the Estonian category of tense.] *Voprosy sopostavitelnogo izutshenija leksiki i grammatiki na materiale estonskogo i russkogo jazykov*. ENSV Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn: 87 – 100.

Rätsep, Huno. 1978. 'Eesti keele lihtlausete tüübid.' [Types of Estonian simple sentences.] *ENSV TA Emakeele Seltsi Toimetised* 12. Valgus, Tallinn.

Rätsep, Huno. 1957. 'Aspektikategooriast eesti keeles.' [On the category of aspect in Estonian]. *Emakeele Seltsi Aastaraamat* III: 72 – 77.

Pusztay, János. 1994. *Könyv az észt nyelvről*. [A book about the Estonian language.] *Folia Estonica*. Tomus III. Savariae, Szombathely.

Rajandi, Henno and Helle Metslang. 1979. 'Määratud ja määramata objekt.' [Defined and undefined object.] ENSV TA KKI, Valgus, Tallinn.

Smith, Carlota. 1991. *The Parameter of Aspect*. Kluwer Academic Publishers, Dordrecht.

Sulkala, Helena. 1996. 'Expression of Aspectual Meanings in Finnish and Estonian.' In Mati Erelt (ed.) Estonian: *Typological Studies 1*. Publications of the Department of Estonian of the University of Tartu: 165–217.

Tamm, Anne. To appear. *Scalar structure underlies telicity and evidentiality: On the aspectual partitive marking of the objects and the evidential partitive of the –vat form in Estonian.* Paper presented at TAMTAM, Nijmegen.

Tamm, Anne. 2006. 'Estonian object and adverbial case with verbs of motion.' In *Proceedings of Grammar and Context - New Approaches to the Uralic Languages*. Budapest.

Tamm, Anne. 2004a. *Relations between Estonian verbs, aspect, and case*. Doctoral dissertation, ELTE, Theoretical Linguistics Program, Budapest.

Tamm, Anne. 2004b. 'On the grammaticalization of the Estonian perfective particles.' *Acta Linguistica Hungarica*. Vol. 51.1-2:143 – 169.

Tamm, Anne. 2003a. 'Estonian transitive verb classes, object case, and the progressive.' In Anne Dahl and Peter Svenonius (eds.) *Proceedings of SCL Working papers of Language and Linguistics*. Nordlyd 31.4.

Tamm, Anne. 2003b. 'Delimitedness, telicity, direct objects and obliques.' In: János Pusztay (ed.) *Specimina Sibirica. Vol. XXI, Materialen der Konferenz Valencia Uralica. Szombathely, 18. – 19. April 2002*. Savaria University Press. Szombathely:115-149.

Tauli, Valter. 1968. 'Totaalobjekt eesti kirjakeeles.' [Total object in Estonian.] In *Suomalais-ugrilaisen Seuran Toimituksia* 145. Helsinki: 216 – 224.

Tauli, Valter. 1983. *Estonian Grammar II. Syntax.* Uppsala.

Tenny, Carol. 1994. 'Aspectual Roles and the Syntax-Semantics Interface.' *Studies in Linguistics and Philosophy,* Vol. 52. Kluwer Academic Publishers. Dordrecht/ Boston/ London.

Toivonen, Ida. 2001. *The phrase structure of non-projectig words*. Doctoral dissertation, Stanford University.

Vainikka, Anne and Joan Maling.1996. 'Is Partitive Case Inherent or Structural?' In: Jack Hoeksema (ed.) *Partitives*. *Studies on the distribution and meaning of partitive expressions*. Mouton de Gruyter, Holland: 179 – 208.

Verkuyl, H. 1993. *A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure*. Cambridge University Press, Cambridge.

OBLIQUE DEPENDENTS IN ESTONIAN: AN LFG PERSPECTIVE

Reeli Torn
University of Tartu

Proceedings of the LFG06 Conference
Universität Konstanz

Miriam Butt and Tracy Holloway King (Editors)
2006

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

Abstract

The status of indirect or oblique dependents in Estonian has long been a matter of controversy. One approach (Kure 1959, Klaas 1988, Nemvalts 2004) classifies them as a class of 'indirect objects', which represent indirectly affected participants. Another approach (Väärі 1959, Erelt 1989, 2004, Erelt et al. 1993) disputes the usefulness of this distinction, and assigns all grammatical dependents other than subjects and direct objects to a large and heterogeneous class of 'adverbials', based on the fact that indirect dependents are similar in form to adverbial modifiers. The present paper takes up this traditional issue from a contemporary theoretical perspective, and argues that Lexical Mapping Theory (Bresnan & Zaenen 1990) clarifies a basic syntactic contrast between oblique functions (the 'object' or 'governed' adverbials in current Estonian grammar) and ungoverned adverbial modifiers. The general dissociation between form and function in LFG also clarifies how a single semantic case form can function syntactically either as a modifying adverbial or as a governed oblique function.

1. Introduction[1]

Estonian has no single case, like the dative, for marking indirectly affected participants. Instead, indirectly affected participants are encoded by the same 'local' case forms that are used with adverbial dependents to express a range of mainly spatial relations.[2] This is illustrated in (1):

(1) a. *Mees istus diivanile.* (**adverbial** allative)
man.NOM sat sofa.ALLA
'A man sat onto the sofa.'

b. *Ema andis lapsele raha.* (**oblique** allative)
mother.NOM gave child.ALLA money.PART
'The mother gave money to the child.'

While an allative dependent such as *diivanile* 'onto the sofa' in (1a) functions as an ungoverned adverbial in construction with a motion verb such as *istuma* 'sit', the same case form expresses an indirectly affected participant in construction with a ditransitive such as *andma* 'give' in (1b). The basic contrast between the role of *diivanile* in (1a) and *lapsele* in (1b) can be expressed in any model that distinguishes a class of 'indirect objects' or governed oblique dependents from subjects and objects on one hand, and from adverbial elements on the other. The present paper takes this traditional issue from a more theoretical perspective, and argues that the classification of grammatical functions in Lexical Mapping Theory (Bresnan & Zaenen 1990) clarifies the grammatical contrast between formally parallel elements. Accordingly, *diivanile* functions as an adverbial and *lapsele* as an oblique.

2. Two traditional views of oblique dependents in Estonian

The status of indirectly affected participants like *lapsele* in (1b) in Estonian has long been a matter of controversy. Some linguists (Kure 1959, Mihkla 1959, Klaas 1988, Nemvalts 2004) analyse such instances as a class of indirect objects, which represent indirectly affected participants. For instance, Klaas (1988) distinguishes three subclasses of indirect objects: 'indirect relative object', 'indirect partner object' and 'indirect possessive object'. In all cases, the 'local'

1 This study was supported by the Doctoral School of Linguistics and Language Technology and by Grant no. TFLEE 2568. I am especially grateful to Jim Blevins and Mati Erelt for their helpful comments on this article.

2 Estonian has 14 cases. Nominative, genitive and partitive are abstract grammatical cases. Illative, inessive, elative, allative, adessive and ablative are 'local' semantic cases, and translative, terminative, essive, abessive and comitative make up the remaining semantic cases.

semantic cases are used to mark an indirectly affected participant, which make them different from the direct object in its morphology as well as in semantics.

Other linguists (Väärі 1959, Erelt 1989, 2004, Erelt et al. 1993) dispute the usefulness of this distinction, and assign all grammatical dependents other than subjects and direct objects to a large and heterogeneous class of 'adverbials', based on the observation that indirect dependents are similar in form to adverbial modifiers. The fact that 'indirect objects' bear the same 'local' cases is one of the main critical arguments against distinguishing them as a separate class of arguments. Moreover, Erelt (2002: 37) states that 'indirect objects' demonstrate no "specific syntactic behaviour". However, the fact that the instances like (1b) actually have some object-like properties is reflected in the names 'object adverbials' or 'government adverbials' used by the opponents of 'indirect objects'. Each of these 'object-like properties', which are described below, distinguishes governed obliques from adverbial elements. A genuinely adverbial dependent is not governed by a predicate but is subject to a looser requirement of 'semantic compatibility'. For example, the adverbial allative case in (1a) can be replaced by other semantically appropriate local expressions, as illustrated in (2).

(2) a. *Mees istus autosse.* (adverbial illative)
man.NOM sat car.ILLA
'A man sat into the car.'

b. *Mees istus diivanil.* (adverbial adessive)
man.NOM sat sofa.ADES
'A man sat on the sofa.'

c. *Mees istus autos.* (adverbial inessive)
man.NOM sat car.INES
'A man sat in the car.'

Which cases are compatible will depend on the type of action expressed by the verb, but also by the physical properties of the dependents that stand in the relation specified by a given case.

Yet, one must agree with the opponents of 'indirect object' that 'indirect objects' do not form a unified class as 'subjects' and 'direct objects' do in Estonian. At the same time the notion of 'indirect object' as a third grammatical argument may reflect a typological bias based on dative dependents in Indo-European, and thus would not apply to Estonian in the same sense.

3. An LFG approach to oblique dependents in Estonian

Building on previous studies, the following sections identify a number of syntactic respects in which 'object adverbials' in Estonian behave like governed grammatical functions and unlike the adverbials in (2). An analysis of Estonian 'object adverbials' within LFG helps to clarify their status by highlighting parallels between their syntactic behaviour and the behaviour of other types of governed grammatical functions. In addition, the separation of form and function in LFG permits a compromise between the two alternatives set out in Section 2. An LFG account can represent the similarity in form and even meaning between oblique dependents and adverbials without assigning the same functional analysis to these elements. Obliques are a type of governed grammatical function, whereas adverbials fall within the class of adjuncts. More specifically, oblique dependents can be treated as thematically restricted obliques (i.e. as [+r, -o] functions in LMT, following Bresnan & Kanerva 1989, Bresnan & Zaenen 1990). Individual predicates may govern particular oblique functions (as in (1b)). Oblique functions may also serve as antecedents in anaphoric control constructions. The case of oblique functions may even alternate with structural subject and object cases.

Although they have distinct syntactic functions, oblique dependents and adverbials are united by the fact that their similarity in form is associated with common thematic properties. While the allative in (1a) represents literal movement in space onto a point of reference, the allative in (1b) expresses metaphorical movement towards a recipient/goal. An LFG analysis can capture the types of syntactic properties that motivated the treatment of 'object adverbials' as indirect objects. At the same time, the analysis can express the formal and semantic similarities that lead others to group object adverbials within a class of formally and semantically similar adverbials. By factoring the properties of these elements, one can avoid the need for the choice in Section 2.

4. Syntactic contrasts between obliques and adverbials

4.1 Verb government

The behaviour of object adverbials differs from that of adverbial modifiers. The case of a particular object adverbial is governed by a verb, just as the case of a subject or direct object is. Verbs such as *mõtlema* 'think' or *rääkima* 'talk' can govern an elative dependent that expresses the propositional content of the verb or an allative dependent that expresses the thing thought about or the person spoken to. Verbs such as *helistama* 'telephone' or *andma* 'give' govern an allative argument, corresponding to the recipient of the call or the goal of the giving. Other classes of verbs govern particular cases for grammatical dependents that express various types of 'indirectly affected' participants (Klaas 1988). Unlike genuinely adverbial uses of 'semantic' cases, these verbs select a particular governed case, not a semantically compatible class of cases.

The examples in (3) and (4) illustrate this contrast. The verbs *helistama* 'telephone', *kirjutama* 'write' and *kuuluma* 'belong to' in (3) all govern the allative.

(3) a. *Ema helistas tütrele.* (oblique allative)
mother.NOM called daughter.ALLA
'The mother called her daughter.'

b. *Sõber kirjutas nendele.* (oblique allative)
friend.NOM wrote them.ALLA
'A friend wrote to them.'

c. *Auto kuulub isale.* (oblique allative)
car.NOM belongs father.ALLA
'The car belongs to the father.'

In (4), *armuma* 'fall in love', *puutuma* 'concern' and *uskuma* 'believe in' govern the illative.

(4) a. *Poiss armus tüdrukusse.* (oblique illative)
boy.NOM fell in love girl.ILLA
'The boy fell in love with the girl.'

b. *See puutub ka temasse.* (oblique illative)
That.NOM concerns also him/her.ILLA
'That also concerns him/her.'

c. *Isa uskus pojasse.* (oblique illative)
Father.NOM believed son.ILLA
'The father believed his son.

Unlike the adverbial uses of allative and illative phrases, the governed dependents in (3) and (4) are 'strictly subcategorized' by particular verbs. Interchanging these cases does not produce a difference in meaning, but leads instead to unacceptability, as the sentences in (5) show.

(5) a. **Ema helistas tütresse.*
(mother called daughter.ILLA)

b. **Sõber kirjutas nendesse.*
(friend wrote them.ILLA)

c. **Auto kuulub isasse.*
(car belongs father.ILLA)

d. **Poiss armus tüdrukule.*
(boy fell in love with girl.ALLA)

e. **See puutub ka temale.*
(that concerns also him/her.ALLA)

f. **Isa uskus pojale.*
(father believed son.ALLA)

The government of oblique dependents is not lexically idiosyncratic or 'quirky' but reflects an abstract or metaphorical sense of their adverbial meaning. This is particularly clear in the case of the allative in (3a) and (3b), which marks the transfer or abstract motion toward a goal dependent. As noted earlier, this metaphorical usage is extended in ditransitive constructions in (6), where the allative expresses the recipient or goal of the action.

(6) a. *Õpetaja andis õpilastele uue ülesande.* (oblique allative)
teacher.NOM gave students.ALLA new assignment.GEN
'The teacher gave the students a new assignment.'

b. *Ma tõin teile hea uudise.* (oblique allative)
I brought you.ALLA good news.GEN
'I brought you good news.'

c. *Sõber saadab sulle raamatu.* (oblique allative)
friend.NOM sends you.ALLA book.GEN
'A friend is sending you a book.' (examples from Klaas 1988: 41)

4.2 Anaphoric control

Some 'object adverbials' may act as antecedents in obligatory control constructions, which provides further confirmation that they are not adverbials but function as obliques. An obligatory anaphoric control construction involves coreference between an argument of the superordinate clause and the controlled position in the subordinate clause (Dalrymple 2001: 324). It might also be possible to treat the examples below as cases of functional control if obliques are allowed to act as functional controllers (as Cook suggests elsewhere in this volume). However, this alternative would strengthen the claim that 'object adverbials' may participate in constructions that allow governed functions but disallow genuine adverbial dependents.

Sentences in (7) illustrate object control constructions. The object control verbs *paluma* 'ask' and *aitama* 'help' take an adessive dependent, which functions as the controller of the implicit subject of the dependent infinitive.

(7) a. *Direktor palus sekretäril asja selgitada.*
director.NOM asked secretary.ADES thing.PART explain.INF
'The director asked the secretary to explain the issue.'

b. *Lapsed aitavad emal nõusid pesta.*
children.NOM help mother.ADES dishes.PART wash.INF
'The children help their mother wash the dishes.'

In these control structures, the adessive dependents *sekretäril* 'secretary' and *emal* 'mother' are again strictly governed by the verb, and other semantic cases cannot be substituted. In addi-

tion, these dependents also play a critical role in a grammatical dependency, by serving as the obligatory antecedent for the subject of the dependent infinitive.

The modal verbs *tarvitsema* and *pruukima* 'need' may occur in subject control constructions where they take an adessive dependent that antecedes the subject of a dependent infinitive. In these constructions, illustrated in (8), the controlling 'object adverbial' shows the same degree of integration in the functional structure of a clause as subject or direct object controllers.

(8) *Sul pruugib/tarvitseb ainult seda raamatut lugeda.*
you.ADES need only this.PART book.PART read.INF
'You need only read this book.'

The control constructions in (7) and (8) clearly distinguish governed uses of semantic cases from their adverbial uses. Semantic case forms in an adverbial function have a much looser connection to the argument structure of the clauses in which they occur and adverbials never serve as obligatory anaphoric (or functional) controllers.

4.3 Case variation with structural cases

The functional similarity between 'object adverbials' and subjects and objects in control constructions is reinforced by systematic patterns of case variation. In one pattern, the case of an 'object adverbial' controller may alternate with a structural partitive. As illustrated in (9), the adessive controller governed by an object control verb like *paluma* or *aitama* may also occur in the partitive, which is the default direct object case in Estonian.

(9) a. *Direktor palus sekretäri asja selgitada.*
director.NOM asked secretary.PART thing.PART explain.INF
'The director asked the secretary to explain the issue.'

b. *Lapsed aitavad ema nõusid pesta.*
children.NOM help mother.PART dishes.PART wash.INF
'The children help their mother wash the dishes.'

In another pattern, the adessive controller governed by a negative modal such as *ei tarvitse* or *ei pruugi* 'need not' alternates with a nominative subject, as illustrated by the examples in (10).

(10) *Sa ei pruugi/tarvitse seda raamatut läbi lugeda.*
you.NOM not need this.PART book.PART through read.INF
'You do not need to read this book through.'

The alternations between 'semantic' cases and governed structural cases are restricted to dependents functioning as governed functions, and are never possible for genuine adverbials.

In short, at least some classes of semantic cases have two quite different syntactic functions. When they occur as unselected modifiers, these cases have an adverbial function. But when they are governed by a particular verb (or construction), exactly the same case forms may function as oblique dependents, representing goals or other indirectly affected arguments of a verb. Other languages show similar formal overlaps, as illustrated, for example by the use of accusative case to mark direct objects or temporal adverbials in many Indo-European languages. However, the pattern is more pronounced in Estonian, due to a much richer inventory of adverbial cases.

5. On 'object adverbials' in LFG

The theoretical perspective of modern syntactic models helps to clarify the status of governed semantic cases. Theories with an explicit focus on grammatical functions are particularly rele-

vant. Relational grammar (RG; Perlmutter 1983) and Lexical Functional Grammar (LFG; Kaplan & Bresnan 1982, Bresnan 2001) are the best-known theories of this kind. LFG makes the status of 'object adverbials' especially clear, because these elements correspond to a type of 'thematically restricted' oblique dependent. The general dissociation between form and function in LFG clarifies how a single semantic case form can be associated with a general case meaning and yet function either as a modifying adverbial or as a governed oblique function.

Unlike RG, which associates different types of indirectly affected participants with a single indirect object relation, 3, the feature system in (10) defines two classes of thematically restricted functions. The first is the class of thematically restricted objects, OBJ_θ, which are assigned the features [+*r*, +*o*]. This class is normally understood to contain structurally case-marked objects that are subject to a semantic restriction, so it is not appropriate for the semantically casemarked arguments in Estonian. The second class of thematically restricted dependents are the oblique functions OBL_θ, which are assigned the features [+*r*, –*o*]. Obliques in this sense are often realised by dependents marked by adpositions, which is appropriate in Estonian, given that that semantic case markers are derived fairly recently from postpositions.

(11) LMT feature analysis of grammatical functions

	(–*restricted*)	(+*restricted*)
(–*objective*)	SUBJ	OBL_θ
(+*objective*)	OBJ	OBJ_θ

The LMT feature analysis of grammatical functions expresses the key insight that the notion 'oblique' does not designate a single grammatical function, like 'subject' or 'direct object', but refers to a family of thematically restricted functions. The table in (12) shows how one and the same semantic case form may function as an adverbial modifier or as a governed dependent. In their governed use, 'object adverbials' serve as oblique grammatical functions. Like 'indirect objects' in other languages, 'object adverbials' in Estonian are often interpreted as goals, recipients or as other types of 'indirectly affected participants'. In their adverbial uses, the same case forms express the more concrete spatial relations usually associated with 'local' cases.

(12) Polyfunctional semantic cases in Estonian

CASE	OBL_θ (OBLIQUE)	ADVERBIAL
ALLATIVE	*helistama, kirjutama, andma, …*	*istuma, hüppama, …*
ILLATIVE	*armuma, puutuma, uskuma, …*	
ADESSIVE	*paluma, aitama, tarvitsema, …*	

5.1 Adverbial uses of semantic cases

The standard LFG analysis of adjuncts applies to adverbial dependents of the sort illustrated by the examples in (2). In the analysis assigned to (2b) in diagram 1, only the subject *mees* 'man' is governed by the verb *istuma* 'sit'. This is indicated by the fact that the PRED value of *istuma* contains only a SUBJ function. The adessive adverbial *diivanil* 'on the sofa' is not governed by the verb, but occurs in the set of ungoverned adjuncts that are the value of the ADJ(UNCT) function.

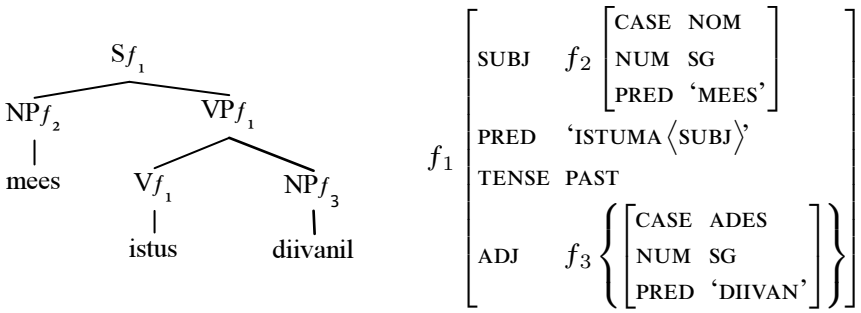


Diagram 1. C-structure and f-structure for sentence (2b) with ungoverned adverbial

The 'compatibility' between the verb and adjunct is not represented in the c(onstituent)-structure on the left or in the f(eature)-structure on the right, because this relation is semantic rather than syntactic, and both of the structures in Diagram 1 are syntactic representations.

5.2 Governed uses of object adverbials

The analysis in Diagram 2 shows how obliques are, like subjects and objects, integrated into the argument structure of a clause. The allative dependent *õpilastele* 'students' is governed by the verb *andma* 'give' and occurs as the value of the thematically restricted function OBL_{alla} in Diagram 2.

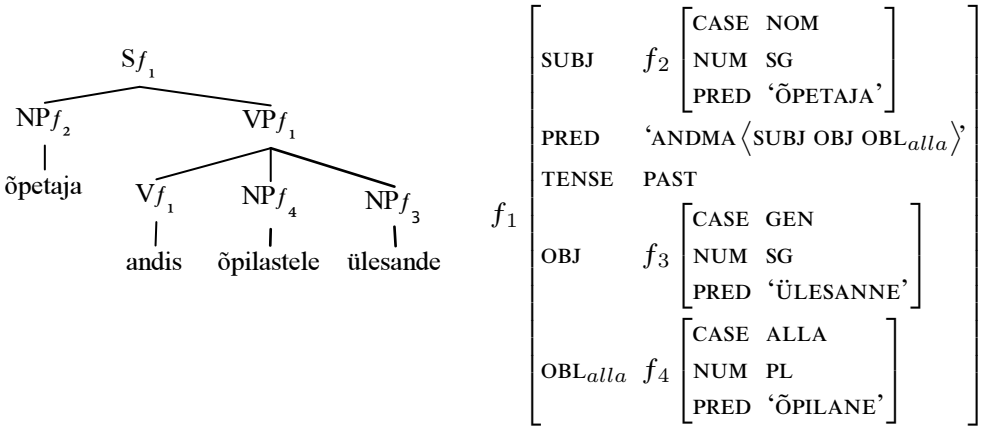


Diagram 2. C-structure and f-structure for sentence (6a) with governed allative oblique

5.3 Anaphoric control by obliques

The analysis assigned to (7a) in Diagram 3 also shows how obliques function as obligatory controllers in anaphoric control constructions. The control relation is represented by the subscripting of the PRED value of the adessive antecedent *sekretäril* 'secretary' and the controlled pronominal subject 'PRO' in the dependent infinitive in Diagram 3.

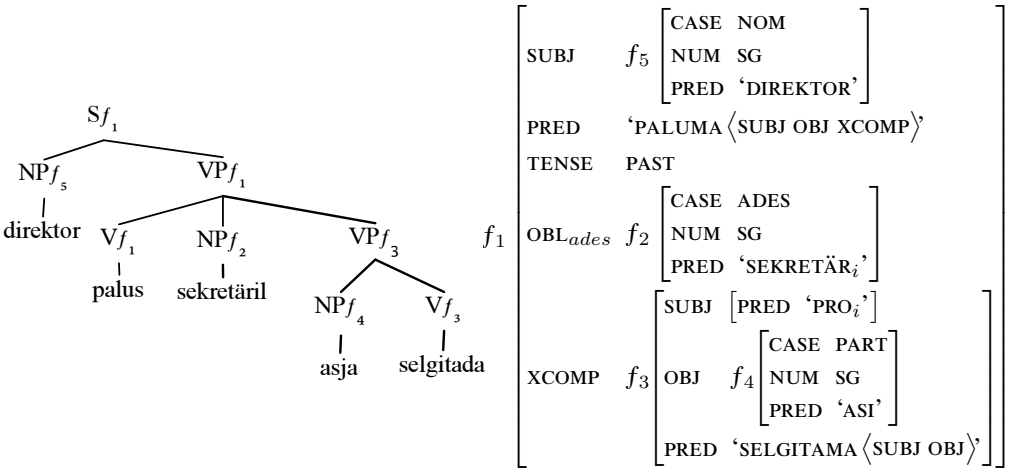


Diagram 3. C-structure and f-structure for sentence (7a) with an adessive oblique controller

5.4 Case variation with oblique and partitive controllers

The grammatical parallel between obligatory oblique controllers and obligatory subject and object controllers is reinforced by the fact that the adessive *sekretäril* in Diagram 3 may alternate with partitive *sekretäri*, as illustrated by example (9a) above and the corresponding Diagram 4.

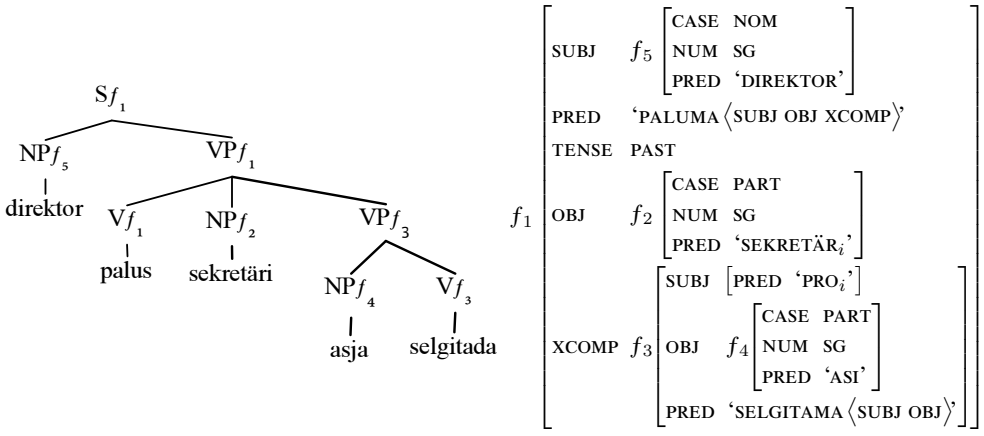


Diagram 4. C-structure and f-structure for sentence (9a) with a partitive controller

Each of the properties illustrated in Diagrams 2–4 distinguishes at least a class of oblique dependents from the type of free adverbials that occur in the structure in Diagram 1.

6. Conclusion

The present article studied a controversial issue in Estonian grammar, that is, the status of so-called 'object adverbials'. There is no agreement in this matter, as some linguists analyse them as 'indirect objects', while others include them in a heterogeneous class of adverbials. The latter view is what is accepted in Modern Estonian Grammar (Erelt et al. 1993).

However, the preceding sections argue that these 'object adverbials' behave more like grammatical arguments than adverbial modifiers, in three main respects. All of them are strictly governed by a particular verb, and some of them may also function as controllers in anaphoric control constructions and demonstrate case variation with structural cases. None of these properties characterise semantically casemarked elements that function as genuine 'adverbials'. Instead, adverbial elements can be integrated into a clause if they show some semantic 'compatibility' with the main verb.

The general dissociation between form and function in LFG clarifies how a single semantic case form can function either as a modifying adverbial or as a governed oblique function. The model of Lexical Mapping Theory (LMT) incorporated within current versions of LFG (Bresnan & Zaenen 1990) offers an elegant analysis of the diverse class of obliques in Estonian. The fact that Estonian 'object adverbials' are governed obliques accounts for their integration into the argument structure of a clause. The thematic restrictions on each type of oblique corresponds to the meanings of semantic cases such as illative, allative or adessive. A given case form may thus function as an oblique dependent when it is governed by a verb, and as an adverbial when it occurs with a compatible motion verb. Furthermore, unlike models such as RG, the fine-grained classification of functions provided by semantic restrictions allows a predicate to govern multiple obliques, provided that each is assigned a different thematic interpretation. In sum, this paper shows how an LFG-based analysis of 'object adverbials' can account for their affinity in form to adverbials while bringing out the syntactic behaviour that identifies them as obliques.

References

Bresnan, J. & J. M. Kanerva. (1989). Locative inversion in Chicheŵa: a case study in factorization in grammar. *Linguistic Inquiry* **20**, 1–50.

Bresnan, J. & A. Zaenen. (1990). Deep unaccusativity in LFG. In K. Dziwirek et al. (eds.), *Grammatical relations: a cross-theoretical perspective*. Stanford: CSLI, 45–57.

Bresnan, J. (2001). *Lexical-functional syntax*. Oxford: Blackwell.

Cook, P. (2006). The German Infinitival Paasive: a Case for Oblique Functional Controllers? In M. Butt and T. Holloway King (eds.), *Proceedings of the LFG06 Conference* (this volume).

Dalrymple, M. (2001). *Lexical Functional Grammar*. New York: Academic Press.

Erelt, M. (1989). *Eesti lauseliikmeist*. Preprint KKI-61. Eesti NSV Teaduste Akadeemia ühiskonnateaduste osakond.

Erelt, M. et al. (1993). *Eesti keele grammatika II. Süntaks*. Lisa: Kiri. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn.

Erelt, M. (2002). Hierarhiatest tüpoloogias. *Teoreetiline keeleteadus Eestis*. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Tartu: Tartu Ülikooli Kirjastus, 34–40.

Erelt, M. (2004). Lauseliigendusprobleeme eesti grammatikas. In L. Lindström (ed.) *Lauseliikmeist eesti keeles*. Tartu Ülikooli eesti keele õppetooli preprindid 1. Tartu, 7–15.

Kaplan, R. M. & J. Bresnan. (1982). Lexical-Functional Grammar: a formal system for grammatical representation. In J. Bresnan (ed.), *The mental representation of grammatical relations*. Cambridge, MA: MIT Press, 173–281.

Klaas, B. (1988). Indirektne objekt. *Keel ja Kirjandus* **1**, 37–42.

Kure, K. (1959). Eesti keele lauseliigenduse alustest. *Keel ja Kirjandus* **1**, 40–49.

Mihkla, K. (1959). Mõningaid märkmeid lauseliikmete ja lauseanalüüsi kohta. *Keel ja Kirjandus* **3**, 171–176.

Nemvalts, P. (2004). Moodustiste toime lausestruktuuris. In L. Lindström (ed.) *Lauseliikmeist eesti keeles.* Tartu Ülikooli eesti keele õppetooli preprindid 1. Tartu, 57-65.

Perlmutter, D. M. (ed.) (1983). *Studies in Relational Grammar 1*. University of Chicago Press.

Vääri, E. (1959). Keele uurimise meetodeist, 'kaudsihitisest' ja öeldistäitest. *Keel ja Kirjandus* **4**, 229–233.