# HPSG'2005
# The 12th International Conference on Head-Driven Phrase Structure Grammar:
# Conference Notes

António Branco
Francisco Costa
Manfred Sailer
(eds.)

DI–FCUL TR–05–10

June 2005

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749–016 Lisboa
Portugal

Technical reports are available at `http://www.di.fc.ul.pt/tech-reports`. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.

# Contents

# Preface

The *12th International Conference on Head-Driven Phrase Structure* took place in Lisbon, at the Faculty of Sciences of the University of Lisbon, in 23-24 August, 2005. It has received a total of 39 submission, out of which the program committee has selected 18 papers for presentation. The local organization managed to provide room for a poster session which includes the alternate papers and six additional posters.

The present conference booklet contains the extended abstracts of the presentations, posters, and invited talks. The contributions are in alphabetic order by the first author.

We are grateful to the Department of Informatics of the University of Lisbon for providing the possibility to publish the conference in their series of technical reports, and to the authors for being willing to re-format their contributions in order to allow for the present homogeneous appearance of this publication.

The present conference booklet is based on the formatting style for the ACL-2005 proceedings by Hwee Tou Ng and Kemal Oflazer. Their style in turn was based, among others, on the formats of earlier ACL and EACL Conference proceedings.

This year's program committee consisted of:
Raúl Aranovich (Davis),
Doug Arnold (Colchester),
Emily Bender (Washington),
Olivier Bonami (Paris),
António Branco (Lisbon),
Berthold Crysmann (Saarbrücken),
Anke Holler (Heidelberg),
Valia Kordoni (Saarbrücken),
Palmira Marrafa (Lisbon),
Tsuneko Nakazawa (Tokyo),
Gerald Penn (Toronto),
Alexander Rosen (Prague),
Manfred Sailer (Göttingen, chair),
Gautam Sengupta (Hyderabad),
Jesse Tseng (Nancy),
Stephen Wechsler (Austin), and
Shuly Winter (Haifa)

We wish to thank the program committee of the conference and Jong-Bok Kim (Seoul), Nurit Melnik (Haifa) and Roland Schäfer (Göttingen) for reviewing the submitted abstracts.

The local organization committee consisted of:
António Branco (chair),
Francisco Costa, and
Filipe Nunes
from the NLX-Group, the *Natural Language Group* of the Department of Informatics, University of Lisbon.

We are grateful to the sponsors of the conference:

- FCT — Fundação para a Ciência e Tecnologia
- Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa

Göttingen and Lisbon, 20 June 2005

António Branco, Francisco Costa, Manfred Sailer

# Contributors


Dora Alexopoulou
ta259@cam.ac.uk

Douglas Ball
dball@stanford.edu

John Beavers
jbeavers@csli.stanford.edu

Emily M. Bender
ebender@u.washington.edu

Núria Bertomeu
bertomeu@coli.uni-sb.de

Francis Bond
bond@cslab.kecl.ntt.co.jp

Denis Creissels
Denis.Creissels@univ-lyon2.fr

Berthold Crysmann

Scott Drellishak
sfd@u.washington.edu

Dan Flickinger
danf@csli.stanford.edu

Frederik Fouvry
fouvry@coli.uni-sb.de

Anette Frank
frank@dfki.de

Danièle Godard
Daniele.godard@linguist.jussieu.fr

Chikara Hashimoto
hasimoto@pine.kuee.kyoto-u.ac.jp

Anke Holler
holler@cl.uni-heidelberg.de

Frank Keller
keller@inf.ed.ac.uk

Christopher Kennedy
ckennedy@uchicago.edu

Jong-Bok Kim
jongbok@khu.ac.kr

Alexander Koller
koller@coli.uni-sb.de

Valia Kordoni
kordoni@coli.uni-sb.de

Yusuke Kubota
kubota@ling.ohio-state.edu

Anna Kup´s´c
kupsc@loria.fr

Sun-Hee Lee
shlee@ling.ohio-state.edu

Takafumi Maekawa
tmaeka@essex.ac.uk

Louise McNally
louise.mcnally@upf.edu

Nurit Melnik
nurit@eyron.com

Stefan M¨uller
Stefan.Mueller@cl.uni-bremen.de

Gertjan van Noord
vannoord@let.rug.nl

Louisa Sadler
louisa@essex.ac.uk

Ivan A. Sag
sag@stanford.edu

Peter Sells
sells@stanford.edu

Jan-Philipp Soehn
jp.soehn@uni-jena.de

Kathrin Spreyer
kathrins@coli.uni-sb.de

Mehran Taghvaipour
matagh@essex.ac.uk

Stefan Thater
stth@coli.uni-sb.de

Beata Trawi´nski
trawinski@sfs.uni-tuebingen.de

Jesse Tseng
tseng@loria.fr

Frank Van Eynde
frank.vaneynde@ccl.kuleuven.be

Aline Villavicencio
avill@essex.ac.uk

# Noun Incorporation in Tongan: A Lexical Sharing Analysis

**Douglas Ball**
Department of Linguistics
Stanford University
dball@stanford.edu

Noun incorporation in Tongan has been generally viewed as having three morphosyntactic properties: (i) it is a syntactic construction where the verb and the incorporated noun are separate words, (ii) it is a construction where only single nouns appear in the incorporated position (Gerdts, 1998) and (iii) the construction's syntactic valency is reduced compared with transitive clauses (Rosen, 1989; Runner and Aranovich, 2003). However, with a closer examination of the data, I propose that none of these claims is quite right for Tongan (and, in some cases, are outright incorrect), and so I propose an alternate analysis of Tongan noun incorporation, one that builds on the mechanism of lexical sharing (Wescoat, 2002; Kim et al., 2004).

In Churchward's (1953) grammar of Tongan, he notes the following minimal pair of an ordinary transitive sentence and a sentence with incorporation.

(1) a. Ordinary Transitive Sentence
Na'e inu 'a e kavá 'e
PAST drank ABS DET kava ERG
Sione
(name)
'Sione drank the kava.'

b. Sentence with Incorporation
Na'e **inu kava** 'a Sione.
PAST drink kava ABS (name)
'Sione drank kava.'

As the second example shows, the sentence with incorporation differs in a number of respects with its ordinary transitive counterpart. First, the incorporated noun is adjacent to the verb, and, in fact, the incorporated noun does not have the word order flexibility that its corresponding phrase does in an ordinary transitive sentence. Second, the prenominal case markers and determiners – including the definitive accent (the *á* in *kavá*) – do not and cannot appear on or with the incorporated noun. Finally, the case of the external argument must be absolutive in an incorporated sentence, contrasting with the ergative marking of external arguments in ordinary transitive sentences.

However, Tongan noun incorporation exhibits several other interesting, yet seemingly conflicting properties. First, incorporated nouns can be accompanied by modifiers and some other phrasal material. Examples of some of the possibilities are shown below:

(2) Conjoined Nouns
Na'e **tō** **manioke** **mo e** **talo** 'a
PAST plant cassava and taro ABS
Sione
(name)
'Sione planted cassava and taro.'

(3) Noun + modificational PP
Na'e **fakama'a** **sea** **'i** **fale** 'a
PAST clean chair in house ABS
Sione
(name)
'Sione cleaned chairs in the house.'

These modifiers also have a rigid placement, the same as their placement within non-incorporated nominal expressions.

Second, though these incorporated expressions can, to a degree, be phrasal, there is also a tight bond between the verb and the incorporated noun.

This can be shown from the behavior of the prenominal adjectives. This class of adjectives cannot incorporate, even though semantically-identical postnominal adjectives can. This is illustrated in (4a-b):

(4) a. *Na‘e **tō** **ki‘i** **manioke** ‘a
PAST plant small cassava ABS
Sione.
(name)
Intended: ‘Sione planted a small amount of cassava.’

b. Na‘e **tō** **manioke** **iiki** ‘a
PAST plant cassava small ABS
Sione.
(name)
‘Sione planted a small amount of cassava.’

Further evidence for this tight bond comes from nominalization – this evidence suggests that verb and incorporated noun should be treated as single lexical unit. Tongan verbs can be nominalized with the suffix -‘*anga*. Single verbs can nominalized, as in (5a), as well as verb-incorporated noun units, as in (5b).

(5) a. nofo-‘anga
dwell-NOM
‘dwelling place’

b. inu-kava-‘anga
drink-kava-NOM
‘place to drink kava’

This nominalization occurs just with the verb and incorporated noun – other parts of the incorporated expression may not be nominalized with verb and the incorporated noun.

Given that there are both phrasal and morphological aspects to noun incorporation in Tongan, how might Tongan noun incorporation be analyzed? I propose that the verb-incorporated noun unit is a single word, but it has two corresponding nodes – two instantiations – in the syntax, following the Lexical Sharing analysis developed by Wescoat (2002). Thus, a unit like *tō-manioke*, ‘plant cassava’, has a lexical entry like as in (6) (following the notation of Kim et al. (2004)):

(6)
$$\begin{bmatrix} \textit{word} \\ \text{PHON} \quad \langle \text{t}\bar{\text{o}}\text{-manioke} \rangle \\ \text{INSTS} \quad \left\langle \begin{bmatrix} \textit{atom} \\ \text{EXPON} \quad \boxed{5} \\ \text{SYN} \mid \text{HEAD} \quad \textit{verb} \end{bmatrix}, \begin{bmatrix} \textit{atom} \\ \text{EXPON} \quad \boxed{6} \\ \text{SYN} \mid \text{HEAD} \quad \textit{noun} \end{bmatrix} \right\rangle \end{bmatrix}$$

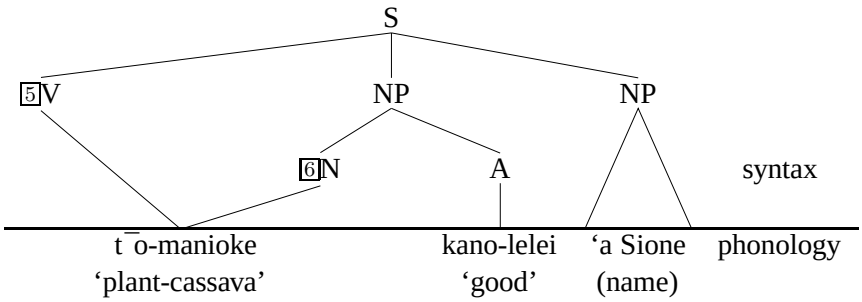
Lexical sharing allows the lexicon to create entries which instantiate more than one preterminal node – each preterminal is a syntactic *atom* (Wescoat, 2002). The entry in (6) shows that one word instantiates two atoms, intuitively, a verb and a noun.

Beyond this, the syntactic component of the language remains essentially unchanged, and in particular, the incorporated noun may head an NP with internal postnominal modifiers. The verb atom combines by the head-argument schema (Schema 3 of Pollard and Sag (1994)) and the postnominal modifers combine by the head-modifier schema (Schema 5 of Pollard and Sag (1994)), and structures like (7) on the following page result.

Such structures account for the configurational properties of noun incorporation in Tongan, but this analysis has yet to deal with the absolutive case marking in noun incorporation. First, let us recognize three types of nominal expression in Tongan: one that has both CASE and DET features, a type I will call *canonical*; one that just has DET features, a type I will call *determined*; and one that lacks both CASE and DET features, which I will call *bare*. These correspond to the phrase types KP (for ‘case phrase’), DP, and NP, respectively, in the analysis of incorporation proposed by Massam (2001) for Niuean, a closely-related language.

Given this typology, and the general patterns of case marking of Tongan, it seems reasonable to assume that Tongan has a case constraint, where, if there are two *canonical* NPs, ergative case is required of the less oblique and absolutive of the more oblique, as in (8):

(7)




(8) $\left[\text{ARG-ST} \left\langle \begin{bmatrix}\text{NP}_{canonical}\\ \text{CASE erg}\end{bmatrix}, \begin{bmatrix}\text{NP}_{canonical}\\ \text{CASE abs}\end{bmatrix} \right\rangle\right]$

If there are not two *canonical* NPs on the ARG-ST, then the first NP is constrained to be in the absolutive.

As the lexical sharing analysis is predicated on the lexical status of the verb-noun combination, it follows that no syntactic atoms may intervene in the syntactic structure between them. In other words, lexical sharing predicts that any prenominal atoms – such as case markers, determiners, and prenominal adjectives – must be absent. This accounts for the properties seen in examples (1)–(4) above without stipulation. As the NP in the syntax cannot host a case marker or determiner (such as the first NP in (7)), it is of type *bare*.

So, while the ARG-ST of transitive verbs is normally instantiated as in (9) below, in noun incorporation structures, this requirement for the second argument to be *bare* causes the default type *canonical* to be overridden and the ARG-ST to be as in (10).

(9) $\left[\text{ARG-ST} \left\langle \text{NP}_{canonical}, \text{NP}_{/canonical} \right\rangle\right]$

(10) $\left[\text{ARG-ST} \left\langle \text{NP}_{canonical}, \text{NP}_{bare} \right\rangle\right]$

With an ARG-ST like (10), incorporating verbs do not meet the case constraint given in (8), and thus must take their first argument in the absolutive. So, on this analysis, noun incorporation does not reduce the number of arguments that combine with the verb, but does reduce their case-marking status.

Further support from this view of case-marking comes from so-called middle objects – objects of verbs with low transitivity. These verbs have their external argument marked in the absolutive case and their second argument – the middle object – is marked only with a determiner, with no prenominal case marker at all, as shown below in (11):

(11) Na‘e heka ‘a Mele he hoosi.
PAST ride ABS (name) DET horse
‘Mary rode on the horse.’

Examples like (11) show that middle object verbs have ARG-STs as in (12).

(12) $\left[\text{ARG-ST} \left\langle \text{NP}_{canonical}, \text{NP}_{determined} \right\rangle\right]$

These structures also do not meet the case constraint in (8), so their ARG-ST initial argument also must be in the absolutive case, as (11) shows it to be.

A critical claim made by previous analyses of this kind of phenomenon in Tongan and related languages (Gerdts, 1998; Massam, 2001) is that the adjacency between the verb and incorporated noun (and its phrase) is coincidental. As the above paragraphs outlined, the Lexical Sharing analysis, in contrast, claims that it is not. Rather, the analysis claims that the lexicon enforces the adjacency, and constrains the syntactic properties (including the case marking) of the construction. Thus, a lexicalist analysis is a superior account of these data, since it offers better predictions about the configuration and the case-marking in noun incorporation in Tongan.

## References

C. Maxwell Churchward. 1953. *Tongan Grammar*. Oxford University Press, London.

Donna B. Gerdts. 1998. Incorporation. In Andrew Spencer and Arnold M. Zwicky, editors, *The Handbook of Morphology*, pages 84–100. Blackwell, Oxford.

Jong-Bok Kim, Peter Sells, and Michael T. Wescoat. 2004. Korean copular constructions: A lexical sharing approach. In M. Endo Hudson, Sun-Ah Jun, and Peter Sells, editors, *Proceedings of the 13th Japanese/Korean Linguistics Conference*, Stanford, California. CSLI Publications.

Diane Massam. 2001. Pseudo noun incorporation in Niuean. *Natural Language and Linguistic Theory*, 19:153–197.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Sara Thomas Rosen. 1989. Two types of noun incorporation: A lexical analysis. *Language*, 65:294–317.

Jeffrey T. Runner and Raúl Aranovich. 2003. Noun incorporation and rule interaction in the lexicon. In Stefan Müller, editor, *Proceedings of the HPSG-2003 Conference, Michigan State University, East Lansing*, pages 359–379. CSLI Publications. http://cslipublications.stanford.edu/HPSG/4/.

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

# Towards A Semantic Analysis of Argument/Oblique Alternations in HPSG

**John Beavers**
Department of Linguistics
Stanford University
Stanford, CA, 94305-2150
jbeavers@csli.stanford.edu

In this paper I outline a semantic analysis of argument/oblique alternations. I argue that when such alternations exhibit semantic contrasts it is always in terms of the relative number of entailments associated with the alternating participant. I sketch a framework for capturing these contrasts in HPSG, using the locative alternation as a case study:[1]

(1) a. John loaded the hay onto the wagon.
b. John loaded the wagon with the hay.

In (1a) the locatum is realized as a direct argument and in (1b) as an oblique, and vice versa for the location participant. The classic semantic observation (Anderson 1971) is that whichever participant is realized as direct object receives a "holistically affected" interpretation (all moved or all loaded up):

(2) a. John loaded the hay onto the wagon, leaving enough space for the grain.
b. #John loaded the wagon with the hay, leaving enough space for the grain.

(3) a. John loaded the wagon with the hay, with enough left over to fill the pick-up.
b. #John loaded the hay onto the wagon, with enough left over to fill the pick-up.

Only the oblique realizations are acceptable in a context where they are not holistically affected. Thus they are **underspecified** for holistic affectedness (i.e. they neither entail nor contradict it). Other properties, however, are **invariant**, e.g. one participant is always a location, the other a locatum, and both are always at least partially affected (loaded/moved). Other realization patterns that are morphosyntactically similar to (1) involve related but distinct differences in interpretation, as in (4).

(4) a. John cut his hand on the rock. (hand affected; rock not necessarily affected)
b. John cut the rock with his hand. (rock affected; hand not necessarily affected)

While the variants in (1) differ in holistic affectedness, (4) exhibits a contrast in simple affectedness. Otherwise, the morphosyntactic and semantic similarities suggest that (1) and (4) are two manifestations of one alternation where the exact contrasts are verb-specific (cf. Fillmore 1977, Dowty 1991).

While the locative alternation has been well studied (see Levin and Rappaport 1988, *inter alia*), few authors have observed that there is a general contrast between alternating direct arguments and obliques in terms of underspecificity (though see Ackerman and Moore 2001, which I discuss further below). For example, in the dative alternation (e.g. *Rich threw Barry the ball/the ball to Barry*) the recipient is invariably a goal (which the theme is intended to reach), but when it is realized as first object it is also an intended possessor, giving rise to the fact that inanimate locations realized as first objects must be construed of as capable of possession (e.g. the London office in *John sent London a package*; Green 1974). Likewise for the reciprocal alternation *The*

[1] This is part of a larger study based on a theory of thematic roles as sets of entailments, following primarily Dowty (1991). I use the term "entailment" in the sense of Dowty's (1989) "lexical entailments", i.e. properties a verb ascribes to an argument due to its role in the event, ignoring their ontological status as e.g. entailments vs. implicatures. See Beavers (to appear) for more details on the English data motivating this analysis and previous literature on the semantic basis of alternations.

*car and the truck collided/The car collided with the truck*, when both entities are realized as a conjoined subject both must be in motion but when one is realized as an oblique it is underspecified for motion. Thus an adequate analysis of alternations must capture the following generalization:

(5) Direct argument variants entail more about the alternating participant than oblique variants.

Previous HPSG analyses have generally failed to capture this, typically by not providing a rich enough semantics to capture the contrasts and not characterizing the argument/oblique contrast in a general way. For example, Koenig and Davis (2004) analyze English locative alternations in terms of UND(ERGOER) assignment. The entity linked to UND is always direct object, and the alternation arises from different choices of UND (resulting from different choices of KEY relations; see Kordoni 2002 for related HPSG work on Greek and Van Valin 2002 for a similar approach in Role and Reference Grammar). However, this does not directly capture the semantics of locative alternations since no specific entailments are associated with either variant. One could stipulate that the entity linked to UND must be associated with more entailments. However, this does not explain what those entailments are on a verb-by-verb basis, and also fails to generalize since recipients in the dative alternation are not necessarily linked to UND (e.g. Kordoni 2004 posits an additional macrorole) and in the reciprocal alternation there is not necessarily an UND feature at all (see also Beavers to appear for discussion of why analyses based on structured semantic representations are generally ill-suited to capture (5)).

Instead, I encode (5) in terms of thematic roles defined as sets of entailments as in Dowty (1989, 1991). For a verb $V$ describing situation $s$, the role a participant $x$ plays in $s$ is defined as a set of verb-specific entailments $R$, which I refer to as an **individual thematic role** (following Dowty 1989). Thus $R$ is the set of all things, from the very general to the quite specific, that $V$ says about $x$'s role in $s$. Individual thematic roles are related to one another in terms of **specificity**. For two individual thematic roles $R$ and $Q$, $R$ is more specific than $Q$ if $Q \subset R$. I characterize (5) in terms of thematic roles as in (6).

(6) **Morphosyntactic Alignment Principle (MAP):** When participant $x$ may be realized as either a direct or oblique argument of verb $V$, it bears role $R$ as a direct argument and role $Q$ as an oblique where $Q \subset R$.

However, (6) fails to explain *which* roles $R$ and $Q$ $x$ will bear for a given verb and alternation, i.e. it misses the generalization that the verb-specific contrasts cross-classify into more general types based on very general notions like degrees of affectedness. For instance, (1) exhibits a contrast in terms of holistic affectedness (however manifested for a given verb, e.g. completely loaded/moved for *load*, completely sprayed/covered for *spray*), whereas (4) exhibits a contrast in simple affectedness (manifested in different ways e.g. for *cut* vs. *break*).

A better solution would derive the contrasts for each verb in terms of a more limited and general notion of possible contrasts. Following Dowty (1989), I propose to do this in terms of smaller, more general sets of entailments called **thematic role types**. Thematic role types are universal sets of non-verb-specific entailments that cross-classify individual thematic roles in terms of properties such as affectedness, possession, motion, etc., relevant for argument linking.[2] For instance, for the alternations in (1) and (4) I propose the thematic role types in Table 1 on the following page (which are also relevant for other object alternations; see Beavers to appear).

Thematic role types form specificity contrasts just as individual thematic roles do, forming general hierarchies reflecting decreasing specificity:

(7) HOL. AFFECTED $\supset$ AFFECTED $\supset$ PARTICIPANT

The alternations of individual thematic roles in (1) and (4) can be described as **minimal contrasts in their thematic role types** along (7):

(8)

| **Role Type** | *load/spray* | *cut/break* |
|---|---|---|
| HOL. AFFECTED | DO | |
| | ↓ | |
| AFFECTED | OBL | DO |
| | | ↓ |
| PARTICIPANT | | OBL |

[2]The thematic role types I propose here are L-thematic roles in the sense of Dowty (1989), defined as linguistically significant intersections of individual thematic roles, i.e. subsets that many individual thematic roles share in common. In light of Dowty (1991) these could be defined instead as sets of proto-role entailments as in Beavers (to appear), though I ignore proto-roles here. Note that the term "type" here is not related to the HPSG notion of "type".

| Thematic Role Type | Example Individual Thematic Roles of this Type |
|---|---|
| HOLISTICALLY AFFECTED | Completely loaded or moved entity ($\text{DO}_{load}$) |
| AFFECTED | Loaded, moved entity ($\text{oblique}_{load}$), or cut entity ($\text{DO}_{cut}$) |
| PARTICIPANT | Entity not known to be affected ($\text{oblique}_{cut}$) |

Table 1: Example Thematic Role Types

This can be characterized via a function from individual thematic roles to individual thematic roles as in (9), by which we can reformulate (6) as in (10).

(9) For thematic role types $\tau_1$ and $\tau_2$, $\tau_1 \supset \tau_2$, forming a minimal thematic role type contrast, and for individual thematic role $R$ of type $\tau_1$, the role $Q = min(R)$ is the maximal subset of $R$ of type $\tau_2$.

(10) **MAP (Revised):** When participant $x$ may be realized as either a direct or oblique argument of verb $V$, it bears role $R$ as a direct argument and role $min(R)$ as an oblique.

For example, if *the wagon* in (1) has individual thematic role $\text{LOCATION}_{load}$ of type HOLISTICALLY AFFECTED as direct object, its role as an oblique is $min(\text{LOCATION}_{load})$ of type AFFECTED, which includes all the entailments in $\text{LOCATION}_{load}$ save those that make it type HOLISTICALLY AFFECTED rather than AFFECTED. To capture (10) in HPSG I first assume a feature ROLES in each verb's CONT value (assuming the MRS semantics of Copestake et al. 2003 but ignoring scoping-related features):

(11) *verb-mrs* $\Rightarrow$ *mrs* & $\left[\text{ROLES } \textit{set(set(entailments))}\right]$

ROLES defines the set of maximal individual thematic roles a verb licenses, i.e. the roles a verb will assign to its direct arguments. Each verb specifies on its RELS lists elementary predications of type *role-rel*, which attribute an individual thematic role in ROLES to a participant:

(12) *role-rel* $\Rightarrow$
*elementary-predication* & $\begin{bmatrix}\text{ARG1 } i\\ \text{ROLE } \textit{set(entailments(i))}\end{bmatrix}$

We can capture (10) as constraints on *v-lxm*, which for expository purposes I present in two parts. First is the linking of direct arguments to maximal roles, done simply by associating each NP argument directly with a role on the verb's ROLES list:[3]

(13) *v-lxm* $\Rightarrow$

$$\begin{bmatrix}\text{ARG-ST} & \left\langle \text{NP}_{i_1}, ..., \text{NP}_{i_n} \right\rangle \bigcirc \textit{list}\left(\textit{non-NP}\right)\\ \text{CONT} & \begin{bmatrix}\text{ROLES} \left\{ \boxed{R_1}, ..., \boxed{R_n} \right\} \cup \textit{set}\\ \text{RELS} \left\langle \begin{bmatrix}\textit{role-rel}\\ \text{ARG1 } i_1\\ \text{ROLE } \boxed{R_1}\end{bmatrix}, ..., \begin{bmatrix}\textit{role-rel}\\ \text{ARG1 } i_n\\ \text{ROLE } \boxed{R_n}\end{bmatrix} \right\rangle \bigcirc \textit{list}\end{bmatrix}\end{bmatrix}$$

The roles assigned to obliques are more complicated. Ideally, they are the output of $min$ for some role on ROLES. However, we also want to restrict which oblique markers occur in which alternations. Following Gawron (1986), Markantonatou and Sadler (1995), and Wechsler (1995) I assume that oblique markers are semantically contentful, contributing individual thematic roles that must be compatible with the role assigned by the verb. For example, the PPs relevant for (1) are given in (14).

(14) a.
$$\begin{bmatrix}\text{ORTH} \left\langle \textit{onto, the, wagon} \right\rangle\\ \text{CONT} \begin{bmatrix}\text{ROLES} \left\{ \text{LOCATION}_{goal} \right\}\\ \text{RELS} \left\langle \begin{bmatrix}\textit{wagon-rel}\\ \text{ARG1 } i\end{bmatrix} \right\rangle\end{bmatrix}\end{bmatrix}$$

b.
$$\begin{bmatrix}\text{ORTH} \left\langle \textit{with, the, hay} \right\rangle\\ \text{CONT} \begin{bmatrix}\text{ROLES} \left\{ \text{CAUSALLY-INTERMEDIATE} \right\}\\ \text{RELS} \left\langle \begin{bmatrix}\textit{hay-rel}\\ \text{ARG1 } i\end{bmatrix} \right\rangle\end{bmatrix}\end{bmatrix}$$

The PPs in (14) correspond to two potential arguments of *load*, where the individual thematic roles supplied by each preposition represent their inherent semantics. For locative prepositions the $\text{LOCATION}_{goal}$ role is the general set of entailments that define a participant as a locational goal (where I assume specific choices of locational prepositions, e.g. *on(to)*, *in(to)*, are pragmatically determined and not part of the thematic role per se). Following Croft (1991), I assume *with* assigns a role CAUSALLY-INTERMEDIATE, representing an entity that is causally intermediate in the event's force-dynamic structure, i.e. acted upon by the agent but force-dynamically antecedent to other participants.

[3] For the remainder of the document I ignore irrelevant features such as SS and LOC in the paths to the features of interest.

This role encompasses both locatums and instruments (see Levin and Rappaport 1988 on *with* as a "displaced theme" marker).

To ensure compatibility between the preposition's and verb's individual thematic roles, the latter must be a superset of the former. I encode this via a function $sup$, where $sup(P, Q) = Q$ if $P \subseteq Q$ and $\perp$ if $P \not\subseteq Q$.[4] The linking constraints are:[5]

(15)
$$\textit{v-lxm} \Rightarrow \Big/ \left[\begin{array}{l} \text{ARG-ST} \left\langle \begin{array}{c} \text{PP}_{j_1}\left[\text{ROLES}\left\{\boxed{P_1}\right\}\right], \\ \ldots, \\ \text{PP}_{j_m}\left[\text{ROLES}\left\{\boxed{P_m}\right\}\right] \end{array} \right\rangle \bigcirc \textit{list}\big(\textit{non-PP}\big) \\ \text{CONT} \left[\begin{array}{l} \text{ROLES}\left\{\boxed{Q_1}, \ldots, \boxed{Q_m}\right\} \cup \textit{set} \\ \text{RELS}\left\langle \begin{array}{l} \left[\begin{array}{l}\textit{role-rel}\\ \text{ARG1}\ j_1\\ \text{ROLE}\ \textit{min(sup(}\boxed{P_1}\textit{,}\boxed{Q_1}\textit{))}\end{array}\right], \\ \ldots, \\ \left[\begin{array}{l}\textit{role-rel}\\ \text{ARG1}\ j_m\\ \text{ROLE}\ \textit{min(sup(}\boxed{P_m}\textit{,}\boxed{Q_m}\textit{))}\end{array}\right] \end{array} \right\rangle \bigcirc \textit{list} \end{array}\right] \end{array}\right]$$

Thus for each PP in (15), its role is a subset of some role Q in the ROLES set of that verb (corresponding to a decrease in thematic role type) and a superset of the role P determined by the preposition:

(16)
$$\begin{array}{ccccc} \textbf{Role}_{Prep} & & \textbf{Actual Role} & & \textbf{Role}_{V} \\ P & \subseteq & min(sup(P, Q)) & \subset & Q \end{array}$$

All *load* need specify is its ARG-ST and a list of maximal roles (including a locatum and locational goal, both holistically affected). No explicit linking needs to be stated (though I stipulate subject linking since I am primarily concerned here with objects):

(17)
$$\left[\begin{array}{l} \text{ORTH}\left\langle \textit{load} \right\rangle \\ \text{ARG-ST}\left\langle \text{NP}_i, \text{NP}, \text{PP} \right\rangle \\ \text{CONT}\left[\begin{array}{l} \text{ROLES}\left\{\boxed{1}\text{LOADER}, \text{LOCATUM}_{load}, \text{LOCATION}_{load}\right\} \\ \text{RELS}\left\langle \left[\begin{array}{l}\textit{role-rel}\\ \text{ARG1}\ i\\ \text{ROLE}\ \boxed{1}\end{array}\right], \ldots \right\rangle \end{array}\right] \end{array}\right]$$

Although (17) stipulates few constraints, its output is restricted by the preposition inventory of English, yielding only two classes of head-complement structures, exemplified by (18) and (19):

(18)
$$\left[\begin{array}{l} \text{ORTH}\left\langle \textit{loaded, the wagon, with the hay} \right\rangle \\ \text{DTRS}\left\langle \text{V, NP}_j\text{, PP}_k\left[\text{ROLES}\left\{\boxed{1}\text{CAUSALLY-INTERMED.}\right\}\right] \right\rangle \\ \text{CONT}\left[\begin{array}{l} \text{ROLES}\left\{\ldots, \boxed{2}\text{LOCATUM}_{load}, \boxed{3}\text{LOCATION}_{load}\right\} \\ \text{RELS}\left\langle \ldots, \left[\begin{array}{l}\textit{role-rel}\\ \text{ARG1}\ j\\ \text{ROLE}\ \boxed{3}\end{array}\right], \left[\begin{array}{l}\textit{role-rel}\\ \text{ARG1}\ k\\ \text{ROLE}\ \textit{min(sup(}\boxed{1}\textit{,}\boxed{2}\textit{))}\end{array}\right] \right\rangle \end{array}\right] \end{array}\right]$$

(19)
$$\left[\begin{array}{l} \text{ORTH}\left\langle \textit{loaded, the hay, onto the wagon} \right\rangle \\ \text{DTRS}\left\langle \text{V, NP}_j\text{, PP}_k\left[\text{ROLES}\left\{\boxed{1}\text{LOCATION}_{goal}\right\}\right] \right\rangle \\ \text{CONT}\left[\begin{array}{l} \text{ROLES}\left\{\ldots, \boxed{2}\text{LOCATUM}_{load}, \boxed{3}\text{LOCATION}_{load}\right\} \\ \text{RELS}\left\langle \ldots, \left[\begin{array}{l}\textit{role-rel}\\ \text{ARG1}\ j\\ \text{ROLE}\ \boxed{2}\end{array}\right], \left[\begin{array}{l}\textit{role-rel}\\ \text{ARG1}\ k\\ \text{ROLE}\ \textit{min(sup(}\boxed{1}\textit{,}\boxed{3}\textit{))}\end{array}\right] \right\rangle \end{array}\right] \end{array}\right]$$

Acceptable structures similar to (19) could also be built with other acceptable locational goal markers (e.g. *in(to)*), while presumably *with* is the only general CAUSALLY-INTERMEDIATE marker in English (*by*, *via*, etc. mark more specific means/manner roles that are not subsets of *load*'s $\text{LOCATUM}_{load}$ role). Any other prepositions, or different linking with the same prepositions, would result in a unification failure. Note that (following Markantonatou and Sadler 1995) no polysemy of the verb is required. Different variants arise from the thematic roles licensed by the verb and the inherent roles of the oblique markers, maintaining the (implicit or explicit) assumption of much recent work cited above that alternations are determined by the lexical semantics of the verbs and the relevant oblique markers.

This approach has two advantages over previous work discussed above. First, the semantics-to-morphosyntax mapping is encoded without intermediate levels of semantic structure such as predicate decompositions or structured elementary predications as in Koenig and Davis (2004) (see Beavers to appear for more discussion). Second, by basing the relevant generalizations on verb-specific individual thematic roles organized by types, it directly links the idiosyncratic semantics of each verb to the more general contrasts alternations exhibits across verbs.

Note that the generalization in (5) differs from the LFG approach in Ackerman and Moore (2001).

[4]The function $sup$ is only for presentational convenience. It simply serves to coidentify every entailment of the preposition's role with an entailment in the verb's role. Spelling this out explicitly reduces the readability of the AVMs.

[5]This constraint is English specific. For a language like Finnish with more elaborate case morphology (13) and (15) could be trivially elaborated by including a distinction between NPs which have a CASE feature with a structural case value vs. those with an oblique case value (which pattern like PPs). Note that the constraints in (15) are defaults; a particular verb can override the general linking of obliques to certain classes of roles if it idiosyncratically selects a particular oblique marker.

Ackerman and Moore propose that obliques deviate more than direct arguments from Dowty's (1991) proto-agent/patient roles ("less prototypical" in their PARADIGMATIC ARGUMENT SELECTION PRINCIPLE). On the approach outlined here, "less prototypical" is given a more specific interpretation as underspecificity of thematic role entailments, making a stronger claim. Furthermore, my approach, though defining thematic roles as sets of entailments, is not wedded to proto-roles and thus may capture a broader set of generalizations. For example, it is not a priori obvious that recipient realization in general needs to be modeled using proto-roles, even if the general principle in (5) nonetheless governs the semantic contrasts the dative alternation exhibits.

However, the analysis presented here is by no means complete; it is instead intended as a proof-of-concept for an entailment-based approach to alternations in HPSG. I have ignored several issues here, for instance verbs that do not undergo alternations (e.g. *put* and *fill* are non-alternating locative verbs) and alternations that exhibit no semantic contrast (e.g. *John blamed Mary for his problems/blamed his problems on Mary*). Likewise I largely ignore Dowty's proto-role theory, which could provide a more principled view of subject/object selection within which this framework could be situated (though see Davis 2001 for a critique of Dowty's approach). For more on these issues, see Beavers to appear. Finally, I make no predictions about which argument structures a given verb may have (having assumed that all locative verbs take one PP and two NP arguments). Presumably this is derivable from some of the same semantic factors discussed above, an issue I leave to future investigation.

## References


Farrell Ackerman and John Moore. 2001. *Proto-Properties and Grammatical Encoding*. CSLI Publications, Stanford, CA.

Stephen R. Anderson. 1971. On the role of deep structure in semantic interpretation. *Foundations of Language*, 7(3):387–396.

John Beavers. To appear. Thematic role specifi city and argument/oblique alternations in English. In *Proceedings of WECOL 2004*, University of Southern California, Los Angeles.

Ann Copestake, Dan Flickinger, Ivan Sag, and Carl Pollard. 2003. Minimal recursion semantics: An introduction. `http://www.cl.cam.ac.uk/~acc10/papers/newmrs.ps`.

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

Anthony Davis. 2001. *Linking by Types in the Hierarchical Lexicon*. CSLI Publications, Stanford, CA.

David Dowty. 1989. On the semantic content of the notion 'thematic role'. In Gennaro Chierchia, Barbara H. Partee, and Raymond Turner, editors, *Properties, Types, and Meaning*. Kluwer, Dordrecht.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Charles J. Fillmore. 1977. The case for case reopened. In Peter Cole and Jerrold M. Sadock, editors, *Grammatical Relations*, pages 59–82. Academic Press, New York.

Jean Mark Gawron. 1986. Situations and prepositions. *Linguistics and Philosophy*, 9:327–382.

Georgia Green. 1974. *Semantic and Syntactic Regularity*. Indiana University Press, Bloomington, IN.

Jean-Pierre Koenig and Anthony Davis. 2004. The KEY to lexical semantic representations. Unpublished ms., the State University of New York at Buffalo.

Valia Kordoni. 2002. Valence alternations in Greek: an MRS analysis. In Jong-Bok Kim and Stephen Wechsler, editors, *Proceedings of the HPSG 2002 Conference*, pages 129–146, CSLI Publications, Stanford, CA.

Valia Kordoni. 2004. Between shifts and alternations: Ditransitive constructions. In Sefan Müller, editor, *Proceedings of the HPSG 2004 Conference*, pages 151–167, Stanford, CA. CSLI Publications.

Beth Levin and Malka Rappaport. 1988. What to do with $\theta$-roles. In Wendy Wilkins, editor, *Thematic Relations*. Academic Press, San Diego, CA.

Stella Markantonatou and Louisa Sadler. 1995. Linking indirect arguments and verb alternations in English. In Francis Corblin, Danièle Godard, and Jean-Marie Marandin, editors, *Empirical Issues in Formal Syntax and Semantics*, pages 103–125. Peter Lang, Berne.

Robert D. Van Valin. 2002. The role and reference grammar analysis of three-place predicates. Unpublished ms., The State University of New York at Buffalo.

Stephen Wechsler. 1995. *The Semantic Basis of Argument Structure*. CSLI Publications, Stanford, CA.

# Integrating Discourse Information in Grammar

**Núria Bertomeu**
Department of Computational Linguistics
Saarland University
Germany
bertomeu@coli.uni-sb.de

**Valia Kordoni**
Department of Computational Linguistics
Saarland University
Germany
kordoni@coli.uni-sb.de

## 1 Introduction

The grammar of a language stipulates what can be uttered and what not. Fragments are felicitous only within a certain context, that is, it is the surrounding context which makes an stand-alone constituent grammatical or ungrammatical. The grammar should, thus, contain contextual constraints, as these are crucial for constituents to be considered root sentences.

In Schlangen's approach (Schlangen, 2003) to the resolution of elliptical fragments, however, it is the context within the discourse model which decides whether a fragment is felicitous, and there is no contextual information in the construction type licensing the fragment. Cooper and Ginzburg (2003), on the other hand, not only include contextual constraints in the type accounting for fragments, but also stipulate in it how the full meaning of the fragment should be recovered from the source.

But there is sometimes some degree of uncertainty about how a fragment should be resolved, especially in the cases where there is no explicit source, like in (1), or in the cases where some information from the source is implicitely overridden by the fragment and it is knowledge which tells us this, like in (2).

(1) > Has Anastacia released any new CDs in the last year?
- Yes, "Left outside alone".
> Any prizes? / > Prizes.

(2) > What is planned for the opening ceremony of the Olympic Games in Beijing?
- Lots of concerts.
> And for the Football World Championship?

Departing from the underspecified semantics of fragments proposed by Schlangen (2003), the main aim of this paper is to constrain the types licensing fragments with information from the preceding discourse. This information must, however, be minimal, since in the resolution process sometimes other extra-linguistic sources of information come into play. Thus, we consider the resolution process to be dependent not only on the grammar, but argue for the need that the grammar ensures that utterances can only be uttered/interpreted given a certain context.

On the other hand, current approaches to the syntax of fragments consider them to be headed, that is, the constituent provided by the fragment is the head-daughter of the sentence. This constituent is raised to a sentence and the GHFP (Ginzburg and Sag, 2001) is, thus, overridden. We will also argue, considering cases of gapping like the one shown in (3), in favour of treating the remnant constituents as non-head-daughters and propose an alternative analysis.

(3) > When did 2-Pac release "All eyez on me"?
> (And) Michael Jackson "Thriller"?

## 2 Contextual Information in the Grammar

The interpretation of fragments is bound to the context. Sometimes the context is just the situation of utterance, like when uttering the following sentence in a restaurant:

(4) A coffee, please.

But more often the contextual information needed for the interpretation of the fragment is in the preceding discourse. This information can be a clause, in which case the resolution involves substitution of a constituents from the sentence by the constituents provided in the fragment, or extension of the sentence by a modifier contributed by the fragment. We will call this type *antecedent ellipsis*. But the anchor within the context can also be some salient entities somehow related to the content contributed by the fragment, like in (1). We will call this type *anchor ellipsis*. To resolve it one has to identify the anchor within the context and infer the relation holding between it and the remnant. Finally, there is one type of ellipsis with a partial linguistic antecedent, like in (5), where the fragment is to be interpreted as a modifier of the antecedent. However, we only know that 'homework' is a reason for not coming but not exactly why[1].

(5) I cannot come. Homework.

Taking into account short answers, clarification requests and sluices, Ginzburg and Cooper (2003) identify the antecedent of the ellipsis with the Maximal Question under Discussion, that is, the question which is currently being discussed. If we take into consideration elliptical questions like in the previous examples we will see that the antecedent cannot be the Maximal Question under Discussion, since the second question opens a new issue which is not dependent on the previous one. At least in those examples, this means that the DP (dialogue participant) has accepted the previous answer, downdating it, thus, from QUD.

In Bertomeu (work in progress) a discourse model is presented which combines a discourse-record containing different level representations of the previous utterances, and their degrees of activation in memory, with plans about actions. Those utterances whose syntactic structure is still available in memory are still accessible as antecedents, as well as those issues which remain open in the conversation. Access to the latter is regulated by the action plan. This model also makes use of a salience ranking to find entities in the context to which the remnant stands in some relation for the cases of *anchor ellipsis*, and of scriptal knowledge for cases of total absence of linguistic source. In this model a fragment can have as antecedent what in (Cooper and Ginzburg, 2003) is called the Maximal Question under Discussion if it is an answer to the question, or a correction, or subquestion of it. Otherwise the fragment can reuse syntactic structure in memory from some previous utterance [2] .

However, sometimes it is not straightforward to decide what counts as an antecedent or anchor, and it may be that a fragment can be resolved upon several of them, resulting in ambiguity. It may also be that there is ambiguity with respect to parallelism, that is, when the fragment shares syntactic and semantic features with more than one element in the source and there is uncertainty about which role it is intended to fill. Finally, adjuncts may not be retained in the resolution as shown in (2). We believe that taking this kind of decisions is the task of a pragmatic module rather than that of grammar, since other sources of information like inferences about goals, knowledge of the world, etc., may play a crucial role. That is why we reject the view that the grammar contains information about how the fragment must be resolved. However, the grammar can include information about the preceding discourse. This permits to restrict well-formed fragments to those which are somehow bound to the preceding discourse context. Without such a constraint every constituent could be raised to a sentence.

Schlangen (2003) assigns an underspecified semantics to fragments using the MRS formalism (Copestake et al., 2001). Fragments have as content a subtype of the type *message*. That is, they have the semantics of a main clause, but the relation instantiating the main event is unknown and the only information available is that the stand-alone constituent instantiates an argument of this relation, which one that argument is remains unknown. We will adopt the semantics proposed by Schlangen (2003), but will introduce another element: contextual constraints, for the reasons argued above. In order to do this, we will introduce the feature ANTECEDENT, a subfeature of CONTEXT, which takes

[1] See (Alcántara and Bertomeu, 2005) for a corpus study of ellipsis in spontaneous spoken language supporting this classification and providing quantitative distributional data.

[2] See (Alcántara and Bertomeu, 2005) for data regarding the availability as antecedents of issues not currently under discussion.

as value an object of type MRS. Our notion of antecedent is a very general one and embraces all possible sources independently of the relation in which they stand to the fragment, i.e. Maximal Question under Discussion, a recently mentioned fact, etc. In section 4 we explain how to constrain by means of this feature the well-formedness of fragments of the both kinds explained above. But first let's turn briefly to speak about the syntax of fragments.

## 3 The Syntax of Fragments.

Regarding the syntax of fragments, both Cooper and Ginzburg (2003) and Schlangen (2003) make the type *head-frag(ment)-phrase* inherit from *h(eade)d-ph(rase)*. They also consider that the stand-alone constituent which constitutes the fragment is the head-daughter of it. The GHFP, which states that mother and daughter must share values for the feature HEAD by default, is, thus, overridden. This is, however, problematic when we want to account for fragments formed by more than one constituent independent from each other, like those shown in (3). Upon which reasons can we decide here which constituent is the head-daughter?

Gregory and Lappin (1997), on the other hand, propose an analysis of intrasentential *gapping* as having a phonetically-null head-daughter. The remnants are, thus, non-head daughters.

A lot of data from German and English suggests that in elliptical constructions it is namely the head that falls away. For example:

(6) My flat is 30 $m^2$ and the neighbour's 40.

Also in the psycholinguistic literature it has been claimed that the most psychologically plausible parsing mechanism is left-corner parsing (Crocker, 1999). This implies that the human parser already begins to build structure as soon as it encounters a new item. For fragments this would mean that the parser analyses remnants as arguments or adjuncts and posits an empty head which is then semantically filled when resolving the fragment. This is less costly than analysing the constituent provided in the fragment as the head and then reanalysing when a sister or the real semantic head is encountered. From the point of view of the syntax-semantics interface it is also desirable that there is parallelism betwen the syntactic and semantic structures, that is, that the semantic heads correspond to the syntactic heads.

For the reasons just mentioned, we argue here for treating remnants of fragments as non-head daughters. However, it remains here an open research question whether the type accounting for fragments should inherit from the type *non-h(eade)d-ph(rase)* in the type hierarchy which states that the phrase doesn't have a head-daughter, but does have non-head-daugthers, or, whether it should inherit from *h(eade)d-ph(rase)* and the head-daughter be phonetically empty.

## 4 Integrating Discourse Information.

We will depart from Schlangen's type hierarchy of fragments but will add two new dimensions: one to account for contextual dependency, **res(olution)-type**, and one to account for adjunct fragments, **frag(ment)-adj(unct)-type**. Both dimensions inherit from the most general type *frag(ment)*. The **res-type** dimension has two subtypes, the *ant(ecedent)-frag(ment)* type and the *anchor-frag(ment)* type. In Figure 1 the representation of *ant-frag* is shown. The type states that one or more constituents, non-head-daughters, are raised to a sentence. The value for the feature HEAD is a finite verb. The value for the feature M(AIN-)C(LAUSE) is positive. The mother gets the REL(ATION)S and H(ANDLE)-CONSTRAINTS from the construction constraints and from the daughters. The G(LOBAL-)TOP has the same value as the label of the elementary predicate containing the message type and this, in turn, has as value for the feature SOA a handle which is *geq*[3] with the label of a soa. The soa's index, in turn, is the main index of the sentence. For the moment we only say that there is at least a non-head-daugther, leaving for the other dimensions to specify its category and function in the sentence. Our contribution at this point is to state that the feature C(ON)T(E)XT contains a subfeature ANTECEDENT which has as value a semantic object of type *mrs* containing at least one elementary predicate. We use the feature REL(ATIO)N as in (Sag and Polard, 1994) and coindex the values of it for the soa-relation and the elementary predicate in the antecedent. We choose to represent the relation with

[3] Greater or equal. See (Schlangen, 2003).

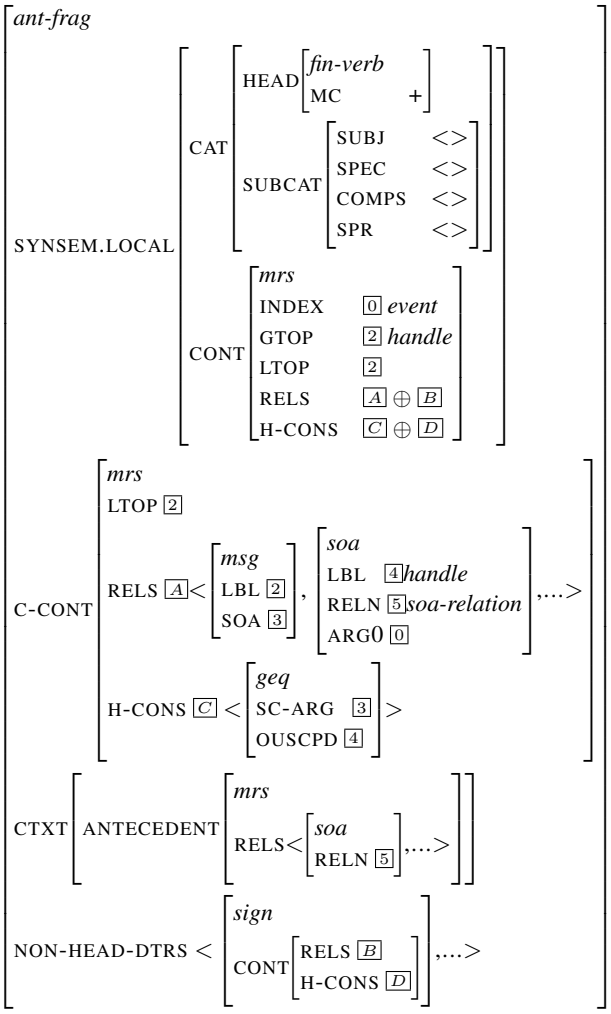



Figure 1: General type for fragments with antecedent

a feature instead of the type of the elementary predicate because this allows to say that both relations are of the same type, without claiming that they are the same event and have the same arguments. Informally, what this type states is that there must be an elementary predicate in the discourse record upon which the fragment is resolved. However, it can be the same event (understood as index) or not. We don't say anything about the arguments because of the possible ambiguity mentioned above.

In Figure 2 the type *anchor-nm-np-int-frag* is shown which inherits constraints from the types *anchor-frag(ment)*, *n(on)m(odified)-frag*, *n(ominal)p(hrase)-frag* and *int(errogative)-frag*. This type accounts for fragments like the one in (1), where the entity provided by the fragment is

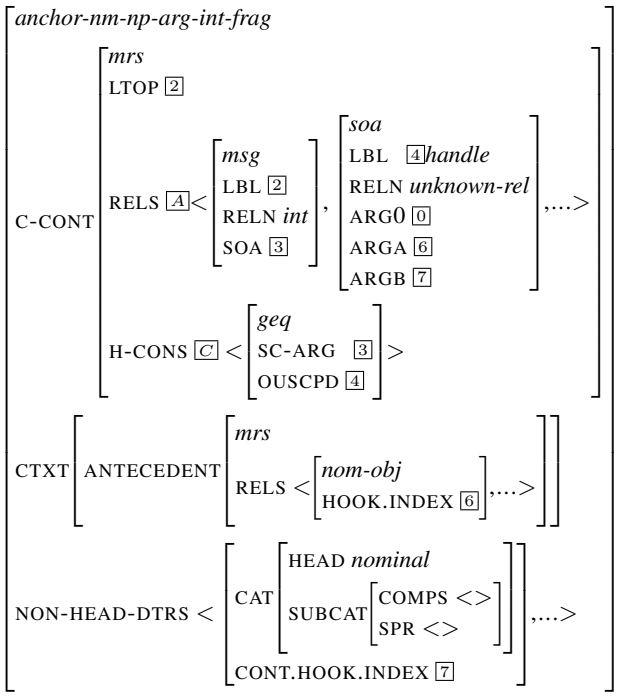



Figure 2: Non-modified interrogative fragment without antecedent (anchor)[5]

related to some salient entity in the discourse record by a relation which must be inferred. Again our contribution to this type is to state that the context provides an entity which, as the one provided by the fragment, is an argument of an unknown relation, leaving underspecified which one. The main characteristic of the dimension **adj(unct)-frag(ment)-type** is that the fragment modifies the soa-relation. This is expressed by the adjunct taking as value for ARG1 the index of the soa-relation. Scopal adverbs will have as value for ARG1 a *handle*, allowing, thus, modals to scope over them. However, there are cases where the fragments do not modify directly the soa-relation like in (5), where the NP provided by the fragment is contained in a modifying PP or subordinated clause. This can be accounted for with the type *ant(ecedent)-n(on)m(odified)-n(oun)p(hrase)-adj(unct)-decl(arative)-cl(ause)*, which is shown in Figure 3. A new semantic object, *undersp(ecified)-mod(ifier)* is added to the type hierachy to account for modifieres in an underspecified way.

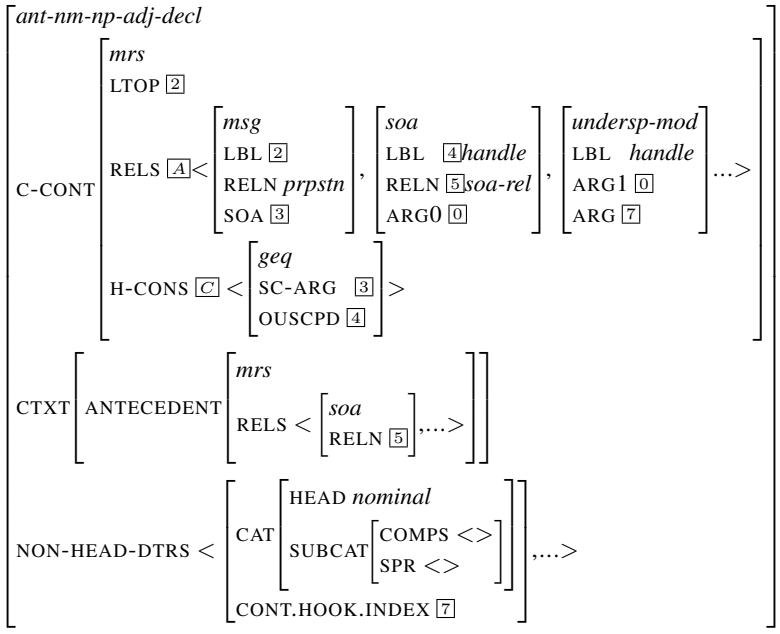



Figure 3: Non-modified adjunct fragment clause with partial antecedent

## 5 Conclusion

In this paper we have proposed a way of integrating discourse information in the grammar. This is an important issue for a theory like HPSG where the representation of a sign contains information from all linguistic levels, assuming, thus, a wider notion of grammar which includes pragmatics.

A very general notion of antecedent is used to constrain the well-formedness of fragments. This approach has advantages over Cooper and Ginzburg's approach (Cooper and Ginzburg, 2003) because the notion of antecedent is wider than the one of Maximal Question under Discussion, and because no constraints fully determine how the fragment must be resolved. Consequently, it covers a much wider range of fragment types. It also has advantages over Schlangen's approach (Schlangen, 2003) in that it reduces felicitous fragments to those which are somehow bound to the context and this happens already in the grammar.

We also have proposed to analyse the remnants of fragments as non-head-daughters, following Howard and Lappin's analysis of intrasentential gapping (Gregory and Lappin, 1997). We believe this is in accordance with psycholinguistically more plausible parsing techniques.

## References

Manuel Alcántara and Núria Bertomeu. 2005. Ellipsis in Spontaneous Spoken Language. In *Workshop on Cross-Modular Approaches to Ellipsis at ESSLLI 2005*.

Robin Cooper and Jonathan Ginzburg. 2003. Clarification, ellipsis, and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, 27(3).

Dan Flickinger, Carl Pollard, Ivan A. Sag and Ann Copestake. 2001. Minimal Recursion Semantics: An introduction. *Language and Computation*, 1(3):1–47.

Matthew W. Crocker. 1999. Mechanisms for sentence processing. In Garrod and Pickering, editors, *Language Processing*. Psychology Press.

Jonathan Ginzburg and Ivan A. Sag. 2001. *Interrogative investigations, the form meaning and use of English interrogatives*. CSLI publications.

Howard Gregory and Shalom Lappin. 1997. A computational model of ellipsis. In *Proceedings of the Conference on Formal Grammar, ESSLLI 1997*, Aix-en-Provence.

Ivan A. Sag and Carl Polard. 1994. *Head-driven Phrase Structure Grammar*. University Chicago Press.

David Schlangen. 2003. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, Institute for Communication and Collaborative Systems, School of Informatics, University of Edinburgh.

# The dual nature of Tswana infinitive forms

**Denis Creissels**
Départment de linguistique
Université Lyon-2 Louis Lumière
Lyon, France
Denis.Creissels@univ-lyon2.fr

**Danièle Godard**
CNRS, Laboratoire de Linguistique Formelle
Université Paris 7, UFRL
Paris 05, France
Daniele.godard@linguist.jussieu.fr

## 1 Introduction

The infinitive in Tswana (a Bantu language) has two types of occurrences, verbal (1) and nominal (2) as shown by the presence of the genitive:[1]

(1) Mosadi o apere mosese
1-woman S3:1-put.on-PFT-V 3-dress
o montle [go ya moletlong]
3.LK 3-pretty INF-go-V 3-fair
'The woman has put on a pretty dress to go to the fair'

(2) Ga ke rate [go nwa bojalwa
NEG-S1S-like-V INF-drink-V 14-beer
ga ba sadi]
15-GEN-2-woman
'I don't like the fact that women drink beer'

We show that the hybrid nature of the Tswana infinitive can be captured nicely in the mixed category approach proposed in Malouf (2000), in spite of a clear incompatibility between most environments in which the two types occur. In fact, we claim that, with its hybrid morphology, the Tswana infinitive bears witness to the (grammatical) reality of mixed categories.

1 We use the accepted orthography; the glosses make it clear what the linguistic segmentation is. The following abbreviations are used: APPL=applicative; CAUS= causative; DEM= demonstrative; FUT=future; GEN= genitive; INF=infinitive; LK= linker; LOC= locative; NEG=negative; O1S= 1pSg object agreeement index, etc. O3:X= 3rdp index agreeing with NCLASS X; PFT= perfect; POT= potential; PRO1S= 1stpSg pronoun, etc.; PRO3:X= 3rdp pronoun, agreeing with NCLASS X; S1S= 1stp subject agreement index, etc.; S3:X= 3rdp subject index, agreeing with NCLASS X; V= final vowel.

## 2 Common properties of the two occurrence types

The two uses of the infinitive share the following properties:

(i) the infinitive has the same morphological variations as a non infinitive V, in both of its occurrence types; it bears tense-aspect markers (TAM), which encompass (a) a VFORM value (which is, leaving aside the infinitive, a choice in {indicative, subjunctive, imperative, relative, circumstantial, sequential1, sequential2}, (b) a TENSE value ({present, perfect, fut., potential, continuative} relevant for indic., circ., relative, inf.), (c) a POLarity (with two values, see NEG in (2)). The infinitive, which is incompatible with the above VFORM values, has the same tense and polarity variations as a finite indic. verb (except for a few combinations). In addition, the final vowel (noted V) depends on the TAM value, as it does with other V forms.

(3)a. [Go kasebale lokwalo lo]
INF-POT-NEG-read-V 11-letter 11.DEM
gago monate
NEG-S3:15 3-pleasant-thing
'It is not pleasant to be unable to read this letter'

b. Ke gakgamalela [go kasebue
S1S-wonder-APPL-V INF-POT-NEG-speak-V
Setswana gabone]
7-tswana 15.GEN-PRO3:2
'I wonder at their inability to speak Tswana'

(ii) the infinitive bears the prefix *go-* , which is the noun class prefix 15. In nominal uses, the dependents show class agreement with this prefix, in

the same way as they do with common nouns (see (5)).

(iii) the subject cannot be realized; the infinitive does not contain the personal subject agreement index which is present on finite forms.

(4)a. O roga batho jalo
S2S-insult-V 2-person this way
'You insult people this way'
b. [Go roga batho jalo]
INF-insult-V 2-person this-way
go tlaa tshwarisa mapodisi
S3:15-FUT-O2S-catch-CAUS-V 6-policeman
'Insulting people this way will make the police arrest you'
c.*[Go o roga batho jalo]
INF-2S2-insult-V 2-person this-way

(iv) the complements of the infinitival are the same as those of the corresponding finite forms. Thus, it can have an object NP (or an adverb such as *jalo*) in both uses (2), (4b).

The last property is usual with hybrid forms such as (verbal) gerunds, but the combination of the other ones is noteworthy. It is not expected that a fully tensed form fails to combine with a subject. In fact, it goes against the universal deverbalization hierarchy proposed by Croft (1991), cited in Malouf (2000, 96). We make the hypothesis that there is a clash between the obligatory presence of a subject agreement index on the verb whose subject is realized, and the fact that *go-* occupies this slot.

## 3 The contrast between nominal and verbal infinitive

The properties of the two occurrence types are organized in two different sets. We have the following correlations.

### 3.1 Nominal infinitive

(i) the infinitive can take a number of dependents characterizing nominals (demonstrative, genitive, adjective, relative clause), in addition to the same complements as the corresponding finite verb. They show class agreement with the infinitive, as do dependents on the noun. The genitive bears the N class 15 in (5b), just as the genitive in (5a) bears the N class 6.

(5)a. madi a basadi
6-money 6.GEN-2-woman
'the women's money'
b. [go bua Setswana ga Lekgoa le]
INF-speak-V 7-tswana 15. GEN-5-euro. 5.DEM
'the fact that this European speaks Tswana'

(ii) an applicative verb has one more complement than the corresponding intransitive verb. If a psychological V such as 'to wonder' takes as its complement an infinitive phrase containing a nominal dependent denoting the source, the applicative morpheme (*-el-*) is obligatory:

(6)a. O gakgamalela bopelokgale
S3:1-wonder-APPL-V 14-courage
jwa mosimanee / *O gakgamala ...
14.GEN-1-boy
'He marvels at the boy's courage'
b. O gakgamalela [go bua Setswana
S3:1-wonder-APPL-V INF-speak-V 7-tswana
ga Lekgoa le] / * O gakgamala ...
15.GEN-5-european 5.DEM
'He marvels at the fact that this European speaks Tswana'

(iii) the argument corresponding to the subject of the fin V is unrealized or realized as a genitive.

(iv) the object NP cannot be separated from the V by an adverb in Tswana, nor can the infinitive in its nominal use.

(7)a. Ke itse monna yo sentle
S1S-know-V 1-man 1.DEM 7-good
/ *Ke itse sentle monna yo
'I know this man well'
b. O rata [go letsa
3S:1-like-V INF-weep.CAUS-V
katara mo ga gago]
9.guitar 15.DEM 15.GEN-PRO2S
c. *O rata thata [go letsa katara mo ga gago]
'He likes (a lot) for you to play the guitar'

(v) they have all the functions of NP (e.g. they can be locative).

(vi) they can be anaphorized like NP; they can co-occur with an object agreement on the V.

### 3.2 Verbal infinitive

(i') there is no nominal dependent.

(ii') an intransitive psychological verb can add an infinitive complement with a source interpretation without being applicative:

(8) O gakgamala [go utlwa Lakgoa
S3:1-wonder-V INF-hear-V 5-european
le mmuisa ka Setswana]
S3:5-O3:1-speak-CAUS-V prep 7-tswana
'He marvels at hearing this Euro. speak T'

(iii') the subject is controlled by, or identified with (raising predicates are common) an NP in the main sentence.

(iv') the infinitive can be separated from the V by an adverb.

(9) O rata thata [go letsa katara]
3S:1-like-V a-lot INF-weep.CAUS-V 9.guitar
'He likes a lot to play the guitar'

(v') in some environments, they alternate with full CP (introduced by comprs such as *gore*).

(vi') they are not anaphorized by pronominals (only by an adverb such as *jalo* 'thus'), and do not give rise to object agreement on the V.

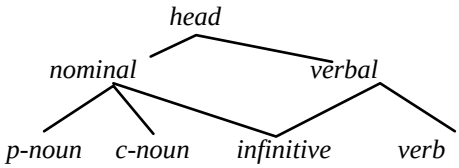
## 4 An analysis in the mixed category approach

### 4.1 Why not a lexical rule (LR)?

Given the clear contrast between the two uses, it is tempting to propose a LR that changes the category of the word, keeping its morphological skeleton, and its argument structure (10). There are at least two morphologically related problems with this analysis. First, the LR is word-to-lexeme, since the output is a lexeme, and the input is a word (it is fully inflected), which would be an isolated case in Tswana. Second, it cannot account for the presence on the verb (the verbal use of the inf) of the N class prefix. Since the Tswana infinitive bears on its morphological sleeve the evidence that such categories exist, we turn to a mixed category analysis.

### 4.2 The lexicon

(i) HEAD value: following Malouf (2000), we propose a mixed category. The partial hierarchy of HEAD values is as follows:




(11) *verbal* => [HEAD|TAM [*tam*, VFORM *vform*, (TENSE *tense*), POL *pol*]]

(12) *nominal* => [HEAD|NCLASS *noun-class*]

Hence, the Tswana infinitive has both a VFORM value and a NCLASS value.

(ii) lexeme and word: the lexeme is defined as being [HEAD *verbal*]; this underspecified value is resolved in the word either as *infinitive*, or as *verb*. Words whose HEAD is infinitive also inherit the feature [NCLASS 15] from *nominal*. They inherit the content (relation) as well as the argument structure from the verbal lexeme.

(iii) morphology: nothing prevents a realizational approach to the sequence given in the description of the LR above (in favor of it, note the dependency of the final vowel on the tam value, and the morphological alternation between the subject index and the nclass prefix *go-*). We use 'base' rather than 'root', since the base, but not the root, can include derivational morphemes (at least in certain traditions), passive, applicative and causative morphemes, in this case. The I-FORM values must take the argument structure into account in order to integrate the possible agreement indices for the complements. See (13).

### 4.3 The constructions

This infinitive can head two constructions, differentiated by their semantics. This accounts for the fact that they do not always alternate, even in the absence of a constraint coming from the environment; this is the case in (1), the verbal infve goal phrase, where one has to use a CP if the subject is realized: the required semantics (presumably 'outcome', a subtype of 'message', Ginzburg and Sag 2000) is incompatible with the presence of nominal dependents. It also accounts for speakers' intuition that, if we have an applicative instead of an intransitive verb in (8), the interpretation is different (glossed by 'the fact that').

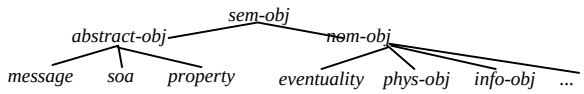
The internal structure of the InfveP in all its uses is 'flat': (i) there is no specifier (an N with its nclass prefix is a full NP); (ii) apart from the adja-

cency of the object with the head which is a general property in Tswana (7), the complements inherited from the lexeme and the nominal dependents can scramble (14), although there is at least a tendency for the nominal dependents to be ordered among themselves (dem< gen< adj< RC). Such 'scrambling' clearly argues against an NP-over-VP analysis, proposed by Mugane (2003) for the related language Kikuyu.

(14) [Go tsena gagwe mo ofising]
INF-enter-V 15-GEN-PRO3:1 in 9.office-LOC
'Her entrance into the office'

We assume that nominal dependents (including the genitive) are [MOD $\boxed{1}$ , $\boxed{1}$ ≥ *nominal* ] (see the above hierarchy of heads, and Sag 2003 for the use of the lattice hierarchy in the formulation of the constraint). Since the infinitive inherits the arguments of the verbal lexeme, we can have, for instance, both an object NP and a genitive (2), or an object NP and an adjective (*go letsa katara mo gontle*, lit. a nice playing the guitar).

We transpose on the *head-comps-cx* the distinction proposed for words in Bouma et al. (2001) between the argument structure defined on the lexeme, and an extended list of dependents. In addition, the two more specific infve-head-comp-cx (coalesced here using an alternative content value) specify the content, relying on the following (partial) hierarchy of sem-objects (where 'message' and 'soa' are as in Ginzburg and Sag 2000):




Fore the constraints on the *infinitive-head-complements-cx*, see (15) and (16).

While an infveP with no nominal dependent is free to have either an *abstract-obj* or a *nom-obj* content, the others must have a content of type nom-obj. The relation which is the content of the verbal lexeme is a building block for two different content types at the level of the construction. 'eventuality' is the primary nom-obj compatible with this relation (this implies that, if an eventuality is associated with the verb-word in an event-based semantics for the S, it may not denote exactly the same object, see Asher 1993). However, the interpretation of the nominal infveP seems to be able to shift towards 'manner of doing sth', as do IE nominalizations. Only nom-objects seem to be anaphorized, or represented by an agreement index on the verb.

Predicates select the so-called verbal or nominal InfveP on a semantic basis. Control and raising verbs select an InfveP with an abstract-obj content, and let one of their argument NP control or be identified with the unrealized subject of the InfveP. Psychological verbs (in our ex., 'to like', 'to marvel at'), which are known not to constrain their source argument, can precisely accept both types. Locative prepositions select a nom-obj, etc. The subject of nominal InfveP is not controlled, since it does not combine with a control predicate. It remains to be seen whether the subject of a (verbal) goal InfveP (1) is syntactically controlled or pragmatically interpreted.

## References


Bouma, G., R. Malouf and I.A. Sag. 2001. Satisfying constraints on extraction and adjunction. *Natural Language and Linguistic Theory*, 19: 1-65.

Cole, D.T. 1955. An Introduction to Tswana grammar. Longman, Cape Town.

Creissels, D. 2003. Présentation du tswana. *Lalies*, 23: 5-128.

du Plessis, J.A. 1982. Sentential infinitives and nominal infinitives. *South African J.of African Languages*, 2/1.

du Plessis, J.A. and M. Visser. 1992. *Xhosa Syntax*. Via Africa, Pretoria.

Ginzburg, J. and I.A. Sag. 2000. *Interrogative Investigations*. CSLI Publ., Stanford.

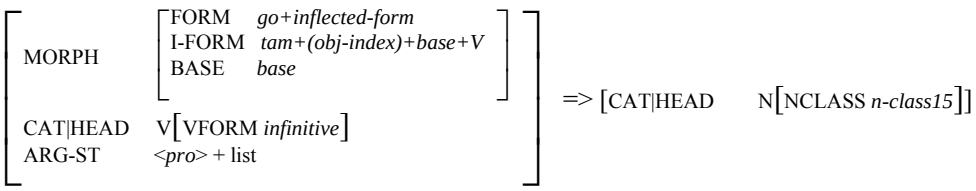
Malouf, R. 2000. *Mixed categories in the hierarchical lexicon*. CSLI Publ., Stanford.

Mugane, J. 2003. Hybrid Constructions in Gikuyu: Agentive Nominalizations and infinitive-Gerund constructions. Butt, M. and T. Holloway King (eds), *Nominals inside and out*, CSLI Publ.., Stanford, 235-265.

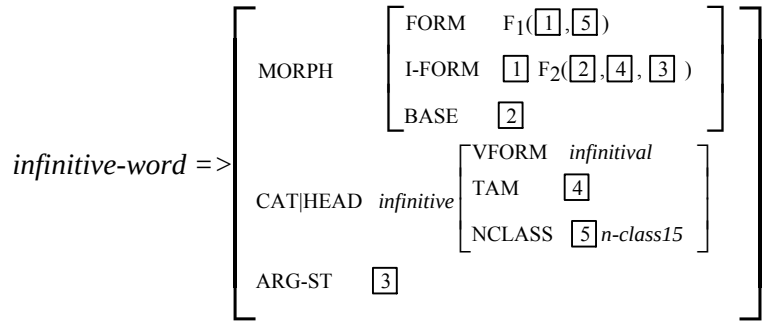
Sag, I.A. 2003. Coordination and underspecification. HPSG Conference (CSLI Publ. web site).

Visser, M. 1989. The Syntax of infinitive in Xhosa. *South African J. of African Languages* 9/4.

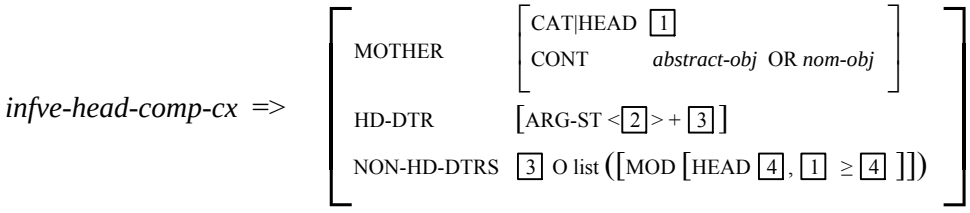
(10)


MORPH
FORM go+inflected-form
I-FORM tam+(obj-index)+base+V
BASE base
CAT|HEAD V[VFORM infinitive]
ARG-ST <pro> + list
=> [CAT|HEAD N[NCLASS n-class15]]

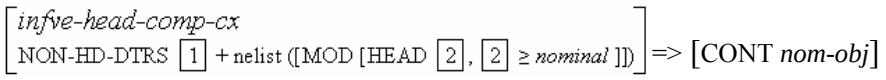
(13)


infinitive-word =>
MORPH
FORM F1(1,5)
I-FORM 1 F2(2,4,3)
BASE 2
CAT|HEAD infinitive
VFORM infinitival
TAM 4
NCLASS 5 n-class15
ARG-ST 3

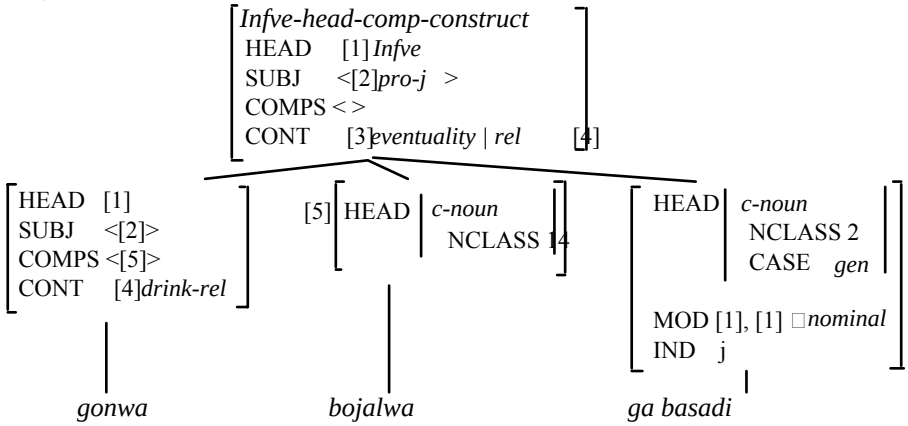
(15)


infve-head-comp-cx =>
MOTHER
CAT|HEAD 1
CONT abstract-obj OR nom-obj
HD-DTR [ARG-ST <2> + 3]
NON-HD-DTRS 3 O list ([MOD [HEAD 4], 1 ≥ 4 ]])

(16)


infve-head-comp-cx
NON-HD-DTRS 1 + nelist ([MOD [HEAD 2], 2 ≥ nominal ]])
=> [CONT nom-obj]

(17)


Infve-head-comp-construct
HEAD [1] Infve
SUBJ <[2]pro-j >
COMPS < >
CONT [3]eventuality | rel [4]
HEAD [1]
SUBJ <[2]>
COMPS <[5]>
CONT [4]drink-rel
[5] HEAD c-noun
NCLASS 14
HEAD c-noun
NCLASS 2
CASE gen
MOD [1], [1] □nominal
IND j
gonwa
bojalwa
ga basadi

# Syncretism in German: a unified approach to underspecification, indeterminacy, and likeness of case

**Berthold Crysmann**
DFKI GmbH & Saarland University
Saarbrücken, Germany

Nouns, adjectives and determiners in German inflect for case, number and gender. However, as is typical for inflectional languages, these morphosyntactic feature dimensions are not expressed by discrete, individually identifiable affixes. Rather, affixes realise complex feature combinations. Although four case, three gender and two number specifications can clearly be distinguished, the morphological paradigms of the language are also characterised by heavy syncretism. Often, syncretism cannot be resolved to disjunctive specification or underspecification within a single feature, but it cuts across the three inflectional dimensions: a German definite determiner, such as *der* can be nominative singular masculine, genitive or dative plural, as well as genitive or dative singular feminine. In the past, this property of German paradigms has provided some motivation for the notion of distributed disjunctions (Krieger, 1996; Netter, 1998). However, since disjunctions are in general much harder to process than type inference, type-based underspecification of case/number/gender specifications appears to be the key towards an efficient and concise treatment of syncretism.

Ambiguous nominal forms in German are also subject to indeterminacy. Again, indeterminacy is not restricted to individual inflectional dimensions, but rather follows the patterns of syncretism. Although the notions of ambiguity and indeterminacy are intimately related, there is currently no analysis at hand that is capable of combining the machinery necessary to cover feature indeterminacy with the benefits of underspecification.

In this paper I will propose an entirely type-based approach to syncretism that will successfully reconcile Daniels (2001)'s approach to feature indeterminacy with morphosyntactic underspecification across features. Furthermore, I will show how list types can be fruitfully put to use to abstract out individual featural dimensions from combined case/number/gender type hierarchies, permitting the expression of likeness constraints in coordinate structures. As a result, the current proposal presents an entirely disjunction-free approach to syncretism, addressing indeterminacy, underspecification and likeness constraints.

## 1 Feature neutrality

It has been argued by Ingria (1990) that the phenomenon of feature neutrality in coordination constitutes a severe challenge for unification-based approaches to feature resolution and concludes that unification should rather be supplanted by feature compatibility checks.

(1) Er findet und hilft Frauen.
he finds.A and helps.D women.A/D
'He finds and helps women.'

(2) * Er findet und hilft Kindern.
he finds.A and helps.D children.D

(3) * Er findet und hilft Kinder.
he finds.A and helps.D children.A

Unification-based frameworks such as LFG or HPSG have taken up the challenge, refining the representation of feature constraints in such a way that neutrality can be modelled without any substantial changes to the underlying formalism. For HPSG, Daniels (2001) proposed to address these problems by means of enriching the type hierarchy to include neutral types, an idea originally due to Levine et al. (2001).

Daniels (2001) has also discussed cases where the potential for feature indeterminacy does not only involve the values of a single feature: as illustrated in (4), a masculine noun like Dozenten can express any cell of the case/number paradigm except nominative singular. Accordingly, one and the same form can be subject to feature indeterminacy regarding number, gender, or even case.

(4) der Antrag des oder der
the petition Def.G.Sg or Def.G.Pl
Dozenten
lecturer.G/D/A+N.Pl
'the petition of the lecturer(s)'

(5) der oder die
Def.N.M.Sg or Def.N.F.Sg
Abgeordnete
representative.N.Sg.M/F
'the male or female representative'

(6) Er findet und hilft Dozenten.
he finds.A and helps.D lecturers.A/D
'He finds and helps lecturers.'

A determiner like *der* is neutral between nominative singular masculine and genitive/dative plural. However, indeterminacy with respect to number is not independent of case, as illustrated by (7), where the unavailability of a nominative singular reading for *Dozenten* is responsible for the illformedness of the sentence.

(7) * der
the.N.Sg.M+G/D.Sg.F+G.Pl
Dozenten ist hier
lecturer.G/D/A+N.Pl is here

To incorporate the issue of neutrality across features, Daniels suggests to combine values of different inflectional features into an overarching type hierarchy, the nodes of which are essentially derived by building the Cartesian product of the types within each inflectional dimension.

## 2 Underspecification

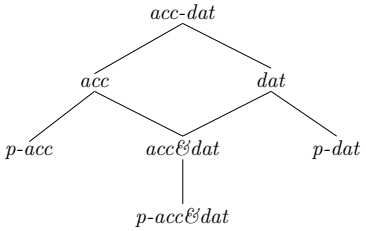
Combined type hierarchies across different inflectional feature dimensions have also been fruitfully put to use in the context of efficient grammar engineering. In the LinGO ERG (Flickinger, 2000), person and number are represented as values of a single feature PNG, permitting the expression of, e.g., non-3rd-singular agreement without the use of negation or disjunction.

In the context of more strongly inflecting languages, such as German, where syncretism is the norm rather than the exception, underspecification of inflectional features across different dimensions is even more pressing: a typical noun such as *Computer* can express any case/number combination, except genitive singular and dative plural, i.e. 6 in total. Using combined case/number/gender hierarchies, the syncretism between nominative/dative/accusative singular and nominative/genitive/accusative plural can be represented compactly as one entry. The very same holds for German determiners and adjectives. Intuitively, it would make perfect sense to try and exploit the combined type hierarchies required for the treatment of neutrality in order to arrive at a more concise and efficient representation of syncretism.

## 3 The Problem

Although both feature indeterminacy and ambiguity do call for type hierarchies combining different inflectional dimensions, these two approaches have not yet received a unified treatment to date: it has been recognised as early as Zaenen and Karttunen (1984) that in unification-based formalisms feature neutrality cannot be reduced to underspecification. The apparent incompatibility of neutrality and underspecification is even more surprising, as these two notions are intimately related: i.e., the ambiguity of a form between two values is a necessary prerequisite for this form to be embeddable in a neutral context.

(8)




Taking as starting point the case hierarchy proposed by Daniels (2001), one might be tempted to assign a case-ambiguous form like 'Frauen' a supertype of both *acc* and *dat*, e.g. *acc-dat*, which can be resolved to *p-acc* ('die Frauen') or *p-dat* ('den Frauen'), depending on

context. However, to include feature-neutrality, it must also be possible to resolve it to the neutral type *acc&dat*. Suppose now that a form like *die* 'the' is itself ambiguous, i.e. between nominative and accusative, representable by a type *nom-acc*, again a supertype of *acc*. Unification of the case values of *die* 'the' and *Frauen* 'women' will yield *acc*, which will still be a supertype of the neutral type *acc&dat*, erroneously licensing the unambiguously non-dative *die Frauen* 'the women' in the neutral accusative/dative context of *findet und hilft* 'finds and helps'.

(9) * Er findet und hilft [die Frauen]
he finds.A and helps.D [the women].A

Thus, under Daniels's account, lexical items are explicitly assigned leaf type values, so-called "pure types". While successful at resolving the issue of indeterminacy, this approach in fact drastically increases the amount of lexical ambiguity, having to postulate distinct entries for type-resolved pure accusative, pure dative, pure nominative, pure genitive, as well as all pairwise case-neutral variants of a single form like *Frauen* 'women'. Ideally, all these different readings should be representable by a single lexical entry, if only underspecification could be made to work together with indeterminacy.

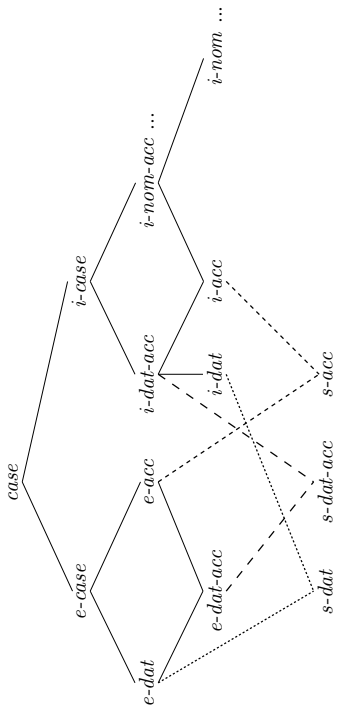
## 4 A Solution

The reason for the apparent incompatibility of underspecification and feature neutrality lies with the attempt to address both aspects within a single type hierarchy. Instead, I shall argue to draw a principled distinction between inherent inflectional feature values, where unification specialises from underspecified or ambiguous types to unambiguous types, and external or subcategorised feature values where unification proceeds from non-neutral, though generally unambiguous to neutral types. As a result we will have two partially independent hierarchies, one for ambiguity (*i-case*) and an inverse one for neutrality (*e-case*).

In order to permit satisfaction of any subcategorised case by some inherent case, all we need to do is define the greatest lower bound for any pair of internal and external case specification.

Thus, underspecified internal cases will unify with a corresponding neutral case, whereas specific internal cases will only unify with their corresponding non-neutral cases. As depicted above, more specific types in one hierarchy will be compatible with less specific types in the other, and vice versa. Thus, disambiguation of *i-case* values will always reduce the potential for neutrality, as required. On a more conceptual level, these cross-classifications between the two hierarchies embody the logical link between underspecification and neutrality.

(10)




## 5 Likeness constraints in coordination

It has been argued by Müller (p.c.) that one of the main obstacles for exploiting combined case-number-gender hierarchies to provide an entirely disjunction-free representation of German syncretism surfaces in certain coordinate structures. It is a well-known fact about German that likeness of category in coordinate structures includes likeness of case specification, but excludes, as a rule, requirements concerning the likeness of gender or number specifications in the conjuncts, a pattern which is quite neatly predicted by HPSG's segregation of HEAD features and INDEX features. However, in free word order languages like German, case arguably serves not only a categorial function, but also a semantic one, thereby

supporting the originally morphological motivation towards organising all agreement features into a single hierarchy (see also Kathol (1999) for a similar proposal). Moreover, the mere existence of indeterminacy across case and index features makes combined hierarchies almost inevitable.

Müller discusses syncretive pronominals in German, such as *der*, which is ambiguous, inter alia, between nominative singular masculine, as shown in (11), and dative singular feminine, as illustrated in (12).

(11) Der schläft.
the.N.S.M sleeps
'That one sleeps.'

(12) Ich helfe der.
I help the.D.S.F
'I help that one.'

This ambiguity could be represented by a type *n-s-m+d-s-f*. Subcategorisation for nominative singular (type *n-s-g*) or dative (type *d-n-g*) will disambiguate these forms accordingly.[1]

In coordinate structures, however, we observe that likeness of case equally eliminates one of the possible gender specifications for *der*, as witnessed by the disambiguation (13). Thus, we must be able to distribute the case requirement over the two conjuncts in such a way that it can exert its disambiguatory potential, without actually unifying the entire case/number/gender specifications of the two conjuncts.

(13) Ich helfe der und dem Mann.
I help the.D.S.F and the.D.S.M man
'I help this one and the man.'

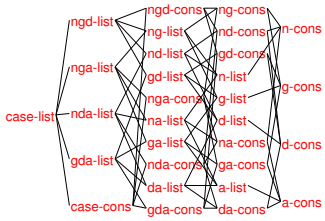
In Daniels (2001), this problem was partly anticipated: he suggests to address the issue of likeness of case by means of a relational constraint same-case/2, which restricts the two arguments to satify identical type requirements. This type equality is essentially imposed by disjunctive enumeration of the four possible subcategorised case values. In typed feature formalisms without relational constraints, his solution may be mimicked by means of unfolding the relevant phrase structure schemata into case-specified variants. In both cases, a greater part of the efficiency gains achieved by underspecification may get eaten up by this disjunctive approach to case similarity.

An alternative, though not fully satisfactory solutiuon would involve retaining a HEAD feature CASE along-side the combined AGR feature. While this move will be at least effective in ruling out unacceptable surface strings, it will fail to impose the disambiguation potential of the subcategorising head onto the individual conjuncts.

What is really needed here is a data structure that may serve to both express the appropriate case-requirements in terms of a combined hierarchy, and permit arbitrarily many specific instantiations of the case constraint. Fortunately, typed feature formalisms do provide for such a data structure, namely typed lists.

To start with, we will set up a hierarchy of case list types, as depicted in figure (14)[2], where each list type immediately subsumes at least one subtype representing a non-empty list of the same case type.

(14)




Types in the combined case-number-gender hierarchy will now restrict their CASE value to an appropriate list type, as given in (15).

(15) *nda-n-g* → [CASE *nda-list*]

Non-empty case lists bear a type constraint restricting the FIRST value to the corresponding agreement type in the combined case/number/gender hierarchy. Actually, thanks to type inference in the hierarchy of case lists, we only need to do this for the 4 immediate subtypes of *case-cons*, namely *ngd-cons*, *nga-cons*, *nda-cons*, and *gda-cons*. In order to propagate the case specification onto all elements of the open list, the tail is constrained to the corresponding list type (see (16)).

[1] For ease of exposition, I am abstracting away from the internal/external distinction, which is immaterial here, since we are only dealing with underspecification, not indeterminacy.

[2] The type hierarchy has been exported from the LKB: supertypes are on the left, subtypes are on the right.

(16) *nda-cons* → ⟨ *nda-n-g* | *nda-list* ⟩

Now that we have a data structure that enables us to encode likeness of case for arbitrary instances of case/number/gender types, all we need to do is refine our existing coordination schemata to distribute the case restriction imposed on the coordinate structure onto the individual conjuncts. In the implemented German grammar we are using, coordinate structures are licensed by binary phrase structure schemata. Thus, all we have to do is to constrain the AGR feature of the left conjunct daughter to be token-identical to the first element on the mother's AGR|CASE list, and percolate the rest of this list onto the (recursive) righthand conjunct daughter's AGR|CASE value:

(17) *coord-phr* →

$$\left[\begin{array}{l} \textsc{ss}\,|\,\textsc{l}\,|\,\textsc{agr}\,|\,\textsc{case}\ \langle \boxed{1}\,|\,\boxed{2}\rangle \\ \textsc{coord-dtrs}\ \left\langle \left[\textsc{ss}\,|\,\textsc{l}\,|\,\textsc{agr}\ \boxed{1}\right], \left[\textsc{ss}\,|\,\textsc{l}\,|\,\textsc{agr}\,|\,\textsc{case}\ \boxed{2}\right] \right\rangle \end{array}\right]$$

Coordinating conjunctions, which combine with a conjunct by way of a head-complement rule, will equate their own AGR|CASE|FIRST value with the AGR value of their complement, percolating the case constraint onto the last conjunct.

(18)
$$\left[\begin{array}{l} \textsc{ss}\,|\,\textsc{l}\ \left[\textsc{agr}\,|\,\textsc{case}\ \langle \boxed{1}, \ldots\rangle\right] \\ \textsc{val}\,|\,\textsc{comps}\ \left\langle \left[\textsc{l}\,|\,\textsc{agr}\ \boxed{1}\right]\right\rangle \end{array}\right]$$

The case of correlative coordinations can be treated exactly analogously.

## 6 Conclusion

In this paper we have argued for an extension to Daniels (2001) original approach to feature indeterminacy in HPSG which makes it possible to combine the empirical virtues of his type-based approach to the phenomenon with the advantages of underspecified representation of syncretism across features, namely generality of specification and efficiency in processing.

We have further shown how likeness constraints abstracting out a particular inflectional dimension from a combined inflectional type hierarchy can still be expressed concisely by means of typed lists.

## References


S. Bayer and M. Johnson. 1995. Features and agreement. In *Proceedings of the 33rd Annual Meeting of the ACL*, pages 70–76.

Ann Copestake. 2001. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.

Michael Daniels. 2001. On a type-based analysis of feature neutrality and the coordination of unlikes. In *Proceedings of the 8th International Conference on Head-Driven Phrase Structure Grammar*, CSLI Online Proceedings, pages 137–147, Stanford. CSLI Publications.

Daniel P. Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

R. J. P. Ingria. 1990. The limits of unification. In *Proceedings of the 28th Annual Meeting of the ACL*, pages 194–204.

Andreas Kathol. 1999. Agreement and the syntax-morphology interface in HPSG. In Robert Levine and Georgia Green, editors, *Studies in Contemporary Phrase Structure Grammar*, pages 209–260. Cambridge University Press, Cambridge and New York.

Hans-Ulrich Krieger. 1996. *TDL — A Type Description Language for Constraint-Based Grammars*, volume 2 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. DFKI GmbH, Saarbrücken.

Robert Levine, Thomas Hukari, and Michael Calcagno. 2001. Prasitic gaps in English: Some overlooked cases and their theoretical implications. In Peter Culicover and Paul Postal, editors, *Parasitic Gaps*, pages 181–222. MIT Press, Cambridge, MA.

Klaus Netter. 1998. *Functional Categories in an HPSG for German*. Number 3 in Saarbrücken Dissertations in Computational Linguistics and Language Technology. German Research Center for Artificial Intelligence (DFKI) and University of the Saarland, Saarbrücken.

Annie Zaenen and Lauri Karttunnen. 1984. Morphological non-distinctiveness and coordination. In *Proceedings of the First Eastern States Conference on Linguistics (ESCOL)*, pages 309–320.

# Coordination Modules for a Crosslinguistic Grammar Resource

**Scott Drellishak**
University of Washington
sfd@u.washington.edu

**Emily M. Bender**
University of Washington
ebender@u.washington.edu

## 1 Background

The Grammar Matrix (Bender et al., 2002) is presented as an attempt to distill the wisdom of existing broad-coverage grammars and document it in a form that can be used as the basis for new grammars. The main goals of the project are: (i) to develop in detail semantic representations and in particular the syntax-semantics interface, consistent with other work in HPSG; (ii) to represent generalizations across linguistic objects and across languages; and (iii) to allow for very quick start-up as the Matrix is applied to new languages. The current Grammar Matrix release includes types defining the basic feature geometry and technical devices (e.g., for list manipulation), types associated with Minimal Recursion Semantics (see, e.g., (Copestake et al., 2003)), types for lexical and syntactic rules, and a hierarchy of lexical types for creating language-specific lexical entries, and links to the LKB grammar development environment (Copestake, 2002). It is, however, completely silent on the topic of coordination.

The next step in Matrix development is the creation of 'modules' to represent analyses of grammatical phenomena which differ from language to language, but nonetheless show recurring patterns. In this paper, we propose a design for a set of modules pertaining to coordination. Coordination is an especially important area to cover early on as coordinated phrases have a relatively high text frequency and thus could pose an important impediment to coverage in the development of Matrix-based grammars. In addition, while the world's languages evince a wide variety of coordination strategies, many of the challenges of providing grammatical analyses of coordination constructions are constant across all of the different strategies. Thus a relatively compact statement of the full set of possible modules is possible and the insights gained in existing work on coordination in the English Resource Grammar (version of 10/04, http://delph-in.net/erg; (Flickinger, 2000)) can be reasonably directly applied to other languages.

In this paper, we restrict our attention to *and* coordination but consider how coordination works for different phrase types as well as both 2-way and n-way coordination.[1] §2 provides a typological sketch of coordination strategies found in the world's languages. §3 motivates design decisions we have taken in this analysis. §4 presents a sample analysis of coordination in Ono. §5 discusses how we encode the information which can be compiled to create the types and instances needed for a particular grammar. Finally, in §6 we discuss further extensions to the grammatical analysis and issues of the user interface.

## 2 Typological Sketch

Across the world's languages, and across the phrase types within those languages, we find a wide variety of coordination strategies. These strategies can be classified along several dimensions; among these are the manner of marking, the location of the marking, and the etymological meaning of the mark.

The manner of marking coordination varies widely, and includes lexical, morphological, and phonological marking, as well as simple juxtaposition. The strategy most familiar from Indo-

[1] We leave for future work issues such as non-constituent coordination or the interaction of syncretism and coordination (e.g., (Beavers and Sag, 2004; Dalrymple and Kaplan, 2000)).

European languages is the use of a separate lexical item (e.g. English *and*). In some languages, coordination is not marked at all: the coordinands are merely juxtaposed. This occurs, for example, in the coordination of noun phrases in Abelam, a Sepik-Ramu language of Papua New Guinea:

(1) wʌny balə wʌny acʌ waryʌ.bər
that dog that pig fight
'that dog and that pig fight' (Laylock, 1965, 56)

Morphological marking generally involves inflecting one or more of the coordinands into some kind of conjunctive or continuative form. For example, in Kanuri (Nilo-Saharan) VPs can be coordinated by placing the earlier one in 'conjunctive form':

(2) kə̀rə̀zə̂ málə̀mrò wálwònò.
studied.CONJ malam became
'He studied and became a malam.'
(Hutchison, 1981, 322)

In a few languages, coordination is marked by what appears to be a phonological alteration of the coordinands. For example, in Telugu (Dravidian), adjective phrases and noun phrases are coordinated by lengthening the final vowels in the coordinands:

(3) kamalaa wimalaa poDugu.
Kamala Vimala tall
'Kamala and Vimala are tall.'
(Krishnamurti and Gwynn, 1985, 325)

Languages which require a special intonation contour to accompany coordination by juxtaposition are arguably using a phonological marking strategy as well. While ideally it would be very interesting to incorporate a model of prosody into grammar implementations, this is currently not feasible. Therefore, for present purposes, we will treat the juxtaposition strategy as though it had no overt marking.

Coordination strategies can also be classified by the location of the marking. In the simple case of two-way coordination, there are three positions where the marking may occur: before the first coordinand (initial), between the coordinands (medial), or after the second coordinand (final). In fact, the medial position is often more clearly associated with either the first or second coordinand, as a postfix or prefix respectively. In addition, languages vary in the number of marks used. If zero marks are used, we have the juxtaposition strategy, also referred to as *asyndeton*; if one mark is used, this is referred to as *monosyndeton*; if each coordinand is marked, this is referred to as *polysyndeton* (Haspelmath, 2000).

Finally, coordination strategies vary in the etymology of the marker. Some languages use an element related to the comitative marker and others an element not clearly related to anything else (Stassen, 2000). Rarer etymological sources include number words (Huánuco Quechua) and pronouns (Sedang).

Our intention with the coordination modules is to provide syntactic and semantic scaffolding powerful enough to deal with most or all of these structures, and flexible enough to be enhanced to cover other esoteric strategies that might be discovered.

## 3 Design Decisions

### 3.1 Category-specific Rules

It may seem desirable at first to have a single rule that covers the coordination of all phrase types. However, experience with detailed work on English (as represented by the English Resource Grammar) suggests that this is not practical, given our formalism and current assumptions about feature geometry. The core generalization[2] is that phrases of the same category can be coordinated to make a larger phrase of that category. Thus a common first-pass attempt at modeling coordination involves a rule that identifies HEAD and VAL values across the coordinands and the mother (see e.g., (Sag et al., 2003)). However, there are features which have been placed inside HEAD for independent reasons which need not be identified across coordinands, such as AUX:

(4) Kim slept and will keep on sleeping.

Further, there are differences in the semantic effects of coordination for individuals and events. In particular, nominal indices must be bound by quantifiers in MRS, leading NP and NOM coordination rules to introduce additional quantifiers. No such constraint holds for event indices.

Finally, there are idiosyncrasies to coordination in certain phrase types. A prime example here is the agreement features on coordinated NPs in English. For NPs coordinated with *and*, at least, the number

[2]This generalization is subject to several well known exceptions, which tend to have low text frequency.

of the conjoined phrase is always plural, and the person is the lesser of the person values of other coordinands (first person and second person give first person, etc.). In the context of our cross-linguistic analysis, we also find languages where the coordination strategy is different for different phrase types.

In light of these facts, the analysis is considerably simplified by positing separate rules for the coordination of different phrase types. These rules stipulate matching HEAD values, rather than identifying them. These rules are, of course, arranged into a hierarchy in which supertypes capture generalizations across all of the different coordination constructions.

### 3.2 Binary branching structure

Whether coordination involves binary branching or flat structure is a matter of much theoretical debate (see e.g., (Abeillé, 2003)). Rather than review those arguments here, we present two engineering considerations which support a binary branching analysis.

First, while the LKB allows rules with any given number of daughters, it does not permit rules with an underspecified number of daughters. This means that a rule like (5a) would have to be approximated via some number of rules with a specific arity (5b):

(5) a. NP → NP+ and NP

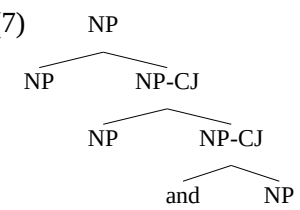
b. NP → NP and NP
NP → NP NP and NP
NP → NP NP NP and NP
...

With binary branching, in contrast, three rules produce an unlimited number of coordinands:

(6)

| | | |
|---|---|---|
| NP-CJ | → and NP | (bottom coord rule) |
| NP-CJ | → NP NP-CJ | (mid coord rule) |
| NP | → NP NP-CJ | (top coord rule) |

(7)




Second, there is the issue of 'promotion' of agreement features in coordinated NPs (and potentially other phrase types). In French, for example, the gender value of a coordinated NP is masculine iff at least one of the coordinands is. In order to state this constraint in this system, we'll need separate rule subtypes which posit [GEND *masc*] on the mother and on one daughter, leaving the other daughter unspecified.[3] In either system, this means doubling the number of rules, but the binary branching system starts out with fewer rules (and in fact, only the top and mid coordination rules need to be doubled, not the bottom coord rule). The flat structure system, on the other hand, potentially has a very large number of rules to start with. When we also consider promotion of person values, the number of rules involved gets larger, and the gain from the binary branching system becomes even clearer.

## 4 Sample Analysis

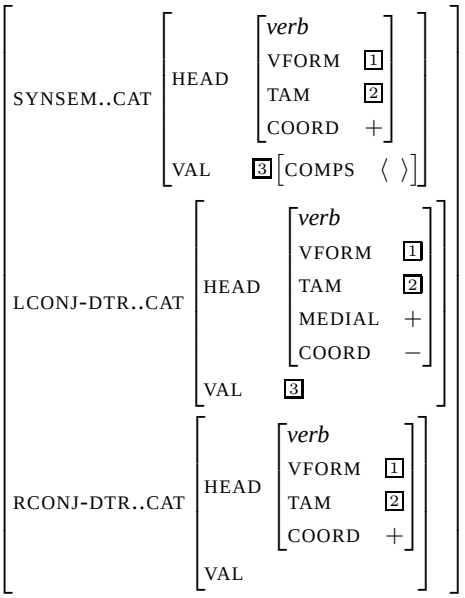
In this section, we provide a sketch of an analysis of coordination of verb phrases and noun phrases in Ono, a Trans-New Guinea language. As described by Phinnemore (1988), Ono verb phrases are coordinated by inflecting non-final verbs into a "medial" form, as in (8), while noun phrases are coordinated with the medial monosyndeton *so*, as in (9).

(8) mat-ine gelig-e taun-go ari
village-his leave-MED town-to go-MED
more zoma ka-ki so ea seu-ke
then sickness see-him-3SDS and there die-fp.-3s
'He left his village, went to town, and got sick and died there.' (Phinnemore, 1988, 109)

(9) koya so kezong-no numa len-gi
rain and clouds-ERG way block-3SDS
'Rain and clouds block the way...'
(Phinnemore, 1988, 100)

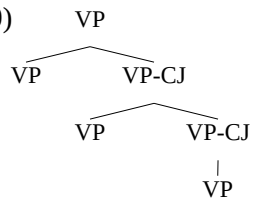
We handle these structures with six rules: *vp_top_coord_rule*, *vp_mid_coord_rule*, *vp_bottom_coord_rule*, *np_top_coord_rule*, *np_mid_coord_rule*, and *np_bottom_coord_rule*. The mid and top coord rules are non-headed rules with two daughters, one for each coordinand, called LCONJ-DTR and RCONJ-DTR. We assume additional boolean HEAD features COORD and (for verbs) MEDIAL. *vp_bottom_coord_rule* simply marks a [MEDIAL −] VP as coordinated (i.e. COORD +). The *vp_mid_coord_rule* will look something like the following:

[3] Dalrymple and Kaplan's (2000) set-based system for succinctly handling such facts is not currently available in the LKB.

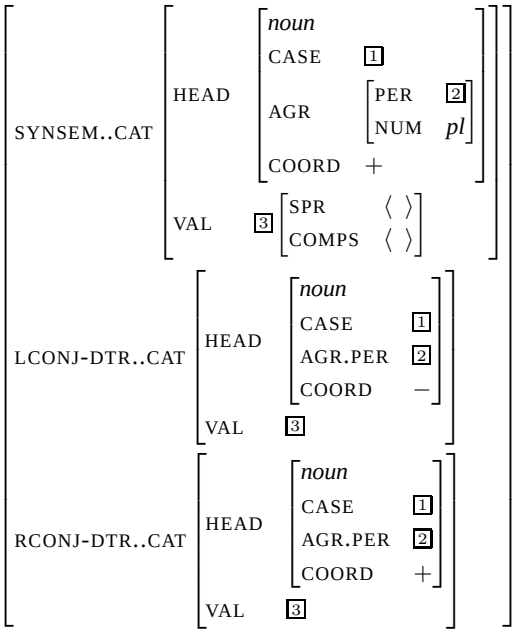



This rule identifies several features of the coordinated VPs, marks the resulting phrase as coordinated, and takes a medial-form, noncoordinate left coordinand. This use of the COORD feature will enforce right-branching structure, so it is not necessary to specify MEDIAL on the mother node, which can only serve as the RCONJ-DTR of any further higher coordination. The Ono *vp_top_coord_rule* differs semantically from the mid rule in how it combines the semantic contributions of the coordinands, and differs syntactically from it only in that the mother node is [COORD −]. The structure assigned the coordination of three VPs, the first two of which are in medial form, is shown in (10), where VP-CJ is a VP marked [COORD +].

(10)




For noun phrases, we will need an additional lexical item *so* of HEAD type *conj*, and the *np_bottom_coord_rule* will combine *so* with an NP into a COORD-marked NP. The *np_mid_coord_rule* will look something like the following:




This rule identifies several features of the coordinated noun phrases, and constrains the mother to be plural, the mother and the RCONJ-DTR to be coordinated and the LCONJ-DTR to be not coordinated. The *np_top_coord_rule* will be similar, except that it combine the semantic contributions of all coordinands slightly differently, and will also mark the mother node [COORD −]. Based on these rules, the structure of a coordinated noun phrase made up of three NPs conjoined with a single *so* will look like (7) above, where NP-CJ is an NP marked [COORD +].

For languages with polysyndeton, the only modification to the rules in (6) is the omission of the mid rule, which results in the marking of coordination on each coordinand, because each additional NP will require one more bottom (and top) node:

(11)

| | | | |
|---|---|---|---|
| NP-CJ | → and NP | (bottom coord rule) |
| NP | → NP-CJ NP-CJ | (top coord rule) |

## 5 Modularization

The intended goal of the coordination modules is to provide a basis for formal analyses for as wide a variety of languages as possible. However, we expect that we will be able to capture this variation based on a more limited set of semantic and syntactic rules. While it is not the case that all languages have the same number of or divisions between word classes, we expect to be able to capture

the semantics of various phrase types in a language-independent way. The Matrix will provide coordination rules for phrases whose semantic contribution consists of individuals (e.g. noun phrases), events (e.g. verb phrases), modification of individuals (e.g. adjectives), modification of events (e.g. adverbs), and so forth.

In addition, we expect to find commonalities among the syntactic rules that can be factored out. For example, the parts of the VP and NP rules for Ono above that deal with the feature COORD can be adapted to deal with general asyndeton, monosyndeton, and polysyndeton coordination. All three strategies will have bottom and top coordination rules (with the mid rule only needed for monosyndeton), but the rules will vary slightly. The monosyndeton rules will look like the rules in (6) above; the polysyndeton rules will look like the rules in (11); and the asyndeton rules will look like (12).

(12)

| | | | |
|---|---|---|---|
| NP-CJ | → NP | (bottom coord rule) |
| NP | → NP NP-CJ | (top coord rule) |

Different manners of marking coordination can be captured by varying the bottom rule. It can be either a rule that combines a separate lexical coordinator with the lowest coordinand, or else a non-branching rule triggered by a morphological feature.

Based on the answers to questions posed to the user about the facts of the language being analyzed, the semantic coordination rules and syntactic/morphological coordination rules will be cross-classified to produce a set of language-specific rules appropriate to the language at hand.

## 6 Conclusion and Outlook

We have presented an overview of an initial set of coordination modules for the Grammar Matrix. We believe that they are suited to providing syntactically and semantically valid analyses of the diverse coordination strategies in the world's languages. Furthermore, the factored representation given to the underlying types used to create language-specific coordination systems provides a means formalizing generalizations across languages.

The next steps for this project include: 1. Testing the coverage of the modules by deploying them in implemented grammars for a diverse range of languages. 2. Expanding the coverage to include other types of coordination (in the first instance, coordination with *or*, *but*, etc.). 3. Working out the user interface and in particular a set of questions and a protocol for presenting them to the linguist which covers the ground necessary to handle any given language while avoiding redundancy in any particular case.

## References


Anne Abeillé. 2003. A lexicalist and construction-base approach to coordinations. In Stefan Müller, editor, *Proceedings of HPSG03*. CSLI, Stanford.

John Beavers and Ivan A. Sag. 2004. Ellipsis and apparent non-constituent coordination. In Stefan Müller, editor, *Proceedings of HPSG04*, pages 48–69. CSLI, Stanford.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix. *Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation*.

Ann Copestake, Daniel P. Flickinger, and Carl Pollard Ivan A. Sag. 2003. Minimal Recursion Semantics. An introduction.

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI, Stanford.

Mary Dalrymple and Ronald M. Kaplan. 2000. Feature indeterminacy and feature resolution. *Language*, 76:759–798.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *NLE*, 6 (1):15 – 28.

Martin Haspelmath. 2000. Coordination. In Timothy Shopen, editor, *Language typology and linguistic description, 2nd edition*. Cambridge University Press, Cambridge.

John P. Hutchison. 1981. *A reference grammar of the Kanuri language*. University of Wisconsin - Madison, Madison, WI.

BH. Krishnamurti and J. P. L. Gwynn. 1985. *A grammar of modern Telugu*. Oxford University Press, Delhi.

D. C. Laylock. 1965. *The Ndu language family (Sepik district, New Guinea)*. Linguistic Circle of Canberra, Series C, No 1. The Australian National Library, Canberra.

Penny Phinnemore. 1988. Coordination in Ono. *Language and Linguistics in Melanesia*, 19:97–123.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*. CSLI, Stanford.

Leon Stassen. 2000. And-languages and with-languages. *Linguistic Typology*, 4:1–54.

# A new well-formedness criterion for semantics debugging

**Dan Flickinger**
CSLI
Stanford University
Stanford, CA
danf@csli.stanford.edu

**Alexander Koller** and **Stefan Thater**
Department of Computational Linguistics
Universität des Saarlandes
Saarbrücken, Germany
{koller,stth}@coli.uni-sb.de

## Abstract

We present a novel well-formedness condition for underspecified semantic representations which requires that every correct MRS representation must be a *net*. We apply this condition to identify a set of eleven rules in the English Resource Grammar (ERG) with bugs in their semantics component, and thus demonstrate that the net test is useful in grammar debugging. In addition, we show that a partly corrected ERG derives 3 % less non-nets on the Rondane treebank and we expect that after completing the correction of the ERG, only 5.5 % non-nets are derived, which we take as support for our initial hypothesis.

## 1 Introduction

A very exciting recent development in (computational) linguistics is that large-scale grammars which compute semantic representations for their input sentences are becoming available. For instance, the English Resource Grammar (Copestake and Flickinger, 2000) is a large-scale HPSG grammar for English which computes underspecified semantic representations in the MRS formalism (Copestake et al., 2004). It is standard to use underspecification to deal with scope ambiguities; apart from MRS, there is a number of other underspecification formalisms, such as dominance constraints (Egg et al., 2001) and Hole Semantics (Bos, 1996).

However, the increased power of the new grammars comes with a new challenge for grammar engineering: How can we be sure that all semantic outputs the grammar computes (through any combination of semantic construction rules) are correct, and how can we find and fix bugs? This problem of *semantics debugging* is an important factor in the 90% of grammar development time that is spent on the syntax-semantics interface (Copestake et al., 2001).

Grammar development systems such as the LKB implement some semantic sanity checks, which are practically useful, but rather shallow, and therefore limited in their power. On the theoretical side, there are attempts to formalise "best practices" of grammar development in a *semantic algebra* (Copestake et al., 2001), but this is quite a far-reaching project that is not yet fully implemented.

One potential alternative method for semantics debugging comes from Fuchss et al.'s recent work on *nets* (Fuchss et al., 2004). They claim that every underspecified description (written in MRS or as a dominance constraint) that is actually used in practice is a *net*, i.e. it belongs to a restricted class of descriptions with certain useful structural properties, and they substantiate their claim through an empirical evaluation on a treebank. If this "Net Hypothesis" is true, we can recognise a grammar rule (or combination of rules) as problematic if it produces only non-nets on a variety of inputs.

In this paper, we show that such a use of nets is indeed possible. We use the ERG to derive MRS representations for all sentences in the Rondane treebank (distributed with the ERG) and the Verbmobil sections of the Redwoods treebank (Oepen et al., 2002). Our first result is a small set of eleven rules which systematically cause the MRS representations to be non-nets for every sentence in which they are used. These rules all have faulty semantics components, i.e. we have identified semantically buggy rules. We are currently correcting the grammar by hand. The partly corrected grammar produces 89.5 % nets and only 8 % non-nets for the syntactic analyses in the Rondane corpus, and we expect that after completing the correction of the problematic rules, only 5.5 % non-nets are derived, which we take as further support of the Net Hypothesis.

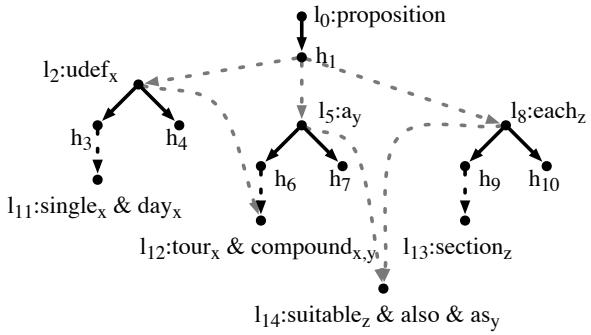



Figure 1: Graphical representation of the MRS for "Each section is also suitable as s single day tour."

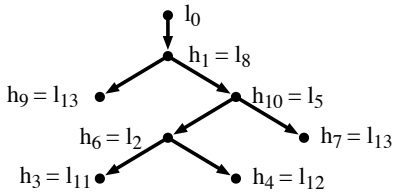



Figure 2: Configuration of the MRS in Fig. 1.

## 2 Minimal Recursion Semantics

We start with a brief overview of Minimal Recursion Semantics (MRS). MRS (Copestake et al., 2004) is the standard scope underspecification formalism used in current HPSG grammars, such as the English Resource Grammar (ERG; (Copestake and Flickinger, 2000)) or grammars derived from the Grammar Matrix (Bender et al., 2002). Its purpose is to separate the problem of resolving scope ambiguities from semantics construction.

Fig. 1 shows a graphical representation of the (slightly simplified) MRS which the ERG derives for the sentence "Each section is also suitable as a single day tour" from the Rondane treebank. It consists of *elementary predications* (EPs) such as $l_2$:udef$(x,h_3,h_4)$, $l_5$:a$(y,h_6,h_7)$, $l_{12}$:tour$(x,y)$, and $l_{12}$:compound$(x,y)$, and of *handle constraints* such as $h_6 =_q l_{12}$. Elementary predications specify the parts that a semantic representation must be made up of, and handle constraints $h =_q l$ specify, approximately, that $h$ must outscope $l$. Terms $l_i$ on the left-hand side of EPs are called *labels*, terms $h_i$ are called *(argument) handles*, and terms $x$, $y$, etc. are ordinary first-order variables. Notice that there are two EPs for the label $l_{12}$; this is called an *EP conjunction*, and interpreted as conjunction of the two formulas labelled by $l_{12}$.

The graph in Fig. 1 can be given an explicit interpretation as a representation of an MRS structure (Fuchss et al., 2004). The nodes correspond to the labels and handles in the MRS, and the solid edges correspond to the EPs. We call the subgraphs that are connected by solid edges the *fragments* of the graph. The dashed *dominance edges* are used to represent handle constraints, the outscoping requirement between a variable and its binder (such as between the quantifier at $l_2$ and the variable in $l_{12}$), and the implicit constraint that the "top" label $l_0$ must outscope all other EPs. Note that we assume that the graph does not contain transitively redundant edges; for instance there is no binding edge between $l_2$ and $l_{11}$. EP conjunctions are represented by explicit conjunction at the graph nodes.

An underspecified MRS structure describes a set of configurations, or *scope-resolved* MRS structures. The scope-resolved MRS structures can be computed by arranging all the fragments of an MRS structure into a tree, in such a way that every label except for the one at the root is identified with a handle, and all the outscoping requirements are respected. One of the five scope-resolved MRSs for the MRS in Fig. 1 is shown in Fig. 2 (we omit EPs for clarity). Note that in general it is possible that more than one label is assigned to the same handle, and that the scope-resolved MRS structure can contain more EP conjunctions than the original MRS structure. In such a case, we call the scope-resolved MRS structure a *merging configuration*.

## 3 MRS-Nets

We say that an MRS structure is a *net* if all the fragments in its graph are of one of the three forms shown in Fig. 3. In a *strong* fragment, every leaf (argument handle) and no other node has exactly one outgoing dominance edge. For example, the nuclear fragments $l_{11}$ and $l_{14}$ in Fig. 1 are strong fragments. In a *weak* fragment, every leaf but one has exactly one outgoing dominance edge, and the root of the fragment has one outgoing dominance edge too. Weak fragments correspond to quantifiers (such as $l_2$ and $l_8$ in Fig. 1) where the dominance edge from

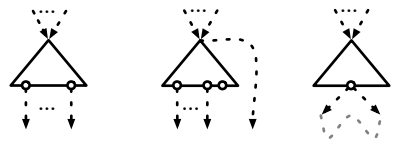



Figure 3: Fragment Schemata of Nets

the root represents the implicit variable binding. Finally, *island* fragments may have leaves with multiple outgoing dominance edges, but then all dominance children must be connected by *hypernormal paths*. A hypernormal path is an undirected path which doesn't use two dominance edges that come out of the same node. An example for an island fragment is the topmost fragment in Fig. 1. Its dominance children are the three quantifier fragments, and there is a hypernormal path between each pair of these fragments – for instance, $l_2, l_{12}, h_6, l_5$ and $l_5, l_{14}, l_8$.

Because all fragments in Fig. 1 are strong, weak, or island, the MRS it represents is a net. By contrast, Fig. 5 shows two MRS structures which are not nets. Both structures violate the *island* condition because the topmost fragment has outgoing edges to quantifier fragments (e.g. in the left-hand graph, the fragments for "a bit" and "two young Norwegians") which are only connected via the top fragment itself, and not by an additional hypernormal path. The left-hand graph also contains a quantifier fragment ("a bit") which violates the *weak* condition, as there is an open argument handle without a corresponding dominance edge out of the root of the fragment.

Nets were introduced in (Niehren and Thater, 2003; Fuchss et al., 2004) as a technical restriction; the key theorem about nets is that they can be translated into normal dominance constraints (Egg et al., 2001) and Hole Semantics (Bos, 1996). This means that nets can be solved efficiently using the solvers for normal dominance constraints (Bodirsky et al., 2004). Nets have other useful properties: For example, nets have no merging configurations, so all EP conjunctions can be resolved to true conjunctions in a preprocessing step.

The crucial restriction that nets impose is that the dominance children of island fragments must be hypernormally connected. Intuitively, hypernormal connectedness means that nets must be "downwards" connected: In the example in Fig. 1, $l_2$ and $l_8$ are "tied together" by the zig-zag path through $l_{12}$ and $l_{14}$, whereas "a bit" and "two young Norwegians" in Fig. 5 have no such connection. A linguistic intuition for this is that quantifiers that are syntactic arguments of the same verb remain hypernormally connected because their variables occur as arguments of this verb.

## 4 Nets in Semantics Debugging

Now we show that nets can indeed be used to identify grammar rules with incomplete semantics components, and that non-nets are so infrequent in practice that it is reasonable to assume that all correct MRS structures are indeed nets.

### 4.1 Previous Work

Recently, Fuchss et al. (Fuchss et al., 2004) presented a first evaluation of whether the MRS structures that can be derived using the ERG are nets or not. They found that about 83% of the MRS structures derived for all syntactic readings of all the sentences in the Redwoods treebank (Copestake and Flickinger, 2000) are in fact nets. Their impression from inspecting some non-nets was that non-nets seemed to be systematically incomplete. They took this as suggestive of what they call the *Net Hypothesis*: that all MRS structures needed in practice (i.e. for the parses of a treebank according to a large-scale grammar) are nets.

### 4.2 Experiment

If the Net Hypothesis is true, the 17% non-nets must be the results of errors in the annotation or the grammar rules, and every MRS that is not a net can be taken as an indicator that the grammar rules used in producing it might be candidates for debugging.

In order to analyse this in more detail, we reran Fuchss et al.'s evaluation, using the October 2004 version of the ERG. As test corpora, we used the Verbmobil sections of the Redwoods 5 Treebank (Jan. 2005) which contains 10503 sentences, and the Rondane Treebank (1034 sentences) distributed with the ERG. Both corpora are annotated with HPSG syntactic structures, for each of which a unique MRS structure can be extracted.

| Treebank | #Sents. | Ill-formed | Non-Nets | Nets |
|---|---|---|---|---|
| Verbmobil | 10503 | 11 % | 17 % | 72 % |
| Rondane | 1034 | 8 % | 11 % | 81 % |

Figure 4: Classification of the sentences in the treebanks.

The table in Fig. 4 shows the results of the experiment. Each sentence in the treebanks was classified into one of three categories: (1) sentences whose MRS structure was not well-formed according to the shallow tests in the LKB system (e.g., structures containing variables that aren't bound by any quantifier, or structures with cycles); (2) sentences whose MRS structures were okay according to the LKB checks, but were not nets, and (3) sentences whose MRS structures were nets. Of all the MRSs that are well-formed according to the test in the LKB, 81 % (Redwoods) and 88 % (Rondane) are nets, and 19 % (Redwoods) and 12 % (Rondane) aren't. Interestingly, the ratio of nets to non-nets is much smaller if we look not only at the annotated syntactic analyses, but at *all* possible analyses (as Fuchss et al. did).

### 4.3 Semantic Debugging

Then we checked which rules were "responsible" for the introduction of non-net structures. We found that there is a group of eleven rules which systematically derive only non-nets for all syntactic analyses of all sentences in the treebanks; these rules account for approx. 55% of the non-nets:

1. Measure noun phrases like "2–3 hours" (MEASURE_NP, BARE_MEASURE_NP)
2. Coordinations of more than two conjuncts like "train, bus or car" (P_COORD_MID, N_COORD_MID)
3. Sentence fragmens like "Delicious!" (rules FRAG_PP_S, FRAG_R_MOD_I_PP, FRAG_ADJ, and FRAG_R_MOD_AP)
4. Other rules: VPELLIPSIS_EXPL_LR, NUM_SEQ, TAGLR.

Indeed, the semantics components of all eleven rules are buggy, in that the MRS graphs that they compute have too few dominance edges or unconnected fragments that should constitute an single fragment (e.g., by forming an EP-conjunction). This is illustrated by the structures shown in Fig. 5. The structure on the left is derived by the ERG for the sentence "a bit further on we meet two young Norwegians". In this structure, the quantifier "a bit" (whose analysis uses the MEASURE_NP rule) introduces a bound variable *x* that is used only in its restriction, but in none of the predicates in its scope ("meet further on"). This is obviously not intended. Because the missing variable binding also relaxes the constraints on how fragments can be plugged together, the underspecified description admits structurally wrong readings, e.g. by plugging "young Norwegian" into the scope of "a bit" (see Fig. 6). If we fix the structure by using *x* in the EPs for "further on", this introduces an additional dominance edge in the graph which makes the structure a net.

A similar bug occurs in the right-hand MRS structure. The EPs "and" and "implicit_conj" are two different components of the same collective "tea, milk, and coffee", and should therefore be connected. Because they aren't, the structure has meaningless scopings such as the one shown in Fig. 6 (and almost 1000 further scopings) where "and" and "drink" have been merged into the same argument handle. If we connect "and" and "drink" either by collecting them into a single EP-conjunction, or by introducing additional material (e.g., an qauntifier fragment) that connects the two nodes, the MRS structure again becomes a net.

| Treebank | #Sents. | Ill-formed | Non-Nets | Nets |
|---|---|---|---|---|
| Rondane | 961 | 2.5 % | 8 % | 89.5 % |

Figure 7: Classification of the sentences in the Rondane treebank for the partly corrected version of the ERG

### 4.4 Re-Evaluating the Net Hypothesis

We are currently working on correcting the semantics components of the eleven faulty rules by hand. If all problematic rules are corrected in a way that only nets are derived, we expect that of the well-formed MRS structures 94.5 % (Rondane) and 91.5 % (Redwoods) of the syntactic structures as annotated in the treebanks derive nets. A first experiment shows that with our partly corrected version of the ERG, almost 92 % of the derivations annoted with well-

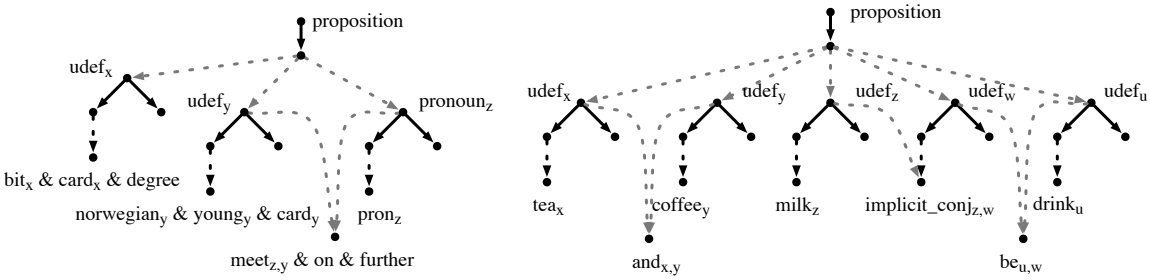



Figure 5: MRS expressions for the annotated derivation for "a bit further on we meet two young Norwegians" (left) and "Drink is tea, milk and coffee" (right) in the Rondane treebank

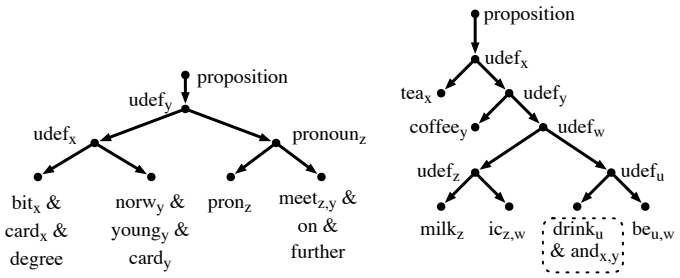



Figure 6: Example solutions for the MRS structures in Fig. 5

formed MRS structures (89.5% of all sentences) in the Rondane treebank produce nets (Fig. 7).[1] It is important to note that in particular measure noun phrases with degree modifications, which are relatively often used, are not yet fully corrected. Note also that the number of ill-formed MRS structures has been considerably reduced.

It is important to note that at least some applications of each of the eleven rules above passed the well-formedness checks in the LKB, which shows that nets can allow us to identify semantically problematic rules which shallower checks can't find. In addition, non-nets make up a larger portion of the MRS structures in the original grammar than the ill-formed structures; so they are likely to capture classes of errors that are at least as prevalent as those that the existing checks do.

[1]For technical reasons, the treebank for the partly corrected grammar contains slightly fewer sentences. Note that if we remove the missing sentences from the classification for the original treebank, we obtain results similar to the ones shown in Fig. 4.

## 5 Conclusion

We have shown that nets can be a useful tool for debugging the semantics component of a large-scale grammar. All eleven rules in the ERG that computed only non-nets turned out to be semantically problematic, typically in that they were missing a variable name coindexation, or some fragments (EPs) were unconnected; also, none of these rules would have been easily found by the existing well-formedness tests in the LKB. A partly corrected version of the ERG derived 89.5 % nets on the Rondane corpus.

In order to make the net criterion practically useful, we have developed an efficient algorithm that checks whether a given MRS is a net in linear time. A portable open source implementation of this algorithm is publically available from `http://utool.sourceforge.net`.

There are various ways in which the work we report here could be extended. On the one hand, it would be interesting to see whether a similar debugging methodology would yield problem rules based on the LKB's well-formedness tests, and it would be natural to look not just for problematic *rules*, but

also for problematic *lexicon entries* this way. On the other hand, we suspect that some semantically problematic MRS structures are derived not by a single rule, but by a combination of rules. One way of finding such rule combinations would be to analyse the MRSs for a corpus with a decision tree learner, which would try to derive rules that capture such combinations.

## References


Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Procedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics.*, Taipei, Taiwan.

M. Bodirsky, D. Duchier, J. Niehren, and S. Miele. 2004. A new algorithm for normal dominance constraints. In *Proc. SODA*.

J. Bos. 1996. Predicate logic unplugged. In *Proc. 10th Amsterdam Colloquium*, pages 133–143.

A. Copestake and D. Flickinger. 2000. An open-source grammar development environment and broad-coverage english grammar using HPSG. In *Proc. LREC*.

A. Copestake, A. Lascarides, and D. Flickinger. 2001. An algebra for semantic construction in constraint-based grammars. In *Proc. 39th ACL*, Toulouse.

A. Copestake, D. Flickinger, C. Pollard, and I. Sag. 2004. Minimal recursion semantics: An introduction. *Journal of Language and Computation*. To appear.

M. Egg, A. Koller, and J. Niehren. 2001. The constraint language for lambda structures. *Journal of Logic, Language, and Information*, 10:457–485.

R. Fuchss, A. Koller, J. Niehren, and S. Thater. 2004. Minimal Recursion Semantics as dominance constraints: Translation, evaluation, and analysis. In *Proc. 42nd ACL*, Barcelona.

J. Niehren and S. Thater. 2003. Bridging the gap between underspecification formalisms: Minimal recursion semantics as dominance constraints. In *Proc. 41st ACL*, Sapporo.

S. Oepen, K. Toutanova, S. Shieber, C. Manning, D. Flickinger, and T. Brants. 2002. The LinGO Redwoods treebank: Motivation and preliminary applications. In *Proc. 19th COLING*.

# Object-to-Subject Raising: An Analysis of the Dutch Passive

**Frederik Fouvry** and **Valia Kordoni**
Department of
Computational Linguistics and Phonetics
Saarland University
Saarbrücken, Germany
{fouvry,kordoni}@coli.uni-sb.de

**Gertjan van Noord**
Center for Language and Cognition
University of Groningen
Groningen, The Netherlands
vannoord@let.rug.nl

## 1 Introduction

This paper focuses on passive constructions in Dutch. Specifically, we focus on *worden*, as well as *krijgen* passives in Dutch, for which we propose a uniform, raising analysis in HPSG. We also show that such an analysis can be carried over to account for passives cross-linguistically. Specifically, we look at corresponding structures in German and show that there is no need for a dual raising and control analysis for the German "agentive" (*werden*) and the German "dative" (*kriegen*) passives, respectively, as has been proposed in Müller (2002) and Müller (2003).

## 2 The data

The following are examples of the main passives in Dutch.[1,2]

(1) a. Peter kust haar.
Peter.subj kisses her.obj1
"Peter kisses her."

b. Zij wordt gekust (door Peter).
she.subj is kissed (by Peter)
"She is kissed (by Peter)."

(2) Het raam is geopend.
the window.subj is opened
"The window is open."

Dutch also exhibits a special kind of passives which are formed with the auxiliary *krijgen* ("to get"; henceforth, *krijgen* passive). The *krijgen* passive is formed from ditransitive verbs in Dutch, which subcategorise for a *primary* (obj1) and a *secondary* (obj2) object. The *secondary* object of the ditransitive verb surfaces as the subject of the *krijgen* passive:

(3) a. Ik stuur hem het boek toe.
I.subj send him.obj2 the book.obj1 to
"I send him the book."

b. Hij krijgt het boek toegestuurd.
he.subj gets the book.obj1 sent-to
"He gets the book sent."

(4) a. We betalen hem zijn salaris door.
we.subj pay him.obj2 his wages.obj1 through
"We continue to pay him his wages."

b. Hij krijgt zijn salaris doorbetaald.
he.subj gets his wages.obj1 paid-through
"He is being paid his wages."

In contrast, when the *primary* object of the ditransitive verb surfaces as the subject of the passive form of Dutch ditransitives, like the one in (3a), for instance, then this passive is formed with the auxiliary *worden*, like the passive form of regular transitive verbs in Dutch (see example (1) above):

(5) a. Ik stuur hem het boek toe.
I.subj send him.obj2 the book.obj1 to
"I send him the book."

b. Het boek wordt hem toegestuurd.
the book.subj is him.obj2 sent-to
"The book is sent to him."

c. *Hij wordt het boek toegestuurd.
he.subj is the book.obj1 sent-to
"He is sent the book."

As can be observed in examples (3) and (4) above, the *primary* objects of the active forms in (3a) and (4a) (*het boek* and *zijn salaris*, respectively) retain

[1]The *zijn* ("stative") passives in (2) above are beyond the scope of this paper.

[2]In the glosses subj = subject, obj1 = object1 (*primary object*), obj2 = object2 (*secondary object*).

their grammatical function (obj1) in the passive sentences in (3b) and (4b). Actually, the absence of the *primary* object of the ditransitive active form from the corresponding *krijgen* passive renders the latter ungrammatical:

(6) *Hij krijgt toegestuurd.
he.subj gets sent-to
"*He was sent."

### 2.1 Some interesting exceptions

The only exception in the passive patterns in Dutch presented in section 2 is observed with the verb *betalen* (to pay) and its derivatives (*doorbetalen* (to continue payment), *uitbetalen* (to pay out), *terugbetalen* (to pay back), etc).

As shown from examples (5a)–(5c) above, in general *secondary* objects (obj2s) in Dutch ditransitives can never passivise with the auxiliary *worden*. That is, the *secondary* object of Dutch ditransitives, like *geven* and *betalen*, can never surface as the subject of a *worden* passive:

(7) *Hij wordt het boek gegeven.
he.subj is the book.obj1 given
"He is given the book."

(8) *Hij wordt zijn salaris doorbetaald.
he.subj is his wages.obj1 paid-through
"He is being paid his wages."

An exception to this pattern is observed in structures like the one in example (9) below. Moreover, when in active sentences headed by the verb *betalen* (to pay) the *primary* object (obj1) is not phonologically realised, then *krijgen* passive structures are also possible (see example (9b) below), in contrast to the behaviour of the rest of the Dutch ditransitives as presented in (6) in the previous section. This last pattern is also to be observed with the verb *uitkeren* (to pay out benefits; see example (10)).

(9) a. Hij wordt doorbetaald.
he.subj is paid-through
"He is being paid."

b. Hij krijgt doorbetaald.
he.subj gets paid-through
"He is getting paid."

(10) a. Hij krijgt uitgekeerd.
he.subj gets paid-out
"He is getting paid out benefits."

b. Hij wordt uitgekeerd.
he.subj is paid-out
"He is being paid out benefits."

But whereas (9a) and (9b) have the same meaning, (10b) does not entail the same as the sentence in (10a). Specifically, *hij* is the secondary object in (9a), (9b) and (10a), whereas it is the primary object in (10b). We will return to examples (9)–(10) in section 5.

## 3 Cross-linguistic evidence and previous analyses

German also exhibits similar passive structures to the Dutch ones we have presented in section 2. Interesting for our purposes here are the passives of German ditransitives shown in the following examples (from Müller (2003)):

(11) a. Der Mann hat den Ball dem Jungen
the man.Nom has the ball.Acc the boy.Dat
geschenkt.
given
"The man gave the ball to the boy."

b. Der Ball wurde dem Jungen geschenkt.
the ball.Nom was the boy.Dat given
"The ball was given to the boy."

c. Der Junge bekam/kriegte den Ball
the boy.Nom got the ball.Acc
geschenkt.
given
"The boy got the ball as a present."

Müller (2002), adapting Heinz and Matiasek's (1994) account of, among others, passivisation in German, proposes a raising analysis for the German *werden* passives (see example (11b) above) and a control-like analysis for the German *bekommen/kriegen* passives, like the one in example (11c) above. The lexical entry for the auxiliary *bekommen* in (12) below is (slightly modified) from Müller (2002, p. 149) and captures the gist of his analysis for the dative *bekommen/kriegen* passives in German.

(12) *bekomm-* (dative passive auxiliary)

$$\left[\begin{array}{l}\textsc{subcat}\left\langle \boxed{1}\ \text{NP}\left[\textit{str}\right]_{\boxed{2}}\right\rangle \oplus \boxed{3} \oplus \boxed{4} \\ \textsc{xcomp}\left\langle \text{V}\left[\begin{array}{ll}\textit{ppp} & \\ \textsc{lex} & + \\ \textsc{subcat} & \boxed{3} \oplus \left\langle \text{NP}\left[\textit{ldat}\right]_{\boxed{2}}\right\rangle \oplus \boxed{4} \\ \textsc{xcomp} & \langle\rangle \end{array}\right]\right\rangle \end{array}\right]$$

The control-like part of the account he proposes lies on the subject of the dative passive auxiliary being coindexed with the dative element of the embedded participle. As mentioned in Müller (2002, p. 149) "all elements from the SUBCAT list of the embedded verb are raised to the SUBCAT list of *bekommen* except for the dative object".

The analysis in (12) above for the German *bekommen/kriegen* passives is somewhat surprising given the fact that passive structures in German headed by *bekommen/kriegen* do not entail that somebody gets something, as the following examples from Müller (2002, p. 132) also aim at showing:

(13) Er bekam zwei Zähne ausgeschlagen.
he got two teeth PART(out).knocked
"He got two teeth knocked out."

(14) a. Der Bub bekommt/kriegt das Spielzeug
the lad gets the toy
weggenommen.
PART(away).taken
"The boy has the toy taken away from him."

b. Der Betrunkene bekam/kriegte
the drunk got
die Fahrerlaubnis entzogen.
the driving allowance withdrawn
"The drunk had his driving license taken away."

As Müller (2002, p. 132) also proposes "the meaning of *bekommen* and *kriegen* is bleached in these constructions. Therefore it is not justified to assume that the subject in such dative passive constructions is a receiver and gets a thematic role from *bekommen/erhalten/kriegen*". In other words, Müller (2002) also disfavours a control analysis for the German *bekommen/kriegen* "dative" passives.

The only reason imposing an analysis like the one presented in (12) we can think of is the realistic technical difficulty to have the lexically case marked dative secondary object (NP $\left[\textit{ldat}\right]$) of the SUBCAT list of the passive participle getting raised to the subject NP of the auxiliary *bekommen/kriegen*, which should bear a structural nominative case. Thus, the analysis in (12) only denotes an index sharing between the structurally case marked subject NP of the auxiliary *bekommen/kriegen* and the lexically case marked secondary object NP of the passive participle, in the spirit of a control analysis, instead of an entire synsem object sharing between these two NPs, which would have been expected under a raising analysis, as would have also, apparently, been favoured by Müller (2002).

## 4 Motivation for a raising analysis of passives in Dutch

The analysis we propose and formalise in the next section for the Dutch passives we have presented in section 2 is a uniform raising analysis. The motivation in favour of such an analysis, especially for the *krijgen* passives, in contrast to a control analysis like the one proposed in (12) in section 3, is based on the general treatment of raising and control phenomena presented in Pollard and Sag (1994).

Specifically, following Jacobson (1990), Pollard and Sag (1994, p. 141) show that whereas equi verbs allow NPs (or PPs) instead of their VP complement, this is never true for raising verbs (the examples are from Pollard and Sag (1994, pp. 141–142)):

(15) Leslie tried/attempted/wants something/it/to win.

(16) *Whitney seems/happens something/it.

Such contrasts between equi and raising verbs, Pollard and Sag (1994, p. 142) comment,"follow directly from the Raising Principle. [3] Since the raising verbs in (16) assign no semantic role to their subject argument, there must be an unsaturated complement on the same SUBCAT list. But NPs like *something* or *it* are saturated, and hence the SUBCAT list required for examples like those in (16) is systematically excluded."

*krijgen*-headed structures in Dutch behave in a similar way to raising structures like the one in example (16) above:

[3] Raising Principle (Pollard and Sag, 1994, p. 140): Let E be a lexical entry whose SUBCAT list L contains an element X not specified as expletive. Then X is lexically assigned no semantic role in the content of E if and only if L also contains a (nonsubject) Y $\left[\textsc{subcat}\ \ \left\langle \text{X}\right\rangle\right]$.

(17) ?Hij krijgt het boek toegestuurd en zijn buurman
he gets the book sent and his neighbour
krijgt dat ook.
gets that too
"*He is sent the book and his neighbour is that too."

(18) *Hij krijgt uitbetaald en Piet krijgt dat ook.
he gets paid and Peter gets that too
"*He gets paid and Peter gets that too."

*krijgen* does not introduce a semantic role (like the auxiliaries *worden* (passive) and *hebben* (perfect tenses)).

## 5 Formalisation of the analysis

Based on the motivation presented in section 4, we formalise our analysis for the Dutch *worden* passive in the lexical entry in (19) below and our analysis for the Dutch *krijgen* passive in the lexical entry in (20) below. Both lexical entries use the function *raise_to_nominative()* (Figure 1).[4]

This function takes a noun synsem, and preserves all values in the output, except for the CASE value, which is set to *nominative*.

As aimed at and expected, in both lexical entries below all the elements of the SUBCAT list of the embedded participle are raised to the SUBCAT list of *worden* and *krijgen*, respectively. In the case of *worden*, the accusative primary object of the embedded participle surfaces as the nominative subject of the auxiliary after raising. In the case of *krijgen*, it is the dative secondary object which surfaces as the nominative subject of the auxiliary after raising.[5]

(19) *worden* (passive auxiliary)

$$\left[\begin{array}{ll}\textsc{subcat} & \left\langle \text{raise\_to\_nominative}\left(\boxed{1}\right)\right\rangle \oplus \boxed{2} \oplus \boxed{3} \\ \textsc{xcomp} & \left\langle \text{V}\left[\begin{array}{ll} \textit{ppp} & \\ \textsc{lex} & + \\ \textsc{subcat} & \boxed{2} \oplus \left\langle \boxed{1}\ \text{NP}\left[\textsc{case}\quad \textit{acc}\right]\right\rangle \oplus \boxed{3} \\ \textsc{xcomp} & \langle\rangle \end{array}\right]\right\rangle \end{array}\right]$$

(20) *krijgen* (dative passive auxiliary)

$$\left[\begin{array}{ll}\textsc{subcat} & \left\langle \text{raise\_to\_nominative}\left(\boxed{1}\right)\right\rangle \oplus \boxed{2} \oplus \boxed{3} \\ \textsc{xcomp} & \left\langle \text{V}\left[\begin{array}{ll} \textit{ppp} & \\ \textsc{lex} & + \\ \textsc{subcat} & \boxed{2} \oplus \left\langle \boxed{1}\ \text{NP}\left[\textsc{case}\quad \textit{dat}\right]\right\rangle \oplus \boxed{3} \\ \textsc{xcomp} & \langle\rangle \end{array}\right]\right\rangle \end{array}\right]$$

The lexical entry in (19) accounts for the examples in (1b) and (5b) in section 2. In the case of example (1b) the value of $\boxed{2}$ in (19) is the empty list, since the verb *kussen* (to kiss) is transitive, and not ditransitive. $\boxed{3}$ may contain PP denoting the logical subject (*door Peter* in example (1b)).

The lexical entry in (20) accounts for the examples in (3b) and (4b) in section 2, where the ditransitive verbs have a primary object. For most ditransitive verbs, the primary object is compulsory, while for *uitkeren* and the *betalen*-family, it is optional. Example (6) demonstrates the former: the primary object is missing, while in (3b) and (4b) it is present (i.e. $\boxed{2}$ in (20) is a list containing the primary object). In examples (9b) and (10a) on the other hand, $\boxed{2}$ is the empty list: the primary object is absent.

This variation is a lexical property of the verbs, and not limited to the passive mood, as the following examples show.

(21) *Ik stuur hem toe.
I.subj send him.obj2 to
"*I send him."

(22) We betalen hem door.
We.subj pay him.obj2 through
"We continue to pay him."

(23) Ze keren het uit.
they.subj pay it. out
"They pay it out benefits."

(21) is (3) without (compulsory) primary object, (22) (4a) without (optional) primary object, and (23) (10) also without (optional) primary object.

As far as example (9) is concerned, we assume that the verb *betalen* (to pay), as well as its derivatives *doorbetalen*, *uitbetalen*, *terugbetalen*, etc., may also have a purely transitive use:

(24) a. Ik betaal de tuinman.
I.subj pay the gardener.obj1

b. De tuinman wordt betaald.
the gardener.subj is paid

[4] There are other ways in which the same effect can be obtained in a formalism. We chose a function because it is compact and easy to understand. Specifically, the function *raise_to_nominative()* (Figure 1) is really only an abbreviatory device, since it only consists of simple unifications. The same effect could be obtained, more verbosely, without functions.

[5] In our analysis, primary objects (obj1) bear accusative case, and secondary objects (obj2) dative case.

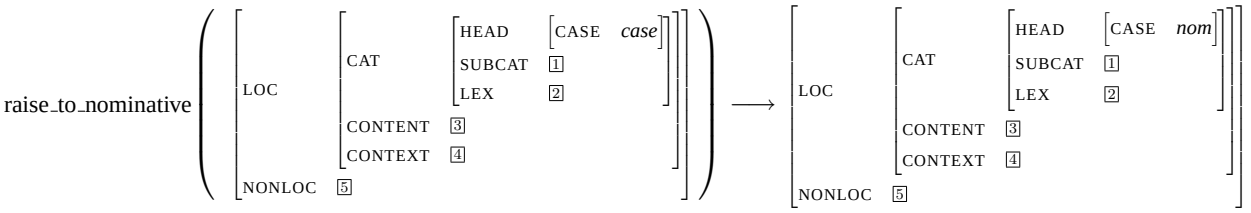



Figure 1: Definition of the function raise_to_nominative()

In such cases, the sole object of the active form of the *betalen*-family verbs is considered to be their primary object, which may, therefore, be accounted for by the auxiliary *worden* in (19). Then the value of 2 in (19) is the empty list, since the verb *betalen* (to pay) is considered to function as transitive, and not ditransitive.

## 6 Conclusion

We have motivated and formalised a uniform raising analysis for the *worden* and *krijgen* passives in Dutch. The analysis accounts for the Dutch data presented in section 2, without needing to find refuge to ad hoc theoretical and technical resorts, like the analysis of Müller (2002) (cf., the control-like analysis of the German *bekommen/kriegen* passives), as presented in section 3. The formalisation of the analysis in section 5 is essentially based on the fact that the information shared in raising constructions may leave out some paths from the SYNSEM information, while still remaining a raising analysis. In the case at hand, the SYNSEM value of the primary object of the embedded participle of the *worden* passive, as well as the SYNSEM value of the secondary object of the embedded participle of the *krijgen* passive, are raised to the subject of their respective auxiliaries, with only their CASE value changing to the nominative case required by the subject. Such a formalisation does not only account in a straightforward way for the behaviour of the Dutch data at hand (see section 2), but it can also offer a solution to the analysis presented in (12) in section 3 for the German *bekommen/kriegen* passives. Finally, such a formalisation also amends naturally the shortcomings of the intended raising analyses of German passives proposed in Kathol (1994) and Pollard (1994), which suggest that what should be raised to the subject of the *werden* and *bekommen/kriegen* passives is not the entire argument NP, but only its INDEX specification, since indices do not contain a specification for CASE, and they can, thus, belong to NPs with *different* case values without giving rise to a conflict. But as was also mentioned in section 3, structure-sharing only among indices points to a control analysis of passivisation in German. Thus, our analysis, which formally captures the fact that passivisation is based on structure-sharing of entire synsem objects, is the most straightforward analysis.

## References

Wolfgang Heinz and Johannes Matiasek. 1994. Argument structure and case assignment in German. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, pages 199–236. CSLI Publications. No. 46 in CSLI Lecture Notes.

Pauline Jacobson. 1990. Raising as function composition. *Linguistics and Philosophy*, 13:423–475.

Andreas Kathol. 1994. Passives without Lexical Rules. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, pages 237–272. CSLI Publications. No. 46 in CSLI Lecture Notes.

Stefan Müller. 2002. *Complex Predicates: Verbal Complexes, Resultative Constructions, and Particle Verbs in German*. Number 13 in Studies in Constraint-Based Lexicalism. Center for the Study of Language and Information, Stanford.

Stefan Müller. 2003. Object-to-subject-raising and lexical rule: An analysis of the German passive. In Stefan Müller, editor, *Proceedings of the HPSG-2003 Conference, Michigan State University, East Lansing*, pages 278–297. CSLI Publications.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Carl Pollard. 1994. Toward a Unified Account of Passive in German. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, pages 273–296. CSLI Publications. No. 46 in CSLI Lecture Notes.

# A Computational Treatment of V-V Compounds in Japanese

**Chikara Hashimoto**
Department of Infomatics
Kyoto University
Japan, 606-8501, Kyoto
hasimoto@pine.kuee.kyoto-u.ac.jp

**Francis Bond**
Machine Translation Research Group
NTT Communication Science Laboratories
Japan, 619-0237, Kyoto
bond@cslab.kecl.ntt.co.jp

## 1 Introduction

In this study, we examine how a large-scale computational grammar can account for the complex nature of Japanese verbal compounds ($V_1$-$V_2$ compounds, hereafter), such as *yomi-owaru* (read-finish) 'finish to read'. It is necessary to develop a linguistically accurate and computationally tractable analysis for $V_1$-$V_2$ compounds, since they are common in written documents and spontaneous speech, and, despite their surface simplicity, they show various complexities. To date, several computational Japanese grammars have been developed, but little attention has been paid to $V_1$-$V_2$ compounds. In fact, their approaches are either enumerating all $V_1$-$V_2$s in the lexicon as if they were single words without internal structures (the exhaustive listing approach) or simply concatenating the $V_1$ and $V_2$ of any kind of $V_1$-$V_2$ without taking into account the differences in their syntactic and semantic composition (the simple concatenation approach). The former suffers from undergeneration since some patterns are very productive and moreover a $V_1$-$V_2$ can embed another one: [[[*nade-mawasi*]-*tuzuke*]-*sobire*]-*kakeru* ([[[stroke-slue]-continue]-fail]-be.about.to). The latter approach leads to overgeneration since not all combinations of two verbs are allowed: **waki-ageru* (boil-raise) (cf. *waki-agaru* (boil-go.up)).

We develop the analysis of $V_1$-$V_2$s that is compatible with the linguistic analyses and observations made by Kageyama (1993) and Matsumoto (1996) while being computationally tractable. The analysis is implemented in JACY (Siegel & Bender, 2002) using the LKB system (Copestake, 2002) and evaluated with the Hinoki corpus (Bond et al., 2004) and the [incr tsdb()] system (Oepen & Carroll, 2000).

## 2 Data

$V_1$-$V_2$s show differences in terms of how productive they are, how their transitivity and case-marking are determined, whether or not they are compositional, and what semantic composition they undergo if they are compositional. First, as for their productivity, some $V_1$-$V_2$s are very productive and allow even a phrase in the $V_1$ position. In (2), for example, the $V_1$-$V_2$ headed by *sobireru* (fail) allows the phrasal $V_1$, *nade-te age* (stroke-TE give), while the $V_1$-$V_2$ headed by *mawasu* (caress) does not.

(1) a. *nade-sobireru* (stroke-fail) 'fail to stroke'
b. *nade-mawasu* (stroke-caress) 'fondle'

(2) a. *nade-te age-sobireru* (stroke-TE give-fail) 'fail to stroke for someone'
b.**nade-te age-mawasu* (stroke-TE give-caress) '?'

Second, some $V_1$-$V_2$s inherit $V_2$'s transitivity and case-marking (3), while others are given those of $V_1$'s (4).

(3) a. Ken-ga huku-o kiru (Ken-NOM clothes-ACC wear) 'Ken wears clothes.'
b. huku-ga kuzureru (clothes-NOM get.out.of.shape) 'Clothes get out of the shape.'
c. huku-ga *ki-kuzureru* (clothes-NOM *wear-get.out.of.shape*) 'Clothes get out of the shape by someone's wearing.'

(4) a. Ken-ga siai-ni katu (Ken-NOM game-DAT win) 'Ken wins games.'
b. Ken-ga siai-o tuzukeru (Ken-NOM game-ACC continue) 'Ken continues games.'
c. Ken-ga siai-ni *kati-tuzukeru* (Ken-NOM game-DAT *win-continue*) 'Ken continues to win games.'

Third, some $V_1$-$V_2$s show semantic compositionality (5), but others are highly lexicalized (6).

(5) a. *kaki-hazimeru* (write-begin) 'begin to write'
b. *naki-sakebu* (cry-shout) 'cry and shout'

(6) a. *uti-kiru* (hit-cut) 'abort'
b. *tori-simaru* (take-fasten) 'police'

Finally, compositional $V_1$-$V_2$s are composed in diverse ways. (7a)–(7b) correspond to (5a)–(5b), respectively.

(7) a. $\exists x\, \exists y$ begin($x$, write($x$, $y$))
b. $\exists x$ and(cry($x$), shout($x$))

## 3 Analysis

### 3.1 Linguistic Analyses

Kageyama (1993)'s insightful analysis claims that different behaviors of different $V_1$-$V_2$s are mostly predictable from how they are composed. He distinguishes two major types: syntactic $V_1$-$V_2$ compounds and lexical $V_1$-$V_2$ compounds. The two component verbs of syntactic $V_1$-$V_2$ compounds are combined in the syntax, while lexical $V_1$-$V_2$ compounds are formed in the lexicon. Accordingly, syntactic $V_1$-$V_2$s are generally as productive and compositional as ordinary phrases, but lexical $V_1$-$V_2$s are often irregular and idiomatic. Table 1 summarizes the characteristics of the two types in more detail.

Kageyama further divides syntactic $V_1$-$V_2$s into three types: Raising (e.g. $V_1$-*kakeru* ($V_1$-be.about.to) 'be about to $V_1$'), Control (e.g. $V_1$-*sobireru* ($V_1$-fail) 'fail to $V_1$'), and $\overline{\text{V}}$ complementation types (e.g. $V_1$-*tukusu* ($V_1$-exhaust) 'work out to $V_1$'). This is supported by, among other things, a contrast in passivizability; Raising and Control types do not allow passivization of $V_1$-$V_2$, as opposed to the $\overline{\text{V}}$ complementation type.

(8) hon-ga Ken-ni
book-NOM Ken-DAT
yomi-{*kake/*sobire/tukus}-rare-ta
read-{*be.about.to/*fail/exhaust}-PASS-PAST

Also, the three kinds show differences in whether $V_2$s thematically restrict their subjects and objects.

(9) a. ame-ga huku-o
rain-NOM clothes-acc
nurasi-{kake/*sobire/*tukusi}-ta
humidify-{be.about.to/*fail/*exhaust}-PAST
'The rain {was about/failed/worked out} to wet the clothes.'
b. Ken-ga atama-o
Ken-NOM head-ACC
hiyasi-{kake/sobire/*tukusi}-ta
cool-{be.about.to/fail/*exhaust}-PAST
'Ken {was about/failed/worked out} to cool off.'

Since $V_2$s of Control (*-sobireru*) and $\overline{\text{V}}$ (*-tukusu*) types put a thematic restriction on a subject, which the subject, *ame* (rain) in (9a), cannot satisfy, only the Raising type (*-kakeru*) is grammatical in the example. In (9b), only the $\overline{\text{V}}$ type is ruled out since it restricts an object to something that can be exhausted, but the object, *atama*, which is a part of the idiom, *atama-o hiyasu* 'cool off,' cannot meet the restriction.

Matsumoto (1996) classifies lexical $V_1$-$V_2$s into seven subtypes according to the semantic relations between $V_1$ and $V_2$. Each subtype, its example and a tentative semantics of the example are depicted in (10).

(10) PAIR $V_1$-$V_2$S: *naki-sakebu* (cry-shout)
→ and(shout($x$), cry($x$))
CAUSE $V_1$-$V_2$S: *yake-sinu* (burn-die)
→ cause(burn($x$), die($x$))
MANNER $V_1$-$V_2$S: *kake-yoru* (run-come)
→ in.manner.of(come($x$), run($x$))
MEANS $V_1$-$V_2$S: *tataki-kowasu* (hit-break)
→ by.means.of(break($x$, $y$), hit($x$, $y$))
$V_1$-$V_2$S WITH DEVERBALIZED $V_1$:
*sasi-semaru* (thrust-close)
→ emphasized.by(close($x$), thrust)

Table 1: Syntactic $V_1$-$V_2$s vs. Lexical $V_1$-$V_2$s

| | **Syntactic** | **Lexical** |
|---|---|---|
| **Productivity** | Very productive; the $V_2$s allow almost any $V_1$. | Not so productive; the combination of $V_1$ and $V_2$ is more restricted. |
| **Transitivity** | The $V_1$'s transitivity and case-marking are passed to the $V_1$-$V_2$. | Either $V_1$ or $V_2$ or both participate in the determination of transitivity and case-marking. |
| **Compositionality** | Compositional. | Some of them show varying degrees of compositionality, but others are highly lexicalized. |
| **Semantics** | The semantics of $V_2$ consistently embeds $V_1$'s semantics. | There are various kinds of semantic composition. |

$V_1$-$V_2$S WITH DEVERBALIZED $V_2$:
*hare-wataru* (clear.up-cross)
→ modified.by(clear.up($x$), cross)

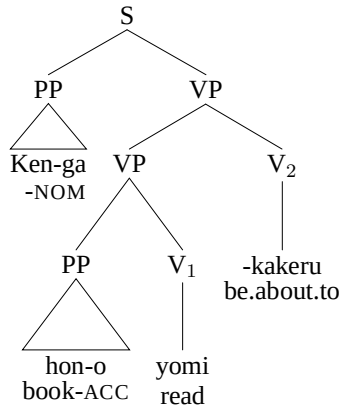
Matsumoto notes how the semantic relation determines the transitivity and the semantic composition of $V_1$-$V_2$ and posits a semantic analysis to deal with the phenomena. Although Matsumoto presents a precise and comprehensive analysis, it assumes fine-grained semantic notions and a complicating mapping theory. To implement this, the grammar would have to recognize which semantic relation holds between the two component verbs. But this depends heavily on world knowledge and pragmatic inference, and hence is not currently computationally tractable.

In sum, Kageyama (1993) and Matsumoto (1996) present useful analyses, but these must be revised to make them computationally tractable.

### 3.2 Computational Analysis — Proposal

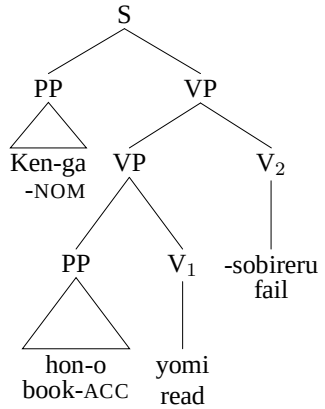
Our analysis of syntactic $V_1$-$V_2$s is mostly compatible with Kageyama (1993) but, as an HPSG analysis, assumes neither PRO nor government. (11) illustrates the analysis. (the V-embedding type corresponds to Kageyama's $\overline{\text{V}}$ complementation type.)

(11) a. **Raising**
'Ken is about to read a book.'



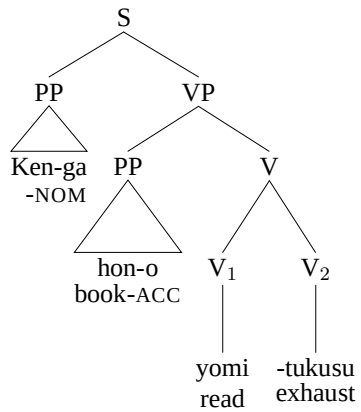

be.about.to(read(Ken, book))

b. **Control**
'Ken fails to read a book.'




fail(Ken, read(Ken, book))

c. **V-embedding**
'Ken reads a book thoroughly.'




exhaust(Ken, book, read(Ken, book))

The Raising and Control structures are almost the same as those of Sag et al. (2003); the subject of Raising type $V_2$ is "raised" from the $V_1$, and the subject of Control type $V_2$ controls that of the $V_1$. The V-embedding type has a structure where the subject and object of the $V_2$ control the subject and object of the $V_1$, respectively. These characteristics of the three are reflected in their semantic representations in (11). That is, the Raising type $V_2$, *kakeru* (be.about.to) in (11a), does not thematically restrict its subject, *Ken*, and object, *hon* (book), while the Control type $V_2$, *sobireru* (fail), puts a thematic restriction on its subject, *Ken*. The V-embedding type $V_2$ assigns thematic roles to both the subject and object. Clearly, these differences account for (9). Note, in addition, that the Raising and Control types have a VP embedding structure, while the V-embedding type does not. The contrast in (8) is accounted for by the difference; only the object of the V-embedding type is selected by both the $V_1$ and $V_2$, thus only this structure allows the passivization of $V_1$-$V_2$ as a whole. Other things to notice are that it is the $V_1$ that determines the $V_1$-$V_2$'s transitivity and, in most cases, case-marking, and that their semantic structures are consistently embedding structures.

One of the divergences from Kageyama (1993) involves the $V_1$ passivization. Kageyama (1993) always accepts the $V_1$ passivization of Control type but necessarily rules out that of his $\overline{\text{V}}$ complementation type, based on the difference in their syntactic configurations: the VP complement vs. the $\overline{\text{V}}$ complement. But this is incorrect as shown in (12).

(12) a.*hon-ga yom-are-sobireru
book-NOM read-PASS-fail
'A book fails to be read.'

b. Ken-ga nagur-are-tukusu
Ken-NOM punch-PASS-exhaust
'Ken endures the successive punches.'

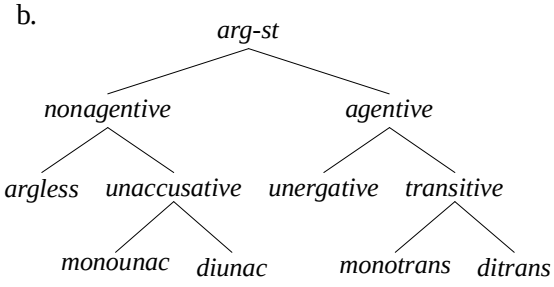
We basically allow all $V_1$ passivizations but semantically restrict them. In (12a), for example, the subject, *hon* (book), cannot be construed as FAILER. In (12b), on the other hand, *Ken* can be interpreted as the one who exhausts himself by being punched a lot.

As for lexical $V_1$-$V_2$s, we classify them into five subtypes roughly following Matsumoto (1996).

(13) a. **Right-headed $V_1$-$V_2$s**
b. **Argument mixing $V_1$-$V_2$s**
c. **$V_1$-$V_2$s with deverbalized $V_1$**
d. **$V_1$-$V_2$s with deverbalized $V_2$**
e. **Non-compositional $V_1$-$V_2$s**

The Right-headed and Argument mixing types jointly cover most of Matsumoto's Pair, Cause, Manner and Means compounds. The Non-compositional type is introduced to distinguish compositional and non-compositional $V_1$-$V_2$s. Unlike Matsumoto's finer grained semantic analysis, our analysis leaves the exact semantic relationship under-specified. The constraints on composition come from an extended ARG-ST. As illustrated in (14), the ARG-ST consists of one EXTernal argument and two INTernal arguments and is classified into six types, following Imaizumi and Gunji (2000).

(14) a.
$$\begin{bmatrix} \textit{arg-st} & \\ \text{EXT} & \textit{index} \\ \text{INT1} & \textit{index} \\ \text{INT2} & \textit{index} \end{bmatrix}$$

b.


c.

| | EXT | INT1 | INT2 |
|---|---|---|---|
| *argless* | × | × | × |
| *monounac* | × | ◯ | × |
| *diunac* | × | ◯ | ◯ |
| *unergative* | ◯ | × | × |
| *monotrans* | ◯ | ◯ | × |
| *ditrans* | ◯ | ◯ | ◯ |

First, the Right-headed $V_1$-$V_2$ obeys the Shared Participant Condition (Matsumoto, 1996), which requires that the two component verbs share at least one argument that is co-indexed with an argument of the other component verb. Any two arguments can be co-indexed between $V_1$ and $V_2$ if the arguments agree in the EXT/INT distinction. The transitivity and case-marking of the $V_1$-$V_2$ are inherited from the $V_2$ (hence Right-headed). The semantics is totally compositional; the two semantic representations of the $V_1$ and $V_2$ are predicated by an underspecified semantic relation, which can be specified as Pair, Cause, Manner or Means by a component outside the grammar. For example, the semantic representations of the first two $V_1$-$V_2$s in (10) can be glossed as *unspec_rel(shout(x),cry(x))* and *unspec_rel(burn(x),die(x))*. The semantic relation cannot be fully specified in a purely syntactic account since it is affected by contexts, pragmatics, and world knowledge, as these become available, the relation can be constrained further. Further, the underspecification greatly simplifies the implementation. The Right-headed $V_1$-$V_2$, formulated in this way, covers most of the lexical $V_1$-$V_2$s (Matsumoto's Pair, Cause, Manner and Means) without making the grammar complicated.

Second, the Argument mixing $V_1$-$V_2$ has a peculiarity; it is ambiguous in that they can take arguments from either the $V_1$ or $V_2$. *nomi-aruku* (drink-walk), for example, can take as the object either something to drink ($V_1$'s argument) or a place to walk ($V_2$'s argument), according to Matsumoto (1996). To account for this, we underspecify the transitivity and case-marking of the $V_1$-$V_2$ such that they can be inherited from either the $V_1$ or $V_2$. Another peculiarity involves the fact that the $V_2$ is restricted to a *monotrans* verb that expresses a spatial motion,[1] while the $V_1$ is *transitive* and must not be a spatial motion verb. As for the semantics, it is the same as that of the Right-headed $V_1$-$V_2$ except that the semantic relation is alway construed as Manner.

Third, the $V_1$-$V_2$ with deverbalized $V_1$ includes a $V_1$ that is deverbalized and only emphasizes the content of $V_2$ in some way (Kageyama, 1993; Matsumoto, 1996). For instance, *sasi-semaru* (thrust-close), in our analysis, represents something like *emphasize(close(x))*. In the sense that the $V_1$ is deverbalized, the $V_1$-$V_2$ is considered not fully compositional. Naturally, as the $V_1$ is deverbalized, it is the $V_2$ that determines the transitivity and case-marking of the $V_1$-$V_2$. As Kageyama (1993) notes, there is no restriction on the possible combinations of the $V_1$ and $V_2$ in terms of ARG-ST.

Fourth, the $V_1$-$V_2$ with deverbalized $V_2$, as the name implies, includes a $V_2$ that loses its original verbal meaning and takes on an adverbial meaning that modifies the $V_1$ (Kageyama, 1993; Matsumoto, 1996). For instance, *hare-wataru* (clear.up-cross) ican be glossed as *cross(clear.up(x))* in our analysis. Similarly to the $V_1$-$V_2$ mentioned in the last paragraph, this type of $V_1$-$V_2$ is also considered not fully compositional, since the $V_2$ has lost its original verbal meaning. Regarding the transitivity and case-marking of the $V_1$-$V_2$, the $V_1$ determines them since the $V_2$ is deverbalized. In addition, according to Kageyama (1993), the $V_1$ and $V_2$ of this type must agree in agentivity, unlike the $V_1$-$V_2$ with semantically deverbalized $V_1$.

The two types with a deverbalized component verb lexically encode an embedding semantic structure, similarly to the lexical treatment of the 'biclausal' nature of Japanese causatives proposed by Manning et al. (1996).

As for productivity, the first two types are more productive than the last two. Actually, we can freely coin a $V_1$-$V_2$ that belongs to the first one, the Right-headed $V_1$-$V_2$, as long as it is semantically and pragmatically plausible. On the other hand, the Non-compositional $V_1$-$V_2$ is absolutely not productive and literally non-compositional; the $V_1$-$V_2$ is totally lexicalized and should be analyzed as a single word.

All in all, even though our analysis might be coarser than Kageyama (1993) and Matsumoto (1996), it is sufficient to account for $V_1$-$V_2$'s complex characteristics summarized in §2 and Table 1 and, what is more, is computationally tractable.

[1] In the JACY framework, a locative accusative argument is considered an object.

## 4 Evaluation

To see if our implementation works well in practice, we conducted a corpus-based evaluation and examined its coverage, the amount of ambiguity, and efficiency. First, we extracted a small evaluation corpus from the Hinoki corpus (Bond et al., 2004). The evaluation corpus consists of 219 sentences, where each sentence contains at least one $V_1$-$V_2$. In addition, we prepared two versions of JACY: JACY-plain and JACY-vv. JACY-plain is given no $V_1$-$V_2$ implementation but contains 1,325 lexical entries in the lexicon, which were added by the developers over the course of its development. In contrast, JACY-vv is equipped with all the $V_1$-$V_2$ implementations but without any compositional $V_1$-$V_2$ entries in the lexicon. Table 2 shows the results of the experiment. We find that JACY-vv gains more coverage and less ambiguity than JACY-plain. The increased coverage is due to the remarkable productivity of the Right headed type. The reduction in ambiguity involves the more restricted nature of our approach to the free word order of Japanese. The table also shows the two versions' processing efficiency: **time** and **space**.[2] Adding the rules and lexical types for $V_1$-$V_2$s slightly degrades JACY-vv's efficiency. However, JACY-vv still works fast enough for practical NLP applications.

Table 2: Experimental results

| | **JACY-plain** | **JACY-vv** |
|---|---|---|
| **Coverage** (%) | 52.1 | 63.5 |
| **Ambiguity** ($\phi$) | 53.41 | 50.78 |
| **time** ($\phi$) | 4.85 | 6.43 |
| **space** ($\phi$) | 816779 | 995681 |

## Acknowledgement

We appreciate many people for helping this research. We especially thank Takao Gunji, Melanie Siegel, Dan Flickinger, Sato Satoshi and NTT Machine Translation Research Group.

[2] **time** shows how long the grammar needs to parse one sentence, and **space** shows how much memory the grammar consumes to parse one sentence.

## References

Bond, F., Fujita, S., Hashimoto, C., Nariyama, S., Nichols, E., Ohtani, A., Tanaka, T., & Amano, S. (2004). The Hinoki Treebank — A Treebank for Text Understanding. In *Proceedings of the First International Joint Conference of Natural Language Processing*, pp. 554–559.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications.

Imaizumi, S., & Gunji, T. (2000). Complex Events in Lexical Compounds. In Itou, T., & Yatabe, S. (Eds.), *Lexicon and Syntax* (in Japanese), pp. 33–59. Hitsuji Shobou.

Kageyama, T. (1993). *Grammar and Word Formation* (in Japanese). Hitsuji Shobou.

Manning, C. D., Sag, I. A., & Iida, M. (1996). The Lexical Integrity of Japanese Causatives. In Gunji, T. (Ed.), *Studies in the Universality of Constraint-Based Structure Grammars*, pp. 9–37. Osaka University.

Matsumoto, Y. (1996). *Complex Predicates in Japanese: A Syntactic and Semantic Study of the Notion 'Word'*. CSLI Publications.

Oepen, S., & Carroll, J. (2000). Performance profiling for grammar engineering. *Natural Language Engineering*, 81–97.

Sag, I. A., Wasow, T., & Bender, E. M. (2003). *Syntactic Theory: A Formal Introduction* (2 edition). CSLI Publications, Stanford.

Siegel, M., & Bender, E. M. (2002). Efficient Deep Processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization* Taipei, Taiwan.

# On Non-Canonical Clause Linkage

**Anke Holler**
Department of Computational Linguistics
Institute of General and Applied Linguistics
University of Heidelberg
69117 Heidelberg, Germany
holler@cl.uni-heidelberg.de

## 1 Introduction

In generative grammar, it is commonly assumed that clauses that can stand alone as complete sentences differ grammatically from ones that are dependent on a matrix clause and are in this respect subordinated. This difference is often expressed by a boolean feature called ROOT (or alike), and by analysing +ROOT-clauses as syntactically highest clauses. The stipulation of a ROOT feature has been motivated by an observation going back to Emonds (1970) whereby clauses vary in admitting of so-called root phenomena. Whereas +ROOT clauses support these phenomena, -ROOT clauses disallow them.[1]

Contrary to this assumption, Green (1996) argues that the best explanation of the acceptability of root phenomena in embedded clauses is not a syntactic, but a pragmatic one, and thus distinguishing dependent clauses from independent utterances can be done ROOT-less. Working within construction-based HPSG, Green (1996) suggests to introduce a new dimension of clauses, called DEPENDENCY, with three partitions *subordinate*, *main* and *indifferent* with most subtypes of clauses being indifferent as to whether they act as main clauses or subordinate clauses. While Green (1996) is correct in assuming that a binary feature is not justified for the distinction of main and subordinate clauses, her approach must be revised to cover dependent clauses that simultaneously behave like main and subordinate clauses with respect to their syntactic form, their interpretation, and their functional usage, and therefore indicate that a pure pragmatic account is not adequate.

The paper is structured as follows: In the next section, several non-canonical clause linkage phenomena occuring in German will be discussed which challenge any approach implementing a twofold differentiation between main and subordinate clause types. Recent HPSG seems well equipped to handle the presented data as will be shown in sec. 3. There, a constraint-based analysis will be sketched that makes use of the idea that feature structures describing clause types can be organized according to the way the respective clause is linked to its syntactic surrounding. Sec. 4 provides some concluding remarks.

## 2 The Data

In German, a typical SOV language, canonical subordinated clauses differ from canonical main clauses by the position of the finite verb. Whereas the finite verb in main clauses is fronted (henceforth called 'V2'), it occurs in clause-final position (henceforth called 'VF') in subordinated clauses. This well-known fact forms the basis of previous HPSGian work on the classification of German clause types, cf. Uszkoreit (1987), Kathol (1995) and Netter (1998), in which the position of the finite verb (i.e. V2 versus VF) is 'hard-wired' to the type or the feature representing main and subordinate clauses, resp. For instance, Kathol (1995) introduces two subtypes of the type *clause*, called *root* and *subordinate*, and partitiones *root* by *v1* and *v2*. Tracing the traditional descriptive model of Topological Fields, cf. Drach (1937), he formulates a set of constraints on constituent order domains, cf. Reape (1994), such that the finite verb is restricted to a par-

[1] For a listing of these phenomena see among many others Hooper und Thompson (1973), Green (1996), Heycock (2002). As for German, an initial position of the finite verb is usally taken as a typical root property.

ticular topological field in dependence of the respective clause type. Thus, for any clause of type *subordinate* the finite verb has to be in clause final position whereas the finite verb of clauses of type *root* always stands in clause initial position. Additionally, Kathol (1995) assumes that clauses of type *root* bear a PHON feature but not clauses of type *subordinate* arguing that *root* clauses only can be uttered independently.[2]

Splitting clause types into root and subordinate depending on the position of the finite verb and the presence of PHON, as Kathol (1995) does it, yields an approach that classifies dependent V2 clauses such as (1a) as root but independent VF clauses such as (1b) as subordinate, predicting contrary to the facts that the respective V2 clause is uttered independently but not the VF one.

(1) a. Ich glaube, er hat recht.
*I think he has right*
'I think that he is right.'

b. Ob er noch kommt?
*Whether he still comes*
'I wonder whether he will still come?'

Reis (1997), however, has demonstrated that dependent V2 clauses like (1a) similarly show properties of clear subordinate clauses *and* clear root clauses, and thus can be assigned to either of them. As evidence she gives inter alia that dependent V2 clauses (i) are information-structurally integrated into their matrix clause signaled by a rising tone at the end of the matrix predicate, (ii) admit variable binding from the matrix clause, (iii) are restricted to a final position within the matrix clause, which means that they must not occur initially or in the so-called middle field, (iv) disallow correlates and *und zwar*-supplements, and (v) disallow extraction.[3] If dependent V2 clauses were the single clausal class exhibiting the listed properties, one might seek for an idiosyncratic explanation closely related to the properties of their matrix clauses. In German, however, there exist several types of clauses showing similar mixed properties in terms of a root-subordinate distinction, albeit occuring in miscellaneous syntactic environments. Reis (1997) provides evidence that the so-called free *dass*-clauses, cf. (2a), have the properties (i) to (v), Gärtner (2001) observes them with a certain class of restrictive relative clauses dubbed V2 relatives, cf. (2b).

(2) a. Er muss im Garten sein, dass er
*He must in the backyard be that he*
nicht aufmacht.
*not opens*
'He must be in the backyard since he does not open.'

b. Das Blatt hat eine Seite, die ist
*The sheet has one side that is*
ganz schwarz.
*completely black*
'The sheet has one side that is completely black.'

Reis (1997) and Gärtner (2001) further show that these clauses are in semantic respects different from their canonical counterparts: In contrast to ordinary complement clauses, dependent V2 clauses and free *dass*-clauses do not realize an argument of the matrix predicate. Also, V2 relatives are interpreted restrictively but differ from restrictive relative clauses in that they are limited to indefinite noun phrases. Thus, the three types of clauses behave all about the same in terms of a restricted licensing by the matrix clause. In addition, dependent V2 clauses share with free *dass*-clauses that they cannot be interpreted in the scope of negation or negative predicates. Similarly, V2 relatives cannot attach to a negated noun phrase, neither.

Pragmatically, the aforementioned clauses have one property in common: They all have illocutionary force. Even though their illocutionary association somehow seems to be related to the matrix clause, cf. Boettcher (1972), Reis (1997), Gärtner (2002) and Meinunger (2004), the fact itself shows that the clauses cannot be ordinary embedded clauses, cf. Green (2000b).

The grammatical properties of the clauses just

[2] Netter (1998) combines verbal position and the root-subordinate distinction by stipulating types of the following kind: *V-2 Declarative Main*, *V-Final Declarative Subordinate*, *V-2 Interrogative Main*, *V-Final Interrogative Subordinate*, etc. Uszkoreit (1987) formulates restrictions relating the value of the boolean feature M(AIN)C(LAUSE) to the value of the boolean feature INV(ERTED) which represents the finite verb's clausal position.

[3] (i) and (ii) are typical properties of subordinate clauses whereas (iii) to (v) usually substantiate root clauses.

considered indicate that their relation to a potential matrix clause is not canonical inasmuch they are not clear-cut subordinated (embedded) clauses.[4] Interestingly, there exists yet another class of dependent clauses that are not canonically linked to their syntactic surrounding in German. This class comprises at least the so-called *weil*-V2 adverbial clauses, cf. (3a), and non-restrictive relative clauses of any kind, in particular *wh*-relatives, cf. (3b).[5]

(3) a. Peter kommt zu spät, weil er hat
*Peter comes too late because he has*
keinen Parkplatz gefunden.
*no parking lot found*
'Peter is late because he could not find a parking lot.'

b. Max spielt Orgel, was gut klingt.
*Max plays organ which good sounds*
'Max is playing the organ, which sounds good.'

It can be shown that the clauses in (3) introduced by *weil* and *was* respectively are prosodically and pragmatically independent from their matrix clause, which is indicated by an independent focus domain and an autonomous illocutionary force. In addition, these clauses are syntactically dispensable, disallow variable binding from outside and occur only at the very end of a complex sentence. Moreover, their semantic interpretation is peculiar. *Weil*-V2 adverbial clauses, for instance, behave differently from canonical *weil*-clauses in that they are able to give reasons for a speaker's attitude.[6] *Wh*-relatives are introduced by an anaphoric pronoun and denote propositions, which is certainly a consequence of their non-restrictiveness and contrasts with restrictive relative clauses which are usally analyzed as denoting properties. Finally, negation does not scope over these clauses, neither.

Looking at the data given so far reveals that three classes of dependent clauses can be distinguished dependending on the way of being linked to their linguistic surrounding. Besides the canonical dependent clauses including all clauses that form directly or indirectly a component part of their matrix clause (such as complement clauses of all kinds, ordinary adverbial clauses, restrictive relative clauses, etc.), two classes of dependent, but non-canonically linked clauses can be identified by means of the grammatical properties afore described. Table 1 gives an overall picture of these facts.[7]

| **Clausal Class** | **I** | **II** | **III** |
|---|---|---|---|
| Typical example | a (VF)<br>b (VF)<br>c (VF) | d (V2)<br>e (V2)<br>f (VF) | g (VF)<br>h (VF)<br>i (V2) |
| Prosodically integrated | yes | yes | no |
| Syntactically attached | yes | yes | no |
| Semantically peculiar | no | yes | yes |
| Independent information struct. | no | no | yes |
| Independent illocutionary force | no | yes | yes |

Table 1: Grammatical properties of three empirically identified clausal classes

It strikes that the position of the finite verb is not appropriate to differentiate between these clausal classes. Rather, the data suggest that the clauses differ in the degree to which they are integrated into a potential matrix clause.

## 3 Accounting for the Data

The sign-based monostratal architecture of HPSG qualifies very well to account for the presented data. The core of the analysis advocated here is the observation that clauses vary with respect to the way they are linked to their linguistic surrounding. Because this originates from syntactic, semantic and pragmatic properties of the clauses involved, it seems to be natural to encode it in grammar. In HPSG, the type hierarchy lends itself to reconstruct the observed distinction. For this reason, it is proposed to partition the type *phrase* in terms of a dimension LINKAGE, and to distinguish between *unlinked* and *linked* objects. The type *unlinked* comprises all independently uttered sentences includ-

[4] On the other hand, they do not show properties of well-defined main (root) clauses, neither.

[5] *Weil*-V2 adverbial clauses are mainly attested for colloquial German, but can be observed in written German as well, cf. Uhmann (1998), who extensively describes this clausal class. Holler (2003) provides a comprehensive analysis of the grammar of *wh*-relatives.

[6] See Haegeman (1984) for a discussion of similar phenomena in English.

[7] For reasons of space, the following abbreviations are used: a = complement clause, b = restrictive relative clause, c = standard adverbial clause, d= dependent V2 clause, e = restrictive V2 relative clause, f = free *dass*-clause, g = non-restrictive *d*-relative clause, h = non-restrictive *wh*-relative clause, i = *weil* V2 adverbial clause.

ing VF-clauses as given by (1b). The type *linked* which describes all objects somehow combined with the linguistic surrounding is further partitioned by the types *integr*(*ated*), *semi-integr*(*ated*) and *non-integr*(*ated*), which represent clausal objects that are fully, partly or not integrated into a potential matrix clause.[8] It is assumed that the newly defined types are cross-classified with subtypes of *phrase* coming from other dimensions such as CLAUSALITY and HEADEDNESS, cf. Sag (1997).

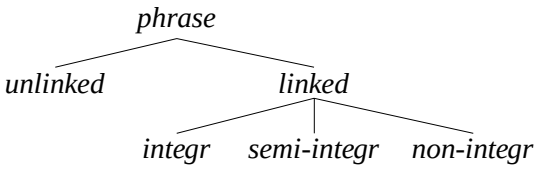



Figure 1: Partition of *phrase* w.r.t. the dimension LINKAGE

Nothing in particular shall be said here about clauses of type *integrated*, since they are analyzed in a standard way. The two remaining clausal classes of type *linked*, i.e. *semi-integrated* and *non-integrated* clauses, are certainly more instructive. Next, an analysis will be sketched which formulates restrictions on these two types and, thus, captures the syntactic, semantic and pragmatic properties of the clauses discussed in sec. 2.

Clauses of type *semi-integrated*: Although *semi-integrated* clauses are less tightly connected to their matrix clause as they show the properties (iii) to (v) presented in sec. 2, it is obvious that they are syntactically attached to it. Thus, they are analyzed as modifiers of a saturated verbal projection. By following Engdahl und Vallduví (1996) in stipulating an INFO-STRUCT attribute that enriches CONTEXT, it can be required that *semi-integrated* clauses identify their INFO-STRUCT value with that of the matrix clause, which easily copes with property (i). In addition, an *psoa* object of type *intend*, cf. Green (2000a), is contained in the BACKGROUND set of a *semi-integrated* clause, thereby accounting for the empirical fact that these clauses have illocutionary force.[9] The constraint on objects of type *semi-integrated* shown in fig. 2 expresses these restrictions.

*semi-integrated* →

$$\left[\begin{array}{l} \text{SS} \mid \text{LOC} \left[\begin{array}{l} \text{CAT} \mid \text{HD} \mid \text{MOD} \left[\text{LOC} \left[\text{CAT} \left[\begin{array}{ll}\text{HD} & \textit{verb}\\ \text{SUBCAT} & \langle\rangle\end{array}\right]\right]\right] \\ \text{CXT} \left[\begin{array}{l}\text{INFO-STR} \boxed{1} \\ \text{BACKGR} \left\{\left[\textit{intend}\right], \ldots\right\}\end{array}\right] \end{array}\right] \\ \text{CXT} \mid \text{INFO-STR} \boxed{1} \end{array}\right]$$

Figure 2: Restricting *semi-integrated* clauses

Clauses of type *non-integrated*: Adapting the approach to peripheral adverbials of Haegeman (1991), it is assumed that clauses of type *non-integrated* are orphan constituents which are syntactically unattached.[10] By providing additional background information, orphaned clauses serve to form the discourse frame against which the proposition expressed in the matrix clause is evaluated. Hence, the modification relation is not established in syntax, but rather at the level of utterance interpretation. This can easily be implemented into the grammar by introducing phrases of type *head-orphan-phrase* as subtype of *headed-phrase*, cf. Sag (1997), and requiring that the CONTENT value of the orphan is unified with the BACKGROUND set of the head as illustrated in fig. 3. The fact that an orphan is not included into the host's information structure and has illocutionary force of its own is again grasped by manipulating the INFO-STRUCT and BACKGROUND values of phrases of type *head-orphan-phrase*. Since it is assumed that *non-integrated* clauses are cross-classified as a subtype of *head-orphan-phrase*, they have to obey the restrictions for orphans. This analysis provides a vanilla account of the properties of *non-integrated* clauses as described in sec. 2.[11]

[8] Unfortunately, it cannot be discussed here to which extent this distinction can be used for constituents other than clauses. At least, there is evidence from German and English that nominal left-peripheral elements also need to be classified regarding their degree of (non-)integrateness into a clause, cf. Shaer und Frey (2004).

[9] Of course, any other analysis of illocutionary force could have been implemented here.

[10] Haegeman (1991) points out that this does not mean that orphans would be syntactically unconstrained.

[11] The fact that negation neither takes scope over *semi-integrated* clauses nor over *non-integrated* ones can easily be implemented in the lexicon by restricting the negation and the negative verbs to clauses of type *integrated*. Further, LP rules are defined which limit clauses of types *semi-integrated* and *non-integrated* to final positions in a complex sentence structure.

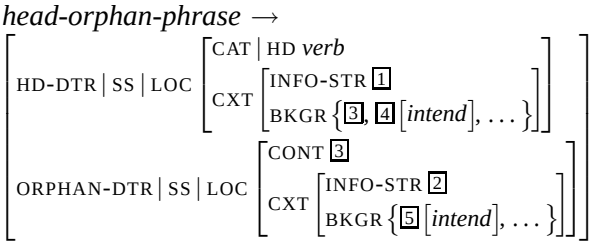



Figure 3: Restricting orphan constituents such as *non-integrated* clauses

## 4 Conclusion

Considering as example German, the present paper has investigated a certain subset of clause linkage phenomena and has developed a constraint-based analysis accounting for the empirical fact that clauses need to be distinguished w.r.t. their degree of integratedness into a potential matrix clause. It has been shown that the generally assumed twofold distinction between main and subordinate clauses (or root and embedded clauses) does not suffice to deal with the presented data. Moreover, it has been argued that the discussed linkage phenomena originate from syntactic, semantic and pragmatic properties of the clauses involved, and should hence be encoded in grammar. By partitioning objects of type *phrase* in terms of a LINKAGE dimension and by constraining the CONTEXT value of these objects, the data are covered without any reference to the position of the finite verb. Additionally, non-integrated clauses are considered as 'orphan' constituents which are unattached in syntax, but provide the context for the interpretation of the matrix clause. Such an approach explains the empirical facts assembled in a straightforward way. Further research must show to what extent the proposed analysis can cope with similar phenomena in other languages.

## References


Boettcher, W. (1972). *Studien zum zusammengesetzten Satz*. Frankfurt/M.: Athenäum.

Drach, E. (1937). *Grundgedanken der deutschen Satzlehre*. Frankfurt: Diesterweg.

Emonds, J. (1970). *Root and Structure-Preserving Transformations*. Doktorarbeit, MIT, Cambridge.

Engdahl, E. und E. Vallduví (1996, May). Information Packaging in HPSG. In C. Grover und E. Vallduví (Hrsg.), *Edinburgh Working Papers in Cognitive Science, Vol. 12: Studies in HPSG*, Chapter 1, 1–32. Scotland: Centre for Cognitive Science, University of Edinburgh.

Green, G. (1996). Distinguishing main and subordinate clause: The root of the problem. unpl. Ms., University of Illinois.

Green, G. M. (2000a). The Nature of Pragmatic Information. In R. Cann, C. Grover, und P. Miller (Hrsg.), *Grammatical Interfaces in HPSG*, Nummer 8 von Studies in Constraint-Based Lexicalism, 113–138. Stanford: CSLI Publications.

Green, M. (2000b). Illocutionary Force and Semantic Content. *Linguistics and Philosophy 23*, 435–473.

Gärtner, H.-M. (2001). Are there V2-Relative Clauses in German? *Journal of Comparative Germanic Linguistics 3*(2), 97–141.

Gärtner, H.-M. (2002). On the Force of V2 Declaratives. *Theoretical Linguistics*.

Haegeman, L. (1984). Remarks on Adverbial Clauses And Definite NP-Anaphora. *Linguistic Inquiry 15*, 712–715.

Haegeman, L. (1991). Parenthetical Adverbials: The Radical Orphanage Approach. In S. Chiba, A. Ogawa, Y. Fuiwara, N. Yamada, O. Koma, und T. Yagi (Hrsg.), *Aspects of Modern English Linguistics*, 232–254. Tokyo: Kaitakusha.

Heycock, C. (2002). Embedded root phenomena. unpl. Ms., University of Edinburgh.

Holler, A. (2003). An HPSG Analysis of the Non-Integrated Wh-Relative Clauses in German. In S. Müller (Hrsg.), *Proceedings of the HPSG-2003 Conference, Michigan State University, East Lansing*, Stanford, 163–180. CSLI Publications.

Hooper, J. und S. Thompson (1973). On the applicability of root transformations. *Linguistic Inquiry 4*, 465–497.

Kathol, A. (1995). *Linearization-Based German Syntax*. Doktorarbeit, Ohio State University.

Meinunger, A. (2004). Verb position, verbal mood and the anchoring (potential) of sentences. In

H. L. und Susanne Trissler (Hrsg.), *The syntax and semantics of the left periphery*, 313–341. Mouton de Gruyter.

Netter, K. (1998). *Functional Categories in an HPSG for German*. Nummer 3 von Saarbrükken Dissertations in Computational Linguistics and Language Technology. Saarbrücken: German Research Center for Artificial Intelligence (DFKI) and University of the Saarland.

Reape, M. (1994). Domain Union and Word Order Variation in German. In J. Nerbonne, K. Netter, und C. J. Pollard (Hrsg.), *German in Head-Driven Phrase Structure Grammar*, 151–197. Stanford University: CSLI Publications.

Reis, M. (1997). Zum syntaktischen Status unselbständiger Verbzweit-Sätze. In C. Dürscheid, K. H. Ramers, und M. Schwarz (Hrsg.), *Syntax im Fokus. Festschrift für Heinz Vater.* Tübingen: Niemeyer.

Sag, I. A. (1997). English Relative Clause Constructions. *Journal of Linguistics 33*(2), 431–484.

Shaer, B. und W. Frey (2004). Integrated and Non-Integrated Leftperipheral Elements in German and English. In W. F. Benjamin Shaer und C. Maienborn (Hrsg.), *Proceedings of the Dislocated Elements Workshop, ZAS Berlin 2003*, Band 2 of *ZAS Papers in Linguistics 35*, 465–502.

Uhmann, S. (1998). Verbstellungsvariation in weil-Sätzen. *Zeitschrift für Sprachwissenschaft 17*(1), 92–139.

Uszkoreit, H. (1987). *Word Order and Constituent Structure in German*. Chicago University Press.

# Gradience and Parametric Variation

**Frank Keller**
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
keller@inf.ed.ac.uk

**Dora Alexopoulou**
Department of Linguistics
University of Cambridge
Sidgwick Avenue
Cambridge CB4 9DA, UK
ta259@cam.ac.uk

## 1 Introduction

Gradient grammaticality has received renewed attention in recent years, which is partly due to a an innovation in experimental methodology, viz., the introduction of the *magnitude estimation* paradigm (Bard et al., 1996; Cowart, 1997) that allows the elicitation of reliable gradient acceptability judgments. The application of this methodology to crosslinguistic variation has revealed a number of interesting results, but these results also pose an important challenge for a parametric approach to variation: often, variation is confined to quantitative differences in the magnitude of otherwise identical principles. Here we approach the issue with particular reference to crosslinguistic studies focusing on superiority and locality violations involved in weak islands (*whether*-islands).

## 2 Experimental Results on Superiority and Relativized Minimality

Two important experimental investigations of gradience and crosslinguistic variation are Featherston's (2005) comparative study of superiority effects and d-linking in English and German, and Meyer's (2003) study of superiority effects in Russian, Polish, and Czech. The main results can be summarized as follows:

- A clear (statistically significant) dispreference for in-situ subjects (English, German, Russian, Polish, Czech, modulo a "reverse animacy" effect in Polish).
- A clear crosslinguistic effect of discourse-linking, where in-situ d-linked subjects are essentially as acceptable as other in-situ phrases.
- Crosslinguistically, the d-linking status of the object is irrelevant (English, German, Polish, Czech).
- No clear interactions between arguments and adjuncts are detected (English, German, Polish, Czech, Russian).
- Not only in-situ subjects are dispreferred, but initial subjects are preferred (marginal effect in German, significant in English).

These studies therefore show that crosslinguistic variation is confined to quantitative differences in otherwise crosslinguistically stable preferences. For example, while initial subjects are clearly preferred in English, only a marginal preference was detected in German.

Our own experimental investigation of the interaction between islands and resumption in English, Greek, and German (Alexopoulou and Keller, 2002, 2003; Keller and Alexopoulou, 2005) shows a pattern of variation consistent with Featherston's (2005) and Meyer's (2003) findings. The results can be summarized as follows:

- A clear crosslinguistic effect of weak island violations (resembling, in all languages a similar drop in the acceptability in that-clauses).
- A strong contrast between weak and strong island violations.
- Resumption is unacceptable in questions in all three languages.
- The acceptability of resumption does improve in embedded *whether*-questions and *that*-clauses in all three languages.

Again, crosslinguistic variation is confined to quantitative variation in the magnitude of the effects. For example, resumption in questions is more acceptable in Greek than in German and English, which is evident from the fact that questions with resumptives are more acceptable than strong island violations in Greek, but as bad as strong islands in English and German. Furthermore, weak islands incur a stronger violation in German (as does ordinary embedding under *dass*).

## 3 Universals and Parameters

The most important aspect of these studies is that they indicate that effects relating to superiority and relativized minimality are present crosslinguistically. This confirms the existence of universals where their status has either been disputed (e.g., superiority in German, Polish, Czech, and Russian, *whether*-islands in Greek and German) or where their existence was not properly acknowledged (e.g., the fact that resumption improves weak islands in English even though such resumptives are less acceptable than gaps). Furthermore, it is probably no accident that such crosslinguistic similarities relate to locality principles.[1]

However, it is not straightforward to account for the type of crosslinguistic variation indicated by these studies, namely quantitative variation in the magnitude of universal principles. An initial response would be to discard such variation as surface noise, of no theoretical interest. However, at the same time it is important to acknowledge that such differences can be responsible for surface contrasts between languages. For example, unlike English, in German and Greek pronominals are acceptable as gaps when embedded in a *whether*-island. Intuitively, this appears to be related to the fact that in Greek pronominals are better in ordinary questions in the first instance, while in German, gaps are much worse in the same structure. Similarly, the stronger preference for initial subjects in English vs. German questions ought to relate to the status of subjects in the two languages. Of course, such facts can be attributed to structural/parametric differences between languages. Indeed, in Alexopoulou and Keller (2003), we attribute the relative tolerance of pronominals in Greek questions to the clitic status of the Greek pronominal and its PF realization as an affix. Similarly, we assume that operations associated with V2 in German incur extra processing costs in both *dass*- and *ob*-clauses that are responsible for the extra drop in acceptability (compared to the reduction in the acceptability of *that*-clauses in English and Greek). If this line of reasoning proves correct, then quantitative differences can indeed be reduced to parametric variation.

[1] But note that Featherston (2005) takes the fact that it is really only subjects that exhibit superiority to corroborate ECP analyses and as counterevidence to economy/locality driven accounts.

## 4 Hard vs. Soft Constraints and Parameters

If we assume that quantitative differences can be reduced to parametric variation, then this means that new questions arise with regard to the status of parameters. When is a given crosslinguistic difference, e.g. the optionality of resumption in Greek and German indirect questions, to be accounted for as the consequence of quantitative differences of interacting principles (related to parameters) and when should a straightforward parametric account be attempted? The literature draws a distinction between *hard* and *soft* constraints. In particular, magnitude estimation studies applied to a variety of phenomena (e.g., agreement, auxiliary selection, binding, information structure, word order) from various languages indicate the existence of these two types of constraints (Sorace and Keller, 2005; Keller, 2000). Hard constraints induce categorical judgments when violated and their acceptability cannot be improved by context (e.g., agreement, case violations); soft constraints induce mild ungrammaticality and they appear to interact with context.

Building on this distinction we hypothesize that the locality principles underlying superiority and weak island effects are governed by soft constraints. They only induce mild ungrammaticality, while factors such as d-linking may improve both superiority and relativized minimality violations (Featherston (2005) has demonstrated the effect of d-linking experimentally for English and German). The consequence of this hypothesis is the admission of two types of “universals”. Those that are not directly related to parametric variation and tend to be con-

stant across languages, inducing only gradient unacceptability, and those that are related to parametric variation and give rise to categorical judgments. Further evidence for this hypothesis is provided by magnitude estimation studies focusing on information structure, binding, and auxiliary selection (see Sorace and Keller 2005 for an overview).

## References


Alexopoulou, Theodora and Frank Keller. 2002. Resumption and locality: A crosslinguistic experimental study. In *Papers from the 38th Meeting of the Chicago Linguistic Society*. Chicago, volume 1: The Main Session, pages 1–14.

Alexopoulou, Theodora and Frank Keller. 2003. Linguistic complexity, locality and resumption. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics*. Cascadilla Press, Somerville, MA, pages 15–28.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1):32–68.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage Publications, Thousand Oaks, CA.

Featherston, Sam. 2005. Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua* 115(22):1525–1550.

Keller, Frank. 2000. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, University of Edinburgh.

Keller, Frank and Theodora Alexopoulou. 2005. A crosslinguistic, experimental study of resumptive pronouns and that-trace effects. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Stresa.

Meyer, Roland. 2003. Superiority effects in Russian, Polish and Czech: Comparative evidence from studies on linguistic acceptability. In *Proceedings of the 12th Conference on Formal Approaches to Slavic Linguistics*. Ottawa.

Sorace, Antonella and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115(11):1497–1524.

# The syntax and semantics of multiple degree modification in English

**Christopher Kennedy**
Department of Linguistics
University of Chicago
1010 E 59th Street
Chicago, IL 60637 USA
ckennedy@uchicago.edu

**Louise McNally**
Dept. Traducció i Filologia
Universitat Pompeu Fabra
La Rambla, 30-32
08002 Barcelona Spain
louise.mcnally@upf.edu

## 1 Multiple degree modification in English

In this paper we offer an integrated syntactic and semantic analysis of various cases of multiple degree modification in English, some examples of which appear in (1).

(1) a. a new tower 10 feet taller than the Empire State Building
b. an old department store a lot less taller than the city hall building than is the new company headquarters
c. a structural engineer very much more afraid of heights than the architect

To our knowledge, no such integrated proposal exists for this kind of modification in the HPSG literature. Pollard and Sag (1994) broadly sketch a syntactic analysis of multiple degree modification. However, because it lacks a semantics, their analysis does not make very specific predictions about the restrictions on various combinations of multiple degree modifiers. Although we show that some of these restrictions are matters of pragmatic or lexical semantic detail, others turn out to involve fundamental aspects of the syntax and semantics of degree modification. In contrast, Abeillé and Godard (2003) present a detailed syntax and semantics for French degree adverbs, but their analysis is situated more in the context of a general analysis of adverbial modification, rather than within the context of a complete treatment of degree modification. As a result, their analysis does not address multiple degree modification or differences in the distributions of different subclasses of degree expressions; on the other hand, nothing in our analysis will conflict in important ways with their proposal.

As the syntax of multiple degree modification is tightly bound up with the semantics of the expressions involved, we begin by presenting our semantic assumptions. We follow Kennedy (1999) in analyzing gradable adjectives and related experssions (such as the vague determiners *many* and *few*) as measure functions, which map individuals to degrees on a scale (type $\langle e, d\rangle$). Measure functions are converted to properties of individuals by degree morphology; in Kennedy's analysis, the category of degree expressions includes measure phrases (e.g. *10 feet*), comparative morphemes (e.g. *-er/more, less, as*), intensifiers (e.g. *very*), and the phonologically null positive degree morpheme *pos* (for the 'positive', unmarked form of a gradable adjective, e.g., *(is) tall*). Such expressions take a measure-function and return a property of individuals that is expressed as a relation between two degrees: one determined by applying the measure function to the argument of the predicate; the other introduced by the degree morpheme (the 'standard value').

For example, the comparative morpheme *more* has the denotation in (2) in Kennedy's analyis.

(2) $[\![more]\!] = \lambda g \in D_{\langle e,d\rangle} \lambda d \lambda x.g(x) \succ d$

The degree argument is expressed by the comparative clause (the constituent introduced by *than*), which denotes a maximal degree (von Stechow, 1984). A simple comparative predicate like (3a) is assigned the denotation in (3b): it is true of an object if it has a degree of height that exceeds the maximal

degree to which the Empire State Building is tall.[1]

(3) a. [[more tall] [than the Empire State Building ~~is tall~~]]
b. $\lambda x.\textbf{tall}(x) \succ max\{d' \mid \textbf{tall}(\textbf{the ESB}) \succeq d'\}$

A problem with this approach is that multiple degree modification facts such as those illustrated in (1) and other data strongly suggest that neither comparative morphemes nor intensifers really belong in the category of degree morphology as defined above. For example, (1b) shows that a comparative can modify another comparative, which is unexpected on Kennedy's analysis, since he treats degree morphemes as type-changing: he would be forced to hypothesize that e.g. *less* can combine not only with measure function-denoting expressions (when it takes a simple adjective) but also with property-denoting ones (when it combines with a comparative+adjective complex). This is not a typical case of type polymorphism.

Similar comments apply to intensifiers. Although it is sometimes claimed to the contrary, a number of combinations of multiple intensifiers are possible (as even a simple Google search will demonstrate):

(4) a. very much alone
b. rather very good
c. rather quite interesting

Again, Kennedy's treatment of intensifiers as type changing forces one to adopt a rather ad hoc type polymorphism to account for the fact that these expressions modify both adjectives and other intensifiers.

In contrast to the comparative morphemes and intensifiers stand a group of degree expressions that 'close off' the predicate they combine with; these include (at least) measure phrases, degree *this/that*, proportional modifiers like *completely* and *half*, and the *wh*-degree morpheme *how*. These expressions can combine with an unmodified adjective or with a comparative (provided a system of measurement is defined for the adjective in the case of measure phrases), as shown in (5) for the measure phrase *2 meters* and degree *that*.

(5) a. 2 meters/that tall
b. 2 meters/that {taller, less tall, too tall}

However, they do not accept further modification (6a), nor can they further modify an intensifier (6b) (we assume the *much* in (5b) is a dummy element; see (Corver, 1997)):

(6) a. *rather 2 meters/that long
b. *2 meters/that very long

## 2 Three classes of degree expressions and one lexical rule

In this paper, we develop an analyis in which degree expressions are divided into three subclasses: (true) DEGREE MORPHEMES, which map gradable adjectives into properties of individuals; INTENSIFIERS, which affect the computation of the standard of comparison for the positive form; and SCALE ADJUSTERS, which modify the measure function expressed by the adjective. In addition, we assume a lexical rule to handle the interpretation of the unmarked positive form.

### 2.1 The positive form

As noted above, Kennedy (1999) assumes that the positive form involvues a null degree morpheme *pos*, which maps a gradable adjective to a property of individuals that expresses a relation to a context-dependent standard of comparison (see also (Bartsch and Vennemann, 1972), (Cresswell, 1977), (Klein, 1980), (von Stechow, 1984), (Kennedy, 1999), (Kennedy and McNally, 2005)). The positive form of an adjective like *tall* is thus analyzed as the predicate [$_{\text{AP}}$ *pos* tall], which denotes the property of having a degree of length that exceeds a standard of length whose value is determined based on features of the context of utterance (what is being talked about, the interests/expectations of the participants in the discourse, etc.; see (Lewis, 1970), (Bogusławski, 1975), (Graff, 2000), (Barker, 2002), (Kennedy and McNally, 2005)). Here we take the (possibly universal) absence of overt morphology in the positive form at face value and instead assume a lexical rule that maps measure functions to

[1] We assume for simplicity here that the comparative clause is an ellipsis structure; this issue is orthogonal to the main concerns of this paper. See (Kennedy, 2002) for a compositional analysis. Likewise, we abstract away from the morphological alternation between *more* and *-er*.

properties of individuals in the absence of overt degree morphology. This rule (whose particular implementation is not crucial for our purposes) is stated in (7), where **stnd** is a context-dependent function from a measure function (a 'basic' gradable adjective meaning) to a degree in the range of the measure function (its scale) that represents an appropriate standard of comparison for the gradable property measured by the adjective in the context of utterance. (Compare Lewis' (1970) and Barker's (2002) DELINEATION FUNCTION.)

(7)
$$\begin{bmatrix} \text{cat} & \text{A} \\ \text{cont} \mid \text{rest} & \begin{bmatrix} \text{reln} & \boxed{1} g_{\langle e,d\rangle} \\ \text{arg1} & \boxed{2} x \end{bmatrix} \end{bmatrix} \Rightarrow$$
$$\begin{bmatrix} \text{cat} & \text{AP} \\ \text{cont} \mid \text{rest} & \begin{bmatrix} \text{reln} & \succ \\ \text{arg1} & \boxed{1} \\ \text{arg2} & \boxed{2} \\ \text{arg3} & \textbf{stnd}(\boxed{1}) \end{bmatrix} \end{bmatrix}$$

With this as our starting point, we now turn to the analysis of degree morphology.

### 2.2 True degree morphemes

This category contains expressions of type $\langle\langle e, d\rangle, \langle e, t\rangle\rangle$; in English: *how*, *that*, and measure phrases. These behave as in (Kennedy, 1999), mapping a measure function onto a property of individuals expressed as a relation between degrees: the degree derived by applying the measure function to the individual argument of the predicate, and a standard degree specified by the degree morpheme itself. For example, in the case of measure phrases, this is the correponding degree of measurement, as illustrated by our analysis of the measure phrase *2 meters* (8).

(8)
$$\begin{bmatrix} \textit{2 meters} & \\ \text{cat} & \text{Deg} \\ \text{cont} \mid \text{rest} & \begin{bmatrix} \text{reln} & \succeq \\ \text{arg1} & g \\ \text{arg2} & x \\ \text{arg3} & \textbf{2 meters} \end{bmatrix} \end{bmatrix}$$

### 2.3 Intensifiers

We analyze intensifiers as traditional predicate modifiers (type $\langle\langle e, t\rangle, \langle e, t\rangle\rangle$), which are restricted to apply only to gradable predicates in the positive form. We derive this restriction from their semantics, treating them as expressions that modify the **stnd** function introduced by the positive form rule in (7) (cf. (Wheeler, 1972), (Klein, 1980)). This proposal is based on two observations. First, the semantic effect of intensification is to 'adjust' the contextually determined standard of comparison. Second, the distribution of degree modifiers is highly sensitive to the type of standard of comparison associated with particular *pos*+adjective combinations (whether the standard is context dependent or lexically determined by the adjectival head; see Kennedy and McNally's (2005) analysis of *very* vs. *much*).

Consider for example the case of *very*. Both *tall* and *very tall* require an object to exceed a contextual standard of height, but the standard of comparison introduced by the latter is greater than that used by the former. Following Wheeler (1972) and (1980), we derive this result by assuming that *very* modifies the **stnd** function associated with its argument (an adjective to which the lexical rule in (7) has applied) so that it computes a standard of comparison based on just the heights of those objects that its argument is true of. That is, [$_{\text{AP}}$ *very tall*] is (syntactically and semantically) just like [$_{\text{AP}}$ *tall*], except that the standard of comparison for the former is computed by considering only those objects that count as tall in the context of utterance. General principles of informativity ensure that the modified **stnd** function will select a new standard of comparison partitions the domain of [$_{\text{AP}}$ *very tall*] into things it is true of and things it is false of, effectively boosting the base standard associated with [$_{\text{AP}}$ *tall*] (i.e., some tall objects will not count as very tall).

This proposal is made explicit in (10) (after the References section). For the puposes of illustration, we adopt Kasper's (1997) treatment of nonintersective modification, where the MOD feature is split up into information about the ARGument of the modifier (including its internal content) vs. the (External) CONTent of the resulting phrase.

Our analysis explains why measure phrases (or rather, measure phrase + adjective combinations)

cannot be intensifed, even though their semantic (and syntactic) type should in principle allow for it. The difference between [$_{\mathrm{AP}}$ MP A] (a type $\langle e,t\rangle$ predicate consisting of a measure phrase plus gradable adjective) and [$_{\mathrm{AP}}$ A] (a positive form gradable adjective to which the rule in (7) has applied) is that the latter is evaluated with respect to the **stnd** function but the former is not. As a result, there is no value for an intensifier to manipulate, so the addition of an intensifier has no semantic effect.

### 2.4 Scale adjusters

This category includes comparatives and *too/enough*, after they have been saturated by their internal (clausal) arguments; their semantic type is that of gradable adjective modifiers ($\langle\langle e,d\rangle,\langle e,d\rangle\rangle$). Specifically, we claim that these expressions modify the measure function they take as input by resetting the maximal or minimal value (depending on the morpheme) to the degree introduced by the comparative clause. For example, *more than CP* (where CP is the comparative clause) takes a measure function and assigns it a new scale whose minimal value is the degree denoted by CP. Thus if *tall* is a function that maps an individual onto whatever part of the height scale corresponds to its height, *taller than the Empire State Building* maps an individual onto whatever region of the height scale represents its 'taller-than-the-ESB-ness': an object whose height is less than or equal to the maximal degree of the Empire State Building's height is mapped onto the zero element of the derived scale, and all others are mapped onto their actual height value. This is made explicit in (11) (after References).

The result of this analysis is that expressions consisting of an adjective plus comparative morphology must ultimately either undergo the positive form rule in (7) or combine with a true degree morpheme (e.g. a measure phrase) in order to derive a property of individuals. Assuming that the positive form of an adjective that uses a scale with a minimal element is true of an object as long as it has a non-minimal degree of the relevant property (Kennedy and McNally, 2005), the result is that *taller than CP* is true of an object if its height exceeds the degree denoted by the CP (the minimal element of the derived scale). In other words, *taller than the Empire State Building* is true of an object just in case its height exceeds that of the Empire State Building, which is exactly what we want.

## 3 Predictions of the analysis

In our presentation, we go through the analysis of complex modification structures like those in (1) in detail; here we outline the predictions about possible combinations of degree expressions made by our proposals:

1. Iteration of comparative expressions and intensifiers should be possible.
2. Iteration of true degree morphemes should not be possible.
3. Measure phrases should be external to all comparative morphology.
4. Under the assumption that intensifiers and scale adjusters are not reanalyzable as intersective (unlike, e.g., what is the case with many adjectives or adverbs), iterations both of comparatives and of intensifiers must be interpreted in a nested right-branching fashion, rather than in a left branching fashion, as predicted on Pollard and Sag's Specifier analysis.

The data presented above illustrate 1-3; 4 is difficult to test because of the rarity of sequences of more than 2 intensifiers, but appears to be borne out by the fact that the interpretation of the string in (9a) corresponds on our intuitions to the bracketing in (9b) rather than that in (9c).

(9) a. Becca was rather very slightly drunk last night (www.elvislovers.fanspace.com/fsguestbook.html)
b. (rather (very (slightly)))
c. *((rather (very))(slightly))

## 4 Concluding remarks

Our HPSG implementation of degree modifiers combines intensifiers and scale adjusters with their semantic arguments in Head-Adjunct structures, while true degree morphemes combine with their arguments in a Head-Specifier structure. Our analysis

thus resembles Abeillé and Godard's insofar as they argue for a Head-Adjunct analysis of French degree adverbs. It refines their proposal in allowing (at least in English) for two types of degree Adjuncts: those that operate on 'bare adjectives' (measure functions), and those that operate on gradable APs (i.e., on the **stnd** function introduced by the positive form). Kennedy and McNally's (2005) comments concerning the semantics of the degree modifier *well* indicate that these two types are clearly justified.

Nonetheless, the analysis also preserves the essence of the insight behind Pollard and Sag's proposal, on which degree expressions are treated as specifiers of adjectives, adverbs or other gradable predicates in a Head-Specifier configuration. It simply reduces the class of expressions that have this specifying function, as a result of having refined the semantics of degree modification.

A question of broader theoretical interest is why the set of degree expressions should be divided up in the way we have proposed here. We claim that this is a natural result of our initial assumptions that gradable adjectives have basic meanings as measure functions, and 'derived' meanings (in the positive form) as context-dependent properties of individuals (where context dependence comes from the **stnd** function). If the basic semantic type of a gradable adjective is $\langle e, d\rangle$ (a measure function), then there should exist overt morphology (in addition to our positive form lexical rule) that converts a gradable adjective to a property of individuals: this is our class of true degree morphemes. Furthermore, if natural language quite generally allows expressions of type $\langle \tau, \tau\rangle$, there should also exist a class of modifiers of measure functions: these are our scale adjusters. By the same token, we also expect to find modifiers of the type $\langle e, t\rangle$ variant of a gradable adjective (the positive form): this is our class of intensifiers.

## References


Anne Abeillé and Danièle Godard. 2003. The syntactic flexibility of French degree adverbs. In Stefan Müller, editor, *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar*, pages 26–46, Stanford, CA. CSLI Publications.

Chris Barker. 2002. The dynamics of vagueness. *Linguistics and Philosophy*, 25(1):1–36.

Renate Bartsch and Theo Vennemann. 1972. The grammar of relative adjectives and comparison. *Linguistische Berichte*, 20:19–32.

Andrzej Bogusławski. 1975. Measures are measures: In defence of the diversity of comparatives and positives. *Linguistiche Berichte*, 36:1–9.

Norbert Corver. 1997. Much-support as a last resort. *Linguistic Inquiry*, 28:119–164.

M. J. Cresswell. 1977. The semantics of degree. In Barbara Partee, editor, *Montague Grammar*, pages 261–292. Academic Press, New York.

Delia Graff. 2000. Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics*, 20:45–81.

Robert T. Kasper. 1997. The semantics of recursive modification. Ms., Ohio State University.

Christopher Kennedy and Louise McNally. 2005. Scale structure and the semantic typology of gradable predicates. *Language*, 81(2):1–37.

Christopher Kennedy. 1999. *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*. Garland, New York. (1997 UCSC Ph.D thesis).

Christopher Kennedy. 2002. Comparative deletion and optimality in syntax. *Natural Language & Linguistic Theory*, 20.3:553–621.

Ewan Klein. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, 4:1–45.

David K. Lewis. 1970. General semantics. *Synthese*, 22:18–67.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Arnim von Stechow. 1984. Comparing semantic theories of comparison. *Journal of Semantics*, 3:1–77.

Samuel Wheeler. 1972. Attributives and their modifiers. *Noûs*, 6(4):310–334.

(10)

$$\left[\begin{array}{ll} \textit{very} & \\ \text{cat}\,|\,\text{head} & \left[\begin{array}{ll} \text{mod} & \left[\begin{array}{ll} \text{arg} & \left[\begin{array}{ll} \text{cat} & \text{AP} \\ \text{cont} & \boxed{1}\left[\begin{array}{ll} \text{reln} & \succ \\ \text{arg1} & g \\ \text{arg2} & x \\ \boxed{2}\text{arg3} & \textbf{stnd}(g) \end{array}\right] \end{array}\right] \\ \text{econt} & \boxed{3}\left(\boxed{1}\right) \end{array}\right] \end{array}\right] \\ \text{cont} & \boxed{3}\left[\begin{array}{ll} \text{reln} & \textbf{recompute-stnd}_{very} \\ \text{arg1} & \boxed{1} \end{array}\right] \end{array}\right]$$

(11)

$$\left[\begin{array}{ll} \textit{more than} & \\ \text{cat}\,|\,\text{head} & \left[\begin{array}{ll} \text{mod} & \left[\begin{array}{ll} \text{arg} & \left[\text{cont} \left[\text{index} \quad \boxed{1}g\right]\right] \\ \text{econt} & \boxed{2}\left(\boxed{1},d\right) \end{array}\right] \\ \text{val} & \left[\text{comps} \quad \left\langle \text{CP}_d \right\rangle\right] \end{array}\right] \\ \boxed{2}\text{cont} & \left[\begin{array}{ll} \text{reln} & \textbf{more-than} \\ \text{arg1} & g \\ \text{arg2} & d \end{array}\right] \end{array}\right]$$

# It-Extraposition in English: A Constraint-Based Approach

**Jong-Bok Kim**
School of English
Kyung Hee University
Seoul, 130-701, Korea
jongbok@khu.ac.kr

**Ivan A. Sag**
Department of Linguistics
Stanford University
Stanford, 94305
sag@stanford.edu

## 1 Basic Issues

According to the Projection Principle (Chomsky 1981), expletives have no semantic content and thus cannot occur in theta-marked positions. However, there seem to exist overt cases where the expletive *it* appears in the object position as in (1) (Postal & Pullum 1988):

(1) a. I dislike it that he is so cruel.
b. They never mentioned it to the candidate that the job was poorly paid.
c. We may depend upon it that we won't abandon him.
d. I should resent it greatly that you did not call.

Many attempts (e.g. the case-based analysis by Authier (1991), the predication analysis by Rothstein (1995), the Spec analysis by Stroik (1996)) have been made to account for such examples, with the common postulation of the expletive in the Spec of CP position and various movement processes. For example, to generate cases like (1d), Stroik (1996) claims that the object expletive *it* is generated in the Spec of CP at base argument structure and then moved into the Spec of an AGR projection to satisfy case checking, together with several movement operations as shown in the following:

(2) [...[$_{PredP}$ resent$_i$ [$_{AGRoP}$ it$_j$ [$_{AGRo'}$ t$_i$ [$_{VP}$ greatly [$_{VP}$ [$_{V'}$ t$_i$ [$_{CP}$ t$_j$ [$_{C'}$ that you did not call]]]]]]]]]

Such analyses may be able to generate cases like (1d) or similar examples, but they still fail to account for the fact that the expletive is obligatory in (3a), optional in (3b), and prohibited in (3c):

(3) a. Type I: I blame *(it) on you [that we can't go].
b. Type II: Nobody expected (it) of you [that you could be so cruel].
c. Type III: John said (*it) to his friends [that we had betrayed him].

## 2 A Proposal

This paper argues that such contrasts, in addition to the distribution possibilities of *it* in the object position, follow naturally from a lexical analysis based on tight interactions between lexical and English-independent constraints, rather than otherwise unmotivated movement operations.

### 2.1 Lexically Controlled Extraposition

The first property of English object extraposition we need to consider is that the overt expletive in direct object position is possible only with certain verbs taking clausal complements, and not with others (Authier 1991). It appears that the verbs taking clausal complements can also take an NP object or allows the expletive object:

(4) a. They didn't even mention his latest promotion/that he was promoted recently.
b. They demanded justice/that he should leave.

(5) a. They never mentioned it to the candidate that the job was poorly paid.
b. They demand it of our employees that they wear a tie.

However, verbs taking either a CP or an NP only disallow object extraposition:

(6) a. I think *(of) you all the time.
b. I wonder *(about) that.

(7) a. I think (*it) that John had an accident.
b. I wondered (*it) how he did on the test.

Another general observation we can make is that many verbs select a complement that can be realized either as an NP or a CP:

(8) a. Tom proved the independence of the hypothesis/the hypothesis was independent.
b. Tom forgot our invitations/that we needed invitations.

As suggested in Sag et al. (2003), one simple way of specifying such a lexical property is to introduce a new part-of-speech type *nominal* that subsumes both *noun* and *comp*. In accordance with the basic properties of type hierarchy, this system then means that if an element is specified with [HEAD *nominal*], it can be realized either as [HEAD *noun*] or as [HEAD *comp*]. English will thus have at least the three lexical types: *v-np-tr* for verbs selecting just NP (*devour, like, pinch, elude,...*), *v-s-tr* for verbs selecting only CPs (*hope, hint, wonder,...)*, and *v-nominal-tr* for verbs selecting both NPs and CPs (*prove, forget, see,...)*. These three types will basically have the following ARG-ST value:

(9) a. $\begin{bmatrix} \textit{v-np-tr} \\ \text{ARG-ST } \langle \text{X, NP,...} \rangle \end{bmatrix}$ b. $\begin{bmatrix} \textit{v-s-tr} \\ \text{ARG-ST } \langle \text{X, CP,...} \rangle \end{bmatrix}$
c. $\begin{bmatrix} \textit{v-nominal-tr} \\ \text{ARG-ST } \langle \text{X, [HEAD } \textit{nominal}\text{],..} \rangle \end{bmatrix}$

Because of the status of the type *nominal*, the present system allows a lexical element with the information (9c) to be realized either as in (9a) or as in (9b).

### 2.2 Some Theoretical Apparatus

As is well-known, there is a systematic alternation between non-extraposed and extraposed sentences as shown in the following pair:

(10) a. That Chris knew the answer occurred to Pat.
b. It occurred to Pat that Chris knew the answer.

Pollard and Sag (1997) and Sag et al. (2003) capture this relationship with a lexical rule that turns the sentential subject (in (10a)) into a sentential 'complement' of the verb (in (10b)). However, this complement approach, as pointed out by Keller (1995), Bouma (1996), and van Frank (1996), suffers from problems for cases like the following:

(11) a. They regret it [very much] [that we could not hire Mosconi].
b. It struck a grammarian last month, [who analyzed it], [that this clause is grammatical].

If the extraposed *that*-clause is the complement of *regret*, we would not expect the intervention of the adverbial elements *very much* or the relative clause in (11). Departing from the traditional complement approach (and following Bouma 1996, among others), we take English extraposition to be a nonlocal dependency and introduce the nonlocal feature EXTRA together with the following lexical rule:[1]

(12) Extraposition Lexical Rule:
$$\begin{bmatrix} \text{ARG-ST } \boxed{A} \oplus \langle \boxed{1}[\textit{nominal}] \rangle \oplus \boxed{B} \\ \text{SEM } \textit{fact} \end{bmatrix} \Rightarrow \begin{bmatrix} \text{ARG-ST } \boxed{A} \oplus \langle \text{NP[NFORM } \textit{it}] \rangle \oplus \boxed{B} \\ \text{EXTRA } \langle \boxed{1}[\text{HEAD } \textit{comp}] \rangle \end{bmatrix}$$

This rule basically turns a *v-nominal-tr* into a word that selects an expletive NP with the CP as its EXTRA value.[2] The input element also has a semantic restriction requiring that the message type of the semantic content be *fact*. As noted by Bolinger (1976), nonfactives or suppositions do not allow object extraposition.

(13) a. *I resent it that she did that, if indeed she did.
b. I resent it that she did that.

The feature EXTRA is discharged when a head combines with the extraposed phrase, in accordance with the Head-Extra Rule:

(14) Head-Extra Phrase:
$$\left[\text{EXTRA } \langle \quad \rangle\right] \rightarrow \text{H}\left[\text{EXTRA } \langle \boxed{1} \rangle\right], \boxed{1}$$

[1] This rule can be applied to subject extraposition, too. Thus, in our analysis, unlike that of Sag et al. (2003), the *that*-clause in (10b) is not the complement of *occurred*, but rather an extraposed element.

[2] One basic constraint that works in the grammar is the Argument Realization Constraint which ensures the ARG-ST elements will be realized as SUBJ and COMPS in syntax.

One additional language-independent constraint relevant in extraposition is that the langauge independently prohibits a CP from having any element to its right (cf. Kuno's (1987) Ban on Non-sentence Final Clause (BNFC)):

(15) a. I believe strongly that the world is round.
b. *I believe [that the world is round] strongly.

This BNFC constraint (an LP rule) basically bars any argument from appearing after a sentential argument. In the present context this means that there exists no word whose COMPS list contains something to its CP complement.

## 3 Explaining the Three Types

Given these independently motivated assumptions, facts concerning English object extraposition then easily follow.

**Type I:** As noted earlier, verbs like *blame* require the obligatory presence of the expletive *it* in the object position:

(16) a. I blame the case on you.
b. *I blame that we can't go.
c. *I blame [that we can't go] on you.
d. I blame it on you that we can't go.
e. *I blame on you that we can't go.

The data imply that *blame* will have the following lexical entry:[3]

(17) $\begin{bmatrix} \textit{v-nominal-tr} \\ \text{ARG-ST } \langle \boxed{1}\text{NP}, \boxed{2}[\text{HEAD } \textit{nominal}], \boxed{3}\text{PP}[\text{FORM } \textit{on}] \rangle \end{bmatrix}$

The verb *blame* selects a *nominal* and a PP as its arguments. If the *nominal* complement is resolved as an NP (as in (18)a), we generate (16a). If it instead is realized as a CP (as in (18b)) then it cannot be realized before the PP, because of the BNFC, thus accounting for the deviance of (16c).

(18) a. $\begin{bmatrix} \textit{v-nominal-tr} \\ \text{ARG-ST } \langle \boxed{1}\text{NP}, \boxed{2}\text{NP}, \boxed{3}\text{PP}[\text{FORM } \textit{on}] \rangle \end{bmatrix}$

b. *$\begin{bmatrix} \textit{v-nominal-tr} \\ \text{ARG-ST } \langle \boxed{1}\text{NP}, \boxed{2}\text{CP}, \boxed{3}\text{PP}[\textit{on}] \rangle \end{bmatrix}$

[3]The boxed integer here is introduced to show the relationships with related lexical entries as in (17) or (18).

The deviance of (16e) is explained by appeal to a further LP Constraint requiring controllers to precede the phrases (whose subject) they control:

(19) $\boxed{1} \prec [\text{SUBJ } \langle \boxed{1} \rangle]$

Note that we know the *on*-phrase here is predicative because of examples like the following:

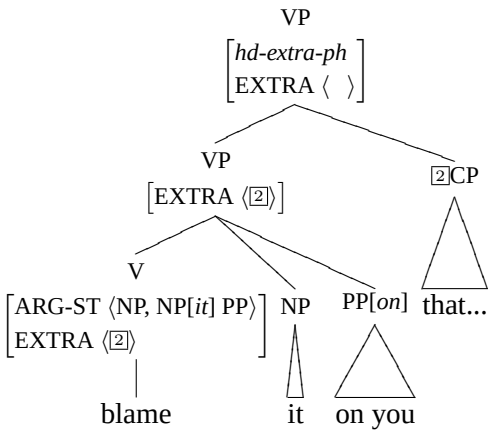
(20) a. They put the blame on Lee.
b. The blame was on us.

One way that (18b) can give rise to a legitimate *word* is to apply the Extraposition Lexical Rule, as shown in the following:

(21) $\begin{bmatrix} \textit{extraposed-w} \\ \text{ARG-ST } \langle \boxed{1}, \text{NP}[\text{NFORM } \textit{it}], \boxed{3}\text{PP} \rangle \\ \text{EXTRA } \langle \boxed{2}[\text{HEAD } \textit{comp}] \rangle \end{bmatrix}$

The output in (21) can then give rise to sentences like (16d) with the following (simplified) structure:

(22)




**Type II:** In the Type II examples, the expletive *it* is optional, as noted earlier:

(23) a. Nobody expected that of you.
b. Nobody expected that you could be so cruel.
c. *Nobody expected [that you could be so cruel] of you.
d. Nobody expected it of you [that you could be so cruel].
e. Nobody expected of you [that you could be so cruel].

The data lead us to assume the following lexical entry for verbs like *expect*:

(24) $\begin{bmatrix} \textit{v-nominal-tr} \\ \text{ARG-ST} \langle \boxed{1}\text{NP}, \boxed{2}\begin{bmatrix}\text{HEAD } \textit{nominal}\end{bmatrix}, \boxed{3}(\text{PP[FORM } \textit{of}])\rangle \end{bmatrix}$

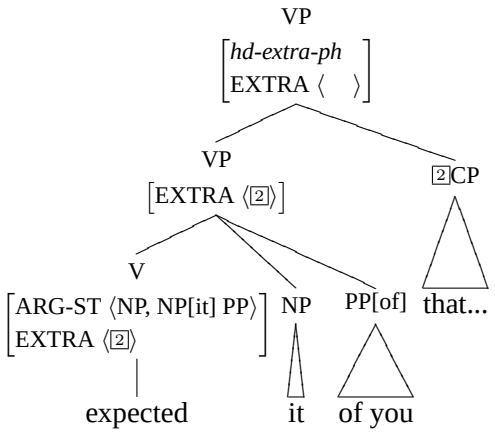
According to the lexical entry in (24), the verb *expect* takes a *nominal* and an optional case-marking (i.e. nonpredicative) PP. Depending on the instantiation of the HEAD value, we will have the following two realizations.

(25) a. $\begin{bmatrix} \textit{v-nominal-tr} \\ \text{ARG-ST} \langle \boxed{1}\text{NP}, \boxed{2}\text{NP}, (\boxed{3}\text{PP[FORM } \textit{of}])\rangle \end{bmatrix}$

b. $\begin{bmatrix} \textit{v-nominal-tr} \\ \text{ARG-ST} \langle \boxed{1}\text{NP}, \boxed{2}\text{CP}, \boxed{3}(\text{PP[FORM } \textit{of}])\rangle \end{bmatrix}$

(25a) will be able to generate sentences like (23a). When the PP complement does not appear, we will have sentences like (23b). However, with the PP argument, the instantiation (25b) cannot give rise to the ungrammatical (23c), because of the BNFC constraint. But because the PP is non-predicative, examples like (23e) are predicted to be well-formed, as no violation of the LP Constraint in (19) is engendered. Extraposition is also possible, generating a word that selects an expletive object and turns the argument CP into the value of the nonlocal feature EXTRA. The output lexical entry will allow sentences like (23d) with the following structure:

(26)




**Type III:** Unlike the other two types, this type of verb appears not to allow object extraposition. That is, it seems that no expletive is possible in such a case even though this type of verb could have a CP complement:

(27) a. John thought (to himself) that Mary was coming.

b. ?*John thought it to himself that Mary was coming.

This fact could be simply be accounted for via an appropriate lexical entry, but, as we will show, there are in fact ample attested examples like (27b) and hence some reason to believe that the entire literature on English object extraposition has made a mistaken assumption about the data in question. We ultimately explore the idea that all CP-compatible complements can in principle give rise to extraposition.

## 4 Conclusion

The analysis presented here implies that lexical specifications — both general constraints on verb classes and individual lexical constraints — play a crucial role in English object extraposition. These specifications interact with one lexical rule and two LP rules to predict that observed patterns of variation.

## Selected References


Authier, J.-Marc . 1991. V-Governed Expletives, Case Theory, and the Projection Principle *Linguistic Inquiry* 22, no.4, 721-740.

Bouma, Gosse. 1996. Complement Clauses and Expletives. *Papers from the sixth CLIN Meeting*.

Iwakura, Kunihiro. 1994. The Distribution of CP-Trace and the Expletive *It*. *Linguistic Analysis* 24, no. 1-2, 122-141.

Keller, Frank. 1995. Towards an Account of Extraposition in HPSG. *Proceedings of the EACL*.

Kuno, Susumo. 1987. *Functional Syntax*. Chicago: The University of Chicago Press.

Postal, Paul, and Geoffrey Pullum. 1988. Expletive Noun Phrases in Subcategorized Positions *Linguistic Inquiry* 19, no. 4, 635-670.

Rothstein,Susan D. 1995. Pleonastics and the Interpretation of Pronouns *Linguistic Inquiry* 26, no.3, 499-529.

Sag, Ivan, Tomas Wasow, and Emily Bender. 2003. *Syntactic Theory: A Formal Approach*. CSLI Publications.

Stroik, Thomas S. 1996. Extraposition and Expletive-Movement: A Minimalist Account. *Lingua: International Review of General Linguistics* 99, no. 4, 237-251.

van Eynde, Frank. 1996. A monostratal treatment of it extraposition without lexical rules. *Papers from the sixth CLIN Meeting 1995*.

# Copy Constructions and their Interaction with the Copula in Korean

**Jong-Bok Kim**
School of English
Kyung Hee University
Seoul, 130-701, Korea
jongbok@khu.ac.kr

**Peter Sells**
Department of Linguistics
Stanford University
Stanford, 94305
sells@stanford.edu

## 1. The Copula

The Korean copula *-i-* forms a phonological word with its preceding N host (see Oh (1991)), and has been treated by some linguists as a syntactic V which happens to be a clitic on a preceding NP complement (in particular Yoon (2003)). (1)a-b illustrate some basic properties:

(1) a. ku salam-i haksayng-i-ta
that person-NOM student-COP-DECL
'That person is a student.'

b. ku salam-i [mikwuk-eyse
that person-NOM America-at
kongpwu ha-n] haksayng-i-ta
study do-PAST student-COP-DECL
'That person is a student who studied in America.'

In this paper we add to a growing body of evidence which shows that N+*i*- is a lexically-formed verb. An apparent problem for this approach is that the N hosting the copula can head a fully-formed syntactic NP (see (1)b). The facts with the copula can be accounted for in HPSG through the adoption of the Lexical Sharing approach of Wescoat (2002). Informally, Lexical Sharing allows words to instantiate *one or more* lexical-category nodes; thus, alongside familiar one-word-to-one-phrase instantiation, exemplified by *salam-i* 'person-NOM', the theory also posits *portmanteau words*, which instantiate two or more adjacent lexical-category nodes. This allows us to accept the lexicality of *haysayng-i-* 'student-COP', a form which may receive verbal inflectional affixation in the lexicon (see Kim et al. (2004)).

## 2. Evidence from the Echo Construction

Our new evidence comes from the 'Echo Contrastive Construction (ECC)', which involves the doubling of Vs, but none of their phrasal arguments and adjuncts. The function of the ECC, whose basic form is V-*ki-nun* V-*ta* as in (2), is to set up a negative implicature in the interpretation of the whole sentence (see Choi (2003), Cho et al. (2004), Kim (2002)). (2)b shows a related construction, what we call the '*Ha* Contrastive Construction (HCC)'. In either example, the event is presented against the background of the negative implicature, indicated by the '*but* . . . ' in our translations.

(2) a. John-i Tom-ul [manna-ki-nun
John-NOM Tom-ACC meet-Nmlz-TOP
manna-ss-ta]
meet-PAST-DECL
'John met Tom, but . . . .'

b. John-i Tom-ul [manna-ki-nun
John-NOM Tom-ACC meet-Nmlz-TOP
hay-ss-ta]
do-PAST-DECL
'John met Tom, but . . . .'

The interaction of the ECC with the copula provides strong support for our claim. The only grammatical form of an ECC with the positive copula *i-ta* also involves doubling the N host of the copula, as in (3)a (see Oh (1991), Kim and Chung (2002)).

(3) a. ku salam-i [mikwuk-eyse
that person-NOM America-at
kongpwu ha-n] haksayng-i-ki-nun
study do-PAST student-COP-NMLZ-TOP

haksayng-i-ta
student-COP-DECL
'That person is a student who studied in America (but he still doesn't speak English well).'

b. *ku salam-i [mikwuk-eyse
that person-NOM America-at
kongpwu ha-n] haksayng-i-ki-nun
study do-PAST student-COP-NMLZ-TOP
i-ta
COP-DECL

Under the clitic analysis, the copula never forms a syntactic unit with its NP complement: thus there is no easy way to make the copied part in (3)a a constituent; it would be the head of the complement NP and the following V selecting for that NP. However, we see clearly that the ECC treats N+Copula as a syntactic constituent, and that the copula cannot function as a pure V in the syntax, from the contrast in (3)a and (3)b. The copula verb alone cannot be copied, as it has no syntactic status by itself.

The facts in (3) contrast directly with the ECC facts with the negative copula *ani-ta*, which takes a nominative-marked complement (see (8)a): the *ani-ta* verbal part can be doubled by itself, as in (4)b, just like a regular verb (cf. (2)a). And while the doubling of N + negative copula as in (4)a is grammatical, this example does not have the 'negative implicature' interpretation typical of the ECC, but rather has a VP-topic interpretation – along the lines of 'as for not being a fool, that person is not a fool'.[1] This asymmetry shows that the ECC targets a verb in the syntax and intuitively copies it, meaning that there is a lexical form *haksayng-i-* for (3)a alongside *ani-ta* for (4)b.[2]

(4) a. ku salam-i papo-ka
that person-nom fool-NOM
ani-ki-nun papo-ka ani-ta
NCOP-NMLZ-TOP fool-NOM NCOP-DECL
'It is true that that person is not a fool.'
(VP-topic)

b. ku salam-i papo-ka
that person-NOM fool-NOM
ani-ki-nun ani-ta
NCOP-NMLZ-TOP NCOP-DECL
'That person is not a fool (but he is not so smart).'

Although the details will come later, the structure we assign to (3)a is given in (5) (next page).

## 3. Further Issues with the Copula

Jo (2004) discusses pairs of examples apparently involving the ECC and the copula, based on the simple example in (6)a:

(6) a. chelswu-ka pwuca-i-ess-e
chelswu-NOM rich-COP-PAST-DECL
'Chelswu was rich (a rich man).'

b. chelswu-ka pwuca-nun
chelswu-NOM rich-TOP
pwuca-i-ess-e
rich-COP-PAST-DECL

c. chelswu-ka pwuca-i-ki-nun
chelswu-NOM rich-COP-NMLZ-TOP
pwuca-i-ess-e
rich-COP-PAST-DECL

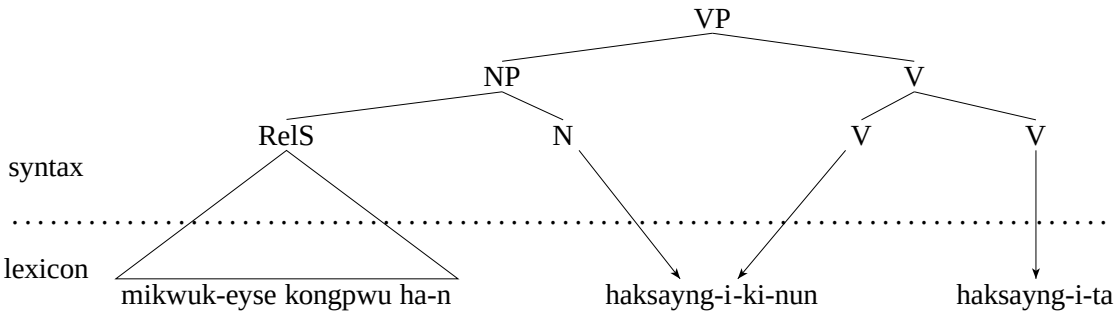
Jo argues that the relation between (6)b and (6)c shows that what is copied is either the N before the copula, for b, or a larger constituent consisting of N and the copula, for c, both coming from the same source in a transformational derivation.

However, there are several types of evidence which show that though (6)c is an instance of the ECC, (6)b is not, and is rather an 'N-Copy Construction' (we will call it 'NCC'), which reinforces the meaning of the N, and we translate it (roughly) as 'truly'. As mentioned above, the pragmatic hallmark of the ECC is that it sets up a negative implicature, without any assistance from other morphemes in the clause which may have adversitive or concessive meanings. This distinguishes (6)b from (6)c, and identifies only c as the ECC. While they both involve copying constructions (which will be related, but not identical, in our analysis), the key difference is that (6)b involves copying Ns, while (6)c involves copying Vs, and only the latter type has the negative implicature. One clear difference can be observed

[1] A caveat: prosodic prominence on the marker *-nun* can also trigger the negative implicature due to its contrastive properties.

[2] The positive copula is one of a class of verbal elements including *-tap-ta* 'is.every.bit' and *-kath-ta* 'seem' (noted by Yoon (2003)), which behave in the same way, including in the ECC.

(5)




from the alternation with the HCC. With noun and copula, the ECC alternates with the HCC (see (2)b), while the NCC does not:

(7) a. chelswu-ka pwuca-nun
chelswu-NOM rich-TOP
pwuca-i-ess-e/*hay-ss-e
rich-COP-PAST-DECL/*do-PAST-DECL
'Chelswu is a truly rich man.'

b. chelswu-ka pwuca-i-ki-nun
chelswu-NOM rich-COP-NMLZ-TOP
pwuca-i-ess-e/hay-ss-e
rich-COP-PAST-DECL/do-PAST-DECL
'Chelswu is a rich man, but . . . .'

The HCC is clearly a V-V complex predicate, so its failure to work with an N first part in a is expected.

Next, the interaction with the negative copula is telling. From the simple example in (8)a, we might expect the following alternatives to be acceptable:

(8) a. chelswu-ka pwuca-ka ani-ta
chelswu-NOM rich-NOM NCOP-DECL
'Chelswu is not a rich man.'

b. chelswu-ka pwuca-ka
chelswu-NOM rich-NOM
ani-ki-nun ani-ta
NCOP-NMLZ-TOP NCOP-DECL
'Chelswu is not a rich man (but he is very generous).' (negative ECC)

c. chelswu-ka pwuca-ka
chelswu-NOM rich-NOM
ani-ki-nun pwuca-ka ani-ta
NCOP-NMLZ-TOP rich-NOM NCOP-DECL
'As for not being rich, Chelswu is not rich.' (negative VP-topic)

d. ??chelswu-ka pwuca-nun pwuca-ka
chelswu-NOM rich-TOP rich-NOM
ani-ta (negative NCC)
NCOP-DECL

However, the last example is essentially unacceptable, showing that while the ECC sets up a negative implicature, the NCC involves N copying and reinforces the positive property of the N. This explains why d is strange – the N-copy part sets up a strong positive assertion, but then the verb negates it.[3]

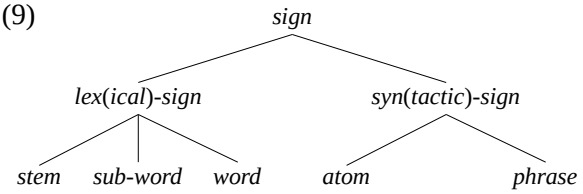
## 4. Lexical Sharing HPSG Analysis

Wescoat (2002) argues that the atomic units of phrase structure are neither words, as claimed by Di Sciullo and Williams (1987), nor morphemes, as assumed in Autolexical Syntax (see Sadock (1991)), but rather lexical-category-bearing *atomic constituents*, each of which maps into a *lexical exponent*, i.e. a word which is said to *instantiate* the atomic constituent. The basic idea of lexical sharing is then that two or more atomic constituents may 'share' the same exponent, or equivalently, that a single word may instantiate multiple atomic constituents. This scheme provides a straightforward model of words that appear to straddle a phrase boundary. Lexical sharing may be simply implemented using the basic machinery of HPSG, in which there is a basic sort of *sign*. Two subtypes of *sign*, namely *phrase* and *word*, have been traditionally employed for representing phrase-structure constituents; thus, standard HPSG is among those

[3]Further differences exist between the ECC and the NCC. Delimiters like *-man* can be used in the ECC but not in the NCC as in *chelswu-ka pwuca-i-ki-man pwuca-ya/*chelswu-ka pwuca-man pwuca-ya*. In addition, a proper noun cannot occur in the NCC as in *ku salam-i John-i-ki-nun John-i-ya/?*ku salam-i John-un John-i-ya*.

theories that regard words as the atoms of phrase structure. In the lexical sharing approach we divorce the type *word* from this role, and have a new, properly syntactic type to represent atomic constituents in phrase-structure, namely *atom*. The modified *sign* hierarchy is shown in (9).

(9)




The type (in *italics*) of an AVM determines what attributes (in SMALL CAPS) and what types of values the AVM may contain. The principal new type declarations are given in (10).

(10) a.
$$\begin{bmatrix} \textit{lex-sign} \\ \text{PHON(OLOGY)} \quad \textit{list(form)} \\ \text{INST(ANTIATE)S} \quad \textit{non-empty-list(atom)} \end{bmatrix}$$

b.
$$\begin{bmatrix} \textit{syn-sign} \\ \text{SYNSEM} \quad \textit{synsem} \end{bmatrix}$$

c.
$$\begin{bmatrix} \textit{atom} \\ \text{EXPON(ENT)} \quad \textit{word} \\ \text{ARG-ST} \quad \textit{list(synsem)} \end{bmatrix}$$

d.
$$\begin{bmatrix} \textit{phrase} \\ \text{D(AUGH)T(E)RS} \quad \textit{non-empty-list(syn-sign)} \end{bmatrix}$$

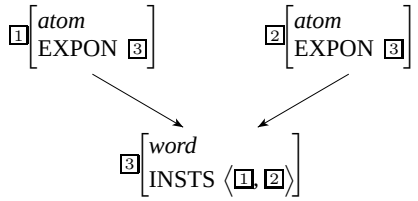
The new attributes EXPON and INSTS in (10) implement lexical sharing: each *atom* is linked via EXPON to a lexical exponent of type *word*; additionally, every *word* contains, as the value of INSTS, an ordered list enumerating each *atom* that the *word* instantiates. We ensure reciprocal linkage between *word* and *atom* with the constraints in (11).

(11) a.
$$\textit{atom} \Rightarrow \boxed{1}\begin{bmatrix} \text{EXPON} & \begin{bmatrix} \textit{word} \\ \text{INSTS} & \langle \ldots, \boxed{1}, \ldots \rangle \end{bmatrix} \end{bmatrix}$$

b.
$$\textit{word} \Rightarrow \boxed{1}\begin{bmatrix} \text{INSTS} & \left\langle \begin{bmatrix} \textit{atom} \\ \text{EXPON} & \boxed{1} \end{bmatrix}, \ldots, \begin{bmatrix} \textit{atom} \\ \text{EXPON} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

The effect of (11) is illustrated by the schematization in (12) of an instance of lexical sharing (compare with the *haksayng-i-ki-nun* part of (5)).

(12)




The tags 1, 2, and 3 reveal the interpenetration of the AVMs which they index: both *atom* 1 and *atom* 2 have the same *word* 3 as value of EXPON, giving rise to lexical sharing; moreover, *word* 3 contains both *atom* 1 and *atom* 2 in its INSTS list, thereby enabling the *word* to determine individually the syntactic features of each *atom*. When there is no actual lexical 'sharing', i.e. when a *word* is exponent of a single *atom*, the INSTS list is simply of length one.

Different from the negative copula verb which selects two nominative arguments, the positive copula *-i* does not exist as a word itself. In morphology, it combines with a noun *sub-word* as input, and returns a *verb-root*. This new lexical item can instantiate two syntactic atoms, the first of which is the noun that was input to the rule, and the second is a two-place predicate which expresses the *be-rel*; see (13) (next page).

This morphological process applies to an N sub-word and creates a form of type *verb-root*. That new form instantiates two atoms in the syntax, an N (which heads NP) and a V (which heads VP), and may be input to further lexical rules. Hence, this is appropriate for the form *haksayng-i-ki-nun* in (5). The lexical rule puts the relevant syntax and semantics of the host N as information about the second argument of the V that the output form instantiates. Nevertheless, this is still a two-place V, an atom which will eventually combine in syntax with a complement NP and then a subject NP.

The first element in INSTS is optional, to capture true incorporation: in one option, the word *haksayng-i-* (the final word in (5)) only instantiates one atom, V, and *haksayng* is truly incorporated.

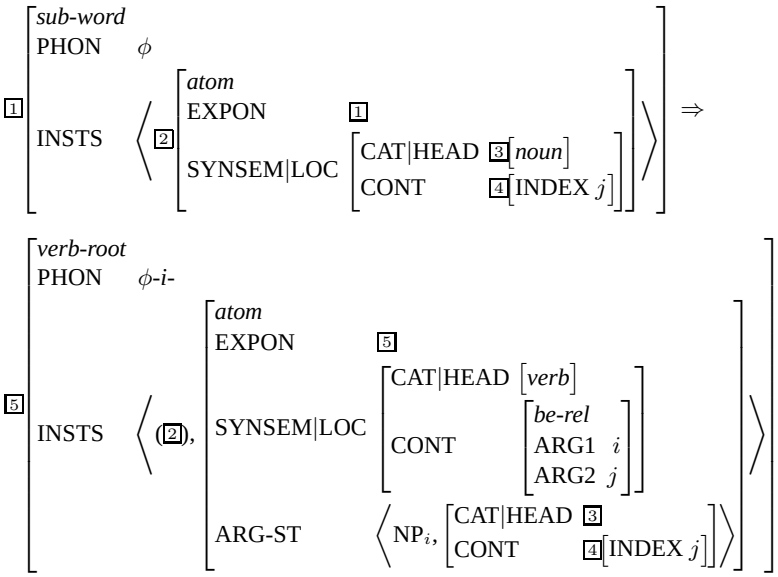
## 5. Copying HPSG Analysis (ECC and NCC)

The similarities and differences of the two constructions can be generalized by the constructional constraint in (14); the two constructions both involve copying the V or N stem (intuitively, within VP or NP, respectively). We specify the stem as the exact inflectional forms of the two 'copies' usually differ, due to other properties of the constructions.

(14) *copy-ph*:

$$[\ \ ] \rightarrow \left[\text{MORPH} \mid \text{STEM}\ \boxed{1}\right], \ \mathbf{H}\left[\text{MORPH} \mid \text{STEM}\ \boxed{1}\right]$$

(13) **Positive Copularization:**




The ECC would inherit from (14) and from a semantic constraint setting up the negative implicature (for details, see Cho et al. (2004). The NCC would also inherit from (14) and but from a semantic constraint expressing a reinforced positive assertion of the property denoted by the (copied) N. The first copy in the NCC is marked with *-nun*, while a verb in the ECC must be first nominalized with *-ki* (which we treat via a FORM feature) before hosting *-nun*, or some other particle (see footnote 3).

In conclusion, the Korean positive copula displays intriguing properties. Closer examination of the relevant data supports our lexical sharing treatment of the precopular element and the copula as a lexical unit. We have shown how a constructional approach can begin to address the similarities and differences of the ECC and the NCC.

## References


Sae-Youn Cho, Jong-Bok Kim, and Peter Sells. 2004. Contrastive verb constructions in Korean. In S. Kuno et al., editor, *Harvard Studies in Korean Linguistics*, volume 10, pages 360–371. Dept. of Linguistics, Harvard University.

Kiyong Choi. 2003. The echoed verb construction in Korean: Evidence for V-raising. In Patricia M. Clancy, editor, *Japanese/Korean Linguistics*, volume 11, pages 457–470. CSLI, Stanford Linguistics Association.

Anna-Maria Di Sciullo and Edwin Williams. 1987. *On the Definition of Word*. MIT Press, Cambridge.

Jung-Min Jo. 2004. Variation in predicate cleft constructions in Korean: Epiphenomena at the Syntax-PF interface. In S. Kuno et al., editor, *Harvard Studies in Korean Linguistics*, volume 10, pages 424–437. Dept. of Linguistics, Harvard University.

Jong-Bok Kim and Chan Chung. 2002. Korean copula constructions: A construction and linearization perspective. *Ene*, 27:171–193.

Jong-Bok Kim, Peter Sells, and Michael T. Wescoat. 2004. Korean copular constructions: A lexical sharing approach. In M. E. Hudson, S.-A. Jun, and P. Sells, editors, *Japanese/Korean Linguistics*, volume 13. CSLI, Stanford Linguistics Association.

Youngsun Kim. 2002. The syntax of Korean verbal focus constructions. In *Proceedings of the 2002 LSK International Summer Conference*, pages 333–344. The Linguistic Society of Korea.

Mira Oh. 1991. The Korean copula and palatalization. *Language Research*, 27:701–724.

Jerrold M. Sadock. 1991. *Autolexical Syntax: A Theory of Parallel Grammatical Representations*. University of Chicago Press, Chicago.

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

James Hye-Suk Yoon. 2003. What the Korean copula reveals about the interaction of morphology and syntax. In Patricia M. Clancy, editor, *Japanese/Korean Linguistics*, volume 11, pages 34–49. CSLI, Stanford Linguistics Association.

# The scope interpretation of the light verb construction in Japanese

**Yusuke Kubota**
Department of Linguistics
The Ohio State University
222 Oxley Hall
1712 Neil Avenue
Columbus, OH 43210, USA
kubota@ling.ohio-state.edu

## 1 The syntax and semantics of the light verb construction in Japanese

In the light verb construction (LVC) in Japanese, an argument of a verbal noun (VN) subcategorized for by a light verb (LV) can sometimes be syntactically realized as an argument of the LV, as was first noticed by Grimshaw and Mester (1988); in (1a), the goal argument *Tookyoo e no* 'to Tokyo' of the VN *yusoo* 'transport' appears with the genitive marker *no*, indicating its status as an argument of a noun, whereas in (1b), the same goal argument appears as an argument of the LV without the genitive marker:

(1) a. Karera wa [Tookyoo e no
They TOP Tokyo GOAL GEN
bussi no yusoo] o si-ta.
goods GEN transport ACC do-PAST
'They transported goods to Tokyo.'

b. Karera wa Tookyoo e [bussi no yusoo] o si-ta.

This phenomenon was termed 'argument transfer' by Grimshaw and Mester (1988). Matsumoto (1996) later discovered that the range of verbs that trigger argument transfer in Japanese is not limited to the prima facie LV *suru* 'do'; a lot of raising and control verbs exhibit patterns of argument realization analogous to that in (1b). In (2), the goal argument of the VN is transferred to the raising verb *hazimeru*:

(2) Karera wa Tookyoo e bussi no
They TOP Tokyo GOAL goods GEN
yusoo o hazime-ta.
transport ACC begin-PAST
'They began transporting goods to Tokyo.' (1996, 77)

The following is a partial list of verbs that trigger argument transfer taken from Matsumoto (1996):[1]

- aspectual verbs: *kurikaesu* 'repeat', *tuzukeru* 'continue', *kaisi suru* 'begin', etc.
- verbs of thinking/planning: *kuwadateru* 'attempt', *wasureru* 'forget', *kangaeru* 'think', etc.
- verbs/nominal adjectives with possibility meaning: *dekiru* 'can', *ari-uru* 'be possible', etc.
- directive and permissive verbs: *meiziru* 'order', *motomeru* 'ask', *mitomeru* 'permit' etc.

Matsumoto (1996) further claimed that not only arguments but also adjuncts can be transferred from the VN to the LV. As convincingly demonstrated by Yokota (1999), however, this assumption is empirically wrong; syntactic dependents of the LV that are unequivocally adjuncts can never allow an interpretation in which it has been 'transferred' from the VN, as shown by the following data:

(3) a. Bussyu wa Koizumi ni tyokusetu
Bush TOP Koizumi DAT direct
no hoobei o mitome-ta.
GEN visit-US ACC permit-PAST
'Bush permitted Koizumi to visit US directly.'

[1] In this paper, I will use the term 'light verb' as a cover term for verbs that trigger argument transfer.

b. Bussyu wa Koizumi ni tyokusetu
Bush TOP Koizumi DAT directly
hoobei o mitome-ta.
visit-US ACC permit-PAST
'Bush permitted Koizumi to visit US in person.'

*Tyokusetu no* 'direct' in (3a) has the form of a nominal modifier whereas *tyokusetu* 'directly' in (3b) has the form of a verbal modifier. If adjuncts could be transferred from the VN to the LV, (3b) should have a reading in which the adjunct *tyokusetu* semantically modifies the embedded VN, from which it has been transferred (i.e. a reading which entails that the visit to US was supposed to be performed in a direct manner). As Yokota correctly points out, however, such an interpretation is only appropriate for sentences like (3a) and not available for sentences like (3b). The only reading available for (3b) is one which the adjunct modifies the LV. This fact is completely unexpected under Matsumoto's (1996) analysis.[2] Thus, the correct generalization is that only arguments can be transferred in the LVC in Japanese.

An important fact that has hitherto been unnoticed in the literature is that quantifiers behave in the same way as adjuncts with respect to the possibilities of scope interpretation in the LVC. As demonstrated by the following data, a quantificational argument of the VN cannot take scope lower than the LV if it has been transferred from the VN to the LV:

(4) a. Zeikan wa gyoosya ni Huransu
customs TOP trader DAT France
kara no wain dake no yunyuu o
from GEN win only GEN import ACC
mitome-ta.
permit-PAST
'The customs permitted the trader to import wine alone from France.' (permit > only)

b. Zeikan wa gyoosya ni Huransu
customs TOP trader DAT France
kara wain dake yunyuu o
from wine only import ACC
mitome-ta.
permit-PAST
'Only as for wine, did the customs permit the trader to import from France.' (only > permit)

In (4a), where the quantified NP *wain dake* 'only wine' syntactically appears as an argument of the embedded VN, the quantifier takes scope lower than the LV *mitomeru* 'permit'. By contrast, in (4b), where the same quantified NP is transferred from the VN to the LV and syntactically realized as an argument of the latter, it has to take scope over the LV.[3]

What the data in (3) and (4) show is that the behaviors of adverbs and quantifiers with respect to the LVC is essentially the same in that their semantic scope is determined simply by their syntactic positions: if they appear as a dependent of the VN, they take scope higher than the VN and lower than the LV; if they appear as a dependent of the LV, they take scope over the LV. As we will see in section 3, this is in contrast with some complex predicates like causatives, for which adverbs and quantifiers can sometimes take scope lower than the positions they syntactically appear.

## 2 Previous analyses and their problems

Matsumoto (1996) employs the mechanism of functional uncertainty (Kaplan and Zaenen, 1989) in LFG to formulate an analysis of the LVC. In his analysis, sentence (2) is assigned the c-structure and f-structure in Figure 1. In the c-structure, the verbally case-marked goal PP *Tookyoo ni* 'to Tokyo' syntactically appears as a sister to the LV. In the f-structure that corresponds to the top S node of this c-structure, this goal argument fulfills the grammatical function $\mathrm{OBL}_{go}$ of the embedded f-structure of the VN. The discrepancy between syntactic sisterhood and semantic head-dependent relation is mediated by functional uncertainty.

[2] Moreover, as I will discuss in detail in the full paper, the alleged cases of adjunct transfer raised by Matsumoto (1996) are dubious in the following two respects: (i) as argued by Yokota (1999), it is doubtful whether what Matsumoto claims to be adjuncts that have been transferred are really adjuncts; (ii) Matsumoto's argument crucially relies on the assumption that the 'transferred' adjuncts semantically modify the VN and not the LV, for which he does not give convincing evidence.

[3] The readings for (4a) and (4b) are clearly distinct from each other with different truth conditions; only the former is compatible with a situation where the customs permitted the trader to import other goods than wine from France.

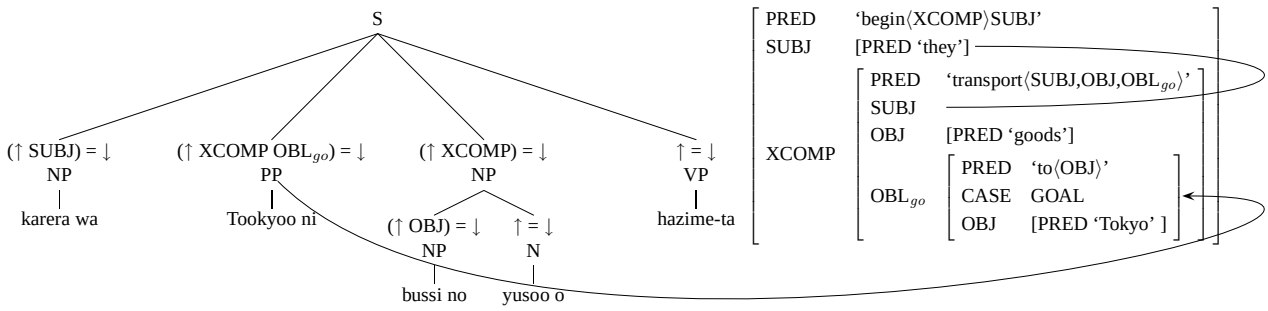



Figure 1: Matsumoto's (1996) analysis of LVC

While this is a simple and elegant analysis, an obvious problem is that it wrongly admits adjunct transfer, since functional uncertainty does not distinguish arguments from adjuncts. Given the observation that adjunct transfer is really impossible, Yokota (1999) proposes an alternative to Matsumoto's analysis in which an explicit stipulation on the function uncertainty schema rules out the possibility of adjunct transfer. Although this modification serves the desired purpose as far as the adjunct data in (3) are concerned, it is implausible in the first place in that it avoids overgeneration by a mere stipulation. More seriously, it is not clear how such an analysis could be extended to account for the fact of quantifier scope. Given the parallelism between adjunct scope and quantifier scope data, what is called for is a mechanism that accounts for these two phenomena in a uniform manner. However, it is unlikely that such an analysis could be developed by extending the proposals of Matsumoto (1996) and Yokota (1999).[4]

## 3 A uniform analysis of the scope interpretation of the light verb construction

Cipollone (2001) recently proposed an analysis of the semantics of the causative construction in Japanese in terms of structured semantic representation, which can potentially account for adverb and quantifier scope phenomena in a uniform manner. I will show in this section that, by extending the proposal of Cipollone, a straightforward analysis of the semantics of the LVC can be constructed, where the parallelism between adjunct and quantifier scope in the LVC is systematically predicted.

The syntax-semantics mismatch in the causative construction in Japanese is well-known at least as early as Shibatani (1976). In the causative construction in Japanese, the sequence of the verb root and the causative suffix behaves as one word, constituting an inseparable syntactic unit. An adverb or quantifier that combines with this complex verb, however, can either take scope over the whole complex predicate or 'inside' the complex predicate, higher than the verb root but lower than the causative predicate. The basic idea of Cipollone (2001) in accounting for this scope ambiguity phenomenon is to allow for slight noncompositionality in the domain of semantics. By doing so, it becomes possible for quantifiers and adverbs to 'look inside' the semantic representation of the phrase they syntactically combine with to pick up the portion they semantically scope over.

Figure 2 illustrates Cipollone's (2001) analysis for one of the readings for the sentence *Gakkoo de hasir-ase-ta* '(I) made (him) run at school', in which the modifier *gakkoo-de* 'at school', which syntactically combines with the whole causative verb *hasir-ase* 'cause to run', semantically modifies only the verb root (i.e. a reading in which what took place at school is the running event). Notice first of all that the CONT feature is list-valued, unlike the standard notation in HPSG (Pollard and Sag, 1994). This list-valued semantic representation can be thought of as a chain of lambda-abstraction, where the value of the LAMBDA feature is a variable index bound by

[4]This point may not be immediately clear given that Matsumoto's (1996) and Yokota's (1999) analyses are not equipped with a mechanism of quantifier scope to begin with. As will be demonstrated in the full paper, however, analyses along the lines of their proposals cannot capture the parallelism between adjunct scope and quantifier scope phenomena straightforwardly, even extended with a suitable mechanism of quantifier scope.

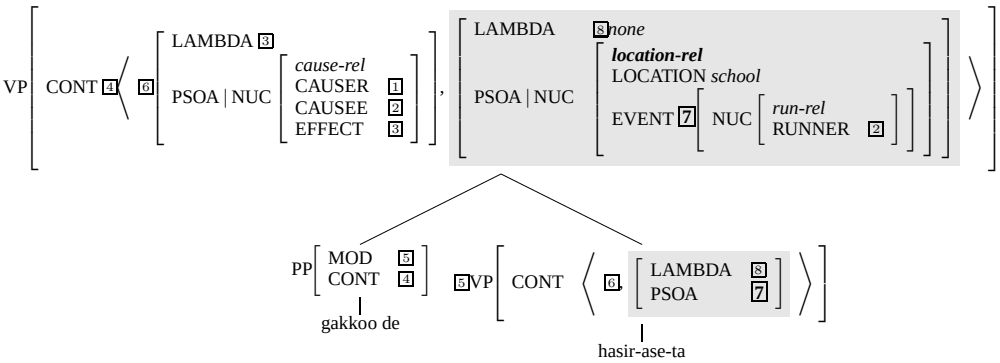



Figure 2: Cipollone's (2001) analysis of narrow scope reading for an adverb in causative

the lambda operator. The semantic interpretation of a phrase is obtained by applying lambda-conversion successively to this list-valued CONT value, where each element of the list is given as an argument to an element immediately to its left. The structure in Figure 2 is licensed by the head-adjunct schema. Thus, the CONT value of the upper VP comes from the CONT value of the adjunct daughter. The adverb, in turn, is specified in the lexicon in such a way that its semantic contribution can be integrated with whichever portion of the list-valued CONT value of the head daughter it combines with; this is formally realized by the constraint in Figure 3. In Figure

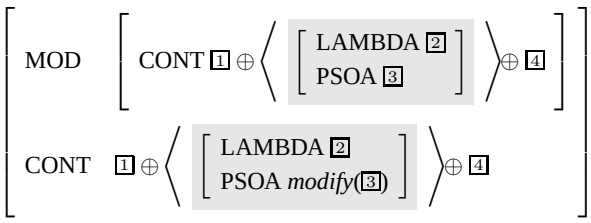



Figure 3: Cipollone's (2001) adjunct schema

2, this constraint is satisfied in such a way that the semantic contribution of the adverb is integrated to the second element of the CONT value of the head daughter, which corresponds to the semantics of the verb root. Thus, the narrow scope reading results. If the semantic contribution of the adverb is instead integrated to the first element of the CONT value of the head daughter, the wide scope reading is obtained, which entails that what took place at school is the causing event.

The system proposed by Cipollone (2001) opens up a way to account for the parallelism of adjunct and quantifier scope phenomena in a uniform manner, since it becomes possible to make the form of the semantic representation of the complex predicate responsible for the availability of scope ambiguity. More specifically, under this scenario, the causative construction exhibits scope ambiguity for adverbs and quantifiers alike since it has a partially transparent semantic representation like the one in Figure 2. By contrast, the LVC does not exhibit scope ambiguity either for adverbs or quantifiers since the internal semantic structure is made invisible to phrases attaching from outside.[5]

The opacity of the semantic representation for the LVC can be guaranteed by lexical specifications on the LV. I propose the lexical entry for the LV *mitomeru* 'permit' in Figure 4.[6,7]

What is crucial here is that the value of the CONT feature is specified as a singleton list. In other words, in the LVC, lambda-conversion of the complex semantic representation, in which the meaning of the VN (tagged as 3 in Figure 4) is embedded under the meaning of the LV, is forced by the lexical

[5]The possibility of accounting for the variability of allowable scope interpretations for different verbs by positing $\beta$-reduced and non-$\beta$-reduced semantic representations is already noted in Cipollone (2001). However, his actual formulation (particularly of the quantifier scope mechanism) would fail to capture the correlation of adverb scope and quantifier scope with respect to different types of complex predicates observed above.

[6]Following Ryu (1993), I model argument transfer in the LVC in terms of the mechanism of argument composition in HPSG (Hinrichs and Nakazawa, 1994), by which unsaturated arguments of an embedded predicate are inherited to the higher predicate by means of structure sharing in the COMPS list.

[7]$\beta$-reduce is a function that produces a *psoa* object from a chain of lambda-abstracted *psoa* objects by lambda-conversion.

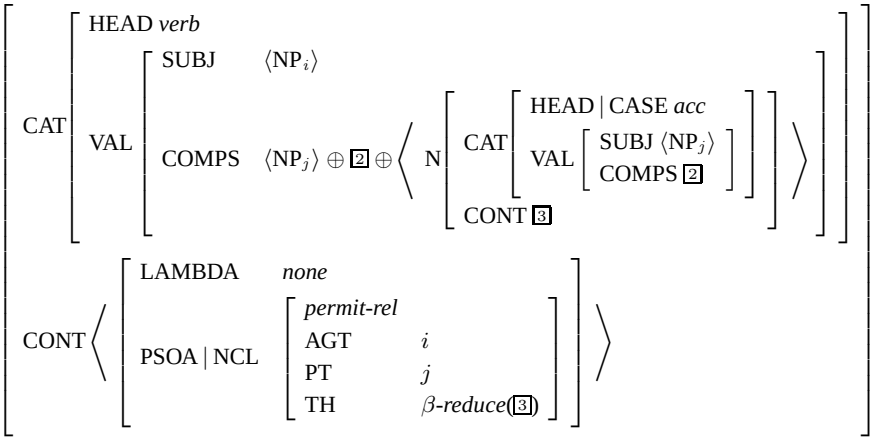



Figure 4: Lexical entry for *tyokusetu* 'directly'

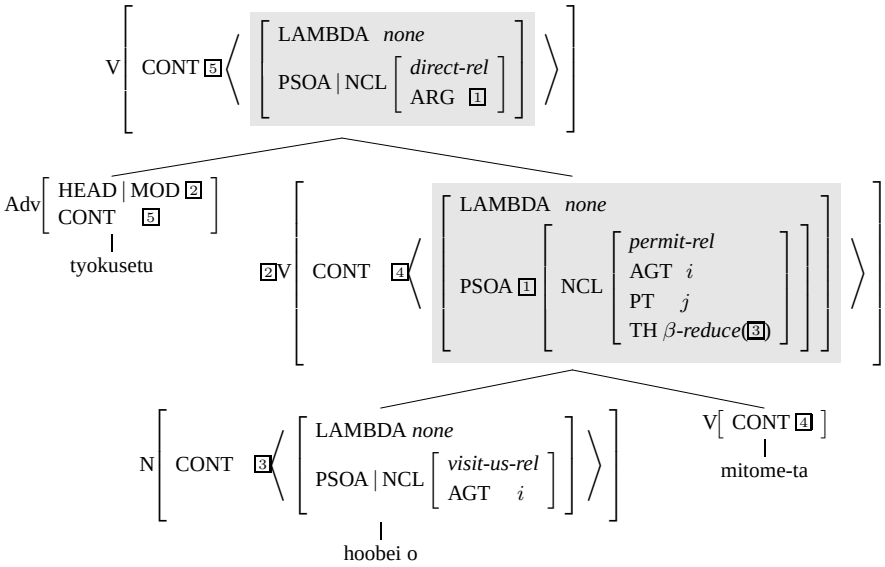



Figure 5: Tree for (3b)

specification of the LV. As a consequence, the internal structure of the complex predicate is no longer visible from outside. Under the present analysis, the structure in Figure 5 is assigned to (3b). It is correctly predicted that the only reading available for this sentence is one in which the adverb semantically modifies the LV. Although the adverb is specified by the constraint in Figure 3 in such a way that it can incorporate its semantic contribution to whichever portion of the list-valued semantic representation of the head daughter it combines with, only one option is available here since the CONT value of the head daughter is made into a singleton list by virtue of the lexical specification of the LV in Figure 4.

Cipollone (2001) accounts for quantifier scope ambiguity of causatives by means of the word-internal quantification mechanism originally proposed by Manning et al. (1999), which is independent of the novel semantic representation he advocates. However, it is trivially easy to revise his analysis in such a way that it crucially makes use of the partially structured semantic representation in determining quantifier scope. By revising his analysis in this way, it becomes possible to capture the parallelism of adverb scope and quantifier scope with respect to different types of complex predicates in a straightforward manner.

In our revised system, quantifier scope is deter-

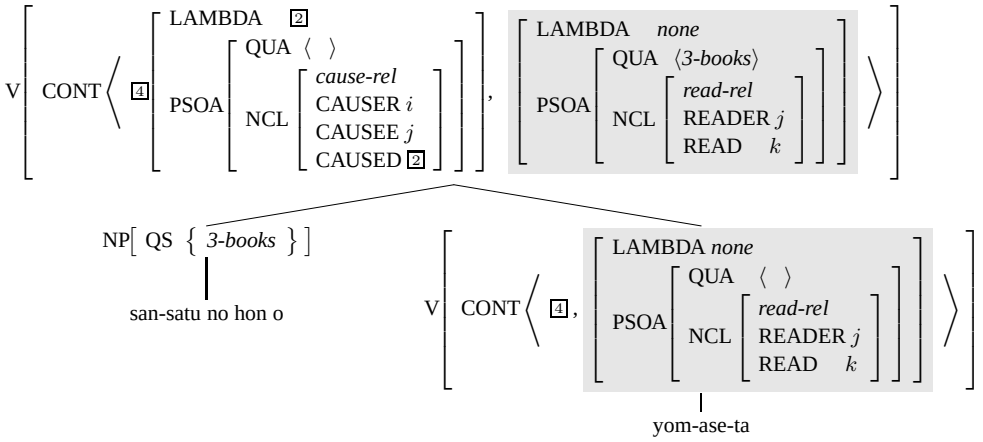



Figure 7: Narrow scope reading for a quantifier in causative

mined in much the same way as adverb scope; quantifiers are allowed to pick up whichever portion of the list-valued semantic representation of the head daughter to scope over. Following Pollard and Sag (1994), I assume that a quantifier takes scope over the *psoa* object to whose QUANTS value it is discharged from QSTORE. The Quantifier Scope Principle in Figure 6 takes care of the determination of quantifier scope along the lines described above.[8] The quantifier inherited via QSTORE is discharged

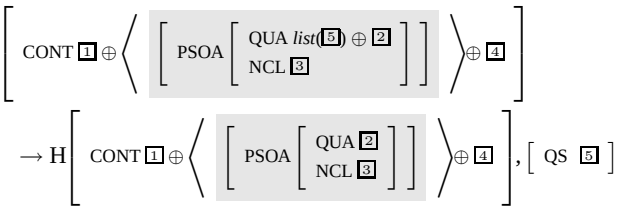



Figure 6: Quantifier Scope Principle

into the QUANTS list of one of the elements of the list-valued CONT value of the head daughter. Figure 7 illustrates an analysis for the narrow scope reading for the sentence *San-satu no hon o yom-ase-ta* '(I) made (him) read three books'.[9] In this tree, the quantifier *san-satu no hon* 'three books' is discharged into the QUANTS list of the second element of the CONT value of the head daughter, which corresponds to the semantics of the verb root. Thus, the narrow scope reading is obtained. If the quantifier is instead discharged into the QUANTS list of the first element of the CONT value of the head daughter, the wide scope reading results. Thus, quantifier scope ambiguity for causatives is correctly predicted.

What is crucial in the present analysis is that adverb scope and quantifier scope are determined with respect to the same information: the (potentially) partially transparent semantic representation of the head daughter. Thus, it is straightforwardly predicted that the narrow scope reading for a quantifier syntactically appearing as an argument of the LV is unavailable in the LVC because of the semantic opacity induced by the LV. Though I omit an analysis here for an LVC sentence involving a quantifier, it should be clear from the discussion so far that quantifier scope is determined uniquely for (4b) in much the same way as adverb scope is determined uniquely as in Figure 5 for (3b). Thus, the fact that quantifier scope ambiguity is unavailable for the LVC is also correctly predicted under the present analysis.

## 4 Conclusion

In the Japanese LVC, adverbs and quantifiers that syntactically appear as dependents of the LV can-

[8] QUA and QS abbreviate QUANTS and QSTORE, respectively.

[9] The narrow scope reading and the wide scope reading for this sentence are distinct from each other: in the narrow scope reading, what the causer did was just to bring about a situation in which the number of books read by the causee amounted to three, where the causer is noncommittal about the choice of specific books; by contrast, in the wide scope reading, the sentence entails that there were three books for which the causer brought about a situation in which the causee read them, in which case the causer knows for sure which books were read by the causee.

not take scope lower than the LV, in sharp contrast to some other complex predicates like causatives, which exhibit the so-called narrow scope readings for adverb and quantifiers. This was first noticed for the case of adverbs by Yokota (1999). However, Yokota's analysis misses the generalization that the same phenomenon is observed for quantifier scope. It was shown in this paper that a theory of semantic interpretation of complex predicates that systematically predicts this parallelism can be constructed by extending the analysis of causatives by Cipollone (2001). The proposed analysis crucially makes use of the partially structured semantic representation introduced by Cipollone in accounting for the availability of adverb and quantifier scope ambiguity for different types of complex predicates in a uniform manner.

## References


Domenic Cipollone. 2001. Morphologically complex predicates in Japanese and what they tell us about grammar architecture. *Ohio State University Working Papers in Linguistics*, 56:1–52. http://ling.osu.edu/osu_wpl/osuwpl56/.

Jane Grimshaw and Armin Mester. 1988. Light verbs and $\theta$-marking. *Linguistic Inquiry*, 19:205–232.

Erhard W. Hinrichs and Tsuneko Nakazawa. 1994. Linearizing AUXs in German verbal complexes. In John Nerbonne, Klaus Netter, and Carl J. Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, number 46 in CSLI Lecture Notes, pages 11–38. CSLI Publications, Stanford.

Ronald Kaplan and Annie Zaenen. 1989. Long-distance dependencies, constituent structure, and functional uncertainty. In Mark Baltin and Anthony Kroch, editors, *Alternative Conceptions of Phrase Structure*, pages 17–42. The University of Chicago Press, Chicago.

Yo Matsumoto. 1996. *Complex Predicates in Japanese*. CSLI Publications/Kurosio Publishers, Stanford/Tokyo.

Carl J. Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. University of Chicago Press, Chicago, London.

Byong-Rae Ryu. 1993. Structure sharing and argument transfer: An HPSG approach to verbal noun constructions. SfS-Report 04-93, Department of Linguistics, The University of Tübingen, Tübingen, Germany.

Masayoshi Shibatani. 1976. Causativization. In Masayoshi Shibatani, editor, *Syntax and Semantics*, volume 5, pages 239–294. Academic Press, New York and Tokyo.

Kenji Yokota. 1999. Light verb constructions in Japanese and functional uncertainty. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG99 Conference*, Stanford. CSLI Publications. http://www-csli.stanford.edu/publications/LFG/lfg4.html.

# A New HPSG Approach to Polish Auxiliary Constructions

**Anna Kupść**
CNRS, Loria (Langue et Dialogue)
*and* Polish Academy of Sciences
Institute of Computer Science
21, Ordona, 01-237 Warsaw, Poland
kupsc@loria.fr

**Jesse Tseng**
CNRS, Loria (Langue et Dialogue)
Campus Scientifique – BP 239
54506 Vandœuvre-lès-Nancy, France
tseng@loria.fr

The "*l*-participle" form of the verb in Polish (for short: *l*-form, so called because it ends in *l* or *ł*, usually followed by a vowel) agrees with the subject in number and gender and is principally associated with the past tense. In this use, it takes personal endings in the 1st and 2nd persons[1] but these endings can also "float" to a position left of the verb—compare (1a) and (1b).

(1) a. Ty widziałeś ten film.
you see$_\ell$-*2sg* this film
'You saw this film.'

b. Tyś widział ten film.
you-*2sg* see$_\ell$ this film

The *l*-form is also used in conditional constructions, in combination with the element *by*. In this case, it is *by* that takes the personal endings, and the inflected forms of *by* either appear immediately after the *l*-form (2a), or they float to its left (2b).

(2) a. Ty widziałby*ś* ten film.
you see$_\ell$.*cond-2sg* this film
'You would see this film.'

b. Ty *byś* widział ten film.
you *cond-2sg* see$_\ell$ this film

And finally, the *l*-form can be used to form the future tense, in combination with future forms of the auxiliary *być* 'be'. In this use, the *l*-form agrees with the subject in gender and number, as usual, but we do not find the 1st and 2nd person endings that characterize the past tense and the conditional. The relative order of the future auxiliary and the *l*-form verb is also much freer than in the other two constructions—see (3).

(3) a. Ty będziesz widział ten film.
you *fut-2sg* see$_\ell$ this film
'You will see this film.'

b. Ty widział będziesz ten film.
you see$_\ell$ *fut-2sg* this film

c. Ty widział ten film będziesz.
you see$_\ell$ this film *fut-2sg*

Some previous accounts of Polish verbal constructions, e.g., Borsley and Rivero (1994), Borsley (1999), Kupść (2000), have attempted to provide a unified analysis of all three uses of *l*-form verbs, although in fact their properties are quite divergent. We will focus on the past tense and the conditional constructions, motivating distinct analyses that account for their particularities more adequately.

## 1 Empirical Observations

There are a number of crucial differences between the conditional particle *by* and the past tense markings that suggest strongly that they do not have the same grammatical status.

First, the forms of conditional *by* can be found after words ending in any segment (i.e., any of the vowels and consonants that appear word-finally in Polish); this is the same behavior we observe for weak (clitic) pronouns. On the other hand, the past tense markings are more particular about the phonological properties of their host, and the different forms have specific constraints (subject to wide variation), as discussed in Bański (2000):

[1] The full inventory of forms is: 1sg *-m*, 2sg *-ś*, 1pl *-śmy*, and 2pl *-ście*.

- the 1sg marking (*-m*) can only be attached to a word ending in a non-nasal vowel (i.e., not *ę* or *ą*);
- the 2sg marking (*-ś*) can additionally (but somewhat marginally) follow a nasal vowel or the glide *j*;
- the 1-2pl forms (*-śmy* and *-ście*) can additionally (but quite marginally) follow *l, r, ł* in a simple coda (e.g., *wór* 'sack' but not *wiatr* 'wind').

Second, the presence of conditional *by* has no morphophonological effect on the preceding material (again, as in the case of pronominal clitics, e.g., Dłuska (1974), Rappaport (1988)). Past tense markings, on the other hand, do induce changes when they follow an *l*-form verb. With a masculine singular subject, the *l*-form ends in *ł*, and so an epenthetic vowel *e* must be inserted before the markings *-m* (4a) and *-ś* (1a). This creates an additional syllable, which results in stress shift, and, with certain verbs, leads to a vowel shift *ó* to *o* (4a).[2] In the plural, the addition of the markings *-śmy*, and *-ście* can, for some speakers or in fast speech, shift the stress one syllable to the right (4b).

(4) a. POmógł → poMOgł**e***m*
help.3sg help.1sg

b. poMOgli → ?pomogLI*śmy*
help.3pl help.2pl

These observations suggest that the past tense endings are much more closely bound to the preceding word than the conditional particle. In fact, their behavior is more typical of morphological suffixes than of independent syntactic items.

Another interesting difference, discussed in Bański (2000), is the interaction of the conditional and past tense markings with coordination. The conditional particle can take wide scope over a coordination of VPs in both preverbal (5a) and postverbal (5b) position. With singular past tense markings, wide scope is possible only in preverbal position (6a) (Bański (2000) overlooks this possibility). The personal ending has to be repeated on all conjuncts if it is realized to the right of the *l*-verb (6b). (For some speakers this requirement is relaxed in the plural).

(5) a. Często *bym* [czytał i pisał].
often *cond-1sg* read and write
'I would often read and write.'

b. Często [czytał*bym* i
often read.*cond-1sg* and
pisał(*bym*)].
write(.*cond-1sg*)

(6) a. Często*m* [czytał i pisał].
often.*1sg* read and write
'I was often reading and writing.'

b. Często [czytał*em* i pisał*(*em*)].
often read.*1sg* and write*(.*1sg*)

According to the criteria of Miller (1992), the obligatory repetition of past tense markings in coordination as in (6b) speaks in favor of their affix status in postverbal positions, whereas optional repetition of the conditional particle in (5b) excludes an affix analysis. On the other hand, the wide scope over coordination in preverbal positions, (5a) and (6a), cannot distinguish between affix and syntactic clitic status.

Finally, there is an important difference in the paradigms of the past tense and conditional markings as there is no 3rd person past tense marking (singular or plural). Compare (7a) and (7b):

(7) a. Tomek czytał książkę.
Tom read book
'Tom read a book.'

b. Tomek *by* czytał książkę.
Tom *cond.3sg* read book
'Tom would read a book.'

This contrast makes it difficult to maintain a parallel treatment of conditional and past tense particles as auxiliaries, because 3rd person past tense constructions would be left strangely 'auxiliary-less'.

The data presented above highlight distinct properties of conditional and past tense constructions and indicate that, despite certain similarities, the two constructions should be analyzed independently. The rest of the paper presents a proposal along these lines.

[2] Capital letters mark lexical stress.

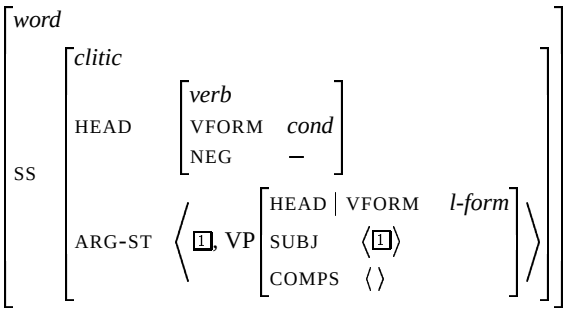



Figure 1: Lexical entry of the conditional clitic

## 2 Proposed Analysis

### 2.1 Conditional Auxiliary *by*

We follow Borsley (1999) and Kupść (2000) in treating all forms of *by* appearing to the left of the *l*-form verb as clitic auxiliaries. They satisfy the lexical description in Fig. 1. As observed in Kupść (2000), there is no direct evidence for the flat structure of conditional auxiliary constructions postulated in Borsley (1999) and we assume VP-complementation here. The resulting entry is that of an ordinary raising verb; the various inflected forms of *by* can make reference to the INDEX of the raised SUBJ element.

The placement of *by* in the sentence field[3] to the left of the *l*-form verb is determined primarily by prosodic structure (see for example Mikoś and Moravcsik (1986) and Bański (2000)). We believe that a DOMAIN-based analysis (cf. Reape (1992)) is the best way to handle the linearization possibilities, although we do not offer a full account here. We simply introduce a shorthand boolean feature CL(ITIC)-HOST to identify words that satisfy (marked [+CL-HOST]) or do not satisfy ([−CL-HOST]) the prosodic and other conditions for hosting a clitic immediately to the right. Non-prosodic conditions on CL-HOST are most apparent in the post-verbal sentence field. All verbs can be [+CL-HOST], so clitics (including conditional *by* and pronominal clitics[4]) can appear immediately to their right. But all (non-clitic) words to the right of the rightmost verb in a clause must be [−CL-HOST] because clitics cannot appear in this field. This is a constraint determined simply by linear order, and one that cannot be overridden by prosodic or syntactic considerations. To account for clitic clusters (including those immediately to the right of the *l*-form verb), we assume that clitics can be [+CL-HOST] and license clitics to their right. As noted in Witkoś (1997), the relative order of pronominal and conditional clitics is very constrained as pronominal clitics tend to follow rather than precede the conditional auxiliary, (8) vs. (9). The same constraint will account for the ungrammaticality of (10) and ensure the correctness of (11).

(8) Ty *byś* go widział.
you *cond-2sg* him.cl see

(9) ?*Ty go *byś* widział.
you him.cl *cond-2sg* see

(10) *Ty widział go *byś*.
you see him.cl *cond-2sg*

(11) Ty widział*byś* go.
you see.*cond-2sg* him.cl
'You would have seen him.'

We do not, therefore, adopt the morphological compound approach of Borsley (1999) for the combination of an *l*-form verb followed by *by*. In our account, *by* is always a clitic, whether it appears somewhere to the left of the *l*-form or immediately to its right. There is no evidence (stress shift or vowel quality alternations, for example) to motivate two distinct types of combination.

### 2.2 Past Tense Agreement Markings

Our analysis of the "floating" past tense elements *-m*, *-ś*, *-śmy*, and *-ście* is motivated by the observation that these elements do not function as auxiliary verbs in modern Polish. In fact, as discussed in section 1, the past tense personal markings are no longer independent syntactic items at all, but suffixes. We take the *l*-form to be the head of a simple past tense construction, requiring the presence of an agreement marking in the 1st and 2nd persons.

Unlike ordinary suffixes, the past tense agreement markings can attach to a variety of hosts, and they do not always appear on the lexical head of a phrase, or in any particular linear position in the phrase (e.g., at

[3] We use the term "field" in a purely descriptive way, without suggesting that any version of the topological fields approach, as used for the analysis of German word order, would be applicable to Polish.

[4] As argued in Kupść (2000), Polish pronominal clitics are syntactic items.

the left or right edge). A special affixation and propagation mechanism is needed to handle this kind of behavior. The realization of the agreement marking is subject to a surface order constraint: it must appear exactly once, somewhere to the left of the *l*-form verb (or on the verb itself). And unlike in ordinary cases of agreement, there are no syntactic restrictions on the constituent that receives the marking: it can be a complement, an adjunct, or a filler, an NP or a PP, or even a complementizer (see, e.g., Borsley and Rivero (1994) for some examples). A special mechanism is also required to handle this aspect of the phenomenon.

Agreement markers appearing to the left of the *l*-form are introduced by the inflectional rule in (12).

$$(12)\quad \begin{bmatrix} \textit{word} \\ \textsc{phon}\ \boxed{1} \\ \textsc{ss}\ \begin{bmatrix} \textit{synsem} \\ \textsc{loc}|\textsc{cat}|\textsc{head}\ \neg\ \textit{verb} \\ \textsc{cl-host}\ + \\ \textsc{agr-mark}\ \langle\,\rangle \end{bmatrix} \end{bmatrix} \mapsto \begin{bmatrix} \textsc{phon}\ \ \mathbf{F}_{\textit{agr-nonvb}}(\boxed{1},\boxed{2}) \\ \textsc{ss}\mid\textsc{agr-mark}\ \ \langle \boxed{2}\ \textit{index}\rangle \end{bmatrix}$$

The input cannot be a verb, but there are no other categorial restrictions on the host. The output of the rule is the same word with the appropriate suffix (identified by its index), with a phonological form determined by the function $\mathbf{F}_{\textit{agr-nonvb}}$. For incompatible combinations (e.g., a word ending in a nasal vowel cannot take the 1st person singular suffix, a word ending in a consonant cannot take any suffix, there are no 3rd person suffixes) no output form can be constructed and the rule fails. When suffixation is possible, the corresponding *index* element appears on the suffixed word's AGR(EEMENT)-MARK(ING) list. The empty AGR-MARK list in the input prevents iteration of the rule. We assume that the AGR-MARK value is introduced lexically, i.e., it is specified in each lexical entry. The only words in Polish with a non-empty AGR-MARK list are those that have undergone rule (12) and, as we will discuss below in sec. 2.3, the 1st and 2nd person forms of the conditional and future auxiliaries.

The suffixed word that introduces the AGR-MARK element participates normally in syntactic combinations, with all possible grammatical functions (head, complement, modifier, etc.). The presence of the agreement affix has no influence on the syntactic properties of the host. The affix does influence the linearization potential of its host: the output of (12) remains [+CL-HOST], so the suffixed word must end up in a surface position that is compatible with this feature. But its exact location within a phrase cannot be specified: it can be the first word, the last word, or somewhere in the middle. But in all cases, information about the affix must be projected. This means that the value of AGR-MARK is amalgamated from all daughters in every phrasal combination. The formalization of this constraint is given in Fig. 2, but first we need to explain the role of the verb in past tense constructions.

The agreement marker is required by the *l*-form verb. We encode this using the feature AGR(EEMENT)-TRIG(GER), which takes a list of *index* objects as its value. For past tense verbs, AGR-TRIG is specified lexically: for *l*-verbs with 1st or 2nd person subject, AGR-TRIG contains the appropriate index (structure-shared with the INDEX value of the subject), whereas *l*-verbs with 3rd person subjects have an empty AGR-TRIG list. All other words have an empty AGR-TRIG list:

$$(13)\quad \begin{bmatrix} \textit{word} \\ \textsc{ss}|\textsc{l}|\textsc{c}|\textsc{h}\ \neg\ \begin{bmatrix}\textsc{vform}\ \ \textit{l-form}\end{bmatrix} \end{bmatrix} \rightarrow \begin{bmatrix}\textsc{ss}|\textsc{agr-trig}\ \langle\,\rangle\end{bmatrix}$$

An element on AGR-TRIG can be discharged by applying a suffixation rule to the verb itself:

$$(14)\quad \begin{bmatrix} \textit{word} \\ \textsc{phon}\ \boxed{1} \\ \textsc{ss}\ \begin{bmatrix} \textsc{l}|\textsc{c}|\textsc{h}|\ \textsc{vform}\ \ \textit{l-form} \\ \textsc{cl-host}\ + \\ \textsc{agr-trig}\ \langle\boxed{2}\rangle \end{bmatrix} \end{bmatrix} \mapsto \begin{bmatrix} \textsc{phon}\ \ \mathbf{F}_{\textit{agr-vb}}(\boxed{1},\boxed{2}) \\ \textsc{ss}|\textsc{agr-trig}\ \ \langle\,\rangle \end{bmatrix}$$

This rule must be distinct from (12) above, because the morphophonological aspects of agreement suffixation (encoded in the function $\mathbf{F}_{\textit{agr-vb}}$) are different for verbal hosts. For example, $\mathbf{F}_{\textit{agr-vb}}$ handles *e*-epenthesis and the *ó* to *o* alternation in masculine past tense forms—recall (1a) and (4a). Moreover, the grammatical effect of the rule is to empty the verb's AGR-TRIG list.

In other cases, AGR-TRIG must be discharged at a distance, by interaction with an AGR-MARK element introduced earlier in the clause. The overall

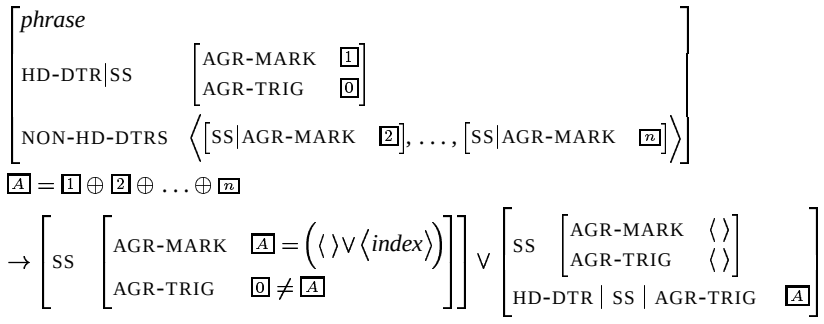



Figure 2: Amalgamation of AGR-MARK from dependents

constraint on phrasal combinations is given in Fig. 2. All the daughters' AGR-MARK elements are amalgamated (to yield the list $\boxed{A}$). If this list does not correspond to the head daughter's AGR-TRIG value, then the first disjunct on the right-hand side is satisfied: the mother's AGR-MARK value is the (maximally singleton) amalgamated list $\boxed{A}$, and the value of AGR-TRIG is taken from the head daughter. But if the amalgamation $\boxed{A}$ of the daughters' AGR-MARK lists does unify with the head's AGR-TRIG list, then they "cancel each other out" and both lists are empty in the phrasal description (second disjunct).[5]

The agreement marking cannot appear to the right of the verb that selects it. To block such structures, we formulate the following linear precedence rule:

$$(15)\quad \left[\text{SS}|\text{AGR-MARK}\ \left\langle \boxed{1} \right\rangle\right] \overset{\text{HD-DTR}}{<} \left[\text{SS}|\text{AGR-TRIG}\ \left\langle \boxed{1} \right\rangle\right]$$

At the clausal level, there can be no unlicensed agreement markers (AGR-MARK elements) and no unsatisfied agreement requirements (AGR-TRIG elements):

$$(16)\quad \textit{clause} \rightarrow \left[\text{SS}\ \begin{bmatrix}\text{AGR-MARK} & \langle\,\rangle \\ \text{AGR-TRIG} & \langle\,\rangle\end{bmatrix}\right]$$

### 2.3 Conditional (revisited) and Future

In addition to the lexical description given in Fig. 1, we assume that the 1st and 2nd person forms of the conditional clitic *by* (*bym*, *byś*, *byśmy*, *byście*) have a non-empty AGR-MARK list, containing the index of their subject. The 3rd person form *by* has an empty AGR-MARK list. This means that the same *l*-form verb appears in past tense and in conditional constructions; we do not need to assume distinct "finite" and "participial" variants. This analysis accounts for the fact that the conditional auxiliary must combine with a "bare" (unsuffixed) *l*-form (17a), and the fact that no other word can carry the agreement marking (17b, 17c). Also, note that the LP rule (15) does not prevent the conditional auxiliary (with non-empty AGR-MARK) from appearing to the right of the *l*-form (with corresponding AGR-TRIG), because in this case the *l*-projection is not the head (17d).

(17) a. *Ty byś widziałeś ten film. (undischarged AGR-MARK)

b. *Tyś byś widział ten film. (undischarged AGR-MARK)

c. *Tyś by widział ten film. (*by* selects a 3rd person subject)

d. Ty widziałbyś ten film.

Recall that the conditional auxiliary cannot appear any further to the right because it is a clitic, and all words in this part of the clause are [−CL-HOST].

The same analysis applies to the forms of the future tense auxiliary—recall examples in (3).[6] This

[5]The second disjunct covers the interesting case where the head's non-empty AGR-TRIG requirement is discharged by presence of the corresponding AGR-MARK element on another daughter, and the less interesting (but much more frequent) case where the daughters' AGR-MARK lists and the head's AGR-TRIG list are all simply empty, and both values remain empty on the mother.

[6]Forms of the future tense auxiliary:

| | SG | PL |
|---|---|---|
| 1st | *będę* | *będziemy* |
| 2nd | *będziesz* | *będziecie* |
| 3rd | *będzie* | *będą* |

auxiliary, however, is quite different from the conditional in other ways: it is an independent word (not a clitic), and it may require a flat structure as proposed for French auxiliary constructions in Abeillé and Godard (2002) (to allow "clitic climbing", cf. discussion in Kupść (2000)).

## 3 Conclusion

We have presented two separate analyses of Polish past tense and conditional constructions based on their distinct morphophonological, inflectional and syntactic properties. As in Borsley (1999), we treat the conditional particle as an auxiliary clitic verb but we do not adopt his morphological compound analysis and explain all positions of the weak auxiliary by linear order. On the other hand, in our analysis the past tense marking is neither a syntactic item nor an auxiliary. We treat it as an agreement suffix which is required by the *l*-verb but can be realized in different places in the sentence.

The past tense in Polish is an interesting case of grammaticalization where competing analyses are available. We have presented an alternative to earlier approaches but a fully adequate analysis would need to be able to model the transition between different grammatical structures which is currently taking place in the grammar of Polish past tense forms.

## References


Abeillé, A. and Godard, D. (2002). The syntactic structure of French auxiliaries. *Language*, **78**, 404–452.

Bański, P. (2000). *Morphological and Prosodic Analysis of Auxiliary Clitics in Polish and English*. Ph.D. thesis, Uniwersytet Warszawski, Warsaw.

Borsley, R. D. (1999). Weak auxiliaries, complex verbs and inflected complementizers. In R. D. Borsley and A. Przepiórkowski, editors, *Slavic in Head-Driven Phrase Structure Grammar*, pages 29–59. CSLI Publications, Stanford, CA.

Borsley, R. D. and Rivero, M. L. (1994). Clitic auxiliaries and incorporation in Polish. *Natural Language and Linguistic Theory*, **12**, 373–422.

Dłuska, M. (1974). *Prozodia Języka Polskiego*. Państwowe Wydawnictwo Naukowe, Warsaw.

Kupść, A. (2000). *An HPSG Grammar of Polish Clitics*. Ph. D. dissertation, Polish Academy of Sciences and Université Paris 7.

Mikoś, M. and Moravcsik, E. (1986). Moving Clitics in Polish and Some Cross Linguistic Generalizations. *Studia Slavica*, pages 327–336.

Miller, P. H. (1992). *Clitics and Constituents in Phrase Structure Grammar*. Garland, New York.

Rappaport, G. C. (1988). On the relationship between prosodic and syntactic properties of pronouns in the Slavic languages. In A. M. Schenker, editor, *American Contribution to the Tenth International Congress of Slavists*, pages 301–327. Slavica Publishers, Columbus, Ohio.

Reape, M. (1992). *A Formal Theory of Word Order: A Case Study in West Germanic*. Ph. D. dissertation, University of Edinburgh.

Witkoś, J. (1997). Polish inflectional auxiliaries revisited. Paper presented at the 30th Poznań Linguistic Meeting, Adam Mickiewicz University, Poznań, Poland, May 1–3, 1997.

# A Trace Analysis of Korean UDCs

**Sun-Hee Lee**
Department of Computer Engineering and Science
Ohio State University
Columbus, OH 43210, USA
shlee@ling.ohio-state.edu

## 1 Introduction

In Korean, there are various grammatical constructions that involve a long-distance dependency between a gap and some constituent that is coreferential with that gap. The dependency is in principle unbounded and can be captured by a feature percolation mechanism within HPSG. However, c ertain properties of gaps in Korean unbounded dependency constructions (hereafter UDCs) raise questions as to whether a syntactic approach to this long-distance dependency is appropriate. In fact, some previous researchers, including (Kang, 1986) and Yoon [1993] have argued that this dependency needs to be handled at the level of semantics, not syntax. In such a semantic approach, UDC gaps are treated as null resumptive pronouns ( so-called *pro*s in GB terms), and syntactic binding between a gap and its antecedent is not required. However, UDC gaps and *pro*s in Korean show different properties with respect to Strong Crossover and Coordination facts. Furthermore, we examine putative resumptive pronouns (RPs), and the resumptive reflexive (RR) *caki* that appear in the same positions of UDC gaps, and argue that these resumptive elements are audible traces. This argument is compatible with resumptive pronoun analyses of (Georgopoulos, 1991) in Palauan and (Vaillette, 2001) in Hebrew. In this paper, we claim that the filler-gap linkage in Korean UDCs needs to be handled at the level of syntax and that unbounded dependencies in Korean can be captured by a feature percolation mechanism within HPSG. We also investigate some controversial issues of island constraints and strong crossover with respect to filler-gap linkage in Korean UDCs.

This paper shows that unbounded dependencies represented by traces, RPs, and the RR *caki* can be simply captured - without posing any extra mechanisms - in the traditional HPSG analysis of UDCs following (Pollard and Sag, 1994). It is because in HPSG traces are not all required to have the same feature, unlike in other movement-basedapproaches including the minimalist program and GB theory. In addition, we conclude that the three kinds of Korean UDC elements appearing in gap positions do not form separate categories from their corresponding forms appearing in non-UDCs based on the same semantic and pragmatic properties such as logophoricity and contrastiveness.

## 2 A Null Pronominal Analysis and Its Problems

Korean has been standardly considered to be a *pro*-drop language. This is a language where a contextually identifiable element or some element introduced in the preceding context can be dropped. (Huang, 1984) argues that "cool" languages, including Chinese and Korean, are different from "hot" languages, like English, in that cool languages license a zero topic that binds a null element. While Huang argues that the phonologically null element *pro* appears only in the subject position in cool languages, it has been argued that there is no subject-object asymmetry in Korean((Cole, 1987)). Since Korean is classified as a *pro*-drop language, it is possible to argue that gaps in UDCs are null resumptive pronouns or *pro*s, and that correspondingly, the long-distance dependencies are not syntactic relations but rather semantic binding relations. The following examples show that a gap can be replaced by an overt pronoun or the long-distance reflexive *caki*, which appears to support the semantic binding analysis.

(1) a. ku namca$_i$-nun [ sacang-i
that man-TOP president-NOM
eps-umyeon, $e_i$ motun il-ul ttemath-aya
absent-if every work-ACC took care
hayssta ].
had to

'As for that man$_i$, if the president were absent, (he$_i$) had to take care of everything.'

b. ku namca$_i$-nun [ sacang-i
that man-TOP president-NOM
eps-umyeon, ku$_i$/caki$_i$-ka motun il-ul
absent-if he/self all work-ACC
ttmath-aya hayssta ].
took care did

'As for that man$_i$, if the president were absent, he$_i$ had to take care of everything.'

As for English Cinque [1990] and (Postal, 1994) propose transformational analyses with null pronominals for English *tough* gaps and parasitic gaps. In Korean, (Chae, 1998) and (Kang, 1986) assumed that *tough* constructions, topicalization, and relativization in Korean license *pro*s, which are phonologically null elements in the gap position. However, in this study we treat those pronouns and the long-distance (LD) *caki* as audible traces and argue that the filler-gap linkages in Korean UDCs need to be captured by a syntactic mechanism of binding and not just by semantic coreference. Three different kinds of traces show the same phenomenon with respect to Strong Crossover and Coordination. This suggests that they belong to the same category of trace.

## 3 Properties of Korean UDC Gaps

A UDC gap needs to have a coreferential element within the given sentence. While the syntactic and semantic connectivity between a gap and its antecedent in Korean UDCs is similar to the corresponding English sentences, Korean UDC gaps are known to be less sensitive to island constraints. The following properties have been pointed out by general properties of Korean UDC gaps.

**[1] Syntactic Connectivity**
There are two natural classies of Korean UDCs: strong UDCs and weak UDCs. In the case of strong UDCs, the filler is accompanied by the morphosyntactic case marker that originated from the gapped position, thus the filler shows a strong syntactic association with its gap. Strong UDCs in Korean include the following topic sentence.

(2) a. Mary-ka *John-eykey* senmwul-ul
Mary-NOM John-to present-ACC
cwuessta.
gave

'Mary gave a present to John.'

b. *John$_i$-eykey*-nun [ Mary-ka $e_i$ senmwul-ul
John-to-TOP Mary-NOM present-ACC
cwuessta].
gave

'As for John$_i$, Mary gave a present (to him$_i$)

The case markers of the topic element in (2) show that it is syntactically connected to the gap; the dative case *eykey* (to) is required by the verb *cwuta* (give).

**[2] Sentence-Internal Binding**
A UDC gap must have a coreferential element within the same sentence. This property distinguishes UDC gaps from *pro*s, which are licensed by various syntactic, semantic, and pragmatic factors. For example, discourse factors allow a repeated or already-known element to be dropped from a sentence in languages like Korean. When this happens, the missing element can be retrieved from the context. However, a UDC gap requires its coreferential element to be present in the given sentence; it cannot be licensed only by context.

**[3] Island Constraints**
With respect to Korean UDCs, it has been argued that some examples of topicalization and relativization are subject to three island constraints: the Complex NP constraint (CNPC), the Sentential Subject constraint, and the Adjunct constraint. This evidence has been used to support the claim that topicalization and relativization involve NP movement out of gap positions in Korean. In contrast, it has been also pointed out that topic and relative clauses in Korean frequently do violate island constraints ((Kang, 1986)). Inconsistency of data with respect to island constraints suggests that unlike most previous analyses in GB theory, island constraints cannot be used as a crucial test for determining whether a particular construction is a UDC or not.

However, some crosslinguistic studies have pointed out that sensitivity to island constraints cannot be used as evidence for the existence of a filler-gap linkage. When dealing with English adjunct extractions, (Hukari and Levine, 1995) argued that island effects are substantially irrelevant to the issue of whether or not adjunct extraction represents a genuine syntactic filler-gap construction. Instead, they argued that adjunct extraction belongs to the same category of UDCs as argument extraction. They based their conclusion on parallel patterns of crossover effects and on cross-linguistic evidence of syntactic binding domain effects. (Szabolcsi and den Dikken, 1999) also argued that some island constraint effects are relevant to the semantic scope that an expression takes over certain operators.

Considering that island constraint violations are driven by semantic and pragmatic factors but not by a syntactic operation like movement, inconsistency of island constraints in Korean UDCs cannot be supporting evidence for semantic binding approaches to Korean. In addition to syntactic connectivity, semantic binding relations between a UDC gap and a constituent are tighter than other binding relations between a pronoun and its antecedent. In the next section, we will examine strong crossover and coordination facts that distinguish the filler-gap linkage

of Korean UDC gaps from semantic binding. Then, later in this paper we will provide a syntactic representation of unbounded dependencies with a simple syntactic tool, which avoids all the problems of island constraint violations that the movement approaches have confronted.

# 4 Characterizing Properties of Korean UDC Gaps

## 4.1 Strong Crossover

The Strong Crossover (SCO) Constraint does not apply to *pro*s in general, as we see in (3).

(3) [ John$_i$-un [ $e_i$ [ Mary-ka ku$_i$-eykey [ $pro_i$
John-TOP Mary-NOM he-to
kayahanta-ko] malhayssta-ko] kiekhanta].
must go-COMP told-COMP remember
'As for John$_i$, (he$_i$) remembers that Mary$_j$ told him$_i$ that (e$_i$) must go.'

In(3), $e_i$ represents a gap directly linked to its antecedent in the position of topic. It contrast with a *pro* that appears in the most deeply embedded clause. In general, *pro*s in Korean occur when their coreferential elements (antecedents) are introduced in the previous context or when their coreferential elements syntactically precede. The *pro*$_i$ takes the preceding pronoun *ku*$_i$ as its antecedent and refers to *John* in (3). This violates the SCO constraint. In contrast with *pro*s, UDC gaps observe the SCO constraint, as in the following example.

(4) * ku ai$_i$-nun Mary-ka ku papo$_i$-eykey
that child-TOP Mary-NOM that idiot-to
[$e_i$/ku$_i$/caki$_i$-lul cal tolpokessta-ko]
/he-/selfACC well take care-COMP
yaksokhayssta.
promised
'As for the child$_i$, Mary promised that idiot$_i$ to take care of him$_i$ well.'

The example (4) shows that SCO is observed for UDCs. Instead of a pronoun an epithet has been used in (4). It is because the use of pronoun *ku* may allow a resumptive pronoun analysis of the intervening pronoun, which follows (Vaillette, 2001). In order to examine the applicability of crossover to Hebrew RPs, (Vaillette, 2001) replaces the upper pronoun by an epithet. The epithet has the same index value as the antecedent, while it retains an independent lexical meaning. Although (what looks like) pronouns and reflexives can be audible (SLASH-bearing) traces, epithets cannot be. Thus, the same strategy can be applied to Korean.

A notable point is that resumptive pronominal elements in Korean UDCs observe the SCO constraint as do inaudible traces. This fact is problematic because previous literature has assumed that SCO violations are triggered by the status of UDC gaps; in general UDC gaps are nonpronominal elements or R(eferring)-expressions. However, RPs in Korean UDCs show the same SCO effects as nonpronominal gaps in spite of their pronominal status. Within Chomskyan approaches, the SCO effects are accounted for by Principle C that requires so-called R-expressions to be unbound. Similarly, within the framework of HPSG, the SCO phenomenon has been explained by the binding condition C that specifies that a nonpronoun must be o-free. However, (Postal, 2004) argues that the SCO phenomenon in English cannot be accounted for by Chomsky's Principle C, and based on his arguemnts it is hard to argue that SCO effects are attributed to the status of UDC gaps as nonpronominal elements.[1] The SCO effects in Korean UDCs are not associated with Principle C (or condition C in HPSG). This argument is supported by the following examples.

(5) a. ku ai$_i$-nun wuli-ka [$_{ADVP}$ John$_k$-ul
the kid-TOP we-NOM John-ACC
thonghay-se] [$_S$ $e_i$ iphak sihem-ey
mediate-by entrance exam-at
hapkyekhayss-um-ul] alkey toyessta.
pass-NML-ACC know became
(lit.)'As for the kid$_i$, we got to know via John that (he$_i$) passed the entrance exam.'

b. * ku ai$_i$-nun wuli-ka [$_{ADVP}$ ku papo$_i$-lul
the kid-TOP wuli-NOM that idiot-ACC
thonghay-se] [$_S$ $e_i$ iphak sihem-ey
mediate-by entrance exam-at
hapkyekhayss-um-ul] alkey toyessta.
pass-NML-ACC know became
(lit.)'As for the kid$_i$, we got to know via that idiot$_i$ that (he$_i$) passed the university exam.'

c. * ku ai$_i$-nun wuli-ka [$_{ADVP}$ ku
that child-TOP we-NOM that
papo$_i$-lul thonghay-se] [$_S$ ku$_i$-ka iphak
idiot-ACC mediate-by he-NOM entrance
sihem-ey hapkyekhayss-um-ul] alkey
exam-at pass-NML-ACC know

[1] (Postal, 2004) points out that the SCO effect cannot be reduced to Chomsky's Principle C that bars anaphoric linkage between pronoun and the nonpronominal trace based on (i) existence of SCO effects in non-NP extraction, (ii) the secondary strong effect, (iii) the Asymmetry Property and (iv) failure of the c-command condition required for Principle C. He claims that even though the Principle C account of the SCO effect is often considered to be supporting evidence of traces as nonpromoninal R-expressions,there is no empirical evidence for any trace-like objects connected with extraction.

toyessta.
became

(lit.) 'As for the kid$_i$, we got to know via that idiot$_i$ that he$_i$ passed the entrance exam.'

In the given examples, the intervening epithets are located in adjunct phrases that do not c-command (or o-command) the gaps in the embedded phrases. Although no violation of Principle C (or condition C) can be induced in (5), anaphoric linkage between a filler and a gap is as impossible as in (5b) and (5c). Moreover, when a gap appears in an adverbial phrase of the embedded clause, the SCO effects appear in spite of the failure of c-command between a pronoun or an epithet and its anaphoric gap. Specifically, the backward linking of a pronoun or an epithet to an antecedent in an adjunct can be licensed as shown in (6a).[2] In contrast, the antecedent in an adjunct cannot be topicalized as in (6b) and (6c).

(6) a. Nay-ka kyay/ku papo$_j$-hanthey [[ John$_j$-i
I-NOM he/that idiot-to John-NOM
ttena-camaca] Mary-ka tochakhayssta-ko]
leave-soon Mary-NOM arrived-COMP
cenhaysse.
told

'I told him$_j$/that idiot$_j$ that Mary arrived right after he$_j$(John)left.'

b. ?* John$_j$-un nay-ka kyay/ku papo$_j$-hanthey [[
John-TOP I-NOM he/that idiot-to
$e_j$ ttena-camaca] Mary-ka
leave-soon Mary-NOM
tochakhayssta-ko] cenhaysse.
arrived-COMP told

'As for John$_j$, I$_i$ told him$_j$/that idiot$_j$ that Mary arrived right after he$_j$ left.

c. ?* John$_j$-un nay$_i$-ka kyay/ku papo$_j$-hanthey [[
JohnTOP I-NOM he/that idiot-to
ku$_j$-ka ttena-camaca] Mary-ka
he-NOM left-soon Mary-NOM
tochakhayssta-ko] cenhaysse.
arrived-COMP told

'As for John$_j$, I$_i$ told him$_j$/that idiot$_j$ that Mary arrived right after he$_j$ left.'

[2]In general, backward linking between a pronoun and its antecedents is often allowed. Postal points out that ungrammatical extractions out of islands can be still used to test binding hypothesis because of the following principle. This principle can be used for examples in (6).

(i) Mere extraction from an island, even when yielding severe ill-formedness, does not inherently block anaphoric linkage if such are licit in the pre-extraction structure itself.

Based on the fact that a pronoun and its anaphoric element do not hold a c-command (or o-command) relation, we conclude that SCO effects in Korean UDCs cannot be reduced to Principle C in GB theory or condition C in HPSG. Thus, there is no factual support for the status of traces as nonpronominal elements, which is why the SCO constraint is observed by both RPs and inaudible traces in Korean UDCs. This accords with SCO effects in English as shown in (Postal, 2004). An RP can be represented in HPSG via the propagation of a non-local feature. In addition to an RP, the long distance reflexive *caki* 'self' can also appear in the position of the trace.

### 4.2 Coordination

In general, it has been argued that the Coordinate Structure Constraint (CSC) is observed in Korean coordinate structures. The constraint disallows asymmetric extraction out of one conjunct. For example, (7b) and (7c) are ungrammatical because only one conjunct has a missing element. However, (7a) is grammatical because the topicalized element is connected to the missing elements in both conjuncts.

(7) a. i chayk$_j$-un [ aitul-i e$_j$ cohaha-ko
this book-TOP kids-NOM like-CONJ
eluntul-to e$_j$ chohahay].
adults-also like

'As for this book$_j$, kids like (it$_j$) and adults also like (it$_j$).'

b. * i chayk$_j$-un [ aitul-i e$_j$ cohaha-ko
this book-TOP kids-NOM like-CONJ
eluntul-i manhwachayk-ul silehay].
adults-NOM comic book-ACC like

'As for this book$_j$, kids like (it$_j$) and adults dislikes comic books.'

c. * i chayk$_j$-un [ aitul-i manhwachayk-ul
this book-TOP kids-NOM comic books-ACC
cohaha-ko elun-i e$_j$ cohahay].
like-CONJ adults-NOM like

'As for this book$_j$, kids like comic books and adults dislike (it$_j$).'

Another fact related to coordination is that a gap in a conjunct is allowed when there is a gap in the other conjunct,or a pronoun, as in (8a) and (8b).

(8) a. i chayk$_j$-un [aitul-i kukes$_j$-ul acwu
this book-NOM kids-NOM it-ACC very
cohaha-ko nointul-to e$_j$ congcong
like-CONJ old people-also often
chassnunta]
ask for

'As for this book$_j$, kids like it$_j$ very much and old people also buy (it$_j$) often.'

b. i chayk$_j$-un [aitul-i $e_i$ acwu
this book-NOM kids-NOM very
cohaha-ko nointul-to kukes$_j$-ul
like-CONJ old people-also it-ACC
congcong chassnunta]
often ask for

'As for this book$_j$, kids like (it$_j$) very much and old people also ask for (it$_j$).'

In particular, the example (8b) shows that the gap in the first conjunct is a trace but not a *pro*. It is supported by the general fact that in Korean a *pro* is not allowed to appear in the first conjunct of coordinated structures.

Given that the CSC operates in Korean UDCs to require a gap in each conjunct and given that the pronominal *kukes* in a conjunct does not cause a violation of the CSC, as in (8a) and (8b), we can argue that those pronouns are RPs and that they behave in the same way as traces. Thus, this favors the UDC approach to RPs.

In summary, we argue that the pronouns appearing in the gap positions are not *pro*s. Instead, we argue that RPs in the gap position work as audible traces. According to the trace approach, RPs and gaps arise from a single mechanism. This argument is crosslinguistically compatible with (Georgopoulos, 1991) and (Vaillette, 2001) with respect to Palauan and Hebrew. The terms for UDC gaps and non-UDC correspondents in Korean are summarized in the following chart. The UDC elements in the left-hand column all triggers a nonzero SLASH feature while the right-hand column cannot.

(9)

| | UDCs | non-UDCs |
|---|---|---|
| zero | trace | *pro* |
| overt | resumptive prn | (ordinary) prn |
| *caki* | resumptive refl | (ordinary) refl |

## 5 The Analysis of RPs and RR *caki*

Korean UDCs always involve the presence of one of three elements that give rise to a nonlocal SLASH feature: trace, resumptive pronoun, and resumptive reflexive. These three elements have certain properties with respect to the SCO constraint and coordination. Each of them shares certain information with a filler that appears in a possibly distant higher node. Furthermore, they share certain properties in common with their corresponding forms in non-UDCs. The occurrences of the reflexive *caki* are associated with semantic and pragmatic properties of logophoricity and constrastiveness, in contrast with neutral occurrences of pronouns.[3] Thus, we claim that RPs and the RR *caki* in UDCs are respectively the same elements of pronouns and the LD reflexive *caki* in non-UDCs. This approach is reminiscent of (Pollard and Xue, 1998) who pointed out that a distinction between structural and discourse binding should not be treated as lexical ambiguity.

Our UDC approach is different from accounts of Chomsky's minimalist program and GB theory, where all traces are considered to be the same category.[4] Chomsky's binding theory requires that fillers be reconstructed to the trace position before binding conditions are applied. Within this kind of approach, it is hard to capture the fact that RPs and RR *caki* work as traces. The HPSG system makes three different kinds of traces possible and captures the fact that traces, RPs, and the RR *caki* in UDCs belong respectively to the subset of *pro*s, pronouns, the LD reflexive *caki* in non-UDCs. In addition, our trace analysis of resumptive elements casts some doubt on traceless approaches proposed by (Sag, 1997) and (Kim, 1998). According to their traceless analyses, gap information is encoded in the lexical entry of a predicate without involving a structural position for an empty category. However, resumptive elements that trigger the SLASH feature need to appear in syntactic structures. Thus, the existence of audible correspondents of traces supports the traditional HPSG analysis of (Pollard and Sag, 1994), which assumes an empty category in a given syntactic structure. One way that a non-local dependency can be bound off is for a local tree to instantiate the filler-gap schema. In line with (Levine et al., 2001)'s unitary analysis of English parasitic gaps, we argue that the non-local feature specification can be used to account for different kinds of Korean UDCs.

## References


Hee-Rahk Chae. 1998. A comparative analysis of *tough*- and comparative constructions in English and Korean. *Language Research*, 34-1:33–71.

Peter Cole. 1987. Null objects in universal grammar. *Linguistic Inquiry*, 18-4:597–612.

Carol Georgopoulos. 1991. *Syntactic Variables: resumptive pronouns and A' binding in Palauan*. Kluwer, Dordrecht.

Cheng-Teh James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry*, 15-4:531–574.

Thomas Hukari and Robert Levine. 1995. Adjunct extraction. *Journal of Linguistics*, 31:195–226.


[3] According to (Sells, 1987), logophoricity refers to subject of consciousness (SELF), the source of reported speech (SOURCE), and deictic perspective (PIVOT)

[4] Within GB theory, noun phrases are classified by the two binary features, a(naphoric) and p(ronominal), and all traces are assumed to be R-expressions with -a and -p features.

Young-Se Kang. 1986. *Korean Syntax and Universal Grammar*. Ph.D. thesis, Harvard University, Cambridge, MA.

Jong-Bok Kim. 1998. A head-driven and constraint-based analysis of Korean relative clause constructions. *Language Research*, 34.4:1–41.

Robert Levine, Thomas Hukari, and Michael Calcagno. 2001. Parasitic gaps in English: Some overlooked cases and their theoretical implications. In Peter Culicover and Paul Postal, editors, *Parastic Gaps*, pages 181–222. MIT Press, Cambridge, MA.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

Carl Pollard and Ping Xue. 1998. Chinese reflexive *ziji*:syntactic reflexive vs. nonsyntactic reflexives. *Journal of East Asian Linguitics*, 7:287–318.

Paul M. Postal. 1994. Contrasting extraction types. *Journal of Linguistics*, 30:159–186.

Paul M. Postal. 2004. *Skeptical Linguistic Essays*. Oxford University Press, New York, NY.

Ivan Sag. 1997. English relative constructions. *Journal of Linguistics*, 33:431–494.

Peter Sells. 1987. Aspects of logophoricity. *Linguistic Inquiry*, 18:445–479.

Anna Szabolcsi and Marcel den Dikken. 1999. Islands. *GLOT International*, 4-6:3–9.

Nathan Vaillette. 2001. Hebrew relative clauses in HPSG. In D. Flickinger and A. Kathol, editors, *The Proceedings of the 7th International Conference on Head-Driven Phrase Structure Grammar*. CSLI.

# An HPSG approach to the *who*/*whom* puzzle

**Takafumi Maekawa**
Department of Language and Linguistics
University of Essex
Wivenhoe Park
Colchester
CO4 3SQ, UK
tmaeka@essex.ac.uk

## 1 Introduction

Numerous attempts have been made to elucidate properties of the case systems of human language.[1] In the case of the example in (1), most of the previous studies on case marking would argue that the English interrogative/relative pronoun *who* manifests nominative case, while accusative case is realized in the form of *whom*.

(1) a. *Who* saw *whom*?
b. someone *whom* you rely on
c. someone *who* relies on you

In this paper, we will first observe that the distribution of *who* and *whom* poses a challenge to any theory of the above sort. Second, we will discuss the analysis by Lasnik and Sobin (2000). Then we will propose an analysis, in which the interaction of small of number of constraints can accommodate the seemingly complex body of data. Section 5 is the conclusion.

## 2 The *who*/*whom* puzzle

The assumption that *who* is nominative and *whom* is accusative is apparently justified by (2), where *who* in the subject position does not alternate with *whom*, which indicates that *whom* is not nominative.

(2) a. *Who*/**whom* wrote the editorial?
b. the man *who*/**whom* came to dinner

However, *whom* is not the only option in non-subject positions. In (3)–(5) *whom* alternates with *who* as object of a verb or preposition in main clauses (3), embedded clauses (4), and in situ (5).

(3) a. those *whom*/*who* we consulted.
b. someone *whom*/*who* we can rely on
c. He didn't say *whom*/*who* he had invited.

(4) a. *Whom*/*who* did you meet?
b. *Whom*/*who* are you referring to?

(5) a. Who is going to marry *whom*/*who*?
b. Who is buying a gift for *whom*/*who*?

Any theory that assumes that *who* is nominative and *whom* accusative would predict that only *whom* is allowed in (3). One might say that this alternation would be predictable if we assumed that *who* can be accusative as well as nominative. (6) shows, however, that the situation is not so simple.

(6) a. To *whom*/**who* are you referring?
b. someone on *whom*/**who* we can rely

A noun phrase as a prepositional complement in a fronted PP take the accusative form, as in *about him*/**he*. This is unexpected if *who* can be accusative.

Another complication about the *who*/*whom* distinction is that *whom* can appear in the syntactic environments where the nominative case is normally expected.

(7) a. We feed children *who*/*whom* we think are hungry.
b. The man *who*/*whom* I believe has left.

[1] For HPSG literature on case, see Heinz and Matiasek 1994; Meurers 2000; Pollard 1994; Przepiórkowski 1999, etc.

c. The man *who*/*whom* it was believed had left.

Any theory that assumes that *who* is nominative and *whom* accusative would only predict the occurrence of *who* in (7); but in fact *whom* is allowed. The above data poses a challenge to any theory of syntax which deals with the *who*/*whom* distinction in parallel to other English pronominal distinctions, such as *they*/*them*.

It is clear that there is a complex body of data here, but the distribution of *who* and *whom* in the above data can be summarized in the following way.

(8) Distribution of *who* and *whom*

| Environments | forms |
|---|---|
| Embedded clause | *who*/*whom* |
| Main clause | *who*/*whom* |
| In-situ | *who*/*whom* |
| Subject | *who* |
| Pied-piping | *whom* |

This summary reveals that *who* and *whom* alternate in all the syntactic environments except for subject of the nearest following finite verb and object of the fronted preposition (pied-piping).

## 3 Lasnik and Sobin's (2000) approach

A recent attempt to provide a theoretical account to the *who*/*whom* distinction is Lasnik and Sobin's (2000).[2] They argue that *who* is the basic form of the *wh*-pronoun, which can check either nominative (NOM) or accusative (ACC) case. The suffix *-m* of *whom* is assumed to be associated with an additional ACC feature and has to be checked independently of the ACC feature associated with the stem *who*. This additional ACC feature carried by the suffix is checked by the following rules with the status of 'grammatical viruses', which serve to license prestige forms.

(9) The Basic 'whom' Rule
(Lasnik and Sobin 2000: 354)
If: [$_{V/P}$] *who-* *-m*
[ACC] [ACC]
1 2 3
then: check ACC on 3

(10) The Extended 'whom' Rule
(Lasnik and Sobin 2000: 359)
If: *who-* *-m* … *NP*, where
[ACC]
1 2 3
a) 3 is the nearest subject NP to 2, and
b) '…' does not contain a V which has 1–2 (a single word *whom*) as its subject,
then: check ACC on 2.

Rule (9) licenses the occurrence of *whom* as object of a verb or preposition, as in (3f), (3g), (6a) and (6b). Rule (10) accounts for the occurrence of initial *whom* in any type of *wh*-construction where the *wh*-pronoun functions as the object of a verb (3a, b, d) or stranded preposition (3c, e), or the subject of an embedded clause (7). The unacceptable occurrences of *whom* in (2) are ruled out by the fact that they are not compatible with the sequential arrangement of (9) nor (10).

However, their approach involves a questionable assumption: it is not clear whether the *who*/*whom* distinction should be treated as a matter of case. Two different forms of a lexeme should not necessarily be seen as two different cases of the lexeme. If they are not realisations of case, it will not be necessary to assume that the stem *who-* and the affix *-m* have two different cases. Other things being equal, it would be preferable not to have such a counter-intuitive assumption.

There are other problems. First, Lasnik and Sobin's (2000) rules are characterised as extra-grammatical devices within Virus Theory. Other things being equal, a theory without such extra devices is preferable to that with them.

Second, the analysis deals with all instances of left-dislocated *whom* in terms of rule (10). This would lead us to expect that *whom* licensed by (10) has the identical distribution in any syntactic environments. Huddleston and Pullum (2002: 465) point out, however, that '[t]he formal feel of *whom* is most apparent in main clause interrogatives'. The following examples from Huddleston and Pullum (2002: 465) illustrate that relative *whom* is not confined to formal style.

(11) a. Hugh wasn't impressed with this ingratiating barman *whom* Roddy had raked up.

[2] See also Kayne (1984) and Radford (1988). For descriptive work, see Jespersen (1924; 1927), Quirk et al (1985), Huddleston and Pullum (2002), etc.

b. Award-winning journalist Nelson Keece (Gary Busey) is coldly detached from his chosen subject, serial killer Stefan (Arnold Vosloo), *whom* he catches in the act of murder.

A satisfactory analysis of the *who*/*whom* distinction should be able to say something about this fact, but it seems that Lasnik and Sobin's (2000) consideration of register is not fine-grained enough to do it.

## 4 An HPSG analysis

A satisfactory analysis of the *who*/*whom* distinction should be able to ensure that there is a position where only *whom* is available (i.e., (6)), a position where only *who* is available (i.e., (2)), and positions where *who* and *whom* alternate (i.e., (3)–(5) and (7)). It should be noted here that *whom* is perceived as formal in style whereas *who* as less formal or informal, so non-existence of *who* in pied-piping will be able to be attributed to the fact that pied piping is confined to formal style (See below for details). Therefore, the apparent complexity of the data is reducible to the following rather simple generalization.

(12) a. Informal style employs *who* in every syntactic environment.
b. In formal style, *whom* is employed in all syntactic environments except when it is subject of the nearest following verb; in the latter cases, *who* is used, as in (2).

Following Wilcock (1999), we represent register variation in terms of the feature REGISTER, which is appropriate for CONTEXT. The REGISTER feature takes a value of sort *register*, which has two subsorts, *formal* and *informal*. The style difference between *who* and *whom* can be formalized in the following lexical constraints of these word forms (cf. Wilcock 1999).

(13) a. $\begin{bmatrix} \text{LME} \langle \text{who} \rangle \\ \text{REGISTER } \textit{informal} \end{bmatrix} \rightarrow \left[\text{PHON} \langle \text{who} \rangle\right]$

b. $\begin{bmatrix} \text{LME} \langle \text{who} \rangle \\ \text{REGISTER } \textit{formal} \end{bmatrix} \rightarrow / \left[\text{PHON} \langle \text{whom} \rangle\right]$

We further assume that a word has the feature L(EXE)ME, which represents the lexeme which the word form instantiates (Ackerman and Webelhuth 1998). (13a) indicates that the lexeme <who> is phonologically realised as <who> in an informal register, while (13b) states that it is realised as <whom> by default in a formal register.

These constraints determine the distribution of *who* and *whom* observed in (3)–(7). The alternation observed in (3)–(5) and (7) is attributed to the difference in register specified in the above constraints: *who* is employed in an informal style while *whom* is employed in a formal style, no matter what syntactic environment they appear in and what sort of semantic role they have. As has been stated earlier, non-existence of *who* in pied-piping in (6) is due to the fact that formal status of pied-piping conflicts with the [REGISTER *informal*] specification of *who* (See Wilcock 1999; cf. Paolillo 2000).

We saw above that *who* appears in a formal register as well as in an informal register when it is a subject of the nearest following verb, as in (2). This means that (13b) should be overridden by a constraint licensing the occurrence of *who* in a formal register. We propose the following constraint.

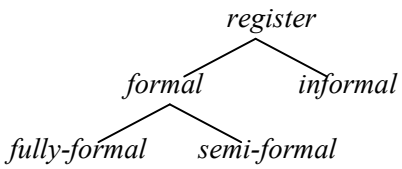
(14)
$$\begin{bmatrix} \text{DOM} \left\langle [1]\begin{bmatrix}\text{LOC } [2] \\ \text{LME } \langle \text{who} \rangle\end{bmatrix}, \begin{bmatrix}\textit{phrase} \\ \text{HEAD verb}[\textit{fin}] \\ \text{SUBJ } \langle [\text{LOC } [2]] \rangle\end{bmatrix}, \cdots \right\rangle \\ \text{REGISTER } \textit{formal} \end{bmatrix} \rightarrow [1]\left[\text{PHON } \langle \text{who} \rangle\right]$$

The value of the attribute DOM(AIN) represents linear order of a clause (see, e.g., Pollard et al. 1993; Reape 1994; and Kathol 2000). (14) involves the assumption that a filler is in the same domain as the following VP. This constraint states that the lexeme <who> takes the phonological form <who> in a formal register if it is subject of the nearest following finite verb phrase. (14) overrides the default constraint (13b), which accounts for ungrammaticality of *whom*, and licences *who* in (2).

We argued above that *whom* is more acceptable in a less formal register in embedded clauses than in main clauses, as illustrated by the not formal but not fully informal examples in (11). To deal with these cases, we need to have a more fine-grained classification of register than that

given earlier. We assume that the type *formal* has two subsorts, *fully-formal* and *semi-formal*. We can formulate the new classification in the following hierarchy.

(15)




There are thus three maximal subtypes for register: *fully-formal*, *semi-formal* and *informal*. We introduce a further constraint in (16).

(16)

$$\begin{bmatrix} ns\text{-}wh\text{-}int\text{-}cl \\ \text{IC} + \\ \text{NON-H-DTR} \left\langle [1] \begin{bmatrix} \text{PHON} \left\langle \text{whom} \right\rangle \end{bmatrix} \right\rangle \end{bmatrix} \rightarrow [1] \begin{bmatrix} \text{REGISTER } \textit{fully-formal} \end{bmatrix}$$

(16) states that the REGISTER value of *whom*, specified as *formal* by constraint (13b), is resolved to *fully-formal* if it is the non-head daughter of independent clauses ([IC +]) of the type *nonsubject-wh-interrogative-clause* (Ginzburg and Sag 2000). This captures the considerably more formal status of *whom* in main clauses. Since constraint (16) does not apply for [IC –] clauses, the REGISTER value of *whom* remains *formal* in embedded clauses. This entails that *whom* can appear in a *semi-formal* as well as *fully-formal* register, which accounts for the occurrence of *whom* in less formal style sentences in (11). This contrast between main clause interrogatives and relatives matches Huddleston and Pullum's (2002) description cited earlier: '[t]he formal feel of *whom* is most apparent in main clause interrogatives'.

## 5 Conclusion

In this paper, we have investigated the *who*/*whom* distinction in English, and provided an alternative analysis to that by Lasnik and Sobin (2000) within HPSG. The register specification via the REGISTER feature and the register hierarchy in (15) allow us a more fine-grained analysis with wider empirical coverage. Another advantage of the present analysis is that constraints are stated in a general formalism available in HPSG, rather than given with extra grammatical devices.

## References


Ackerman, F. and Webelhuth, G. 1998. *A Theory of Predicates*. Stanford: CSLI Publications.

Ginzburg, J., and Sag, I. A. *Interrogative Investigations*. Stanford: CSLI Publications.

Heinz, W. and Matiasek, J. 1994. Argument structure and case assignment in German. In Nerbonne, J., et al (eds.), 199–236.

Huddleston, R. and Pullum, G. K. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Jespersen, O. 1924. *The Philosophy of Grammar*. Chicago: University of Chicago Press.

Jespersen, O. 1927. *A Modern English Grammar on Historical Principles, Part III*. London: Allen & Unwin.

Kathol, K. 2000. *Linear Syntax*. Oxford: Oxford University Press.

Kayne, R. S. 1984. *Connectedness and Binary Branching*. Dordrecht: Foris Publications.

Lasnik, H. and Sobin, N. 2000. The *who*/*whom* puzzle: on the preservation of an archaic feature. *Natural Language and Linguistic Theory* 18, 343–371.

Meurers, W. D. 2000. Raising spirits (and assigning them case). *Groninger Arbeiten zur Germanistischen Linguistik* 43, 173–226.

Nerbonne, J., Netter, K., and Pollard, C. (eds.). 1994. *German in Head-Driven Phrase Structure Grammar*. *Stanford: CSLI Publications*.

Paolillo, J. Formalizing formality: an analysis of register variation in Sinhala. *Journal of Linguitics* 36, 215–259.

Pollard, C. 1994. Toward a unified account of passive in German. In Nerbonne, J., et al (eds.).

Pollard, C., Kasper, R., and Levine, R. 1994. Studies in constituent ordering: Towards a theory of linearization in Head-Driven Phrase Structure Grammar. Research Proposal to the National Science Foundation, Ohio State University.

Przepiórkowski, A. 1999. On case assignment and "adjuncts as complements". In Webelhuth, G., et al (eds.), 231–245.

Quirk, R.,. Greenbaum, S., Leech, G., and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Radford, A. 1988. *Transformational grammar*. Cambridge : Cambridge University Press.

Reape, M. 1994. Domain union and word order variation in German. In Nerbonne, J., et al (eds.), 151–98.

Wilcock, G. 1999. 'Lexicalization of context'. In Webelhuth, G., et al (eds.), 373–387.

Webelhuth, G., Koenig, J-P. and Kathol, A. (eds.). 1999. *Lexical and Constructional Aspects of Linguistic Explanation*. Stanford: CSLI Publications.

# From "hand-written" to computationally implemented HPSG theories

**Nurit Melnik***
Caesarea Rothschild Institute
for Interdisciplinary Applications of Computer Science
Haifa University
Haifa, Israel 31905
nurit@eyron.com

## 1 Overview

HPSG has logical and mathematical foundations which make it amenable to computational implementation. Yet it is seldom the case that this potential is in fact fulfilled, although there exist a number of platforms for implementing HPSG grammars. Thus, most descriptions and analyses of linguistic phenomena in the literature are not substantiated by a working computational grammar.

Two leading implementation platforms are available for implementing HPSG grammars. The Linguistic Knowledge Building (LKB) system (Copestake, 2002) is the primary engineering environment of the LinGo English Resource Grammar (ERG) at Stanford. LKB is developed not particularly for implementing HPSG grammars, but rather, as a framework independent environment for typed feature structures grammar. TRALE, an extension of the Attribute Logic Engine (ALE) system, is a grammar implementation platform that was developed as part of the MiLCA project (Meurers et al., 2002), specifically for the implementation of theoretical HPSG grammars that were not explicitly written for language processing.[1] The two platforms are based on different approaches, distinct in their underlying logics and implementation details.

This paper adopts the perspective of a computational linguist whose goal is to implement an HPSG theory. It is based on the implementation of a "hand-written" grammar proposed by Melnik (2002) to account for verb initial constructions in Modern Hebrew. A representative subset of the grammar, including word order, agreement, and valence alternation phenomena, serves as a test case.

The paper focuses on different dimensions, relevant to HPSG grammar implementation: type definition, grammar principles, lexical rules, exhaustive typing, definite relations, non-binary grammar rules, semantic representation, grammar evaluation, and user-interface. It examines, compares, and evaluates the different means which the two approaches provide for implementation, by referring to examples from a "hand-written" grammar fragment that was implemented in the two systems. The paper concludes that the approaches occupy diametrically opposed positions on two axes: FAITHFULNESS to the "hand-written' theory and COMPUTATIONAL ACCESSIBILITY. The findings of this paper are valuable to linguists who are interested in implementing their grammar, as well as to those who develop implementation platforms.

## 2 Type Definition

Types in a typed feature-structure framework are defined by determining (i) the type's hierarchical relation to other types, (ii) appropriateness conditions, (iii) constraints on the values of embedded features, and (iv) path equations.

TRALE separates the SIGNATURE, where the first two properties are defined, from the THEORY, in which the latter are stated. In the signature file, types are entered in a list format, where subtypes appear

*This research was supported by the Israel Science Foundation (grant no. 136/01) and by The Caesarea Edmond Benjamin de Rothschild Foundation Institute for Interdisciplinary Applications of Computer Science.

[1] See http://milca.sfs.nphil.uni-tuebingen.de/A4/HomePage/English/beschr.html

indented under their respective supertype(s). Features and values are introduced following the type. Constraints on embedded features and path equations are entered separately from the signature in the theory file as implicational constraints in which the type is the antecedent.

LKB, on the other hand, takes a centralized bottom-up approach, where all the information related to a type is defined in one location, in the TYPES file. The definition of each type, then, includes a list of its immediate supertype(s) and introduced features, as well as all other type-related constraints. This approach facilitates the task of defining the type inventory and accessing this information while developing the grammar.

Although the hierarchies are defined differently in the two systems, they are both subject to the glb condition, which requires that the hierarchy be a bounded complete partial order (BCPO). Thus, when a non-BCPO hierarchy is defined, TRALE enforces the condition by producing an error message during compilation. LKB, on the other hand, automatically creates a glb type in each case of violation and restructures the hierarchy accordingly.

On the one hand, by automatically fixing the violation, LKB enables the grammar writer to maintain ignorance regarding a potentially confusing issue. This ignorance, however, turns into confusion once the grammar writer views the type hierarchy diagram. The automatic restructuring of the hierarchy, including the addition of generically named types, may be incomprehensible to the naive grammar writer. Moreover, the resulting hierarchy is reflected only in the display and not in the actual definitions, rendering the automatically created glb types, along with their generic names, inaccessible. A possible solution is to modify the hierarchy definition to reflect the corrected hierarchy, thus allowing the grammar writer to give the glb types more meaningful labels.

Multi-dimensional type hierarchies are widely used in the HPSG literature, yet multi-dimensionality is not a part of the formal type system itself (Penn and Hoetmer, 2003). Neither LKB nor TRALE provide the grammar writer with a way to define partitions (or dimensions) in the hierarchy. Consequently, if partition labels are implemented as types in the hierarchy, they are not distinguished formally from other types, nor do LKB and TRALE prevent the grammar writer from defining types that inherit from two subtypes under one pseudo-partition. Moreover, a multi-dimensional inheritance hierarchy in which partitions are defined as types does not respect the glb condition, and is therefore subjected to the systems' distinct treatments, described above. Although this omission does not prevent grammar writers from implementing their grammars, the result clearly does not reflect the source and the intention of the grammar writer.

## 3 Principles

Principles in HPSG are often defined as implicational constraints. Thus, for example, the Head Feature Principle (HFP), which states that the value of the HEAD feature of the headed-phrase is structure-shared with that of its head-daughter, is defined as a type constraint on the *hd-ph* type.

$$\textit{hd-ph} \rightarrow \begin{bmatrix} \text{HEAD } \boxed{1} \\ \text{HD-DTR} \begin{bmatrix} \text{HEAD } \boxed{1} \end{bmatrix} \end{bmatrix}$$

In LKB principles are necessarily linked to types and are stated as part of the type definition. Thus, the HFP is implemented as part of the definition of the type *hd-ph*. In TRALE, on the other hand, principles such as the HFP are stated as part of the theory, in the form of implicational constraints where the type is the antecedent, similarly to the definition above. TRALE, however, extends implicational constraints to express principles which do not target a particular type. More specifically, the antecedent of implicational constraints can be arbitrary function-free, inequation-free feature structures .

Consider, for example, the following complex-antecedent principle (Meurers, 2001).

$$\begin{bmatrix} \textit{word} \\ \text{SYNSEM} \mid \text{LOC} \mid \text{CAT} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \textit{verb} \\ \text{VFORM } \textit{finite} \end{bmatrix} \\ \text{VAL} \mid \text{SUBJ} \left\langle \text{LOC} \mid \text{CAT} \mid \text{HEAD } \textit{noun} \right\rangle \end{bmatrix} \end{bmatrix}$$
$$\rightarrow \left[ \text{SYNSEM} \mid \text{LOC} \mid \ldots \mid \text{SUBJ} \left\langle \left[ \text{LOC} \mid \text{CAT} \mid \text{HEAD} \mid \text{CASE } \textit{nominative} \right] \right\rangle \right]$$

The principle expresses the generalization that NP subjects of finite verbs are assigned nominative

case. The complex antecedent singles out the relevant class of verbs without requiring there to be a corresponding type.

The ability to use implicational constraints with complex antecedents provides the grammar writer with additional means to express generalizations. When the given dimensions in the type hierarchy do not group together a particular set of objects to which a certain generalization applies, the grammar writer can choose not to expand the hierarchy, but rather to use a complex feature structure as an antecedent to an implicational constraint expressing the generalization. This solution can cut down on the size of the type hierarchy and its complexity.

## 4 Lexical Rules

The main issue that is pertinent to the implementation of lexical rules (LRs) is the "carrying over" of information from the input to the output of the rule. The descriptions of the input and output of lexical rules generally include only the features and values that are relevant for the particular rule; either those which constrain the types of objects on which to apply the rule or those which provide "information handles" (Meurers, 1994). All information which is not changed by the lexical rule is assumed to be copied over from the input to the output. An implementation platform thus has to implement the explicit as well as implicit copying of values.

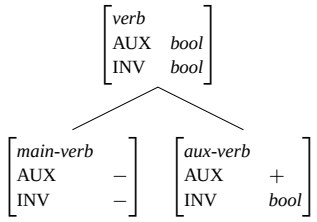
LKB views lexical rules as unary grammar rules which relate a mother structure (the output) to its daughter (the input). Similarly to grammar rules, the description of the daughter is included in the ARGS feature of the mother. This provides a partial solution to the "carrying over" problem — the descriptions of both the mother and daughter are a part of a single feature structure. Nevertheless, the grammar writer is required to explicitly specify by structure-sharing the information that is copied over. Aside from deviating from HPSG conventions, this solution may result in a loss of generality.

TRALE provides two mechanisms for implementing lexical rules: the traditional ALE mechanism and a mechanism referred to as 'description-level lexical rules' (DLRs) which encodes the treatment proposed in Meurers & Minnen (1997). Unlike the format of the rules in LKB, the TRALE syntax for both types of LRs is similar to the familiar 'X $\Rightarrow$ Y' notation. More importantly, from the perspective of the grammar writer, the main distinction between the two approaches is in the "carrying over" mechanism. ALE LRs, similarly to the LKB mechanism, require explicit specification of "carried over" information. The DLR version provides an automatic "carrying over" mechanism which implements the intuitions behind the "hand-written" version of lexical rules. This is a clear advantage in terms of approximating written theories and maintaining generality.

## 5 Exhaustive Typing and Subtype Covering

'Exhaustive typing' refers to a particular interpretation of the signature according to which subtypes exhaustively cover their supertypes. Consequently, if an object is of a certain non-maximal type *t* then it is also of some more specific subtype subsumed by *t*.[2]

A simple example is the HPSG analysis of subject-auxiliary inversion in English. In order to restrict the licensing of inversion to auxiliary verbs, verbs are defined as having two features: INV and AUX. Furthermore, the general type *verb* is assumed to have two subtypes: *main-verb* and *aux-verb*.




Under an exhaustive typing interpretation, objects of type *verb* which are not compatible with either *main-verb* or *aux-verb* (e.g., verbs specified with $[\text{AUX} -]$ and $[\text{INV} +]$) are rejected. This is the interpretation which TRALE employs. In LKB such feature structures are accepted.

In addition, TRALE employs a subtype covering strategy whereby if the system recognizes that the values of a feature structure of a non-maximal type

[2]This interpretation is also referred to in the literature as 'closed world'. However, as one reviewer pointed out, the terms 'closed/open world' have a different meaning in the study of programming languages and should therefore be avoided.

are consistent with the values of only one of its subtypes, it will promote those values to the values of the compatible subtype. This is justified only under an exhaustive typing interpretation, and is therefore not a part of the LKB system.

One advantage to TRALE's approach is that it implements an implicit assumption in "standard" HPSG (e.g., (Pollard and Sag, 1994)) and is thus appropriate if the goal is to narrow the gap between "hand-written" theories and their implemented counterparts. Second, Meurers (1994) notes that "while both interpretations allow the inference that appropriateness information present on a type gets inherited to its subtypes, we can now additionally infer the appropriateness specifications on a type from the information present on its subtypes". Moreover, in addition to increasing the expressive power, such a system facilitates syntactic detection of errors and increased efficiency in processing (Meurers, 1994).

The main reasons that are given for adopting the alternative approach, often referred to as 'open-world reasoning', are not theoretical, but rather, motivated by engineering considerations. This type of reasoning allows the grammar writer to be non-committal regarding the complete inventory of types needed to account for the language. This is particularly helpful during incremental grammar/lexicon development.

## 6 Definite Relations

"Hand-written" HPSG makes use of various relations which are external to the description language, many of which apply to lists and sets. One such relation is APPEND. LKB and TRALE differ greatly in the solutions that they offer for implementing "hand-written" analyses which make use of definite relations. LKB takes a conservative stance and adheres to the description language, while TRALE augments the description language with a programming language for implementing definite relations and incorporating them into type constraints and rules.

Programming definite relations in the TRALE environment is very similar to programming in Prolog, with the exception that first-order terms in Prolog are replaced with descriptions of feature structures. Thus, a list in this case is not a list of terms, but rather a list of descriptions of feature structures.

A thorough discussion of the benefits of adding recursive relations to the description language of implementation platforms for HPSG grammars is found in Meurers *et al.* (2003), which compares the treatment of unbounded dependencies and optional arguments in the ERG, implemented in LKB, with that of TRALE. They conclude that the ability to express relational goals increases the grammar's modularity and its ability to express generalizations, and reduces the gap between "hand-written" theories and their implemented counterparts. This conclusion is echoed in the following section.

## 7 Non-binary Grammar Rules

Grammar rules in the HPSG literature are not restricted to binary rules. A prime example is the head-complement phrase, one of the most basic phrase structures in the grammar. In addition to being non-binary, the head-complement phrase rule is designed to account for phrases with a varying number of daughters. Implementing a rule for such a phrase type poses a number of challenges for a computational system, challenges which are handled differently by the two systems.

The assumption in LKB is that the number of daughters associated with each rule is fixed. Thus, for grammars which are not restricted to binary branching trees the grammar writer needs to define phrase types and grammar rules for each arity. TRALE provides a special `cats>` operator to express rules with daughters lists of unspecified length. This, combined with the ability to incorporate definite recursive relations into the grammar provides the grammar writer with a way to implement non-binary grammar rules, such as the head-complement rule, in a concise and elegant manner, which closely approximates "hand-written" grammars. This, however, does require from the grammar writer the programming skills needed to be able to code using the definite logic programming language.

## 8 Semantic Representation

LKB contains a module for processing Minimal Recursive Semantics (MRS) representations. The module is independent from the rest of the LKB and provides tools for manipulating MRS structures

in feature structure representations (Copestake and Flickinger, 2000). TRALE provides an alternative module which is an implementation of Lexical Resource Semantics (Penn and Richter, 2004). A comparison and evaluation of the two systems will be given in the full paper.

## 9 Evaluating Competence and Performance

Implemented grammars can be evaluated according to two dimensions: competence and performance. The competence of a grammar refers to its coverage and accuracy, that is the ability to account for all and nothing but sentences which are assumed to be grammatical. Performance relates to the resources — mainly processor time and memory space — that are used during processing.

Both LKB and TRALE provide a way for defining a test suite which can be used as a benchmarking facility. A batch parse returns for each sentence in the test suite the number of parses and passive edges. In terms of performance, TRALE indicates for each sentence the CPU time in seconds that it took to process the sentence. In LKB only a total figure for all sentences is given. More sophisticated tools for evaluating competence and performance of grammars are available in both systems through the `[incr tsdb()]` package (Oepen, 2001).

## 10 User-Interface Issues and Features

Aside from major design differences between the two systems, LKB and TRALE are distinguished by other more superficial user-interface type of differences.

- LKB provides an interactive display of the grammar's type hierarchy. The user can click on types and examine their immediate and expanded definitions. TRALE produces static images of the hierarchy.
- Both systems provide ways for displaying and inspecting feature structures and syntactic trees. TRALE's Grisu graphical interface displays feature structures in AVMs that are identical to those of "hand-written" HPSG. The LKB display is less compact and more difficult to navigate.
- Parametric macros in TRALE are used as a shorthand for descriptions that are used frequently. Macros are especially useful for defining the lexicon when it is structured to minimize lexeme-specific information.
- LKB is a graphic-user-interface system where commands are invoked through drop-down menus. In TRALE the user interacts with the program by using commands entered at the Prolog prompt.
- LKB uses the same syntax to define types, lexical rules, grammar rules, and words in the lexicon. In TRALE distinct formats, similar to "hand-written" HPSG, are used for each of the grammar components.
- LKB comes with the Matrix (Bender et al., 2002), an open-source starter-kit for rapid prototyping of precision broad-coverage grammars. TRALE grammars need to be implemented from scratch, or based on existing grammars.

## 11 Conclusion

Generally speaking, the characterization of HPSG as an implementable grammatical theory is justified, due to the computational effort that was put into designing and developing the two implementation platforms discussed in this paper. The major gap that was identified between "hand-written" HPSG and its implemented counterpart was in the multi-dimensional inheritance mechanism, which is not incorporated into neither implementation platforms.

LKB and TRALE can be compared and evaluated along two different axes: FAITHFULNESS and ACCESSIBILITY. Faithfulness is the extent to which the implemented grammar resembles the original "hand-written" one. Accessibility, on the other hand, is the degree of computational skills that is required from a linguist in order to implement a grammar.

In some way, LKB can be viewed as a simplified TRALE. Thus, when implicational constraints with complex antecedents, DLR lexical rules, the `cats>` operator, definite clauses, and macros are eliminated, one can implement an LKB-like grammar in TRALE. Of course, one LKB feature that cannot be assimilated is the automatic correction of glb condition violations.

The gap between the LKB-like TRALE grammar and a grammar implemented using the entire collection of tools provided by TRALE character-

izes the differences between the systems. The 'true' TRALE grammar is positioned much higher on the faithfulness axis than the LKB-like TRALE grammar. The TRALE tools needed in order to elevate the LKB-like grammar on this axis require from the linguist more computational skills. This is especially true when writing (and debugging) Prolog definite clauses to express relational constraints.

In terms of accessibility, the menu-driven user interface of LKB is more user-friendly than TRALE's command-line interface, making LKB more attractive to the less computationally savvy linguist. However, tipping the balance a little on the accessibility scale towards TRALE is its AVM display, which is much easier to process than LKB's. Consequently, a computational linguist interested in implementing an HPSG theory must consider these dimensions when choosing an implementation platform.

## References


Emily M. Bender, Daniel P. Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Ann Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI publications, Stanford, CA.

Nurit Melnik. 2002. *Verb-Initial Constructions in Modern Hebrew*. Ph.D. thesis, University of California at Berkeley.

Detmar Meurers and Guido Minnen. 1997. A computational treatment of lexical rules in HPSG as covariation in lexical entries. *Computational Linguistics*, 23(4):543–568.

W. Detmar Meurers, Gerald Penn, and Frank Richter. 2002. A web-based instructional platform for constraint-based grammar formalisms and parsing. In Dragomir Radev and Chris Brew, editors, *Effective Tools and Methodologies for Teaching NLP and CL*, pages 18 – 25, New Brunswick, NJ. The Association for Computational Linguistics.

Detmar Meurers, Kordula De Kuthy, and Vanessa Metcalf. 2003. Modularity of grammatical constraints in HPSG-based grammar implementations. In *Proceedings of the ESSLLI '03 workshop "Ideas and Strategies for Multilingual Grammar Development"*, Vienna, Austria.

Detmar Meurers. 1994. On implementing an HPSG theory – Aspects of the logical architecture, the formalization, and the implementation of head-driven phrase structure grammars. In Erhard W. Hinrichs, Detmar Meurers, and Tsuneko Nakazawa, editors, *Partial-VP and Split-NP Topicalization in German – An HPSG Analysis and its Implementation*, pages 47–155. Eberhard-Karls-Universität Tübingen, Tübingen, Germany.

Detmar Meurers. 2001. On expressing lexical generalizations in HPSG. *Nordic Journal of Linguistics*, 24(2):161–217. Special issue on 'The Lexicon in Linguistic Theory'.

Stephan Oepen. 2001. [incr tsdb()] — competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany. in preparation.

Gerald Penn and Kenneth Hoetmer. 2003. In search of epistemic primitives in the English Resource Grammar. In *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar*, East Lansing, Michigan.

Gerald Penn and Frank Richter. 2004. Lexical resource semantics: From theory to implementation. In Stefan Müller, editor, *Proceedings of the HPSG-2004 Conference, Center for Computational Linguistics, Katholieke Universiteit Leuven*, pages 423–443. CSLI Publications, Stanford.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. CSLI Publications and University of Chicago Press.

# Phrasal or Lexical Resultative Constructions?

**Stefan Müller**
Theoretische Linguistik/Computerlinguistik
Universitt Bremen/Fachbereich 10
Postfach 33 04 40
D-28334 Bremen

Stefan.Mueller@cl.uni-bremen.de

## Abstract

Goldberg and Jackendoff (2004) argue that a phrase structure rule like (1a) is not sufficient to account for resultative constructions like (1b) and suggest a family of related constructions that make explicit the idiosyncratic properties of resultative secondary predications.

(1) a. VP → V NP AP/PP
b. They drank the pub dry.

In my talk I show that the data discussed by Goldberg and Jackendoff can be captured in a lexicon-oriented analysis and that syntactic structures may be even more abstract than (1a).

Furthermore, I show that it is difficult or impossible to capture the interaction between other phenomena and resultative constructions in an insightful way and that it is not obvious how a phrasal approach can capture cross-linguistic generalizations.

The paper will be published elsewhere and therefore cannot be included here. If you are interested in the paper, please contact the author.

## References

Adele E. Goldberg and Ray S. Jackendoff. 2004. The English resultative as a family of constructions. *Language*, 80(3):532–568.

Adele E. Goldberg. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture. University of Chicago Press, Chicago/London.

# Adverbial Extraction: A defense of tracelessness

**Ivan A. Sag**
Department of Linguistics
Stanford University
Stanford, 94305
sag@stanford.edu

## 1 Introduction

Ever since the 'adverbs-as-complements' [A-as-C] analysis was first proposed by Bouma and van Noord, it has been controversial. Even though a number of people have offered extensive motivation for this view in various languages (e.g. Przepiórkowski 1999, Manning et al. 1999), there are various issues of adequacy that have been raised by researchers in the HPSG research community. Specifically, in a penetrating study, Levine (2002) raises important questions about how the A-as-C analysis developed by Bouma et al. (2001) [BMS] can be reconciled with examples like (1):

(1) In how many seconds flat did Robin find a chair, sit down, and whip off her logging boots?

Because in BMS's analysis, an adverb selected by a verb identifies its MOD value's KEY value with the verb's KEY value, (1) poses a dilemma: if the extracted adverb is associated with a dependent of each verb (*find*, *sit*, and *whip*), then three contradictory KEY values must be equated. Intuitively, (1) requires that the adverb modify the coordinate structure (since this sentence has a cumulative reading and its meaning is a question about the duration of a tripartite event), yet BMS's analysis assumes that all postverbal adverbials are complements, and hence lacks any way to associate the adverb with the appropriate modifier position and no way to assign it the correct scope. On the other hand, Levine argues, if there are adverbial traces that can appear wherever adverbs can appear, then these examples are unproblematic – the adverbial trace is in a position adjoined to the coordinate structure, and hence outscopes the conjunction.

In this paper, I explore a small modification of the BMS analysis that resolves this problem without introducing traces of the sort that Levine argues would provide an alternative account of data like (1). The goal of saving the BMS analysis is worthwhile, it should be noted, as it is the only extant HPSG extraction analysis which is immediately consistent with the extensive evidence cited for A-as-C and which also provides a straightforward account of the fact, documented extensively by Hukari and Levine (1995), that adverb extraction triggers the same morphosyntactic repercussions as complement extraction in languages (e.g. Chamorro, Palauan, Thompson River Salish, Irish,...) that register extraction dependencies locally. Under BMS's proposal, all verbs and complementizers within an extraction domain are distinguished by having a nonempty SLASH value. Under an approach where adverbial traces terminate filler-gap dependencies, there is no motivation for extraction information (a nonempty SLASH value) to be registered on the verb modified by an adverbial trace.

## 2 Analysis

In unpublished work, Bouma et al. (1998) observe that the BMS analysis requires a stipulation based on a binary relation they call `successively-out-modify` in order to ensure that the linear order of postverbal modifiers determines their relative scope:

(2) a. Robin reboots the Mac [frequently] [intentionally]. **intnl(freq(reboot..))**

b. Robin reboots the Mac [intentionally] [frequently]. **freq(intnl(reboot..))**

This stipulation can be eliminated by returning to a lexical-rule (LR)-based analysis like that originally proposed by van Noord and Bouma(see also Przepiórkowski 1999). For convenience, I will formulate this as lexical rule as a unary schema that simply extends a verb's ARG-ST list, i.e. as in (3), where the daughter is the 'LR input' and the mother is the 'LR output':[1]

(3) Adverb Addition Schema (AAS):
Mother:

$$\begin{bmatrix}\text{PHON } \boxed{C} \\ \text{SS|LOC|CONT } \begin{bmatrix}\text{LTOP } \boxed{4} \\ \text{HCONS } \boxed{B} \oplus \langle \boxed{1} \leq \boxed{2} \rangle \\ \text{RELS } \boxed{D}\end{bmatrix} \\ \text{ARG-ST } \boxed{A} \oplus \left\langle \begin{bmatrix}\text{LTOP } \boxed{4} \\ \text{MOD } \begin{bmatrix}\text{LOC } \begin{bmatrix}\text{CAT|HEAD } \textit{verb} \\ \text{CONT|LTOP } \boxed{2}\end{bmatrix}\end{bmatrix}\end{bmatrix} \right\rangle\end{bmatrix}$$

Dtr:

$$\begin{bmatrix}\text{PHON } \boxed{C} \\ \text{SS|LOC } \begin{bmatrix}\text{CAT|HEAD } \textit{verb} \\ \text{CONT } \begin{bmatrix}\text{LTOP } \boxed{1} \\ \text{HCONS } \boxed{B} \\ \text{RELS } \boxed{D}\end{bmatrix}\end{bmatrix} \\ \text{ARG-ST } \boxed{A}\end{bmatrix}$$

The AAS requires that the local top (4) of the selected adverb is also the verb's local top. It also ensures that the local top (1) of the daughter verb is less than or equal to the adverb's MOD value's local top (2). This means that when a verb combines with a scopal adverbial complement, the verb's predication will always be within the scope of that adverbial, as shown in (4). In addition, selected adverbials must be able to modify verbal expressions (hence the [HEAD *verb*] specification in the adverbial's MOD value[2]):

(4)
$$\begin{bmatrix}\text{PHON } \langle \text{ found } \rangle \\ \text{SS|LOC|CONT } \begin{bmatrix}\text{LTOP } h_4 \\ \text{RELS } \left\langle \begin{bmatrix}\textit{find-rel} \\ \text{LBL } h_1 \\ \text{ARG1 } i \\ \text{ARG2 } j\end{bmatrix} \right\rangle \\ \text{HCONS } \langle\, h_4 \leq h_2\,,\, h_4 =_q h_1 \,\rangle\end{bmatrix} \\ \text{ARG-ST } \left\langle \text{NP}_i, \text{NP}_j, \begin{bmatrix}\text{LTOP } h_4 \\ \text{MOD } \begin{bmatrix}\text{...HEAD } \textit{verb} \\ \text{...LTOP } h_2\end{bmatrix}\end{bmatrix} \right\rangle\end{bmatrix}$$

Here the selected adverb, if scopal, will have to include the verb's local top, and hence the verb's predication, within its scope. The use of $\leq$, rather than $=_q$, is crucial to my analysis.

Notice that the mother in (3) (the 'LR output') says nothing about the KEY value of the verb or that of the MOD value. In addition, when a verb selects two adverbials, the first adverbial's local top enters into an $\leq$ relation with the local top of the second adverbial's MOD value. This ensures that subsequent scopal adverbials will always outscope prior adverbials (and that all such adverbials will include the verb's predication in their scope).

The only two resolved mrs-s that satisfy the constraints imposed by (4) for an example like (5a) are shown in (5b,c):

(5) a. Kim found a chair in 30 seconds.

b. $\begin{bmatrix}\text{LTOP } h_0 \\ \text{RELS } \langle\, h_1\textit{:found(k,y)},\ h_2\textit{:a(y,}h_3\textit{,}h_1\textit{)}, \\ \qquad h_3\textit{:chair(y)},\ h_0\textit{:in-30-secs(}h_2\textit{)} \,\rangle\end{bmatrix}$
**in-30-secs(a (y, chair(y), found(k,y)))**

c. $\begin{bmatrix}\text{LTOP } h_0 \\ \text{RELS } \langle\, h_4\textit{:found(k,y)}\,,\ h_0\textit{:a(y,}h_3\textit{,}h_1\textit{)}, \\ \qquad h_3\textit{:chair(y)}\,,\ h_1\textit{:in-30-secs(}h_4\textit{)} \,\rangle\end{bmatrix}$
**a (y, chair(y), in-30-secs( found(k,y)))**

The handle ($h_0$) of the quantifer *a* is within the preposition's scope in (5b), but outside it in (5c).

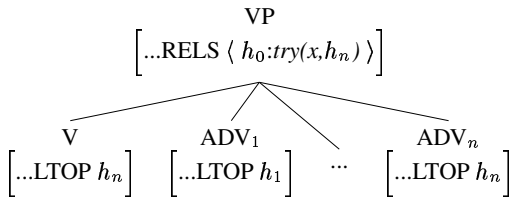
It is important to understand that the adverbial complement's scope remains 'clause-bounded' under this proposal. A verb like *believe* or *try* selects a verbal phrase as complement and lexically identifies the local top of the relevant complement with the appropriate semantic argument (the second argument

[1] I am aware that by eliminating DEPS, I raise controversial issues about the role of binding theory in the treatment of Principle C effects, but these are orthogonal to the matters at hand. I follow Copestake et al.'s presentation of MRS throughout. In particular, lexical constraints are assumed to ensure that the local top (a handle) of a verb or a scopal adverb is equal to that of its predication, modulo quantifiers ($=_q$).

[2] Note that no further LOC, CAT, SUBCAT or HEAD identity is enforced.

of *believe-rel* or *try-rel*). Since a VP's local top will be identified with that of the rightmost adverbial in an example like (6), all of the adverbs must be within the scope of the embedding handle-embedding relation:

(6)

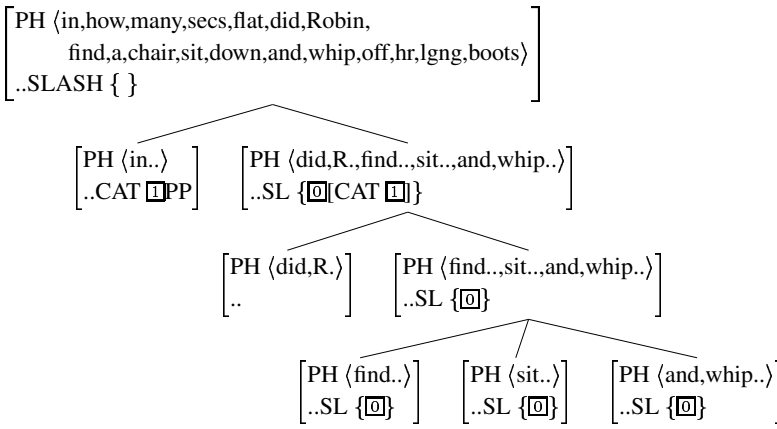



In short, this proposal entails that the scope interactions of selected scopal adverbials parallel that of true modifiers, but in the opposite order. (see Copestake et al. (to appear) discussion of *Kim apparently almost succeeded*. (has only an **apparently(almost(succeeded(k)))** reading).

## 3 Extracted Adverbials Scope over Conjunctions

In head-filler constructions of all sorts, it is reasonable to assume that the filler daughter's CAT and INDEX values are identified with those of the head daughter's SLASH member.[3] Now reconsider Levine's example in (1) above. In this case, the CAT and INDEX values of the adverbial filler (the PP *in how many minutes flat*) will be identified with those of the SLASH member, which will in turn (via standard HPSG principles governing the inheritance of SLASH specifications) be identified with the SLASH members of the selected adverbials, as sketched in (7):

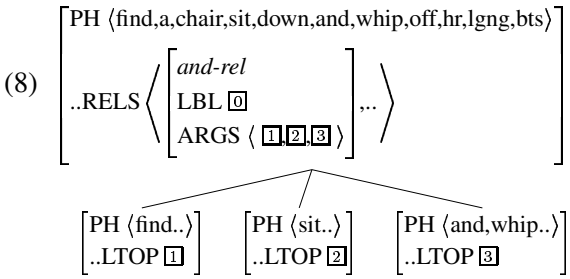
(7)




The SLASH values also make their way down to the verbs *find*, *sit*, and *whip*, where they are 'amalgamated' from the selected adverbial, as in the BMS analysis. Making normal assumptions about gaps, the CAT value of each selected adverbial is identified with the CAT value of its SLASH value. Since MOD is within CAT, it follows that the filler's MOD value must outscope each verbal predication.

Following Copestake et al. (to appear), I assume that conjunctions embed the local tops of the conjuncts as their arguments, roughly as in (8):

(8)




Since each conjunct's local top is embedded as an argument of the conjunction, the only way the filler adverbial can simultaneously outscope *find-rel*, *sit-rel*, and *whip-rel*) is for that adverbial to outscope the *and-rel* (since, given the nature of MRS, the adverbial's relation can only appear once in a resolved *mrs* structure). The correct result thus results from the resource-sensitive nature of MRS. Assuming a variant of *and-rel* that provides the appropriate cumulative event interpretation discussed by Levine, his example (1) is properly analyzed, as sketched in (9):

[3] Given MRS, it would be an unwanted complication to identify the entire CONT value of filler and the gap in a UDC. Identifying the LTOP of the filler daughter with that of the SLASH value would also impose unwanted scope restrictions when the filler is scopal.

(9) $\left[\begin{array}{l}\text{LTOP } h_0 \\ \text{RELS } \langle\ h_0\text{:}\textit{how-many(x,}h_1\textit{,}h_2\textit{)},\ h_1\text{:}\textit{second(x)}, \\ \quad h_2\text{:}\textit{in(}h_3\textit{,x)},\ \text{h}_3\text{:}\textit{and}(\text{h}_4,\text{h}_5,\text{h}_6), \\ \quad h_4\text{:}\textit{a(y,}h_7\textit{,}h_8\textit{)},\ h_7\text{:}\textit{chair(y)}, \\ \quad h_8\text{:}\textit{found(k,y)}\ ,\ h_5\text{:}\textit{sit-down(k)}, \\ \quad h_6\text{:}\textit{whip-off-h-l-boots(k)}\ \rangle\end{array}\right]$

Note that the use of $\leq$, rather than $=_q$ (as in Copestake et al. to appear), is crucial, as this is what allows the *and-rel* to 'slip in' to the resolved *mrs* structure.

## 4 Further Issues

The question remains of how to deal with other examples involving adverbs that follow a coordinate-structure, e.g. (10)[from Levine 2002]. Exactly the same analysis developed above extends to these examples if they are analyzed in terms of a rightward extraction scheme of the sort that would also treat examples like (11a), where a left-adjoined (true) modifier is within the scope of the right adjoined PP modifier:

(10) Robin [found a chair, sat down, and whipped off her logging boots] [in twenty seconds flat].

(11) a. Sandy [[rarely visited a friend] because of illness].

b. Sandy [rarely [visited a friend because of illness]].

The **because(rarely...)** reading associated with (11a) involves rightward extraction of the *because*-phrase. This should be contrasted with the **rarely(because...)** reading associated with (11b), where the *because*-phrase is directly realized as a complement of *visited*, with *rarely* modifying the resultant VP.

## 5 Conclusion

The traceless adverb-as-complement analysis is alive and well. It gives a principled answer to the important questions raised by Levine about the interaction of adverbial extraction and cumulative conjunction, while at the same time providing a coherent, unified approach for systematizing the massive evidence for the A-as-C approach provided by van Noord and Bouma (1994), Przepiórkowski 1999, Manning et al., and others. Although I have modified the BMS analysis in three ways, by eliminating DEPS, returning to van Noord and Bouma's lexical rule analysis of adverb addition, and introducing $\leq$ constraints, I preserve the elegant account that BMS provide of the Hukari/Levine (1995) observation that adverb and complement extraction are both morphosyntactically registered in all languages that locally register extraction dependencies. No analysis with '*wh*-traces' has achieved a comparable result.

**References**


Bouma, Gosse, Robert Malouf, and Ivan A. Sag. 1998. Adjunct Scope and Complex Predicates. Paper presented at the 20th Annual Meeting of the DGfS. Section 8: The Syntax of Adverbials – Theoretical and Cross-linguistic Aspects. Halle, Germany.

Bouma, Gosse, Robert Malouf, and Ivan A. Sag. 2001. Satisfying Constraints on Extraction and Adjunction. *Natural Language and Linguistic Theory*. 19.1: 1–65.

van Noord, Gertjan, and Gosse Bouma. 1994. The Scope of Adjuncts and the Processing of Lexical Rules. *Proceedings of Coling: Kyoto, Japan*.

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan A. Sag. To appear. Minimal Recursion Semantics: an Introduction. To appear in *Research on Language and Computation*.

Hukari, Thomas E., and Robert D. Levine. 1995. Adjunct Extraction. *Journal of Linguistics* 31: 195–226.

Levine, Robert D. 2002. Adjunct valents: Cumulative scoping adverbial constructions and impossible descriptions. In J. Kim and S. Wechsler, eds., *Proceedings of the 9th International Conference on Head-Driven Phrase Structure Grammar, Kyung-Hee University, Seoul*. Stanford: CSLI Publications. http://cslipublications.stanford.edu/HPSG/3/. Pp. 209–232.

Manning, Christopher, Ivan A. Sag, and Masayo Iida. 1999. The Lexical Integrity of Japanese Causatives. In R. D. Levine and G. Green, eds., *Readings in Modern Phrase Structure Grammar*. Cambridge University Press. Pp. 39–79.

Przepiórkowski, Adam. 1999. *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Doctoral dissertation, University of Tübingen.

# Selectional Restrictions in HPSG: I'll eat my hat!

**Jan-Philipp Soehn**
Department of German Linguistics, University of Jena
Fürstengraben 28, D-07743 Jena, Germany
jp.soehn@uni-jena.de

## 1 Introduction

The phenomenon of selectional restrictions, first described by Chomsky (1965, pp. 114ff), is part of almost every introduction to linguistics. A violation of selectional restrictions is the explanation for the oddity of the following examples:[1]

(1) !*Kim ate a motor-bike.*

(2) !*There is an apple bathing in the water.*

The verb *eat* requires an *edible* object and the action of *bathing* can be fulfilled only by an *animate* actor. Even though the view about the role of selectional restrictions is rather diversified, there is general agreement about the central point of compatibility between verbs and their arguments.[2]

Implemented in a natural language processing system, selectional restrictions help with parsing, word-sense disambiguation and the resolving of anaphora. The word *star* in the sentence "*The astrologer married a star*" is ambiguous between "famous person" and "celestial body". However, the example can be disambiguated because we know that the object of *marry* must be *human.*

A characteristic of selectional restrictions is that they are language-specific. This can be illustrated by the verbs *drive* and *ride* and their German counterparts *fahren* and *reiten*. Consider the following data:[3]

(3) a1) *Kim drives a truck/car/* !*motor-bike/* !*bike/* !*horse*
a2) *Kim rides a* !*truck/* !*car/motor-bike/ bike/horse*
b1) *Ute fährt ein(en) Lastwagen/Auto/ Motorrad/Fahrrad/* !*Pferd*
b2) *Ute reitet ein(en)* !*Lastwagen/* !*Auto/* !*Motorrad/* !*Fahrrad/Pferd*

Whereas in English *drive* means a locomotion by operating a motorized vehicle having more than three wheels, the German *fahren* is not sensitive to the number of wheels of the vehicle. The English word *ride* denotes a locomotion while sitting on a saddle or seat like on a horse, the German counterpart *reiten* can be said only for riding on the back of an animal. Thus, selectional restrictions are part of language-dependent lexical information.

Does violation of selectional restrictions always result in an ungrammatical utterance? The answer is no. In metonymic, metaphoric or idiomatic utterances, selectional restrictions may be violated:

(4) *She puts the wine on the table, right next to the glasses.*

A metonymy can be found in example (4), for the object of *put* is the container (e. g. a bottle), rather than the substance.
As a *book* is not *edible*, violating the selectional restriction of *devour*, we understand (5) as being metaphoric:

(5) *He devoured the book in one single night.*

Thus, the violation of selectional restrictions allow us to recognize a nonliteral meaning.

Information from selectional restrictions mark sentences as odd only if one has in mind the lexical meaning of the words and a "normal" context of utterance. This means that there is nothing inherently wrong with a sentence such as (1), because the reader only has to imagine a suitable context (e. g. eating chocolate motor-bikes). In addition, there are certain contextual features that render expressions like *ate a motor-bike* perfectly grammatical. These "repairing contexts" (cf. Chomsky, 1965, p. 158 and Androutsopoulos and Dale, 2000, p. 1) neutralize violations of selectional restrictions and the sentence is fully interpretable:

(6) a) !*Kim ate a motor-bike.*

[1] A superscript exclamation mark indicates a violation of selectional restrictions.

[2] Selectional restrictions play a role with adjectives and nouns, too. In this contribution we will confine ourselves with the discussion of verbs.

[3] The German examples are a nearly word-by-word translation, therefore they are not glossed.

b) *Kim did not eat a motor-bike.*

c) *One cannot eat motor-bikes.*

d) *Kim tries to eat a motor-bike./Kim believes/dreames that she can eat motor-bikes.*

e) *I'll eat my hat if Kim ate a motor-bike.*

f) *Did Kim really eat a motor-bike?*

The repairing contexts are negation (6 b), modals and negation (c), epistemic verbs as *believe*, *try*, etc. whose arguments introduce a state-of-affairs in a possible – not the actual – world (d), conditionals (e) and questions (f).[4] Thus, a violation of selectional restrictions is highly context sensitive. Therefore, Androutsopoulos and Dale argue that selectional restrictions are a pragmatic phenomenon.

To sum up, we have so far seen that, on the one hand, selectional restrictions are part of the lexical information. On the other hand, a violation of selectional restrictions does not mean that the expression becomes totally uninterpretable, but some context features may repair the violation or a suitable context-of-utterance even renders the expression perfectly inconspicuous. In our view, one can account for these facts best when regarding the phenomenon of selectional restrictions as part of the semantics-pragmatics-interface.

## 2 Selectional Restrictions in HPSG

### 2.1 Previous Approaches

There are not many publications about selectional restrictions in HPSG. We only know about those of Nerbonne (1996) and Androutsopoulos and Dale (2000).

In his article, Nerbonne focuses on topics which are related to the processing of semantic information. In order to disambiguate the sense of *chair* in the example "*The chair decided on Mary*" he introduces a new feature M-AGT for "mental agent" within the semantics module. Thus one can distinguish between the two meanings "piece of furniture" and "head of organization". However, the author does not make clear what other features would be necessary and a worked-out concept of selectional or sortal constraints is far beyond the focus of Nerbonne's contribution.

A more concrete proposal for handling selectional restrictions is described by Androutsopoulos and Dale (op. cit.). The authors describe two alternative approaches. In their first proposal Androutsopoulos and Dale adopt a pragmatic point of view, putting all relevant information about a verb's selectional restrictions on the BACKGROUND set of the verb. They argue that selectional restrictions belong to the non-literal information, which is always situated in CONTEXT BACKGROUND, in contrast to literal information, which is to be handled in the CONTENT. For this approach the authors need an inferencing component which compares the relevant psoas to rule out signs corresponding to readings that violate a selectional restriction. This "constraint-satisfaction reasoning" would have to be pipe-lined after the parser of a natural language processor, because the information comes from a semantic hierarchy and has to be compared with the arguments present.

In their alternative approach Androutsopoulos and Dale treat selectional restrictions exclusively within CONTENT. They introduce a sortal hierarchy below *index*. So the INDEX value of the object of *eat* can be constrained to be of sort *edible*. This approach is more efficient for NLP applications. However, it yields an immediate failure of analysis when there is a violation of selectional restrictions and so does Nerbonne's proposal. Neither approach takes into account the effect of a repairing context. Only the first alternative by Androutsopoulos and Dale seems to be capable of being sensitive to contextual effects but, regrettably, the authors do not explain how this might work.

### 2.2 Our Proposal

As we have argued above, the phenomenon of selectional restrictions can be best accounted for by regarding it as part of the semantics-pragmatics-interface. The idea is to put the relevant information into the BACKGROUND set (BGR) of the CONTEXT of a sign and use structure-sharing with respective semantic indices. Contrary to the first proposal by Androutsopoulos and Dale (op. cit.) we introduce a semantic hierarchy with new sorts and relations as part of every *unembedded-sign*. Thus, we avoid the need for a separate inferencing component.

Unembedded signs are potential stand-alone utterances. According to Richter (2004, ch. 2.1.2), they are empirical objects and central to linguistic research. Richter argues already in (1997, ch. 5.2) that a more fine-grained distinction of *sign*s is necessary. In the signature which he develops, every subsort of *sign* can occur as an embedded and as an unembedded version. Major differences between embedded and unembedded signs are that the latter do not contain any unbound traces (if one assumes that traces exist) and that they have illocutionary force.

As a first step, we define two new elements to figure on the BGR set. These are, following standard assumptions, subsorts of *psoa*.

$$\begin{bmatrix} \textit{sel-restr-imp} & \\ \text{ARG} & \textit{index} \\ \text{MUST-SATISFY} & \textit{selection-sort} \end{bmatrix} \begin{bmatrix} \textit{sel-restr-stf} & \\ \text{ARG} & \textit{index} \\ \text{SATISFIES} & \textit{selection-sort} \end{bmatrix}$$

The first psoa can be introduced to BGR by signs which impose a selectional restriction.[5] A verb, e. g. *eat*, can

[4] Chomsky (1965, p. 158) also mentions meta-linguistic expressions like *It is not a good idea to eat motor-bikes.*

[5] *sel-restr-imp* for *imposed*

subcategorize for a noun with a certain restriction. Nouns such as *apple* satisfy this restriction.[6] They have also included this information in their BGR set.

The phrase „... *eats apples*“ is sketched in Fig. 1. The collection of all elements in all BGR sets is guaranteed by the CONTEXTUAL-CONSISTENCY-PRINCIPLE, which exists independently of our proposal.

As a second step we introduce a principle which ensures that the values of MUST-SATISFY (M-STF) and SATISFIES (STF) in the CTXT BGR set are compatible. To be compatible means that the STF value of the argument of *eat* is either identical to the M-STF value of the verb itself, or that the STF value is a sub-element of the M-STF value in a semantic ontology. In other terms, the verb only requires an edible object, whereas the object itself can be more concrete – a pancake or a banana.

The principle should license only phrases which have compatible values of M-STF and STF – but only if the argument or the whole proposition is outside the scope of a negational, a conditional or a question-operator. As stated above, these contexts “repair” the effect of a violation of selectional restrictions.

(7) VALIDITY-PRINCIPLE OF SELECTIONAL RESTRICTIONS (VPSR, preliminary version):
If in a phrase $x$ there is a sign $s$, a verb $v$ ($s$ is an argument of $v$) and a proposition $p$, which is formed by $v$ and its arguments, and
if neither the meaning associated with $s$ nor the meaning associated with $p$ are within the scope of a negational operator, a conditional operator or a question-operator or an epistemic verb,
then the STF value of a *sel-rest-stf* element in the CTXT BGR set of $x$ and the M-STF value of a *sel-restr-imp* element that shares the ARG value with *sel-rest-stf* must be compatible.

How can we capture this compatibility formally? The values of M-STF and STF are a subsort of the newly-introduced *selection-sort*, cf. Fig. 2. This sort has a finite number of subsorts such as *abstract, physical, artifact, animate, edible,...* which correspond to units of a semantic ontology as in WordNet[7] or GermaNet[8]. In Fig. 3, we roughly sketch such a semantic ontology, including multiple inheritance (subunits inherit from more than one superunit). In such an ontology the units are related to each other, indicated by the tree-structure. We want to establish such relations between the subsorts of *selection-sort*, too.

A sort hierarchy, as used for the normal HPSG sort inventory, cannot be adopted here. An HPSG formalism for Pollard/Sag-style grammars (as RSRL e.g. Richter et al., 1999) requires that objects be sort-resolved. This allows us to talk about objects having maximally specific sorts on the one hand and about underspecified descriptions (among them lexical entries) on the other. If we had a sort hierarchy for *selection-sort* analogous to the one in Fig. 3, we could not capture generalizations such as, e. g., that *eat* takes something *edible* as its object, for *edible* is not maximally specific. To clarify this point, we still take our example of *eat* which the lexical constraint to have an *edible* object. Consider a concrete utterance “*She eats pancakes.*” where there is a noun-object with $\left[\text{STF } \textit{pancake}\right]$, which is the argument of a verbal object *eat* with an arbitrary, maximally specific value $\left[\text{M-STF } \textit{banana}\right]$. Even though *banana* is a subsort of *edible* (the constraint in the lexical entry of the verb thus is fulfilled), the two sorts *banana* and *pancake* are still incompatible and the selectional restriction seems to be violated. This shows that we need sorts such as *edible*, which are somewhere in the middle of the hierarchy, as values in sort-resolved objects.

Thus we insert the subsorts of *selection-sort* into the signature as depicted in Fig. 2. The relations have to be defined separately, e. g. they can be collected in a list. This list is the value of a new attribute HIERARCHY, which we define for all unembedded signs. It contains pairs of subsorts of *selection-sort* being in an “is a”-relation. Formally this is a partial order of the elements below *selection-sort*. The following principle describes the list and defines it as the value of HIERARCHY for every unembedded sign.

(8) SELECTION-HIERARCHY-PRINCIPLE (outlined):
*unembedded-sign* →

$$\left[\text{HIERARCHY}\left\langle \begin{bmatrix}\textit{is\_a}\\ \text{ARG1}\ \textit{animate}\\ \text{ARG2}\ \textit{animate}\end{bmatrix}, \begin{bmatrix}\textit{is\_a}\\ \text{ARG1}\ \textit{animate}\\ \text{ARG2}\ \textit{person}\end{bmatrix}, \begin{bmatrix}\textit{is\_a}\\ \text{ARG1}\ \textit{animate}\\ \text{ARG2}\ \textit{animal}\end{bmatrix},\ldots \right\rangle\right]$$

We do not mean that the HIERARCHY, which can easily get quite big, is a genuine “linguistic” part of every unembedded sign. We only want to express the fact that every speaker has access to this kind of knowledge when formulating or hearing an utterance. Technically but not conceptually, this amounts to the same. Defining HIERARCHY as a feature of *unembedded-sign* allows us to determine the grammaticality of each unembedded sign without additional context. Thus we do not have to postpone the treatment of selectional restrictions to a separate inferencing component but we can recognize the semantical ill-formedness immediately for each unembedded sign.

Returning back to our selectional restriction approach, we recapitulate: compatibility of *selection-sort*s means

[6] *sel-restr-stf* for *satisfies*

[7] cf. Christiane Fellbaum, ed. (1998): *Wordnet: An Electronic Lexical Database*. Bradford Books, The MIT Press.

[8] cf. http://www.sfs.nphil.uni-tuebingen.de/lsd/

that there is an "is-a"-relation between the values of MUST-SATISFY and SATISFIES. This relation can contain one or more intermediate sorts; it is transitive.

(9) *She drank a sip of the Cabernet Sauvignon 2001.*

This example is about a special kind of wine. *Cabernet Sauvignon* **is** wine, which **is** an alcoholic beverage, which **is** a beverage, which **is** drinkable. The example shows that such an ontology becomes remarkably complex. At this point we have to admit that it is very easy to postulate and outline such ontologies. However, the implemention requires a lot of work, particularly when accounting for all the theoretical and empirical problems such a project raises (for a successful project cf. the one mentioned in footnote 7).

Having formalized the notion of compatibility, we can now reformulate the VPSR in the following way.

(10) VALIDITY-PRINCIPLE OF SELECTIONAL RESTRICTIONS (VPSR, final version):
If in an unembedded sign $x$ there is a sign $s$, a verb $v$ ($s$ is an argument of $v$) and a proposition $p$, which is formed by $v$ and its arguments, and
if neither the meaning associated with $s$ nor the meaning associated with $p$ are within the scope of a negational operator, a conditional operator or a question-operator or an epistemic verb,
then the STF value of a *sel-rest-stf* element in the CTXT BGR set of $x$ and the M-STF value of a *sel-restr-imp* element that shares the ARG value with *sel-rest-stf* must be in a relation on the HIERARCHY list of $x$.

## 3 Summary

We have proposed a way to integrate selectional restrictions into HPSG which includes the effects of repairing contexts. Restrictions are imposed by the verbs in their lexical entries and have to be satisfied by the verbs' arguments. If the argument is within the scope of a repairing operator, the whole sign is not ungrammatical – it is licensed by the VPSR.[9]

A further application of our approach might be the handling of metonymy (see e. g. Egg, 2004). It requires a certain amount of world knowledge to understand a metonymic utterance. For example, one has to know that wine, like every other drinkable liquid, is normally stored in a container, which can be placed on a table, cf. (4). Thus, for a metonymic utterance to be felicitous, a certain relation must hold between an element in the utterance and another object, as e. g. *in_container*, *has_part* or *consists_of*. These relations could be defined for all sorts in the HIERARCHY list. As we have already implemented the *is_a*-relation there, some generalizations can be captured in an elegant way.

[9] One argument we have disregarded is that a violation of selectional restrictions gets repaired by a certain kind of contexts like fairy tales of science fiction stories. To account for this kind of contextual shift one would have to assume a more fine-grained structure in the CONTEXT and distinguish between a standard context and an active context. Moreover, one would need relations which can take over standard assumptions (footballs are not edible) to the actual context or which can introduce new scenarios (starships can travel faster than light).

## Acknowledgements

The research to the paper was funded by the *Deutsche Forschungsgemeinschaft*. I am grateful to Frank Richter, Manfred Sailer, Janina Radó, Adrian Simpson and the reviewers of HPSG'05 for comments and Michelle Wilbraham for help with English.

## References

Androutsopoulos, Ion and Dale, Robert (2000). Selectional Restrictions in HPSG. In *Proceedings of COLING 2000*, Saarbrücken, pp. 15–20.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Egg, Markus (2004). Metonymie als Phänomen der Semantik-Pragmatik-Schnittstelle. *metaphorik.de (Online-Journal), ISSN 1618-2006 6*, 36–53.

Nerbonne, John (1996). Computational Semantics – Linguistics and Processing. In S. Lappin (Ed.), *Handbook of Contemporary Semantic Theory*, pp. 459–482. Blackwell Publishers, London.

Richter, Frank (1997). Die Satzstruktur des Deutschen und die Behandlung langer Abhängigkeiten in einer Linearisierungsgrammatik. Formale Grundlagen und Implementierung in einem HPSG-Fragment. In E. Hinrichs, D. Meurers, F. Richter, M. Sailer, and H. Winhart (Eds.), *Ein HPSG-Fragment des Deutschen, Teil 1: Theorie*, Number 95 in Arbeitspapiere des SFB 340, pp. 13–187. Universität Tübingen.

Richter, Frank (2004). *Foundations of Lexical Resource Semantics*. Professorial dissertation (version of 09-24-2004), Eberhard-Karls-Universität Tübingen.

Richter, Frank, Sailer, Manfred, and Penn, Gerald (1999). A Formal Interpretation of Relations and Quantification in HPSG. In G. Bouma, E. Hinrichs, G.-J. M. Kruijff, and R. Oehrle (Eds.), *Constraints and Resources in Natural Language Syntax and Semantics*, pp. 281–298. Stanford: CSLI Publications.

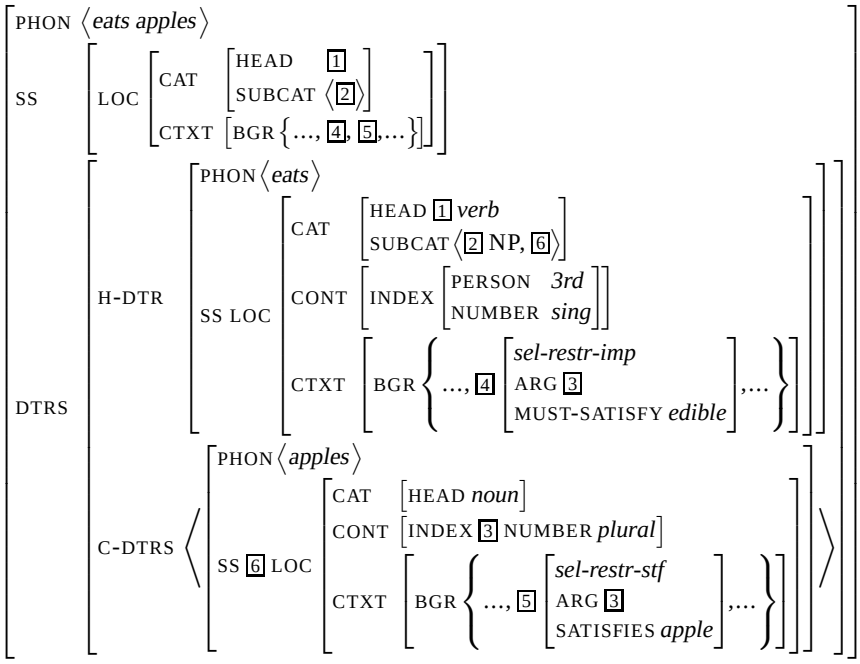



Figure 1: Phrase including selectional restrictions

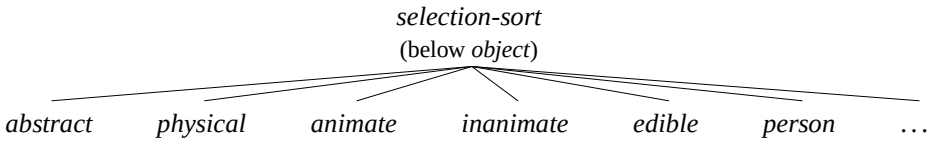



Figure 2: The sort *selection-sort*

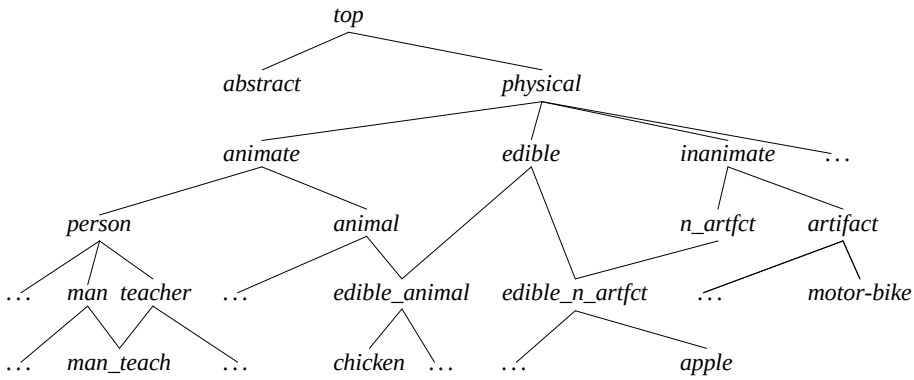



Figure 3: A semantic ontology

# Projecting RMRS from TIGER Dependencies

**Kathrin Spreyer**
Computational Linguistics Department
Saarland University
D-66041 Saarbrücken, Germany
kathrins@coli.uni-sb.de

**Anette Frank**
Language Technology Lab
DFKI GmbH
D-66123 Saarbrücken, Germany
frank@dfki.de

## 1 Introduction

Since the successful exploitation of treebanks for training stochastic parsers, treebanks are under development for many languages. Treebanks further enable evaluation and benchmarking of competitive parsing and grammar models. While parser evaluation against treebanks is most natural for treebank-derived grammars, it is extremely difficult for hand-crafted grammars that represent higher-level functional or semantic information, such as LFG, HPSG, or CCG grammars (cf. Carroll et al., 2002).

In a recent joint initiative, the TIGER project provides dependency-based treebank representations for German, on the basis of the TIGER treebank (Brants et al., 2002). Forst (2003) applied treebank conversion methods to the TIGER treebank, to derive an f-structure bank for stochastic training and evaluation of a German LFG parser. A more general, theory-neutral dependency representation is currently derived from this TIGER-LFG treebank, to enable cross-framework parser evaluation (Forst et al., 2004). However, while Penn-treebank style grammars and LFG analyses are relatively close to dependency representations (cf. Crouch et al., 2002; Kaplan et al., 2004), the situation is different for grammar formalisms that deliver deeper semantic representations, such as HPSG or CCG.

In order to provide a closer evaluation standard and appropriate training material for German HPSG grammars, we propose a method for the semi-automatic construction of an RMRS treebank for German on the basis of the LFG- resp. TIGER-Dependency Bank. In contrast to treebanks constructed from analyses of hand-crafted grammars, the RMRS treebank constitutes a standard for comparative parser evaluation where the upper bound for coverage is defined by the corpus (here, German newspaper text), not by the grammar.

Our treebank conversion method effectively implements RMRS semantics construction from dependency structures, and can be further developed to a general method for RMRS construction from LFG f-structures, similar to recent work in the LOGON project.[1]

## 2 The TIGER Dependency Bank

The input to our treebank conversion process consists of dependency representations of the TIGER Dependency Bank (TIGER-DB). The TIGER-DB is derived from (a subset of) the TIGER treebank. It abstracts away from constituency in order to remain as theory-neutral as possible. The TIGER-DB is derived semi-automatically from the TIGER-LFG Bank of Forst (2003), by defining various normalisations. The dependency format is similar to the Parc 700 Dependency Bank (King et al., 2003). So-called dependency triples are sets of two-place predicates that encode grammatical relations. The arguments represent the head of the dependency and the dependent, respectively. The triples further retain a number of morphological features from the LFG representations, such as agreement information for nominals and adjectives, or tense information. Figure 1 displays a sample dependency representation.

[1] See the online demo for LFG-based MRS semantics construction for Norwegian, as currently used in the LOGON project: http://decentius.aksis.uib.no:8010/logon/xle-mrs.xml

```
sb(müssen~0, Museum~1)
case(Museum~1, nom)
gend(Museum~1, neut)
num(Museum~1, sg)
mod(Museum~1, privat~1001)
cmpd_lemma(Museum~1, Privatmuseum)
oc_inf(müssen~0, weichen~3)
mood(müssen~0, ind)
tense(müssen~0, pres)
sb(weichen~3, Museum~1)
```

Figure 1: TIGER-DB structure for *Privatmuseum muss weichen* – Private museum deemed to vanish.

However, dependency structures are difficult to match against the output of HPSG parsing. HPSG analyses do not come with an explicit representation of functional structure, but directly encode semantic structures, in terms of MRS (Copestake et al., 2005) or RMRS (Copestake, 2003). This leaves a gap to be bridged in terms of normalisation of diathesis, the encoding of arguments vs. adjuncts, the representation of constructions like relative clauses, and the representation of quantifiers and their scoping relations.

In order to provide a gold standard that can be matched against the output of HPSG parsing for evaluation, and further, for training stochastic grammar models, we propose a method for treebank conversion that essentially performs RMRS construction from LFG-based dependency representations.

For the purpose of semantics construction, the triples format has both advantages and disadvantages. On the one hand, the LFG-derived dependencies offer all the advantages of a functional as opposed to a constituency-based representation. This representation already filters out the semantically inappropriate status of auxiliaries as heads; their contribution is encoded by features such as `perf` or `fut`, which can be directly translated into features of semantic event variables. Most importantly, the triples localize dependencies which are not locally realized in terms of phrase structure (as e. g. in control structures, coordination, or long-distance constructions), so that when constructing the semantics from the dependency format, we do not need additional mechanisms to identify the arguments of a governing predicate.

The challenges we face mainly concern the lack of constituency information in the dependency representations. While standard definitions of the principles for (R)MRS construction refer to constituency information, we now have to define RMRS composition on the basis of dependency relations.

## 3 RMRS Construction from TIGER Dependency Structures

### 3.1 Treebank Conversion by Term Rewriting

Similar to Forst (2003) we are using the term rewriting system of Crouch (2005) for treebank conversion. Originally designed for Machine Translation, the system is a powerful tool for structure rewriting that is also applied to other areas of NLP, such as induction of knowledge representations (Crouch, 2005).

The input to the system consists of a set of facts in a prolog-like term representation. The rewrite rules refer to these facts in the left-hand side (LHS), either conjunctively (expressed by separating conjuncts with a comma '`,`') or disjunctively (expressed by '`|`'). Expressions on the LHS may be negated by a prefixed '`-`', thereby encoding negative constraints for matching. A rule applies if and only if all facts specified on the LHS are satisfied by the input set of facts. The right-hand side (RHS) of a rewrite rule defines a conjunction of facts which are added to the input set of facts if the rule applies. The system further allows the user to specify whether a matched fact will be consumed (i. e., removed from the set of facts) or whether it will be retained in the rule's output set of facts (marked by the prefix '`+`').

The processing of rules is *strictly ordered*. The rules are applied in the order of textual appearance. Each rule is tested against the current input set of facts and, if it matches, produces an output set of facts that provides the input for the next rule in sequence. Each rule applies concurrently to all distinct sets of matching facts, i.e. it performs parallel application in case of alternative matching facts.

The system offers powerful rule encoding facilities: Macros are parameterized patterns of (possibly disjunctive) facts; templates are parameterized abstractions over entire (disjunctive) rule applications. These abstraction means help the user to define rules in a perspicious and modular way.

### 3.2 RMRS Construction

Within the formal framework of HPSG, every lexical item defines a complete RMRS structure. Semantics composition rules are defined in parallel with syntactic composition. In each composition step, the RMRSs of the daughters are combined according to strict semantic composition rules, to yield the RMRS representation of the phrase (cf. Copestake et al., 2001). Following the scaffolding of the syntactic structure in this way finally yields the semantic representation of the sentence.

For our task, the input to semantics construction is a dependency structure. As established by work on Glue Semantics (Dalrymple, 1999), semantics construction from dependency structures can in similar ways proceed recursively, to deliver a semantic projection of the sentence. However, the resource-based construction mechanism of Glue Semantics leads to alternative derivations in case of scope ambiguities.

In contrast to Glue, we target an underspecified semantic representation. Although defined on phrasal configurations, the algebra for (R)MRS construction as defined in Copestake et al. (2001) can be transposed to composition on the basis of dependency relations, much alike the Glue framework.

Yet, the rewriting system we are using is not suited for a recursive application scheme: the rules are strictly ordered, and each rule simultaneously applies to all facts that satisfy the constraints. That is, the RMRS composition cannot recursively follow the composition of dependents in a given input structure.

**The RMRS Skeleton.** RMRS construction is thus designed around one *global RMRS*, featuring a TOP label, a RELS set containing the *elementary predications* (EPs), a set HCONS of *handle constraints* which state restrictions on possible scopes, and a set of ING constraints that represent the *in-group* relation.[2]

Instead of projecting and accumulating RMRS constraints step-wise by recursive phrasal composition rules from the lexical items to the top level of the sentence, we directly insert all EPs, ING and HCONS constraints into the global RMRS, i.e. the RMRS with the top handle. The semantics composition rules are thus reduced to the inherent semantic operations of the algebra of Copestake et al. (2001): the binding of argument variables and the encoding of scope constraints. These basic semantic operations are defined by appropriate definitions and operations on the HOOK features in the composition rules.

**Lexical RMRSs.** The notion of *lexical RMRSs* as it is defined here slightly differs from the standard one. If semantic composition proceeds along a tree structure, lexical RMRSs are constructed at the leaf nodes. In our scenario, a lexical RMRS is projected from the PRED features in the dependency structures, irrespective of any arguments, which are considered by subsequent composition rules.

We define the lexical RMRSs in two steps: First, the hook label is (freely) instantiated and thus available for reference to this RMRS by other rules. Second, the hook variable and the basic semantics (EPs for the relation and the ARG0, at least) are introduced on the basis of the predicate's category. This category information is not explicit in the dependencies, but it can be induced from the grammatical function borne by the predicate, as well as the presence or absence of certain morphological features.

Figure 2 shows a sample lexical RMRS and the rule that yields it: The rule applies to predicates, i.e. to `pred` features, with a value `Pred` and a hook label `Lb`. In the RHS, one EP is added for the relation represented by `Pred`, and one for the ARG0, which is identified with the hook variable.[3]

**Composition.** The semantic composition of arguments and functors makes use of an attribute `arg()` which encodes the argument structure of the predicates.[4] Given a predicate `arg(Fctor,N,Arg)`,

[2] Whenever two handles are related via an ing constraint, they can be understood to be conjoined. This is relevant, e.g., for intersective modification, since a quantifier that outscopes the modified noun must also take scope over the modifier.

[3] In fact, for modifiers and specifiers we define lexical RMRSs in a special way, in that we immediately bind the semantic argument. The motivation for this is that whenever one of the dependency relations `mo` or `spec` are encountered, no matter what their exact `Pred` value may be, the semantics contributed by the head of this dependency can be unambiguously related to the semantic head, and is thus recorded already at the "lexical" level.

[4] As explained below, the information about subcategorized arguments is reconstructed from the triples, in the predicate `arg(Fctor,N,Arg)`, where `N` encodes the argument position, `Fctor` and `Arg` are indices of functor and argument, re-

```
(a) add_ep(Lb,Type,Feat,Val) ::
        +rels(_,Rels)
    ==> ep(Rels,EP), type(EP,Type), lb(EP,Lb),
        complex_term(Feat,EP,Val).
(b)     +pred(X,Pred), -mo(_,X), -spec(_,X),
        +'s::'(X,SemX), +hook(SemX,Hook), +lb(Hook,Lb)
    ==> var(Hook,Var)
    &&  add_ep(Lb,ep_rel,rel,Pred)
    &&  add_ep(Lb,ep_arg0,arg0,Var).
```

(c)

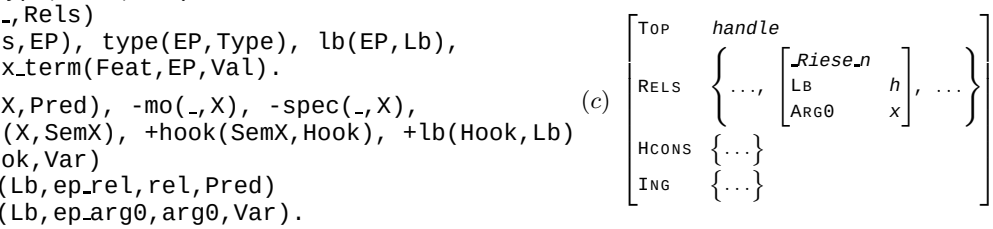



Figure 2: (a) Expansion of `add_ep` template, (b) a rule with a template call, (c) the output lexical RMRS.

```
(a) +arg(X,2,Arg), +g_f(Arg,'oc_fin'),
    get_lb(X,LbX), get_lb(Arg,LbArg),  ==> sort(Lb,h), sort(LbPrpstn,h)
    +comp_form(Arg,dass)                && add_ep(LbX,ep_arg2,argx,LbPrpstn)
                                        && add_ep(LbPrpstn,ep_rel,rel,'prpstn_m_rel')
                                        && add_ep(LbPrpstn,ep_arg0,arg0,Lb)
                                        && add_qeq(Lb,LbArg).
```

(b)

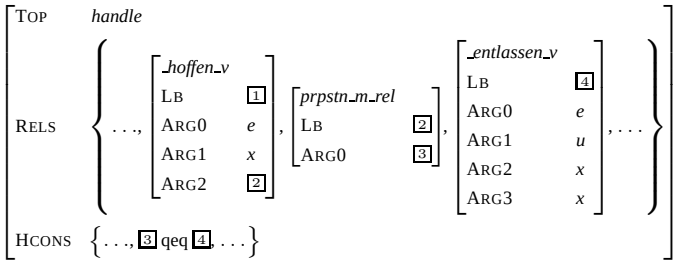



Figure 3: (a) Sample argument binding rule and (b) output RMRS.

the binding of the argument to the functor is steered by the previously defined hooks of the two semantic entities in that the matching rule attaches an EP with an attribute ARG*N* to the externalized label in the functor's hook. The value of the attribute ARG*N* is the hook variable of the argument. A slightly more complicated example is shown in Figure 3, it features the introduction of an additional proposition and a scope constraint. This rule binds a declarative (marked by the complementizer *dass*) finite clausal object (`oc_fin`) to the verb it is an argument of. To achieve this binding, a proposition relation is assigned as the value of the verb's ARG2, and this proposition in turn has an ARG0, which takes scope over the hook label of the matrix verb in the object clause (for the definition of the template `add_ep`, see Figure 2; the template `add_qeq` works similarly: It adds a qeq constraint to the set of handle constraints). In general, the binding of arguments does not depend on the order of rule applications. That is, the fact that the system performs concurrent rule applications in a cascaded rule set is not problematic for semantics construction. Though, we have to make sure that every partial structure is assigned a hook, prior to the application of composition rules. This is ensured by stating the rules for lexical RMRSs first.

**Scope constraints.** In having the rules introduce handle constraints, we define restrictions on the possible scoped readings. These are defined maximally restrictive in the sense that they must allow for all and only the admissible scopes. This is achieved by gradually adding qeq relations to the global HCONS set. Typically, this constraint relates a handle argument of a scopal element, e. g. a quantifier, and the label of the outscoped element. However, we cannot always fully predict the interaction among several scoping elements. This is the case, inter alia, for the modification of verbs by more than one scopal adverb. This type of ambiguity is modeled by means of a UDRT-style underspecification, that is, we leave the scope among the modifiers unspecified, but restrict each to outscope the verb handle.[5]

spectively.

[5] This is in accordance with the German HPSG grammar, and will also be adapted in the ERG (p.c. D. Flickinger).

### 3.3 Challenges

Some aspects of semantic composition crucially depend on lexical and phrase structural information which is not available from the dependencies. Here we briefly point out the problems and how we solved them.

**Argument Structure.** Although LFG grammars explicitly encode argument structure in the semantic form of the predicate, the derived dependency triples only record the atomic PRED value. We recover the missing information by way of preprocessing rules. The rules make reference to the local grammatical functions of a predicate, and test for features typically borne by non-arguments, for instance, expletives can be identified via the feature `pron_type(_,expl)`. In the composition step, the resulting `arg` predicates will be interpreted as the *slots* that a functor needs to fill.

The TIGER-DB does not provide information about control properties of equi-verbs, nor do they mark scopal modifiers. We extracted lexical entries from the broad-coverage German HPSG, and interleave them with the rules for semantics construction, to ensure their proper representation.

**Constituency.** It is often assumed that there is a crucial difference between the semantics of VP-modification and that of S-modification. Thus, we are faced with the problem that no distinction whatsoever is drawn between heads and their projections in the dependency structures. Hence, we restrict scope with respect to the verb, but do not exclude the proposition-modifying reading.

Similarly, coordination is represented as a set of conjuncts in the triples, but to meet the binary branching coordination analysis of HPSG, we must construct a recursive semantic embedding of partial coordinations. The rules process the conjuncts in a right-to-left manner, each time combining the partial coordination to the right with the conjunct on the left, thereby building a left-branching coordination.

### 3.4 Treebank Construction and Quality Control

**TIGER 700 RMRS Treebank.** Our aim is to construct a treebank of 700 sentences from the TIGER dependency bank. Instead of selecting a random sample of sentences, we opt for a block of consecutive sentences. In this way, the treebank can be further extended by annotations for intersentential phenomena, such as co-reference relations, or discourse relations.

However, we have to accommodate for gaps, due to sentences for which there are reasonable functional syntactic, but (currently) no sound semantic analyses. This problem arises for sentences involving, e.g., elliptical constructions, or else ungrammatical or fragmented sentences. We will include, but explicitly mark such sentences for which we can obtain partial, but no fully sound semantic analyses. We will correspondingly extend the annotation set to yield a total of 700 correctly annotated sentences.

**Automatic Conversion and Quality Control.** Currently, we have covered the main body of rewrite rules for converting dependency structures to RMRSs. The grammar comprises about 70 rules, 15 macros and templates. In the next step we will implement a cascaded approach for quality control, with an initial feedback loop between (i) and (ii):

(i) Manual phenomenon-based error-detection. In the construction process, we mark the application of construction rules by inserting phenomenon-specific identifiers, and use these to select sample RMRSs for phenomenon-based inspection, both in the development phase and for final quality control.

(ii) Investigation of detected errors can result in the improvement of automatic RMRS construction (feedback loop to (i)). Errors that cannot be covered by general rules need to be adjusted manually.

(iii) Manual control. Finally, we need to perform manual control and correction of errors that could not be covered by automatic RMRS construction. In this phase, we will mark and separate the structures or phenomena that are not covered by the state-of-the-art in RMRS-based semantic theory.

## 4 Conclusion

We have presented a method for semantics construction which converts dependency structures to (R)MRSs as they are output by HPSG grammars. This approach allows cross-framework parser evaluation on a broad-coverage basis, and can be applied to existing dependency banks for English (e. g. King et al. (2003)).

## References


Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

Carroll, J., Frank, A., Lin, D., Prescher, D., and Uszkoreit, H., editors (2002). *Beyond PARSEVAL – Towards Improved Evaluation Measures for Parsing Systems*, Workshop Proceedings of the *Third International Conference on Language Resources and Evaluation*, LREC 2002 Conference, Las Palmas, Gran Canaria.

Copestake, A. (2003). Report on the Design of RMRS. Technical Report D1.1a, University of Cambridge, University of Cambridge, UK. 23 pages.

Copestake, A., Flickinger, D., Sag, I., and Pollard, C. (2005). Minimal Recursion Semantics. to appear.

Copestake, A., Lascarides, A., and Flickinger, D. (2001). An Algebra for Semantic Construction in Constraint-based Grammars. In *Proceedings of the ACL 2001*, Toulouse, France.

Crouch, R. (2005). Packed Rewriting for Mapping Semantics to KR. In *Proceedings of the Sixth International Workshop on Computational Semantics, IWCS-06*, Tilburg, The Netherlands.

Crouch, R., Kaplan, R., King, T., and Riezler, S. (2002). A comparison of evaluation metrics for a broad coverage parser. In *Beyond PARSEVAL. Workshop at the LREC 2002 Conference*, Las Palmas.

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

Forst, M. (2003). Treebank Conversion – Establishing a testsuite for a broad-coverage LFG from the TIGER treebank. In *Proceedings of LINC'03*, Budapest, Hungary.

Forst, M., Bertomeu, N., Crysmann, B., Fouvry, F., Hansen-Schirra, S., and Kordoni, V. (2004). Towards a Dependency-Based Gold Standard for German Parsers: The Tiger Dependency Bank. In Hansen-Schirra, S., Oepen, S., and Uszkoreit, H., editors, *Proceedings of LINC 2004*, Geneva, Switzerland.

Kaplan, R., Riezler, S., King, T., Maxwell, J., Vasserman, A., and Crouch, R. (2004). Speed and accuracy in shallow and deep stochastic parsing. In *Proceedings of HLT-NAACL'04*, Boston, MA.

King, T., Crouch, R., Riezler, S., Dalrymple, M., and Kaplan, R. (2003). The PARC 700 Dependency Bank. In *Proceedings of LINC 2003*, Budapest.

# Free Relatives in Persian

**Mehran Taghvaipour**
Department of Language and Linguistics
University of Essex
Colchester CO4 3SQ
matagh@essex.ac.uk

## 1 Data

Like in English, free relatives (FRs) in Persian are Unbounded Dependency Constructions (UDCs). Persian FRs may contain a gap or a resumptive pronoun (RP) which is linked to and licensed by the FR word. Example (1) shows a Persian FR in brackets[1].

(1)
*Yasmin* [*hærči* *Amy* ____ *xærideh.bud*]
Yasmin whatever Amy ____ had.bought

*ra*[2] *bærdašt.*
RA took-3sg

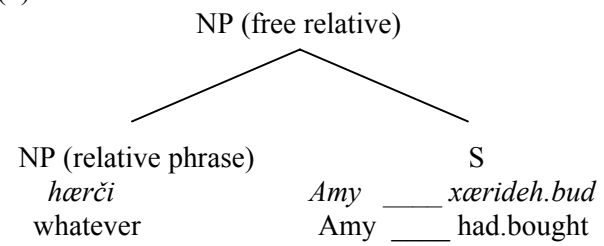
'Yasmin took whatever Amy had bought.'

Example (2) shows a Persian FR with a RP.

(2)
[*hærki* *beš* *pul* *dadim*] *umæd.*
whoever to-**him** money gave-1sg came-3sg
'Whoever we gave money to came.'

The structure of FRs is bipartite, containing a phrasal part ('relative phrase') and a sentential part. This is illustrated in (3).

(3)




Persian FRs come only in non-specific form. That is, the relative phrase always contains a word which is accompanied by the prefix *hær*, '-ever'. Also, as in most languages (e.g., English) the range of *wh*-words that can appear in Persian FRs is restricted. However, what is more interesting in Persian is that free relative words are not restricted to *wh*-words. For example, for *hærki*, 'whoever', we can have '*hærkæs'*, which is a combination of *hær*, '-ever', and the noun *kæs*, 'person'. Therefore, what is always present in a Persian FR word is the prefix *hær-*, 'ever'.

The sentential part of a Persian FR is an incomplete finite sentence that contains either a gap or a RP. The pattern of distribution of RPs in Persian FRs is different from that of the ordinary relative clauses (RCs) in this language. While in ordinary RCs, it is possible to have gaps in object position, in FRs this is not possible. Examples in (4) illustrate.

[1] The FR in (1) is nominal. Other types of FRs, such as adjectival or adverbial, are also possible in Persian; however, I shall focus on nominal FRs here.
[2] This particle (whose colloquial form is *ro* or simply *-o*) is a specificity marker (Karimi, 1989). It comes after an NP when the NP is specific and is not in the position of subject of object of preposition.

(4)

a. *zæn-i* *ke* *mæn* ***u***
Woman-RES[3] COMP I **her**

*ra* *dust.daræm* *injast.*
RA love-1sg here-is

'The woman that I love is here.'

b.
[*hærki-o* *mæn* (***u***)
whoever-RA I (her)

*dust.daræm*] *injast*
love-1sg here-is

'Whoever I love is here.'

Moreover, the complementizer *ke* which is obligatory in ordinary RCs is optional in FRs. While the examples in (1), (2) and (4b) above show that we do not need *ke* in FRs, example (5) shows that it is possible to have *ke* in a FR. It is indeed possible to have *ke* in any of the abovementioned above examples.

(5)

*hærkæs* *(ke)* ___ *pirhæn*
every/any + person COMP ___ shirt

*pušide* *tu tim-e* *mast.*
wear-3sg in team-EZ we+be-1pl

'Whoever is wearing a shirt is in our team.'

Another property of Persian FRs is that they allow pied piping. Examples in (6) illustrate.

(6)

a. *hærki* *mæn* ***baš*** *hærf+zædæm* *æz*
Whoever I with-**he** tak-1sg from I

*mæn* *bištær* *midunest.*
more knew-3sg

'Whoever I talked to knew more than me.'

b. *ba* *hærki* *mæn* *hærf+zædæm*
with whoever I talk-1sg

*æz* *mæn* *bištær* *midunest.*
from I more knew-3sg

'To whoever I talked knew more than me.'

Persian data also show that FRs in this language are subject to categorial matching. Examples in (7) illustrate.

(7)

a. *Yasmin* [***æz hærči***] *Setareh*
Yasmin from whatever Setareh

*bædeš.miyad* *xošeš.miyad.*
dislike-3sg like-3sg

'Yasmin dislikes whatever Setareh likes.'

b. *Yasmin* [***æz hærči***] *Setareh*
Yasmin from whatever Setareh

*bædeš.miyad* *dust.dareh.*
dislike-3sg like-3sg

'Yasmin dislikes whatever Setareh likes.'

The matrix verb in (7a) is the compound verb *xošeš.miyad*, 'like' which subcategorizes for a PP, starting with the preposition *æz*, 'from'. The relative verb in this sentence is *bædeš. miyad*, 'dislikes', which also subcategorizes for a PP, starting with the preposition *æz*, 'from'. Both requirements are met and the result is grammatical. However, (7b) is ungrammatical as the categorial matching is not observed. The matrix verb in (7b) is the compound verb *dust.dareh*, 'likes', which subcategorizes for an NP. The relative verb in (7b) is the compound verb *bædeš.miyad,* 'dislikes', which subcategorizes for a PP, starting with the preposition *æz*, 'from'. The requirements of the two verbs are not satisfied and the sentence is, therefore, ungrammatical.

## 2 An HPSG Analysis

I shall provide a unified HPSG approach to take care of the dependency between the relative phrase

[3] Particle *–i* attaches to the NPs modified by restrictive relative clauses. I will show it by –RES in gloss.

and the gap or the RP with a truly single mechanism, using only the SLASH feature. A variety of evidence from coordination, parasitic gaps, crossover, and island constraints shows that Persian gaps and RPs are strikingly similar. A coordination example is given in (8) below. One conjunct contains a gap whereas the other conjunct contains a RP. The gap and the RP are both licensed by and linked to *hærki*, 'whoever'.

(8)
[*hærki umæd va beš*
whoever came-3sg and to-**him**

*pul dadim*] *ræqsid.*
money gave-1sg dance-3sg

'[Whoever ___ came and we gave money to ____(*him)] danced.'

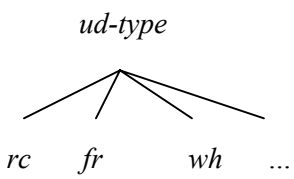
To capture the similarities of RPs and gaps, I suggest that gaps and RPs should be treated similarly (unlike Vaillette's (2001) analysis). If gaps are treated as traces[4] (as in Pollard and Sag's (1994), Levine and Hukari (2003), and Lee (2004)), then RPs will be similar to traces except in two respects. Firstly, RPs will have phonological content whereas traces will not. Secondly, the value of their GAPTYPE features is different. GAPTYPE is a feature that I have introduced in order to capture the distributional properties of RPs and traces. GAPTYPE is a non-local feature whose value can be either *trace* or *rp*, for traces and RPs, respectively. The reason for distinguishing traces and RPs with a NONLOCAL feature is that this is not reflected within the value of SLASH and hence it is possible for a single unbounded dependency to be associated with a trace and an RP. This makes the inheritance of the nonlocal feature easy and possible in the middle of those UDCs which involve coordination of two NPs where one contains a RP and the other a gap. Other analyses (e.g., Vaillette (2001)) which utilise more than one nonlocal feature (SLASH and RESUMP) do not seem to be able to handle the inheritance of the features in such coordinate structures that contain gap in one conjunct and RP in the other. The lexical entry for trace is shown in (9).

(9)
$$\begin{bmatrix} \text{PHONOLOGY} & \langle\,\rangle \\ \text{SYNSEM} & \begin{bmatrix} \text{LOCAL} & \boxed{1} \\ \text{NONLOC} & \begin{bmatrix} \text{SLASH} & \boxed{1} \\ \text{GAPTYPE} & \textit{trace} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

In addition, the impossibility of RPs in object positions in FRs requires a more complex value for SLASH. By doing so, the encoded information in SLASH will show the type of the unbounded dependency (e.g., *wh*-interrogative, relative clause, free relative, etc.) as well; therefore, making it accessible not only at the bottom of the dependency but also at the top. I shall assume that the value of SLASH is a set of *ud-object* elements, for which two features are appropriate: LOCAL and UD-TYPE. The value of LOCAL is a set of *local* structures, and the value of UD-TYPE is *ud-type,* which can be for instance *rc* (for relative clauses)*, fr* (for free relatives)*,* or *wh* (for *wh*-interrogatives). The complex value of SLASH is shown in (10). The hierarchy in (11) shows three of the possible instances of *ud-type*.

(10)
$$\text{SLASH} \begin{bmatrix} \textit{ud-object} & \\ \text{LOCAL} & \textit{local} \\ \text{UD-TYPE} & \textit{ud-type} \end{bmatrix}$$

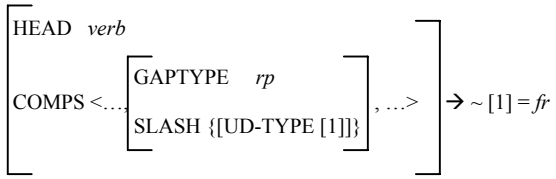
(11)




We noted above that RPs are not allowed in object positions. To prevent RPs from appearing in object positions, I shall propose the following constraint.

[4] It is also possible to develop a traceless approach by, for instance, treating *synsems* of RPs (*rp-ss*) a mixed category: a subtype of *gap-ss* and *canon-ss* at the same time. I shall not follow this approach here.

(12) RESUMPTIVE OBJECT CONSTRAINT

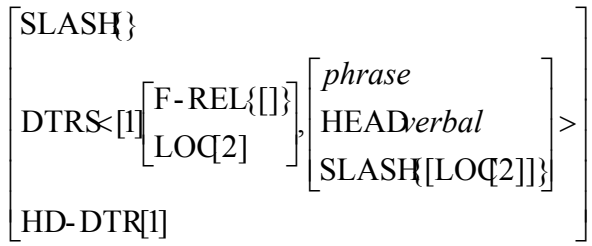



The effect of this constraint is that if a complement of a verb is a resumptive pronoun, then the value of its UD-TYPE cannot be *fr*. In other words, a pronoun which is resumptive by having a *rp* value for its GAPTYPE feature cannot be used in unbounded dependencies of the type free relative (*fr*), if it is going to be a verbal complement.

At the top of the dependency, the SLASH feature needs to be bound off at an appropriate point. Similar to Wright and Kathol's (2003) analysis, I assume that this appropriate point is the relative phrase which acts as the filler. However, if the relative phrase is the filler, then naturally, we expect to have the sentential part as the head. Persian data do not support this idea and suggest that it is the relative phrase that acts as the head in determining the external distribution of the phrase. For example, categorical matching comes from the relative phrase (see example 7 above). Therefore, I am assuming that the relative phrase in Persian FRs is the head and the filler at the same time. I propose the following constraint on Persian FRs.

(13) Free Relatives Constraint

*free-relative* →




There are two points noteworthy here. Firstly, the filler is the head daughter in this constraint. Secondly, the value of HEAD is *verbal* and not *v*. Following Sag (1997), I will assume that *verbal* is a supertype of both verb (*v*) and complementizers (*c*). This assumption will allow me to handle the optionality of complementizer *ke* in Persian FRs.

The standard HPSG constraints that operate on headed phrases will suffice for the inheritance of SLASH. There is also another nonlocal feature which is originated at the relative phrase: the F-REL feature. We noted earlier that not all *wh*-words can come in relative phrases. Likewise, not all words with the prefix *hær-* are allowed to appear in FR constructions. In order to differentiate phrases that are eligible to occur as fillers in the FR constructions, I will use, following Kim's (2001), the nonlocal feature F-REL which takes a set of referential indices as its value (Jacobson (1977), Kim and Park (1997) as cited in Kim (2001: 42)). FR words will have a nonempty specification for this feature. Other instances of *hær-* combinations or *wh*-words in any context other than the FR will have empty F-REL features. The F-REL generated from a lexical entry is subject to two constraints: Lexical Amalgamation of F-REL (Kim 2001: 43) and Ginzburg and Sag's (2000) Generalised Head Feature Principle (GHFP).

**Selected References**


Ginzburg J. and I. Sag, 2000. *Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives*. CSLI Publications, Stanford, California.

Lee, S-H. (2004) *A Lexical Analysis of Select Korean Unbounded Dependency Constructions.* Doctoral dissertation in Ohio State University, USA.

Levine, R. D. and T. Hukari. 2003. *The Unity of Unbounded Dependency Constructions*. University of Chicago Press. USA.

Karimi, S. 1990. "Obliqueness, Specificity, and Discourse Functions: *râ* in Persian," *Linguistic Analysis*, Vol. 3&4. PP 139-191

Kim. J.B. 2001. Constructional Constraints in English Free Relative constructions. *Korean Society for Language and Information* 5(1): 35-53.

Pollard C., and I. Sag, 1994. Head-Driven Phrase Structure Grammar. The University of Chicago Press, USA.

Sag, I., 1997. English Relative Clause Constructions. Journal of Linguistics 33:431-474.

Vaillette, N. 2001. Hebrew Relative Clauses in HPSG. Proceedings of the 7th International Conference on Head-Driven Phrase Structure Grammar, CSLI Publications.

Wright, A. and A. Kathol. 2003. When a Head is not a Head: A Constructional Approach to Exocentricity in English. In Jong-Bok Kim and Steve Wechsler (eds.) *Proceedings of the Ninth International Conference on Head-Driven Phrase Structure Grammar*. Pages: 370-387.

# Plural Comitative Constructions in Polish

**Beata Trawiński**
Eberhard-Karls-Universität Tübingen
Sonderforschungsbereich 441
Nauklerstraße 35
D-72074 Tübingen
trawinski@sfs.uni-tuebingen.de

## 1 Introduction

This paper deals with Polish comitative constructions (CCs) involving the preposition *z* 'with', as appears in (1).

(1) Jan z Marią odjechali.
Jan.NOM with Maria.INSTR left.PLUR
'Jan and Maria left.'

The CC in (1) consists of the nominative singular NP *Jan* and the *z*-PP, and combines with the plural predicate. Because of the plural agreement, we will refer to this type of CC as the plural comitative construction (PCC).[1]

PCCs have previously been treated by linguists in terms of coordination[2] (cf. (Vassilieva and Larson, 2001) for corresponding Russian expressions and (Dyła, 1988) and Dyła and Feldman (to appear) for Polish), complementation (cf. (Feldman, 2002) and Dyła and Feldman (to appear) for Russian) and adjunction (cf. (McNally, 1993) and (Ionin and Matushansky, 2002) for Russian). However, most of these analyses remain problematic in some respects. For example, the analysis of (Vassilieva and Larson, 2001) fails to explain the case assignment on the second NP, cannot rule out the ungrammatical inversion of the first and the second NPs and cannot account for grammatical structures involving a conjunction. The approach of (Dyła, 1988) and Dyła and Feldman (to appear) cannot prohibit the inversion of the first and the second NPs, nor ungrammatical iteration. In addition, it is both conceptually and formally incompatible with HPSG, which provides the underlying grammatical framework. (Feldman, 2002) analyzes the Russian comitative preposition *s* 'with' as a noun which selects two complements. However, by treating *s* 'with' as a noun, neither the vocality alternation (cf. *s* vs. *so*), which is typical for prepositions but not for nouns, nor the modification by the adverb *vmeste* 'together', can be explained.

This paper offers an HPSG adjunction-based analysis of PCCs that accounts for their syntactic, semantic and pragmatic properties by providing a special lexical entry for the preposition *z*. Consequently, no additional constraints on phrase structure or semantic constraints will be needed in order to license PCC.

In the following section we will present the results of an examination of PCCs with respect to various linguistic phenomena such as number and gender resolution, control phenomena, distributivity, extraction, case assignment, iteration, recursion and pro-drop phenomena. The aim of our tests was to provide an empirical basis for generalizations about the syntactic and semantic properties of PCCs. At the same time, we have investigated coordinate, adjunct and complement structures with respect to the same phenomena. The objective was to determine which of these structures shares the most syntactic

[1] Note that CCs in Polish, as well as in many other languages, can also involve singular agreement on the verb. The treatment of CCs with singular agreement on the verb seems relatively straightforward. However, this is not the case for PCCs. Therefore, PPCs will be the exclusive focus of this paper.

[2] Note, however, that there has been no uniform treatment of coordination. Thus, coordination might correspond to other syntactic structures, such as adjunction, depending on the analysis used.

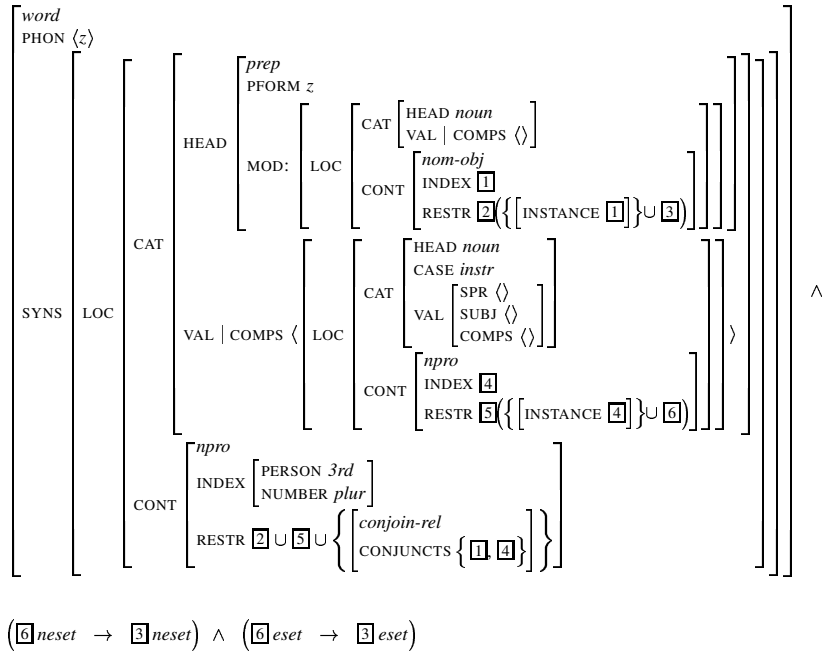



Figure 1: The relevant part of the lexical entry of the preposition *z* 'with'

and semantic properties with the PPC.[3]

## 2 Results of the Empirical Observations

Based on a number of linguistic tests, we have been able to observe that PCCs behave in the same way as does ordinary coordination with regard to number resolution, gender resolution, control of pronouns, PRO subjects and distributive interpretation. However, with respect to case assignment and the grammatical category of the constituents involved, PCCs share its properties with both NP-adjuncts and NP-complements. Since PCCs also show the same behavior as NP-adjuncts with respect to the control of pronouns within *z*-PPs, the occurrence of pronouns within PCCs, iteration, and recursion, we consider it plausible to analyze PPCs syntactically as an instance of adjunction.

Based on these empirical observations, two generalizations can be made: (1) PCCs share their semantic properties with ordinary coordination, (2) PCCs share their syntactic properties NP-adjunction. Furthermore, PCCs show several idiosyncratic features, e.g., with respect to the distribution of pronouns within PCCs, or concerning requirements for definiteness, number or restrictiveness of the NPs involved.

In the next section we will provide our HPSG analysis for PCC in Polish.

## 3 The Analysis

We have adopted the proposal by (McNally, 1993), thereby treating PCCs as adjunct-structures.[4] The core component of our analysis is the lexical entry for the preposition *z* in Figure 1.

The lexical entry in Figure 1 licenses the preposition *z* 'with', which selects one non-pronominal complement and modifies an NP. In this respect, the description in Figure 1 does not differ from descriptions of other modifying prepositions. However, the CONTENT value in Figure 1 differs from that of ordinary modifying prepositions. The value of the attribute CONTENT is a nominal object of the usual form. Note that the NUMBER value is assumed to be *plural*. The GENDER value depends on the GENDER values of the selected NP and the modified NP. Since PCCs show the same gender resolution pattern as coordination, we assume that PCCs are subject of

[3] In the full version of the paper, we will provide appropriate examples for each of the tests.

[4] Note that (McNally, 1993) has not provided a description of gender and number resolution, or of any control-related phenomena.

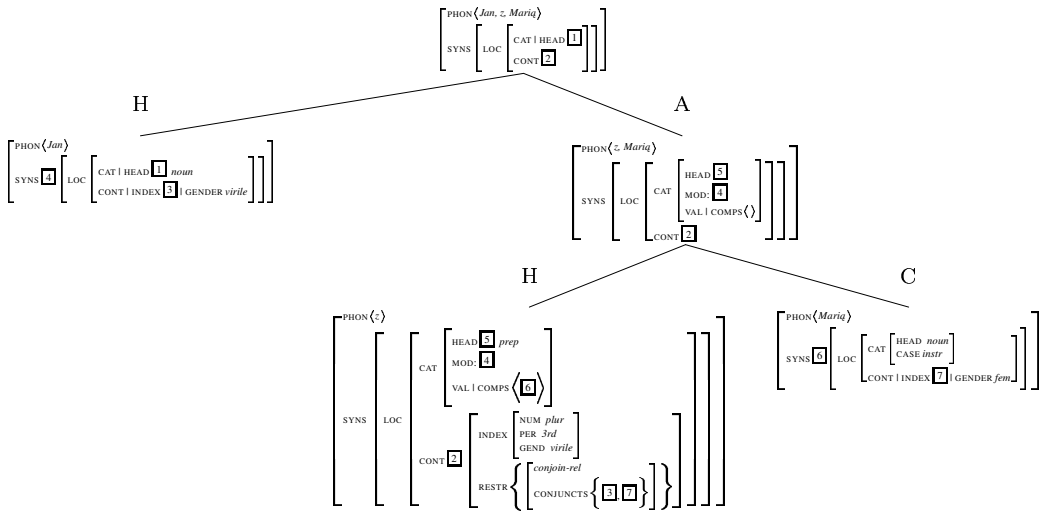



Figure 2: The structure of the PCC *Jan z Marią* 'Jan and Mary'

general constraints on gender resolution.

The RESTR feature of the preposition *z* 'with' provides information on the relation between the object denoted by the selected NP and the object denoted by the modified NP. This involves a conjoin relation. The value of the CONJUNCTS feature, which is appropriate for the sort *conjoin-rel*, is a set of *index* objects identified as the INDEX values of the modified and the selected NPs. Note that the proposed architecture of the CONTENT value of the preposition *z* 'with' which occurs in PCCs makes it possible to account for a distributive and collective reading of PCCs.[5]

The last two conjuncts of the description in Figure 1 ensure that the selected NP and the modified NP are either both modified or both not modified. PCCs require the same level of modification from both constituents. However, an additional constraint is needed which will ensure that the complement NP and the modified NP agree with respect to definiteness.

The structure in Figure 2 describes the PCC *Jan z Marią* 'Jan and Mary', using the lexical entry in Figure 1.

By virtue of the description in Figure 1, the preposition *z* selects first the non-pronominal NP *Marią* as its complement, assigning to it the instrumental case. Then *z* combines with the NP *Jan*. The CONJUNCTS set in the semantic representation of the preposition *z* 'with' contains the *index* values of the selected and the modified NPs. This reflects the fact that the both NPs are interpreted as conjuncts.

Note that *z* provides its own INDEX value, which percolates to the mother node according to the common semantics principle of HPSG in the tradition of (Pollard and Sag, 1994). Thus, the entire PCC can control third person plural virile pronouns as well as PRO subjects.

## 4 Summary and Outlook

We have provided a lexicalist analysis of Polish PCCs assuming PCCs are head-adjunct-structures. In our future work, we will examine whether other types of CCs in Polish involving the preposition *z*, such as plural pronoun CCs (cf. (2) with the plural pronoun *my* 'we') and verb-coded CCs (cf. (2) with *pro*), can be described in a similar way.[6]

[5]For more details on a collective and distributive reading associated with (Russian) comitative constructions see (Dalrymple et al., 1998).

[6]As indicated by R1-R3, the CC in (2) has three possible interpretations. According to the first interpretation (see the translation R1), the first person plural pronoun *my* ('we') and *pro* denote a set of individuals including the speaker but not including the individual denoted by the NP selected by the preposition *z*, that is, *Maria*. In contrast, the meaning of the pronoun *my* ('we') and *pro* according to the interpretation indicated by the translation R2, includes both the denotation of *Maria* and the speaker. It does not include any further individuals, and thus carries the meaning *Maria and I*. Finally, the pronoun *my* ('we') and *pro* according to the third interpretation (see the translation R3) refer to a set of individuals including the speaker, the individual denoted by the argument of *z*, i.e., *Maria*, and some further individuals.

(2) My / *pro* z Marią odjechaliśmy.
we / *pro* with Maria left
R1: 'We left with Maria.'
R2: 'Maria and I left.'
R3: 'Maria and the rest of us left.'

We will attempt to provide a uniform treatment of all CC types in Polish that can license each of the interpretations indicated in R1-R3 and that accounts for the idiosyncratic properties of CCs.

## References


Mary Dalrymple, Irene Hayrapetian, and Tracy Holloway King. 1998. The Semantics of the Russian Comitative Construction. *Natural Language and Linguistic Theory*, 16:597–631.

Stefan Dyła and Anna Feldman. to appear. On Comitative Constructions in Polish and Russian. In *Proceedings of the Fifth European Conference on Formal Description of Slavic Languages*, Lepizig.

Stefan Dyła. 1988. Quasi-Comitative Coordination in Polish. *Linguistics*, 26:383–414.

Anna Feldman. 2002. On NP-Coordination. In Maaike Schoorlemmer Sergio Baauw, Mike Huiskes, editor, *Yearbook 2002*, pages 39–67. Utrecht Institute of Linguistics OTS.

Tania Ionin and Ora Matushansky. 2002. DPs with a Twist: A Unified Analysis of Russian Comitatives. In *Proceedings of FASL 11*, Amherst, MA.

Louise McNally. 1993. Comitative Coordination: A Case Study in Group Formation. *Natural Language and Linguistic Theory*, 11:347–379.

Carl J. Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

Maria B. Vassilieva and Richard K. Larson. 2001. The Semantics of the Plural Pronoun Construction. In R. Hastings, B. Jackson, and Z. Zvolenszky, editors, *Proceedings of Semantics and Linguistic Theory (SALT) XI*, Ithaca. CLC Publications, Dept. of Linguistics, Cornell University.

# Phrasal prenominals with peculiar properties

**Frank Van Eynde**
Centre for Computational Linguistics
University of Leuven – Belgium
frank.vaneynde@ccl.kuleuven.be

## Abstract

Drawing on data from Dutch I will demonstate that there are phrasal prenominals which cannot plausibly be analysed in terms of any of the usual subtypes of *headed-phrase*. To model them I propose a new type, called *head-independent-phrase*. Its properties are spelled out in formal detail and its range of application is illustrated with various examples, including cases of asymmetric coordination.

## 1 Introduction

In many languages, the prenominals show morpho-syntactic agreement with the nouns they modify. In Dutch, for instance, the adjectival prenominals take the nondeclined (or base) form if the noun is singular neuter and in standard case,[1] as in *elk zwart paard* 'each black horse'. Otherwise, they take the declined form, as in the singular nonneuter *elke zwarte ezel* 'each-DCL black-DCL donkey'.[2] If the prenominal is a phrase, rather than a single word, then it is the adjectival head of the prenominal which hosts the declension affix, as in the plural *zeer snelle paarden* 'very fast-DCL horses' and the singular nonneuter *een van Rusland afhankelijke staat* 'a from Russia dependent-DCL state'.

Given this generalization it is somewhat unexpected that the declension affix in the singular nonneuter *een zo groot mogelijke winst* 'an as large (as) possible-DCL profit' is not hosted by the adjective which is intuitively the head of the AP (*groot*), but by a word which is intuitively its dependent (*mogelijke*). The same phenomenon can be illustrated with superlative and comparative adjectives, as in *de grootst mogelijke verwarring* 'the largest possible-DCL confusion' and *de lager dan verwachte beurskoers* 'the lower than expected-DCL rating'.

The objective of the paper is to find out how such prenominals can best be modeled in terms of the HPSG framework.

## 2 Adjectival and participial prenominals

To pave the way I first spell out a general format for the treatment of prenominals. Following (Allegranza, 1998) and (Van Eynde, 2003), I treat the prenominals as functors. Functors are non-head daughters which select their head sister.[3] The selection is modeled in terms of a *synsem* valued feature SELECT, which is part of the functor's HEAD value.[4]

$$\left[\begin{array}{l} \textit{head-functor-phr} \\ \text{DTRS} \left\langle \left[\text{SYNSEM} \mid \text{LOC} \mid \text{CAT} \mid \text{HEAD} \mid \text{SELECT} \ \boxed{1}\right], \boxed{2} \right\rangle \\ \text{HEAD-DTR} \ \boxed{2}\left[\text{SYNSEM} \ \boxed{1}\textit{synsem}\right] \end{array}\right]$$

As demonstrated in (Van Eynde, 2003), the SELECT feature can be used to model NP-internal

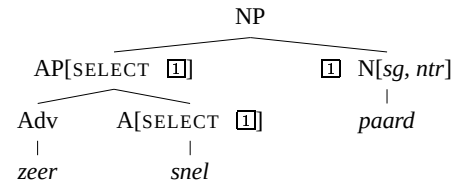
[1] Standard case is either nominative or accusative.

[2] The distinction is neutralized in NPs with a definite determiner. In such NPs, the adjectives are also declined if the noun is singular neuter, as in *het zwarte paard* 'the black-DCL horse'.

[3] The notion 'functor' is also used in a broader sense. In (Reape, 1994, 154), for instance, it covers all kinds of selectors, i.e. adjuncts, specifiers and complementizers as well as heads in head-complement combinations. In my use the term covers adjuncts, specifiers and markers, but not heads.

[4] The HEAD|SELECT feature is a generalization of the HEAD|MOD and HEAD|SPEC features of (Pollard and Sag, 1994). Non-head daughters which do not select their head sister have the SELECT value *none*. Predicative adjectives, for instance, are complements, rather than functors, and therefore have the SELECT value *none*.

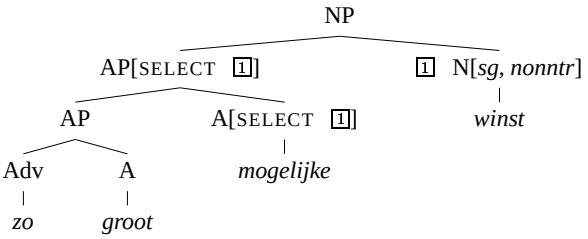
agreement. The Dutch nondeclined prenominal adjectives, for instance, can be treated as functors which select a singular neuter nominal in standard case. Since the SELECT feature is part of the HEAD value, it is shared between the mother and its head daughter. The prenominal *zeer snel* 'very fast', for instance, selects a singular neuter noun, since its head *snel* 'fast' selects a singular neuter noun.[5]



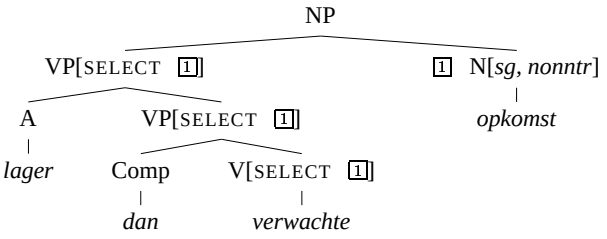

The internal structure of the AP can be modeled along the same lines: the degree marker *zeer* 'very' is an adverbial functor which selects an adjectival head.

### 2.1 A new type of headed phrases

Turning now to the exceptional AP in *een zo groot mogelijke winst* 'an as large (as) possible-DCL profit', we can apply the same approach and treat the declined AP as a functor selecting a singular nonneuter or plural noun. Moreover, since it is the rightmost word of the AP which hosts the affix, it is that daughter which is the syntactic head of the AP.




The rightmost daughter cannot only be a word, but also a phrase, as in *een lager dan verwachte opkomst* 'a lower than expected-DCL turn-out'. In this phrase, the participle is the syntactic head of a prenominal VP[*ptc*].




[5] I use the notation XP for all phrasal signs, no matter whether they are fully saturated or not.

Having granted head status to the rightmost daughter, we are now left with the problem of identifying the role of the leftmost daughter. For a start, notice that we cannot plausibly treat it as selected by the head: *lager*, for instance, is clearly not a complement or a specifier of *verwachte*. More plausible would be a functor treatment, as in the case of *zeer snel* 'very fast'. The comparatives, for instance, could be treated as functors which select a *dan* phrase as their head. The problem with this treatment, though, is its lack of generality. The same comparative would be treated as selecting a nominal in *een lager rendement* 'a lower return' and as selecting a *dan* phrase in *een lager dan verwachte opkomst* 'a lower than expected-DCL turn-out'. Similarly, in the prenominal *een lager dan verwacht rendement* 'a lower than expected return', the *dan* phrase would be the head of the comparative, while in the predicative AP of *het rendement is lager dan verwacht* 'the return is lower than expected', it would be a dependent of the comparative. This suggests that the functor treatment is not very appropriate either.

For the development of a more plausible alternative, I start from a proposal, originally made in (Van Eynde, 1998), to postulate a separate type for headed phrases in which neither daughter syntactically selects the other. As a name for the phrases of that type, I employed the term *head-independent-phrase*, but its properties were not spelled out in any detail. At this point, I will keep the name, but add a definition.

$$\left[\begin{array}{l} \textit{head-independent-phr} \\ \text{DTRS} \left\langle \left[\text{SYNSEM} \mid \text{LOC} \mid \text{CAT} \mid \text{HEAD} \mid \text{SELECT} \;\; \textit{none}\right], \boxed{2} \right\rangle \\ \text{HEAD-DTR} \;\; \boxed{2} \end{array}\right]$$

The rightmost daughter is the head and, hence, shares its HEAD value with the mother. Given the feature geometry this includes the SELECT value. As a consequence, if the phrase as a whole is a prenominal AP which selects a plural or a singular nonneuter noun, then it follows that its rightmost daughter must be declined, and this is only possible if that daughter is a word which can host the declension affix. This implies that it can be an adjective or a participle, but not an adverb or a pronoun. Combinations, such as * *de grootst ooite winst* 'the largest ever-DCL profit' and * *een groot genoege keuken* 'a large enough-DCL

kitchen' are, hence, correctly excluded.

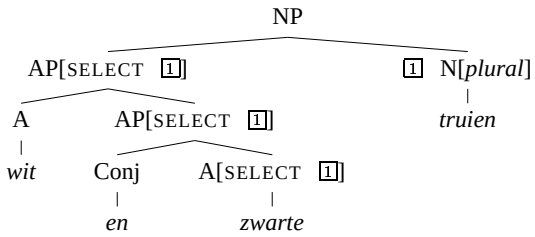
The leftmost daughter is required to have the SELECT value *none*. This constraint not only captures the difference with the functors, it also accounts for the fact that the leftmost daughter must be nondeclined. This follows from the fact that Dutch adjectives and participles are invariably nondeclined when their SELECT value is *none*. Adjectives in predicative position, for instance, never take the declension affix.

As pointed out in the introduction, the combinations which this new phrase type is intended to handle are unusual, in the sense that the syntactic head does not coincide with what is intuitively taken to be the semantic head. This discrepancy, though, is not included in the general definition of the type. The reason for this omission is that it would inadvertently exclude its application to a number of constructions which are, syntactically speaking, of the same type. The prenominal AP in *een meer dan behoorlijke opbrengst* 'a more than decent-DCL yield', for instance, has the same syntactic structure as the one in *een lager dan verwachte beurskoers* 'a lower than expected-DCL rating', but its syntactic head does coincide with what is intuitively taken to be the semantic head. Similarly, the syntactic structure of the prenominal AP in *een zo goed als nieuwe keuken* 'an as good as new-DCL kitchen' is nearly identical to the one of *een zo groot mogelijke winst* 'an as large as possible-DCL profit', but in the former the declined adjective is also the semantic head; its meaning can be paraphrased as 'an almost new kitchen'. It would, hence, be overly restrictive to require the nonhead daughter to be the semantic head. Further evidence against such a move is provided in the next paragraph.

### 2.2 Asymmetric coordination

When prenominal adjectives are coordinated they canonically take the same form. In *witte en zwarte truien* 'white-DCL and black-DCL sweaters', for instance, both conjuncts are declined. This follows from the strong version of the Coordination Principle, which requires the conjunct daughters to share the CATEGORY value of the mother (Pollard and Sag, 1994, 202). Given this requirement, one would expect the combination in *wit en zwarte truien* 'white and black-DCL sweaters' to be ill-formed, but it is not. Apparently, it is possible to limit the declension to the rightmost conjunct. Limiting it to the leftmost conjunct, however, is not possible: * *witte en zwart truien* 'white-DCL and black sweaters' is plainly ungrammatical.

What distinguishes the asymmetric coordination from the canonical symmetric coordination, is not only its form, but also its meaning. While the phrase with the symmetric coordination denotes a set of sweaters which includes both white exemplars and black ones, the phrase with the asymmetric coordination denotes a set of bi-colored sweaters. In more general terms, the symmetric coordination is distributive, whereas the asymmetric one is not. To model the latter's syntactic properties, we need a phrase type which is right-headed. Moreover, since conjuncts do not select one another, it should be a phrase type in which neither daughter syntactically selects the other. Finally, since none of the conjuncts can be claimed to be the semantic head, we need a phrase type without constraints on semantic headedness. In sum, what we need is exactly what is provided by the type head-independent-phrase. Applying it to the example yields the following structure:




Another example of this kind is *een of andere kerel* 'one or other-DCL guy'. Also here, the first conjunct lacks the affix, even though it does have a declined counterpart (*ene* 'one-DCL'), and also here the coordination is not distributive.

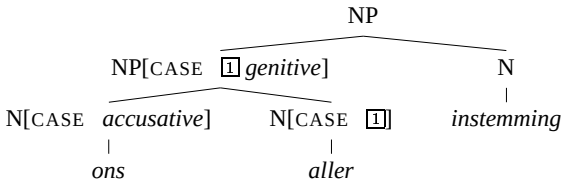
## 3 Summing up

To model the unusual prenominals I have introduced a new type of headed phrases, called *head-independent-phrase*. In such phrases the rightmost daughter is the head and neither daughter syntactically selects the other; the syntactic head may but need not coincide with what is intuitively the semantic head. The examples given so far all concern APs and nonfinite VPs, but they can also be NPs, as will be demonstrated in the next section.

## 4 Genitive and numeral prenominals

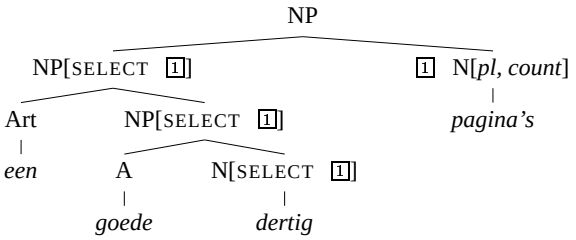
NPs in prenominal position often take a genitive affix, as in *wiens hoed* 'who-GEN hat' and *Peters vrienden* 'Peter-GEN friends'. In contrast to the adjectives and the participles, the genitive (pro)nouns do not show morpho-syntactic agreement with the nouns they modify: *Peters*, for instance, is singular and genitive, whereas its head *vrienden* 'friends' is plural and in standard case. If the genitive is a phrase, the case is marked on the head noun. In *mijn vaders vrienden* 'my father-GEN friends', for instance, the genitive affix is hosted by the noun. This is just what one expects, since CASE is a HEAD feature of nominal signs.

What is less expected, though, is that the genitive affix is sometimes hosted by a word which is intuitively not the head of the NP. Some relevant example are the prenominals in *met ons aller instemming* 'with us all-GEN consent' and *in u beider voordeel* 'in you both-GEN advantage'. In these combinations, the personal pronoun and the quantifier form a phrase which modifies the common noun.[6] The pronoun is intuitively the semantic head of the phrase, but it is the quantifier which has the genitive affix, while the pronoun is in the unmarked accusative case. To model this, we need a phrase type in which the rightmost daughter is the syntactic, but not necessarily the semantic, head. Moreover, since the relation between the pronoun and the quantifier is of a rather loose quasi-appositional kind, it should be a type of phrase in which neither daughter syntactically selects the other. In sum, what we need is again what is provided by the type *head-independent-phrase*.

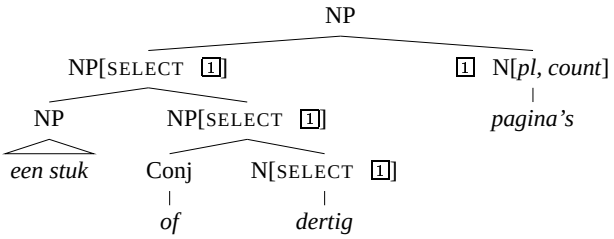



Interestingly, the general constraint that the SELECT value of the leftmost daughter be *none* has a nontrivial consequence, for since it is typical of genitive (pro)nouns that they select a nominal head and that their SELECT value is, hence, of type *synsem*, the constraint accounts for the fact that the personal pronoun cannot host the genitive affix.

Prenominal NPs are often genitive, but not always. Let us, for instance, take the numeral in *een goede dertig pagina's* 'a good-DCL thirty pages'. As all numerals, different from *één* 'one', *dertig* selects a plural count noun. At the same time, the numeral is itself a singular noun, as demonstrated by its compatibility with the indefinite article.[7] The lack of number agreement between the numeral and the noun it modifies is not exceptional: it simply follows from the fact, already illustrated above, that prenominal NPs are not subject to morpho-syntactic agreement. Employing the functor treatment the structure of the NP can be spelled out as follows.




The article and the adjective select the singular numeral as their head, and the latter in turn selects a plural count noun as the head of the prenominal. The combinations in this example are, hence, all three of the head-functor type. More complex, though, is the prenominal in *een stuk of dertig pagina's* 'a piece or thirty pages'. This prenominal, meaning something like 'around thirty', is headed by the numeral and selects a plural count noun, but cannot plausibly be analysed in terms of the head-functor type of combination: it would, for instance, make little sense to treat the numeral as selected by *een stuk*. Instead, the prenominal is a typical instance of asymmetric coordination and, hence, a candidate for treatment in terms of the head-independent phrase type.




[6] The modifier can also be realized as a postnominal PP, as in *met de instemming van ons allen* 'with the consent of us all-PL' and *in het voordeel van u beiden* 'in the advantage of you both-PL'.

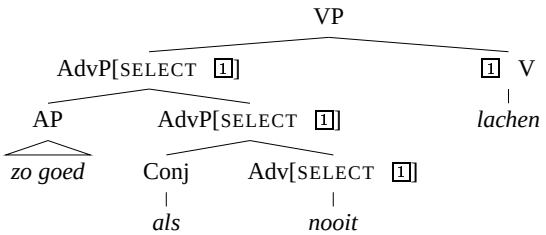
[7] For arguments that also the English numerals are nouns, see (Jackendoff, 1977).

The asymmetry is clear a.o. from the fact that the numeral selects a plural count noun, whereas the indefinite NP in the left conjunct is not even compatible with plural count nouns, as illustrated by the ungrammaticality of * *een stuk pagina's* 'a piece pages'. Further evidence for the head-independent analysis is provided by the fact that the coordination is not distributive.

## 5 An extension

So far, I have only made use of prenominal phrases to motivate and illustrate the use of the head-independent-phrase type. This limitation, though, is only due to presentational considerations. Since the prenominals tend to show more inflectional variation and to impose more detailed syntactic restrictions on their heads than most other phrases, they provide more salient and striking evidence for or against certain possible treatments; in the absence of affixation and syntactic selection restrictions, the evidence would be less direct and, therefore, less convincing.

At this point, though, where a treatment has emerged that fits the facts, we can leave the safe shore of the prenominals, venture into the open sea of clausal syntax, and discover that also there the phrases of type head-independent abound. Let us, for instance, take the frequency adverbial in *zo goed als nooit lachen* 'as good as never laugh'. The adverbial has the same internal structure and the same approximating sense as the prenominal AP in *een zo goed als nieuwe keuken* 'an as good as new-DCL kitchen'. The relation between the AP *zo goed* and the adverb *nooit* is, hence, of the same type (head-independent) as in the case of the prenominal:




Notice that it is the adverb, and not the AP, which selects the verb. XPs of type head-independent can, hence, select a verbal projection, as well as a nominal one. Moreover, they can also be used as complements. The predicative AP in *die keuken is zo goed als nieuw* 'that kitchen is as good as new', for instance, is a complement of the copula, and has the same internal structure as the adverbial above.

In sum, the phrases of type head-independent can belong to any of the usual categories (AP, VP, NP, AdvP, ...) and can have any kind of syntactic function.

## 6 Conclusion

Some prenominals, such as the AP in *de grootst mogelijke verwarring* 'the largest possible-DCL confusion', pose a challenge for a head-driven framework, since their syntactic head does not coincide with what is intuitively taken to be the semantic head. To model such combinations, I have employed a type of headed phrases, called head-independent-phrase, building on a proposal in (Van Eynde, 1998). Typical of head-independent phrases is that they are right-headed and that neither daughter syntactically selects the other.

## References

V. Allegranza. 1998. Determiners as functors: NP structure in Italian. In S. Balari and L. Dini, editors, *Romance in HPSG*, pages 55–107. CSLI Publications, Stanford.

R. Jackendoff. 1977. *X-bar syntax: a study of phrase structure*. MIT Press.

C. Pollard and I. Sag. 1994. *Head-driven Phrase Structure Grammar*. CSLI Publications and University of Chicago Press, Stanford/Chicago.

M. Reape. 1994. Domain union and word order variation in German. In J. Nerbonne, K. Netter, and C. Pollard, editors, *German in HPSG*, pages 151–197. CSLI Publications, Stanford.

F. Van Eynde. 1998. The immediate dominance schemata of HPSG. In P.-A. Coppen, H. van Halteren, and L. Teunissen, editors, *Computational Linguistics in the Netherlands 1997*, pages 119–133. Rodopi, Amsterdam/Atlanta.

F. Van Eynde. 2003. Prenominals in Dutch. In J.-B. Kim and S. Wechsler, editors, *On-line Proceedings of HPSG 2002*, pages 333–356. CSLI Publications, Stanford University.

# An HPSG Account of Closest Conjunct Agreement in NP Coordination in Portuguese

**Aline Villavicencio**
Department of Language and Linguistics
University of Essex
Wivenhoe Park
Colchester, CO4 3SQ, UK
avill@essex.ac.uk

**Louisa Sadler**
Department of Language and Linguistics
University of Essex
Wivenhoe Park
Colchester, CO4 3SQ, UK
louisa@essex.ac.uk

## 1 Introduction

This paper discusses agreement strategies inside the Noun Phrase observed in an empirical study of Portuguese. Agreement phenomena in general have received considerable attention in recent years from the HPSG community (Pollard and Sag, 1994; Wechsler and Zlatić, 2003), among others. Pollard and Sag (1994) propose what can be thought of as a multilayered theory of agreement that allows different agreement relations to hold between different objects simultaneously. This proposal, further developed by Wechsler and Zlatić (2003) is centred on the notions of CONCORD and INDEX agreement attributes. CONCORD, is closely related to the noun's inflected form and includes information about case, number, and gender, which are relevant for agreement between e.g. determiners and nouns. INDEX, is related to semantic characteristics of the noun like male or female, aggregate or non-aggregate and are defined in features such as number, gender and person, for agreement between e.g. subjects and verbs. Languages vary as to which features are used for which agreement processes. Using this framework it is possible to account for many agreement phenomena, as for example, hybrid nouns (Corbett, 1991), which can be exemplified by titles like *Majesty* in languages like Spanish and French, since they trigger different agreements on different targets within the same clause, as in sentence (1) in Spanish, where the noun, which refers to a male referent, triggers feminine agreement on the attributive adjective and masculine in the predicative adjective. They can be analysed in terms of a mismatch of INDEX and CONCORD values by specifying that a title like *Majestad* has a feminine CONCORD and a masculine INDEX and assuming that in Spanish predicate adjectives show INDEX agreement while NP-internal attributive adjective shows CONCORD agreement.

(1) Su Majestad$_i$ suprema esta contento. Él$_i$ ...
his majesty supreme.F is happy.M. He ...
'His supreme majesty is happy. He ...'

A somewhat similar situation can be found in NP (and plural noun) coordination in Portuguese, where a mixed gender coordinate structure can trigger different agreement patterns in different targets in the NP. For example, in (2), a masculine noun and a feminine noun are coordinated, and they trigger masculine agreement with the determiner and feminine with the postnominal adjective.[1]

(2) Esta canção anima os corações e
This song animate the.MPL heart.MPL and
mentes brasileiras.
mind.FPL Brazilian.FPL
'This song animates Brazilian hearts and minds.'

(3) Esta canção anima os corações e
This song animates the.MPL heart.MPL and
mentes brasileiros.
mind.FPL Brazilian.MPL

(4) Esta canção anima as/*os mentes
This song animate the.FPL/MPL mind.FPL
e corações *brasileiras/brasileiros.
and heart.MPL Brazilian.FPLMPL

[1] In these examples the adjectives and the determiners scope over the coordinate structure as a whole.

One problem with this kind of construction lies in deciding which gender the coordinate structure as a whole should have. If one decides for a given gender for the coordination, e.g. masculine, this would capture the agreement with the determiner, but not with the adjective (even though it would also account for sentences like (3) where the adjective is resolved to masculine). Therefore a second problem is how to account for the agreement with the adjective in (2), given that is has a strong correlation with the gender of the closest noun, so that sentence (4) would be ungrammatical.

These are examples of closest conjunct agreement (CCA) in noun coordinations, where a modifier/specifier of a coordinate structure agrees with only one conjunct: the one that is closest to it. Closest conjunct agreement has been discussed by Corbett (1991), Sadler (1999), Moosally (1999), Abeillé (2004) and Yatabe (2004) *inter alia*, and it is a strategy of partial agreement that can be found in many languages such as Ndebele (Moosally, 1999) and Welsh (Sadler, 1999). Moosally (1999) proposes an HPSG formalisation for capturing partial agreement in Ndebele, where agreement constraints are defined in a multiple inheritance hierarchy, and the CCA constraint is defined as:

$$\left[\begin{array}{l} \text{CONT.INDEX.GENDER} : \boxed{1} \\ \text{CONJ-DTRS} : \langle \ldots \left[ \text{INDEX.GENDER} : \boxed{1} \right] \rangle \end{array}\right]$$

capturing agreement with the last conjunct. Yatabe (2004) formalises CCA as part of a unified treatment which also deals with coordination of unlike categories. However, in order to capture cases like that in (2), it is essential to take into account information about the conjuncts in both extremities. In this paper we discuss a possible formalisation for capturing agreement patterns found in NP coordinations in Portuguese. The discussion is based on an empirical study of different agreement strategies and the requirements they pose for an HPSG treatment. Portuguese presents an interesting case study, since a number of different agreement strategies triggered by a given source can be employed at the same time, such as those in (2), where a coordination can have closest conjunct agreement between the determiner and the first coordinated noun and between the last noun and the postnominal adjective. In order to cover these cases, we propose an HPSG analysis that has access to the agreement features of the first and last conjuncts. To present that we start with a discussion of nominal agreement patterns found in NP (and noun) coordinations in Portuguese, section 2. We then look at an empirical study of closest conjunct agreement, in section 3. Finally, the account proposed in HPSG that captures these cases is presented, and the implications of adopting this approach are discussed.

## 2 Agreement in Portuguese

In Portuguese determiners and adjectives straightforwardly agree in gender and number with the noun they scope over:

(5) a/*as parede colorida/*vermelhas
the.FS/the.FPL wall.FS coloured.FSG/red.FPL
the coloured/red wall

On the other hand coordinate structures present a much wider range of agreement patterns, since coordinated nouns often jointly control agreement on determiners, adjectives and other dependents within the NP. A crosslinguistically very common pattern in a two gender system involves the syntactic principles of resolution summarized as:

(6) If all conjuncts are GEN = FEM, resolve to FEM
else, resolve to MASC

Although valid, this generalisation fails to address cases of CCA. A more complete picture is given by Moosally (1999), where agreement strategies in coordinations in Ndebele are classified in 3 types, which can also be applied to Portuguese. The first one, **Regular Agreement**, is adopted when all the coordinated NPs have the same gender, and specifiers and modifiers of the coordinate structure follow that gender (7). **Resolution Agreement** can be adopted for a conjunction of mixed gender nouns, whereby agreement is triggered by a specific feature in one of the conjuncts. For Portuguese, if there is at least one masculine noun in the coordinate structure, it can trigger agreement with a postnominal adjective, regardless of the gender of the other conjuncts (sentence 3). The third agreement strategy is that of **Closest Conjunct Agreement**, when agreement with dependents like determiners and adjectives is

triggered by the conjunct that is closest to each of them (sentences 2 and 8).

(7) a parede e a janela
the.FS wall.FS and the.FS window.FS
coloridas/vermelhas/*vermelhos
coloured.FPL/red.FPL/red.MPL
the coloured/red wall and window

(8) ... que o professor possa recontextualizar
... that the teacher may recontextualise
o aprendizado e a experiência
the.MS learning.MS and the.FS experience.FS
vividas durante a sua formação ...
lived.FP during the his training ...
from www.seed.pr.gov.br/evento_fust/

Sentence (8) in particular raises a number of interesting issues because it is a clear case where number resolution is also involved: despite the fact that the postnominal adjective scopes over the NP coordination as a whole, the feminine gender on the adjective indicates gender agreement with the closest conjunct, while plurality on the adjective indicates a resolved feature, since each NP is actually singular.

Determiners and adjectives differ as to which of these strategies they employ for agreement with NP/noun coordinations. Both **determiners** and **prenominal adjectives** scoping over all coordinated nouns, follow CCA, e.g. sentences (4) and (9), respectively. For a **postnominal adjective** modifying a coordination of mixed gender, agreement can either follow a resolution strategy, with the adjective in the masculine form (3) or it can agree with the closest conjunct (2 and 8).

(9) ...as assustadoras colinas e
...theFPL frightening.FPL mounds.FPL and
morros de argila do Parque Nacional...
hills.MPL of clay of the National Park...
from www.nationalgeographic.pt/revista/0404/wallpaper.asp

A complex picture emerges in which the three sorts of agreement coexist in the NP, triggered by different targets: determiners, prenominal adjectives and postnominal adjectives. In many cases coordinated nouns trigger agreement between the determiner and the leftmost conjunct and the postnominal adjective and the rightmost conjunct (2). This picture is confirmed in a corpus based investigation of NP internal agreement patterns in Portuguese, discussed in the next section.

## 3 A Corpus Study

To estimate the approximate frequency with which the agreement strategies are used in coordinate NPs modified by postnominal adjectives, a corpus-based investigation was performed. Of particular interest are cases that combine agreement strategies (closest conjunct agreement of gender, but semantic resolution of number).

In order to perform this analysis we searched the Web (using Google) for occurrences of coordinated NPs followed by plural adjectives. The searches used the following pattern: "<ART> * e <ART> * <ADJ>", where ART refers to Portuguese (definite and indefinite) articles, and ADJ to adjectives, which were extracted from the 1,528,590 entry NILC Lexicon (http://www.nilc.icmc.usp.br/nilc/index.html). As we want to test the correlation between the gender of each of the NPs and the gender of the adjective, only adjectives that overtly reflect gender distinction were used (9,915 masculine and 9,811 feminine adjectives). The results found are displayed in tables 1 and 2, where **Frequency** indicates the number of pages returned by Google for the searches, and **NP1, NP2** and **Adj** refer to the gender of the first conjunct, second conjunct, and adjective, respectively. The number of the NP conjuncts is indicated in each table, but the adjective is plural in both cases.

Table 1: Frequency of Adjective modifying an NP Conjunction - Conjunction of Plural NPs

| Case | Frequency | NP1 | NP2 | Adjective |
|---|---|---|---|---|
| (a) | 0 | FEM | MASC | FEM |
| (b) | 489 | FEM | MASC | MASC |
| (c) | 468 | MASC | FEM | FEM |
| (d) | 2317 | MASC | FEM | MASC |

These results give an indication of how widespread the adoption of the closest conjunct agreement strategy is, in cases (b) and (c). Case (c), in particular, provides clear evidence for CCA, where the adjective scopes over both nouns,

Table 2: Frequency of Adjective modifying an NP Conjunction - Conjunction of Singular NPs

| Case | Frequency | N1 | N2 | Adjective |
|---|---|---|---|---|
| (a) | 0 | FEM | MASC | FEM |
| (b) | 137 | FEM | MASC | MASC |
| (c) | 90 | MASC | FEM | FEM |
| (d) | 1737 | MASC | FEM | MASC |

but agrees in gender with the closest. Simultaneous CCA of gender and number resolution are shown in table 2, case (c) . These constitute unambiguous cases of number resolution, as both NPs are singular and the adjective is plural (8). Such data provides empirical evidence for the complex interrelation of agreement strategies in NP coordinate structures that need to be accounted for in a theory of agreement.

## 4 An HPSG Formalisation

To account for these cases we propose an analyses that stores agreement information about the leftmost and rightmost noun conjuncts, introducing two additional agreement attributes: LAGR, for the leftmost conjunct, and RAGR for the rightmost conjunct.

For non-coordinated nouns LAGR, RAGR and CONCORD share the same values.

$$\begin{bmatrix} \text{VAL.SPR.HEAD.CONCORD} : \boxed{1} \\ \text{HEAD} : \begin{bmatrix} \text{LAGR} : \boxed{1}/_p\boxed{2} \\ \text{RAGR} : /_p\boxed{2} \\ \text{CONCORD} : /_p\boxed{2} \end{bmatrix} \end{bmatrix}$$

Determiners and prenominal adjectives agree with nouns via LAGR, while postnominal adjectives agree with nouns via RAGR, all having the same value but via different attributes. For a coordinate structure, the values of LAGR, RAGR and CONCORD may differ. As these reentrancies between them are defined as persistent default specifications (Lascarides and Copestake, 1999), noun coordinate structures override these defaults and instead define that LAGR values are reentrant with those of the leftmost conjunct and RAGR with those of the rightmost conjunct.

We assume that the CONCORD of the coordinate structure reflects the resolved gender for the whole conjunct, adopting a resolution approach like that of Wechsler and Zlatić (2003) or of Dalrymple and Kaplan (2000). Kaplan and Dalrymple, for example, use marker sets to represent gender information and this approach gives the desired result that, if there is at least a masculine noun in the coordinate structure, CONCORD.GENDER is masculine. As determiners and prenominal adjectives in Portuguese agree with the leftmost noun closest to them, the agreement is by coindexation with LAGR. Postnominal adjectives, on the other hand, agree with the rightmost noun closest to them (via RAGR), or adopt a resolved agreement (via CONCORD).

For sentences like (2 and 3), both LAGR and CONCORD are masculine and RAGR is feminine and the correct agreement values are observed, since the adjective can either agree with RAGR or CONCORD, but it will correctly rule out sentence (4) as ungrammatical. This formalisation can also capture a sentence like (8), which has CCA for gender, but resolved number agreement for the postnominal adjective.

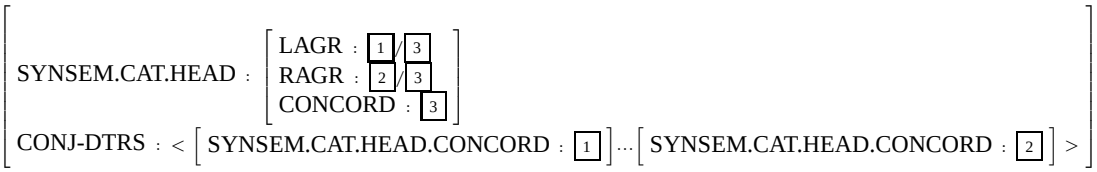
The coordination of mixed gender singular nouns sharing a determiner seems to be much more constrained. However, acceptable occurrences can be found as the sentence below, from one of the classic exponents of Brazilian literature, *O Guarani* by Machado de Assis, exemplifies, where the relevant parts are in bold font:

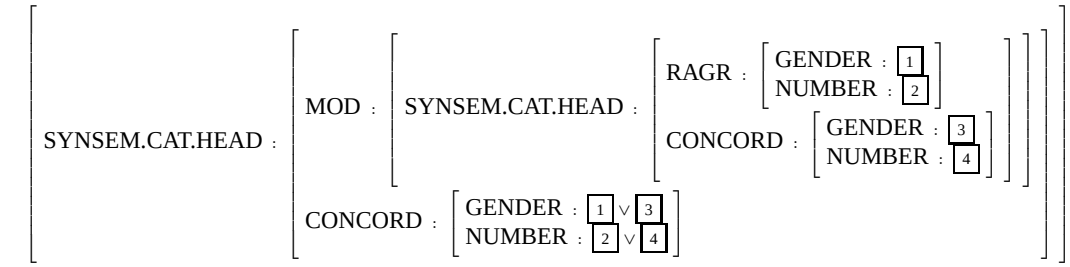
*D. Antônio tinha ajuntado fortuna durante os primeiros anos de sua vida aventureira; e não só por capricho de fidalguia, mas em atenção à sua família, procurava dar a essa habitação construída no meio de um sertão, todo* ***o luxo e comodidade possíveis****.*

The principles that allow cases like this need to be further investigated, and will not be addressed in this paper.

## 5 Conclusions

In this paper we discussed agreement processes found in NP/noun coordinations in Portuguese, further investigated through an extensive empirical study. Although we concentrated on gender and number agreement between nouns and their dependents in Portuguese, some similar strategies can also be found in other languages like Spanish and Arabic (Camacho, 2003), and this proposal could be used as the basis for a cross-linguistic formalisation of these agreement processes. In order to cap-




ture these cases, we proposed an HPSG formalisation that stores information about the leftmost and the rightmost conjuncts, which, together with the resolved CONCORD feature, control agreement between nouns, prenominal (determiners and adjectives) and postnominal (adjectives) dependents. This formalisation successfully captures the cases found in the empirical study, and correctly rules out ungrammatical combinations.

## Acknowledgements

This research was supported by the Noun Phrase Agreement and Coordination AHRB Project MRG-AN10939/APN17606.

## References

Abeillé, A., 2004. A lexicalist and construction-based approach to coordinations. In: Müller, S. (Ed.), Proceedings of the HPSG04 Conference. CSLI Publications, Katholieke Universiteit Leuven.

Camacho, J., 2003. The Structure of Coordination: Conjunction and Agreement Phenomena in Spanish and Other Languages. Kluwer Academic Publishers, Dordrecht.

Corbett, G. G., 1991. Gender. Cambridge University Press, Cambridge, UK.

Dalrymple, M., Kaplan, R. M., 2000. Feature indeterminacy and feature resolution. Language 76 (4), 759–798.

Lascarides, A., Copestake, A., 1999. Default representation in constraint-based frameworks. Computational Linguistics 25 (1), 55–105.

Moosally, M. J., 1999. Subject and object coordination in Ndebele: and HPSG analysis. In: Bird, S., Carnie, A., Haugen, J. D., Norquest, P. (Eds.), Proceedings of the WCCFL 18 Conference. Cascadilla Press.

Pollard, C., Sag, I. A., 1994. Head-Driven Phrase Structure Grammar. The University of Chicago Press, Chicago, IL.

Sadler, L., 1999. Non-distributive features and coordination in Welsh. In: Butt, M., King, T. H. (Eds.), On-line Proceedings of the LFG99 Conference.

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

Yatabe, S., 2004. A comprehensive theory of coordination of unlikes. In: Müller, S. (Ed.), Proceedings of the HPSG04 Conference. CSLI Publications, Katholieke Universiteit Leuven, pp. 335–355.