Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

http://csli-publications.stanford.edu/HPSG/2015

Contents

1 Editor's Note

The 22th International Conference on Head-Driven Phrase Structure Grammar (2015) was held at the Nanyang Technological University (NTU), Singapore.

The conference featured 3 invited talks, 15 papers, and 4 posters selected by the program committee (Anne Abeillé, Farrell Ackerman, Doug Arnold, Emily M. Bender (chair), Francis Bond, Gosse Bouma, George Broadwell, Rui Chaves, Philippa Cook, Ann Copestake, Kordula De Kuthy, Elisabet Engdahl, Dan Flickinger, Antske Fokkens, Danièle Godard, Petter Haugereid, Fabiola Henri, Anke Holler, Jong-Bok Kim, Jean-Pierre Koenig, Anna Kupsc, Bob Levine, Janna Lipenkova, Rob Malouf, Nurit Melnik, Philip Miller, Tsuneko Nakazawa, Joanna Nykiel, Gerald Penn, Adam Przepiorkowski, Frank Richter, Louisa Sadler, Pollet Samvellian, Sanghoun Song, Jesse Tseng, Steve Wechsler, Shûichi Yatabe and Eun-Jung Yoo).

A workshop on *Verb Classes and the Scale of Change in Affected Arguments* was attached to the conference. The workshop had six invited speakers and five regular papers. The workshop program was put together by František Kratochvíl and Joanna Ut-Seong Sio.

We want to thank the respective program committees for putting this nice program together.

Thanks go to Francis Bond, who was in charge of local arrangements, and his assistants Sanghoun Song, Michael Goodman, Luis Morgado da Costa.

As in the past years the contributions to the conference proceedings are based on the five page abstract that was reviewed by the respective program committees, but there is no additional reviewing of the longer contribution to the proceedings. To ensure easy access and fast publication we have chosen an electronic format.

The proceedings include all the papers except the one by Sanghoun Song, Chen Bo, Joanna Sio Ut Seong, and Francis Bond titled *An HPSG-based Analysis of Resultative Compounds in Chinese*, the one by Frank van Eynde, which is published in his book on predication, and the ones by I Wayan Arka, Kazuko Yatsuhiro, Juwon Lee, and Hans Uszkoreit. Most of the workshop contributions will be published in a separate volume and therefore are also not included.

Part I
Contributions to the Main Conference

An Analysis of Simple and Construct-State Noun Phrases in Modern Standard Arabic

Issa S. AlQurashi
Taif University

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 6–26

http://csli-publications.stanford.edu/HPSG/2015

AlQurashi, Issa S. 2015. An Analysis of Simple and Construct-State Noun Phrases in Modern Standard Arabic. In Müller, Stefan (Ed.), *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, 6–26. Stanford, CA: CSLI Publications.

Abstract

This paper aims to propose an HPSG analysis for simple and construct-state noun phrases in Modern Standard Arabic (MSA). To the best of my knowledge, there are no major HPSG analyses of MSA noun phrases (NPs). A parallel phenomenon in Hebrew has been discussed quite extensively in the same framework by Wintner (2000). Most of the discussion will be devoted for the construct-state noun phrase in which the order of the elements within it is NP AP PP. Three analyses will be outlined within the HPSG framework: the extra complement analysis, the special complement analysis, and the head-adjunct-complement analysis. These analyses will be evaluated and it will be concluded that the last analysis seems to be the best and the most promising approach to Arabic NPs.

1. Data

Simple MSA noun phrases can be definite or indefinite. Definite nouns are prefixed with the definite article (*al-*) -glossed ‘DEF’- (see, for example, Ouhalla, 1991; Fassi Fehri, 1993; Ryding, 2005; Benmamoun, 2006, among others), and indefinite nouns are suffixed with the indefinite marker (-n) - glossed ‘INDEF’- (see, for example, Ryding, 2005, among others) - as in (1).

(1) a. ʔal-kitaab-u
DEF-book-NOM[1]
‘The book’
b. kitaab-u-n
book-NOM-INDEF
‘a book’

MSA also has construct state nouns consisting of a head noun directly followed by a possessor. The head/construct noun can carry neither the definite article (*al-*) as in (2a), nor the indefinite marker (-n) as in (2b) (Ouhalla, 1991; Fassi Fehri, 1993; Benmamoun, 2006; Ryding, 2005), but the form of a modifying adjective (which follows the possessor) shows that the nouns agrees with the possessor in definiteness.

† I am grateful to reviewers and audience at HPSG22 in Singapore for their helpful comments and discussion. Remaining flaws are purely my fault.

[1] The nominative case is the citation form in MSA.

(2) a. [(*al)-kitaab-u T-Taaliba-t-i]
DEF-book.SG.MASC-NOM DEF-student.SG-FEM-GEN
l-jadiid-u
DEF-new.SG.MASC-NOM
'the female student's new book'
b. [kitaab-u-(*n) Taaliba-t-i-n]
book.SG.MASC-NOM-INDEF student.SG-FEM-GEN-INDEF
jadiid-u-n
new.SG.MASC-NOM-INDEF
'a (female) student's new book'

Adjectives in MSA agree in definiteness, gender, number, and case with the noun they modify. The form of the adjective in (2a) shows that the noun is definite although it does not bear the definite article, and the form of the adjective in (2b) shows that the noun is indefinite although it does not have the indefinite suffix. It should also be noted that the adjective in both examples modifies the head noun but not the possessor. This is because of the gender agreement between the adjective and the head noun.

An adjective cannot precede the possessor as the following example demonstrates:

(3) kitaab-u (*l-qayyim-u) l-muʔallifa-t-i
book.SG.MSAC-NOM DEF-valuable DEF-author.SG-FEM-GEN
l-qayyim-u
DEF-valuable.SG.MASC-NOM
'the author's valuable book'

In addition to the attributive adjective and the possessor, the construct-state noun can have a PP or a clause as a complement. Consider the following example showing a PP complement:

(4) kitaab-u l-muʔallifa-t-i
book.SG.MSAC-NOM DEF-author.SG-FEM-GEN
l-qayyim-u fii n-naHw-i
DEF-valuable.SG.MASC-NOM in DEF-syntax-GEN
(*l-qayyim-u)
DEF-valuable.SG.MASC-NOM
'the erudite author's valuable book about syntax'

Any such complement appears after the possessor and the adjective. This means that the order has to be NP AP PP. If a relative clause is used, it will occur after the ordinary complement as in (5) below:

(5)	kitaab-u	siibawayh-i	l-qayyim-u	fii	n-naHw-i
	book-NOM	Siibawaih-GEN	DEF-valuable-NOM	in	DEF-syntax-GEN
	[ʔallaðii	ʔahdayta-nii	ʔiyyaah]		
	that.SG.MASC	give present-me	it		

‘Siibawaih’s valuable book about syntax which you gave me as a present’

The examples in (4) and (5) show the most important facts in this paper and hence they will be the central focus of the analysis.

As for the complement selection possibilities of definite and indefinite nouns, they both allow a complement (PP) following the attributive adjective (just like construct-state nouns above) as shown in the following examples:

(6)	a.	qaraʔ-tu	kitaab-a-n	jadiid-a-n	fii
		read.PAST-1SG	book-ACC-INDEF	new-ACC-INDEF	in
		n-naHw-i			
		DEF-syntax-GEN			
		‘I read a new book about syntax’			
	b.	qaraʔ-tu	l-kitaab-a	l-jadiid-a	[fii
		read.PAST-1SG	the-book-ACC	DEF-new-ACC	in
		n-naHw-i]			
		DEF-syntax-GEN			
		‘I read the new book about syntax’			

These differences between definite and indefinite nouns on the one hand, and construct state nouns on the other hand will be captured by appropriate constraints in the following section.

2. Analysis

2.1. Basics

I will begin with the treatment of possessors, and the constraints on the three types of noun (def, indef, and construct). After that, I will discuss the status and position of attributive adjectives.

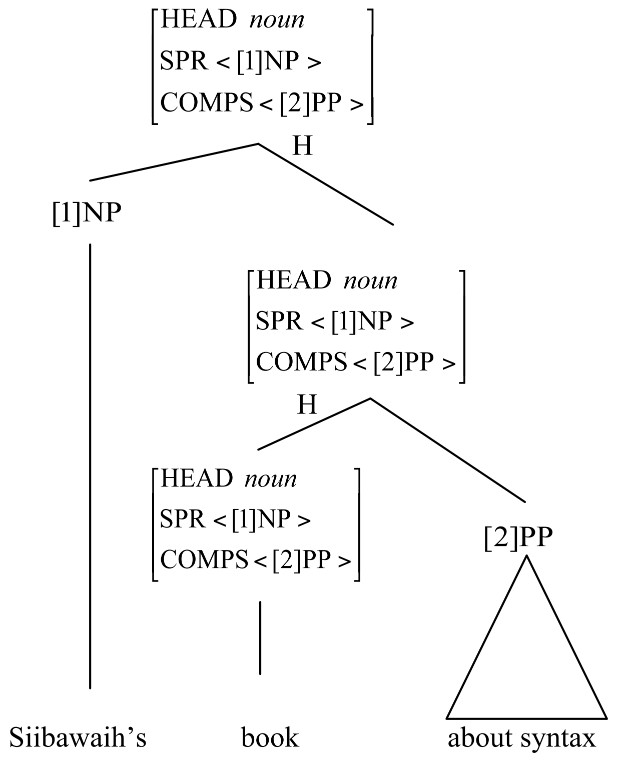
2.1.1 The possessor

In HPSG analyses (Sag, Wasow and Bender, 2003), possessors in English are analysed as realisations of the SPR (SPECIFIER) feature, giving categories like (7) and structures like (8).

(7)

$$\begin{bmatrix} \text{HEAD } \textit{noun} \\ \text{SPR} < [1] > \\ \text{COMPS} < [2] > \\ \text{ARG-ST} < [1]\text{NP}, [2]\text{PP} > \end{bmatrix}$$

(8)



Unlike English, I treat the possessor in MSA as an extra complement of the head noun rather than a realisation of the SPR feature, as is clear from the COMPS' list of the head daughter.This position is taken by Borsley (1989, 1995) for Welsh and Arabic, and by Wintner (2000) for Hebrew. Borsley based his arguments on the fact that possessors always follow the associated noun and can be realised as clitics like the objects of verbs (9a) and prepositions (9b).

(9) a. fahd-un ra?aa-haa
Fahd-NOM see.PAST.3SG.MASC-her
'Fahd saw her.'
b. maʕa-haa
with-her
'with her'

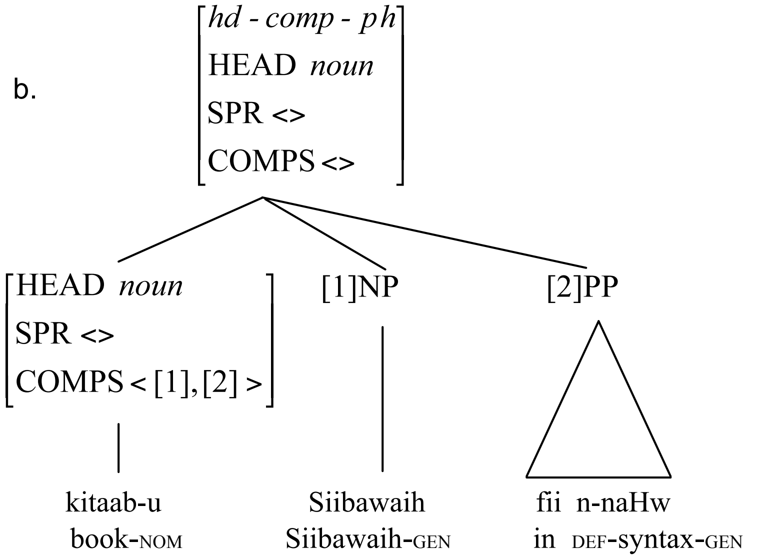
With verbs and prepositions clitics realise what is an uncontroversial complement. This suggests they also realise a complement with nouns and hence that possessors are complements. An example where a possessor is realised as a clitic is shown below:

(10) kitaab-u-hu fii n-naHw-i
book-NOM-her in DEF-syntax-GEN
'his book about syntax'

The following tree represents the structure of an example with an ordinary possessor.

(11) a. kitaab-u siibawayh-i fii n-naHw-i
book-NOM Siibawaih-GEN in DEF-syntax-GEN
'Siibawaih's book about syntax'

b.

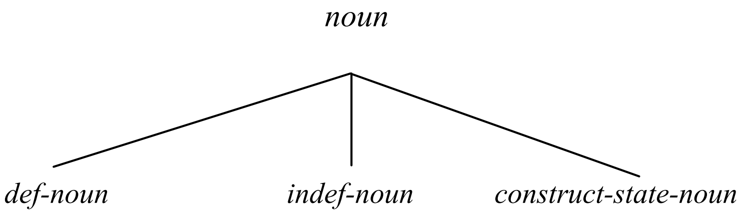


Next, I discuss the constraints to which the subtypes of the type *noun* are subject in the following section.

2.1.2 Constraints on subtypes of nouns

The following is the type hierarchy of some nouns (the type hierarchy not only includes *def-noun* and *indef-noun*, but also *construct-state-noun* subtype as seen in (12):

(12)



We can have the following constraints for each subtype:

(13)
def-noun → [DEF +]
indef-noun → [DEF -]
construct-state-noun → [DEF *boolean*]

The type *noun* has the subtypes *def-noun, indef-noun,* and *construct-state-noun*. Each subtype is associated with some features. The subtype *def-noun* is [DEF +], which means that the noun is definite and marked with the definite article. As for the subtype *indef-noun*, it has the feature [DEF -], which means that the noun is indefinite and marked with the indefinite marker. The last subtype of the type *noun* is *construct-state-noun*, which is associated with the features [DEF *boolean*]. The feature [DEF *boolean*] indicates that the construct noun is unspecified for definiteness and it could be [DEF +] or [DEF -], but this does not mean that the noun is morphologically marked as such. The morphological rules that introduce the definite prefix and the indefinite suffix must not apply to construct-state-nouns. The construct noun can be definite without the attachment of the definite article or can be indefinite without the indefinite marker as the data show in (2) above.

The type *noun* is subject to the following constraints:

(14)

$$\textit{noun} \rightarrow \begin{bmatrix} \text{SYNSEM} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{noun} \\ \text{DEF}\,[1] \\ \text{NUM}\,[2] \\ \text{GEND}\,[3] \\ \text{CASE}\,[4] \end{bmatrix} \\ \multicolumn{2}{l}{\text{ARG-ST}\,[5] < (\text{NP}) > \oplus [6]} \\ \multicolumn{2}{l}{\text{DEPS}\ [5] \oplus \underset{\begin{bmatrix} \text{DEF}\,[1] \\ \text{NUM}\,[2] \\ \text{GEND}\,[3] \\ \text{CASE}\,[4] \end{bmatrix}}{\text{AP}\,*} \oplus [6]} \end{bmatrix} \end{bmatrix}$$

$$\textit{noun} \rightarrow \begin{bmatrix} \textit{COMPS}\ [2] - \textit{list}\,(\,\textit{noncanon-ss}\,) \\ \text{DEPS}\ [2] \end{bmatrix}$$

The AGR-ST list involves two lists. The first list consists of NP or nothing (since the possessor is optional) and the second list is a (real) semantic argument such as PPs or clauses. The optionality of the possessor is indicated by the use of the parentheses. Only construct-state nouns have possessors as shown in (2) above.

The other member of the ARG-ST list is tagged by number [6], which can be a prepositional phrase or a clausal complement. The possessor and the other member of the ARG-ST list also appear on the DEPS list. In addition, following Bouma, Malouf, and Sag's (2001) approach to adverbials as will be discussed in § 2.2 below, APs appear on the DEPS list after the possessor (if there is one) and before any ordinary complements. This constraint plus the one, discussed below, which says that the value of COMPS is DEPS ensure that optional adjectives are complements. The asterisk sign (*) on AP means that we can have any number of APs (including none).

The second constraint says that the value of COMPS is DEPS minus any noncanonical-*synsem* objects in the DEPS list.[2] I am assuming the Miller and Sag's (1997) approach to clitics in which they are affixes realizing an affixal *synsem* object. The view in Miller and Sag and elsewhere is that *synsem* objects may be canonical, in which case they will appear in COMPS lists, or noncanonical, in which case they will not appear

[2] This would have to go if affixal *sysnems* appear in COMPS lists. Instead, the constraint of head-comp-phrases will have to ensure that only canonical *synsems* are realized as complements.

there. Noncanonical synsem objects include unbounded dependency gaps and arguments realized as affixes. If a DEPS list contains a canonical possessor, it will also appear in the COMPS list. If it contains an affixal possessor, it will not appear in the COMPS list, but the noun will have the appropriate suffix. I will assume the same sort of approach as Miller and Sag (1997), as I will do for definite and indefinite suffixes below.

The first constraint in (13) above also says that the value of HEAD feature is a feature structure that has a number of agreement features: DEF, NUM, GEND, and CASE. The constraint guarantees that the values of those features are identical to the values of the similar features of the modifying adjectives. This is— as I mentioned in §1 above —because adjectives in MSA agree in number, gender, case, and definiteness with the noun they modify as the following examples demonstrate:

(15) a. ʔT-Taalib-u l-mujtahid-u
DEF-student.SG.MASC-NOM DEF-diligent.SG.MASC-NOM
'the diligent (male) student'
b. ʔT-Taalib-aat-u l-mujtahid-aat-u
DEF-student-PL.FEM-NOM DEF-diligent-PL.FEM-NOM
'the diligent (female) students'

The subtype *indef-noun* is subject to the following constraint:

(16)

$$\textit{indef-noun} \rightarrow \begin{bmatrix} \text{MORPH} \begin{bmatrix} \text{FORM } \mathrm{F}_{\textit{indef}}\,[1] \\ \text{I-FORM}\,[1] \end{bmatrix} \\ \text{SYNSEM} \begin{bmatrix} \text{HEAD}\,[\text{DEF} -] \\ \text{ARG-ST } \neg < \text{NP}... > \end{bmatrix} \end{bmatrix}$$

The constraint in (16) contains MORPH and SYNSEM features. The MORPH feature has two features: FORM and I-FORM, which are taken from Miller and Sag (1997). The I-FORM is the inflectional form of the noun without the indefinite marker. A noun will have various values for I-FORM depending on its case and whether it is singular or plural. The value of FORM is the noun suffixed with the indefinite marker. The function F_{indef} adds the indefinite marker to the inflectional form of the noun. As for the SYNSEM feature, it has the indefinite marker because it is indefinite. The ¬ <NP…> stipulation ensures that a noun bearing the indefinite marker does not have an ARG-ST list whose first member is a possessor. This means that the indefinite noun can have an ARG-ST list which may contain other members such as PPs and clausal complements but not a possessor.

The subtype *def-noun* is subject to the following constraint:

(17)

$$\textit{def-noun} \rightarrow \begin{bmatrix} \text{MORPH} \begin{bmatrix} \text{FORM F}_{\textit{def}}[1] \\ \text{I-FORM}[1] \end{bmatrix} \\ \text{SYNSEM} \begin{bmatrix} \text{HEAD}[\text{DEF} +] \\ \text{ARG-ST } \neg < \text{NP}... > \end{bmatrix} \end{bmatrix}$$

Again the features FORM and I-FORM in (17) are not identified. The function F_{def} adds the definite article to a basic form of the noun which marks it as definite. Hence, the value of DEF feature is [+]. The ¬ <NP…> stipulation ensures that a noun bearing the definite article does not have an ARG-ST list whose first member is a possessor. This means that the definite noun can have an ARG-ST list which may contain other members such as PPs, but not a possessor as shown in (6b) and (2a) above.

The subtype *construct-state-noun* is subject to the following constraint:

(18)

$$\textit{construct-state-noun} \rightarrow \begin{bmatrix} \text{MORPH} \begin{bmatrix} \text{FORM}[1] \\ \text{I-FORM}[1] \end{bmatrix} \\ \text{SYNSEM} \begin{bmatrix} \text{HEAD}[\text{DEF}[1]] \\ \text{ARG-ST} < \underset{\begin{bmatrix} \text{DEF}[1] \\ \text{CASE } \textit{gen} \end{bmatrix}}{\text{NP}} \; ... > \end{bmatrix} \end{bmatrix}$$

The constraint in (18) says that the values of the FORM and I-FORM features are identified. This ensures that a construct-state-noun has neither a definite prefix nor an indefinite suffix. Furthermore, the constraint guarantees that the construct-state noun has an ARG-ST list whose first member is a possessor, which is genitive and has the same value for DEF as the head noun. It thus requires definiteness agreement between the head noun and the possessor.

In the following sections. I will be concerned with how attributive adjectives should be analyzed and especially how they can be correctly positioned after possessors and before ordinary complements.

2.2. Attributive adjectives as complements

Attributive adjectives are standardly analysed as modifiers combining with a nominal constituent to form a larger nominal constituent. It is fairly easy to apply this approach to Welsh and Persian (see Samvelian, 2007, for more details in Persian) in which attributive adjectives precede both possessors and ordinary complements. Take the following example for Welsh in (19):

(19) llyfr newydd Megan am gystrawen
book new Megan about syntax
'Megan's new book about syntax' Borsley (pc)

Therefore, it can be assumed that adjectives modify nouns and that the result combines with whatever complements it requires**.**

If we propose the adjunct/modifier analysis for MSA, it will run into the problem of ordering the adjective between the possessor and the ordinary complements as in the following example:

(20) maqaal-u l-kaatiba-t-i l-jayyid-u
article-NOM DEF-writer-FEM-GEN DEF-good-NOM
ʕani l-ʔirhaab-i
about DEF-terrorism-GEN
'the writer's good article about terrorism'

It is not clear how the adjective *l-jayyid* 'the good' can be ordered in between the possessor *al-kaatibati* 'the writer' and the PP complement *ʕani l-ʔirhaabi* 'about the terrorism' in an adjunct/modifier analysis. If attributive adjectives are noun modifiers they will precede possessors. If they are NP modifiers they will follow ordinary complements, which are not the right positions of attributive adjectives in MSA as demonstrated in examples (3) and (4) above.

Consequently, a different approach is necessary for MSA. One possibility is that attributive adjectives are optional extra complements since they are preceded and followed by elements which are analyzed as complements (possessors and ordinary complements, respectively). Treating adjectives as extra complements is rather like the approach taken to verbal adjuncts (particularly postverbal adverbs) in Bouma, Malouf, and Sag (2001). They argue that in English, postverbal adjuncts are extra complements of the verb. However, to distinguish them from ordinary arguments such as PP, we suggest that adjectives like English postverbal adverbs in Bouma et al.'s analysis do not appear in ARG-ST lists, but appear in DEPS lists and COMPS lists, as I indicated in § 2.1.2 above.

To ensure that attributive adjectives do not appear as adjuncts modifying N or NP in head-adjunct structures, we could impose a restriction on the type *head-adjunct phrase* excluding a nominal head, as in the following constraint:

(21) $$\textit{head-adjunct-ph} \rightarrow \begin{bmatrix} \text{HEAD} \neg \begin{bmatrix} \textit{noun} \\ \text{LEX} + \end{bmatrix} \end{bmatrix}$$

This says that a *head-adjunct-ph* cannot be a noun that is [LEX +]. Thus, a nominal head is excluded. However, this will only prevent adjectives from modifying a noun and coming before the possessor. We also need to prevent adjectives from modifying NP and coming after a complement. Probably the best thing to do is to assume that adjectives are [MOD *none*] and hence they don't modify anything.

There is one important objection to this analysis. Treating attributive adjectives as extra complements makes them different from relative clauses (assuming the latter are adjuncts). However**,** they are like relative clauses in reflecting the definiteness of the modified noun. To remind the reader of how adjectives reflect the definiteness of the modified noun, as shown in examples (15) and (2) above, I give the following examples:

(22) a. ʔal-walad-u ð-ðakiyy-u
DEF-boy.SG-NOM DEF-clever.MASC.SG-NOM
'the clever boy'
b. walad-u-n ðakiyy-u-n
boy.SG-NOM-INDEF clever.MASC.SG-NOM-INDEF
'a clever boy'

Adjectives modifying a definite NP appear with the definite article while adjectives modifying an indefinite NP appear with an indefinite marker. The definiteness agreement of relative clauses with the associated nominal is shown on the head of the relative clause (the complementizer). Relative clauses modifying a definite NP are introduced by a complementizer whereas relative clauses modifying an indefinite NP lack a complementizer as the following examples show:

(23) a. raʔay-tu r-rajul-a *(llaðii) qaabal-tu-hu
see.PAST.1SG DEF-man-ACC that.SG.MASC meet.PAST-1SG-him
bi-l-ʔams
in-DEF-yesterday
'I saw the man whom I met yesterday'
b. raʔay-tu rajul-a-n (*llaðii) qaabal-tu-hu
see.PAST.1SG man-ACC-INDEF that.SG.MASC meet.PAST-1SG-him
bi-l-ʔams
in-DEF-yesterday
'I saw a man whom I met yesterday'

In the following section, I will propose a different approach in which a possessor is treated differently.

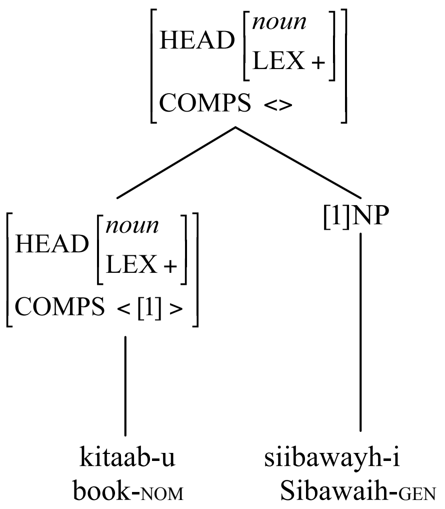
2.3. Possessors as special complements

A second way to ensure the correct positioning of adjectives is to assume that they modify a noun but to treat possessors as special complements with which the noun combines to form a complex noun. This requires a special type, which might be called a construct-state-noun, subject to the following constraint:

$$(24)\qquad c\text{-}s\text{-}n(oun)\rightarrow \begin{bmatrix} \text{HEAD}[\text{LEX}+] \\ \text{COMPS}[3] \\ \text{DTRS} < [1]\begin{bmatrix}\text{HEAD}\ \textit{noun} \\ \text{COMPS} < [2] > \oplus [3]\end{bmatrix}, [2]\text{NP}[\text{CASE}\ \textit{gen}] > \\ \text{HD-DTR}\ [1] \end{bmatrix}$$

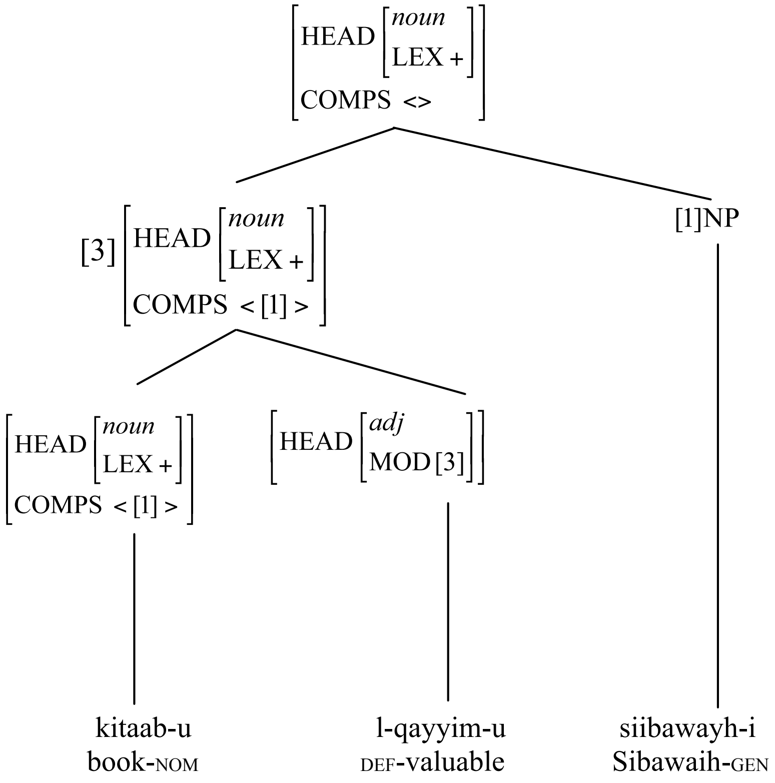
The constraint states that a construct state noun is [LEX +], and has a nominal head daughter and a genitive NP non-head daughter which is the first item on the COMPS list of the head and that the COMPS value of the phrase is identical to the remainder of the head's COMPS list. This will give structures like the following in (25) for the example in (5) above (the tree shows the structure of the head noun and the possessor only):

(25)



If there is an adjective modifying the head noun, it will be able to combine with a noun either before (as in the structure in (26) which is grammatical for Welsh and Persian, but not for MSA) or after the possessor.

(26) *


HEAD [noun LEX +] COMPS <>
[3] HEAD [noun LEX +] COMPS < [1] >
[1]NP
HEAD [noun LEX +] COMPS < [1] >
HEAD [adj MOD [3]]
kitaab-u
book-NOM
l-qayyim-u
DEF-valuable
siibawayh-i
Sibawaih-GEN

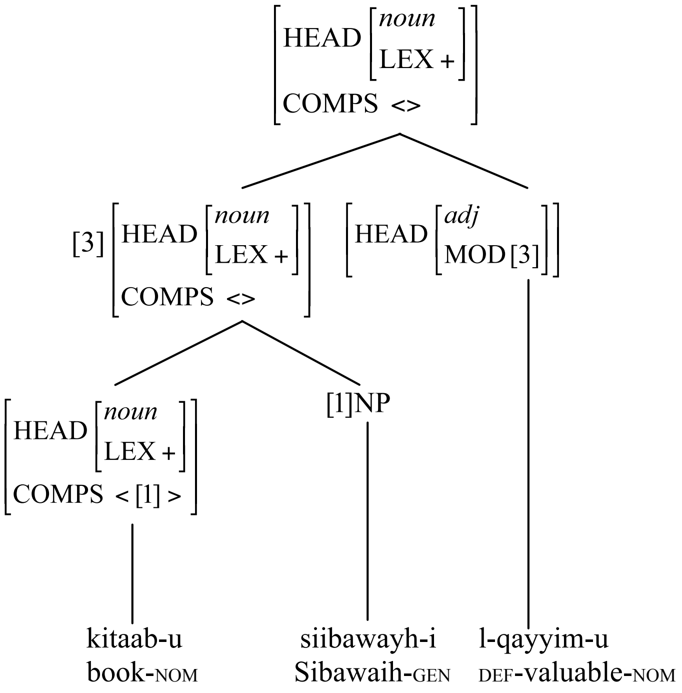
To prevent an adjective modifying the head noun and intervening between the head noun and the possessor, we could stipulate that adjectives are [MOD N [COMPS ¬ <NP, …>]] so that they can only modify nouns which do not require a nominal complement (possessor). So, the constraint on adjectives will look like the following:

(27)

$$
\textit{adj} \rightarrow \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{adj} & \\ \text{DEF}\,[1] & \\ \text{NUMB}\,[2] & \\ \text{GEND}\,[3] & \\ \text{CASE}\,[4] & \\ \text{MOD} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{noun} \\ \text{LEX} + \\ \text{DEF}\,[1] \\ \text{NUMB}\,[2] \\ \text{GEND}\,[3] \\ \text{CASE}\,[4] \end{bmatrix} \\ \multicolumn{2}{l}{\text{COMPS} \neg < \text{NP}, \ldots >} \end{bmatrix} \end{bmatrix} \end{bmatrix}
$$

With the constraint in (27), the grammatical version of the structure in (25) can be licensed in (28):

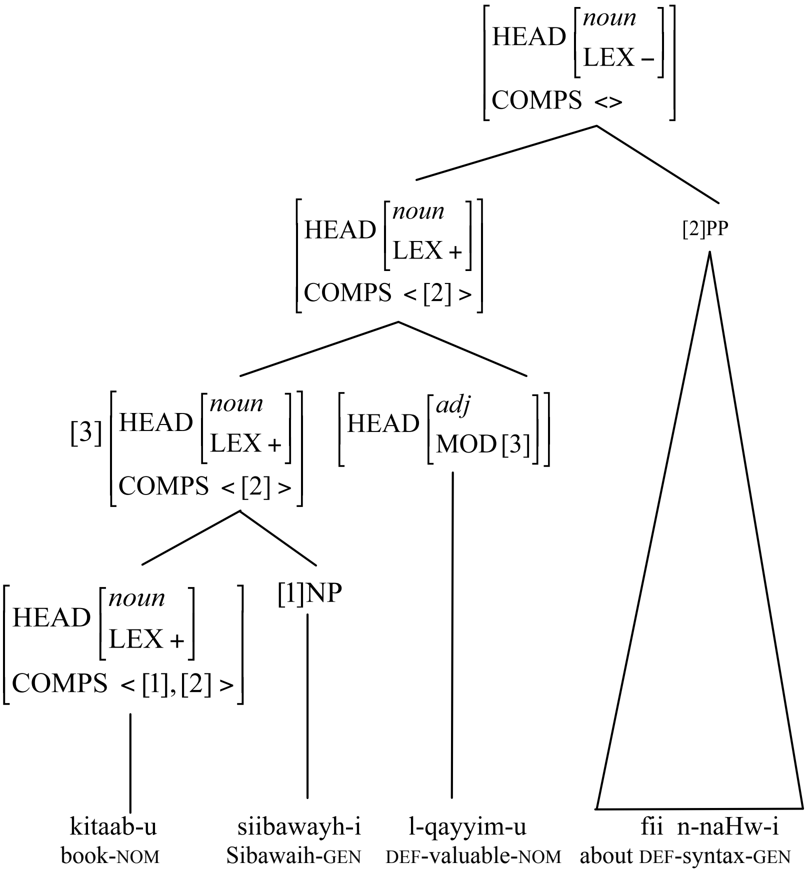
(28)



Given the treatment of the possessors, we will have structures like the following in (30) for an example with a possessor, an adjective, and a PP complement given in (5) above repeated here for convenience (without the relative clause) in (29):

(29) kitaab-u siibawayh-i l-qayyim-u fii n-naHw-i
book-NOM Siibawaih-GEN DEF-valuable-NOM in DEF-syntax-GEN
'Siibawaih's valuable book about syntax'

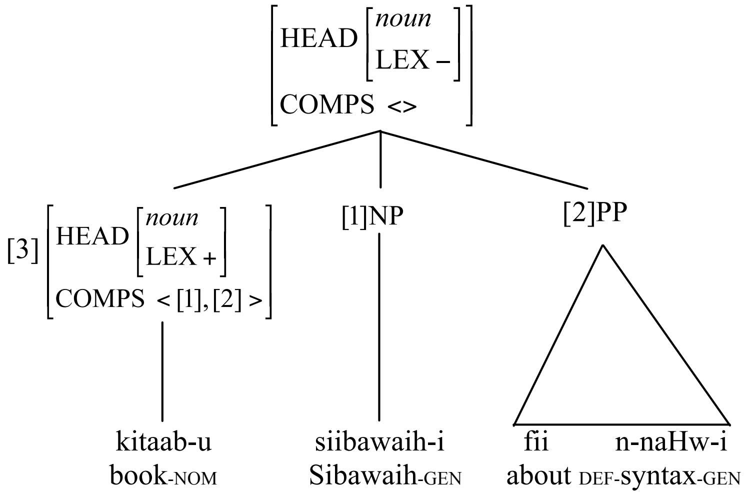
(30)



The combination of the head noun and the possessor is licensed by the constraint in (24) above, and the combination of the head noun and the PP complement is licensed by the head-complement-phrase. The combination of construct state phrase and adjective is licensed by the constraint on head-adjunct structures.

However, a question arises as to what rules out a structure like the following (without the adjective):

(31)



We should rule out (31) because we want to avoid two structures for unambiguous expressions. Since (31) is an ordinary head-complement-phrase in which the noun is [COMPS <NP, …>], We can stipulate that a nominal head of a head-complement-phrase is [COMPS ¬ <NP, …>]. This ensures that the first member of the COMPS list is not a possessor. Thus, possessors are not analysed as ordinary complements and (31) is ruled out.

This analysis is quite complex since it not only needs the special treatment of possessors but also needs a stipulation on adjectives to prevent them combining with a noun before it combines with a possessor and a stipulation to prevent possessors being analysed as ordinary complements. So, I reject this analysis, and I will go on to suggest a third approach in the next section.

2.4. Head-adjunct-complement analysis

Kasper (1994) has proposed that heads, adjuncts, and complements may be sisters. This permits a simple account of examples in which a head and a complement are separated by an adjunct.

(32) a. He [went **last night** to the cinema].
b. She [talked **incessantly** about syntax].
c. Sandy [said **yesterday** that he would be here].

In (32), we see in all the bracketed VPs that the verbs and their complements are separated by an adjunct. In (32a), *Last night* is an adjunct and *to the cinema* is a complement. In (32b), *incessantly* is an adjunct and *about syntax* is a complement. In (32c), *yesterday* is an adjunct and *that he would be here* is a complement. MSA can have similar examples where the verbs and their complements are separated by an adjunct, as shown in the following examples:

(33) a. takallam-tu biwuDuH-i-n ʕani l-muškilat-i
talk.PAST.1SG clearly-GEN-INDEF about DEF-problem-GEN
'I talked clearly about the problem'
b. ðahab-tu bi-l-ʔams ʔilaa l-maʕraD-i
go.PAST.1SG in-DEF-yesterday to DEF-gallery-GEN
'I went yesterday to the gallery'

In this approach, I will propose that nouns appear in head-adjunct-complement structures, in which the head has both adjuncts and complements as sisters. These require something like the following constraint:

(34) *head-adjunct-complement-phrase* →

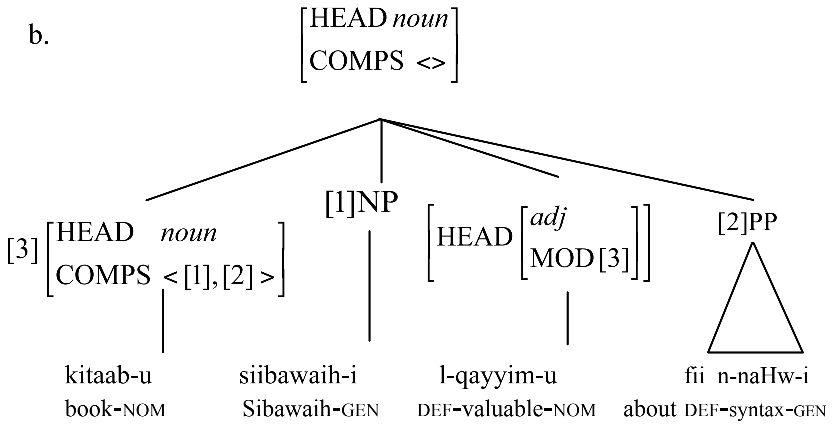
$$\begin{bmatrix} \text{DTRS} < [1] > \oplus \textit{list}\ ([\text{SS}[\text{MOD}[2]]) \oplus < [\text{SS}[3]],..[\text{SS}[n]] > \\ \text{HD - DTR}\ [1]\begin{bmatrix}\textit{word} \\ \text{SS}[2][\text{COMPS} < [3],...[n] >]\end{bmatrix} \end{bmatrix}$$

This says that the head-adjunct-complement-phrase has a head daughter and two lists of non-head daughters. The first list is optional adjunct daughters whose MOD value is identical to the value of SYNSEM in the head daughter. The second list is complement daughters whose SYNSEM values are identical to those in the COMPS value of the head daughter. It should be noted that (34) is not only relevant to NPs. Probably it is relevant to VP's as well given examples like (32) for English and (33) for MSA above.

The constraint in (34) will allow structures like the following in (35b) for the example in (35a):

(35) a. kitaab-u siibawayh-i l-qayyim-u fii n-naHw-i
book-NOM Siibawaih-GEN DEF-valuable-NOM in DEF-syntax-GEN
'Siibawaih's valuable book about syntax'

b.



The order NP AP PP can be ensured by LP constraints since these elements are sisters.

Having allowed nouns to appear in head-adjunct-complement structures, we need to exclude them from head-adjunct structures in order to avoid structures where an adjective appears between the head noun and the possessor. The obvious approach to do this is with the following constraint:

(36) $$\textit{head-adjunct-ph} \rightarrow \neg \begin{bmatrix} \text{HEAD} \begin{bmatrix} \textit{noun} \\ \text{LEX} + \end{bmatrix} \\ \text{COMPS} < \text{NP}, \ldots > \end{bmatrix}$$

This says that a head-adjunct-phrase cannot be a noun that requires an NP complement (i.e. a possessor). It is [LEX +] because we need to allow the head to be an NP (a [LEX -] constituent); this is what we have with relative clauses as they appear after the ordinary complement as in the example given in (5) above and repeated here for convenience in (37).

(37) kitaab-u siibawayh-i l-qayyim-u fii n-naHw-i
book-NOM Siibawaih-GEN DEF-valuable-NOM in DEF-syntax-GEN
[ʔallaðii ʔahdayta-nii ʔiyyaah]
that.SG.MASC give present-me it
‘Siibawaih’s valuable book about syntax which you gave me as a present’

The analysis in § 2.4. above seems simpler as it only needs one stipulation. The *head-adjunct-complement-phrase* is needed anyway for the examples in (30). We just need to stipulate that nouns cannot appear in head-adjunct structures. Therefore, I conclude that it is the best approach for Arabic NPs.

3. Conclusion

In this paper, I have presented the facts about MSA simple and construct-state noun phrases**.** I have provided an account of the definite and indefinite affixes, capturing the fact that they do not appear with construct nouns although the latter may be definite or indefinite. I have shown that the order of the elements within the construct-state noun phrases is NP AP PP. In addition, I have outlined three analyses within HPSG. The first analysis treats possessors and attributive adjectives as extra optional complements. However, there is an objection to this analysis as it treats adjectives differently from relative clauses and thus misses the similarities, one of which is that both adjectives and relative clauses reflect the (in)definiteness of the associated nominal. Therefore, assuming that relative clauses are adjuncts selecting the nominal that they combine with through their MOD feature suggests that adjectives should be analysed as adjuncts as well. The second analysis treats possessors as special complements with which a noun combines before it combines with anything else to form a complex noun. I reject this analysis as it has a number of stipulations. It needs the special treatment of possessors. It also needs a stipulation on adjectives to prevent them combining with a noun before it combines with a possessor, and a stipulation to prevent possessors being analysed as ordinary complements. In the third analysis, I have proposed that nouns appear in head-adjunct-complement structures, in which head has both adjuncts and complements as sisters. This is not only needed for noun phrases but it is also needed for verb phrases. I have only stipulated that head-adjunct-phrases cannot be headed by a noun that requires an NP complement (i.e. a possessor). As the third analysis has only one stipulation, it makes it simpler, and therefore, I conclude that it is the best approach for Arabic NPs.

References

Benmamoun, E. (2006). Construct state. In K. Versteegh, M. Eid, A. Elgibali, M. Woidich & A. Zaborski (Eds.), *Encyclopedia of Arabic language and linguistics* (Vol. I, pp. 477–482). Leiden: Brill.

Borsley, R. D. (1989). An HPSG approach to Welsh. *Journal of Linguistics, 25*(2), 333–354.

Borsley, R. D. (1995). On some similarities and differences between Welsh and Syrian Arabic. *Linguistics, 33*(1), 99–122.

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

Fassi Fehri, A. (1993). *Issues in the structure of Arabic clauses and words*. Dordrecht: Kluwer Academic Publishers.

Kasper, R. T. (1994). Adjuncts in mittelfeld. In J. Nerbonne, K. Netter & C. Pollard (Eds.) (39–69).

Miller, P. H., & Sag, I. A. (1997). French clitic movement without clitics or movement. Natural Language & Linguistic Theory, 15(3), 573–639.

Ouhalla, J. (1991). *Functional categories and parametric variation*. London: Routledge.

Ryding, K. (2005). *A reference grammar of modern standard Arabic*. Cambridge: Cambridge University Press.

Sag, I. A., Wasow, T., & Bender, E. M. (2003). *Syntactic Theory: A formal introduction*. Stanford: CSLI Publications.

Development of Maximally Reusable Grammars: Parallel Development of Hebrew and Arabic Grammars

Tali Arad Greshler
University of Haifa

Livnat Herzig Sheinfux
University of Haifa

Nurit Melnik
The Open University of Israel

Shuly Wintner
University of Haifa

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 27–40

http://csli-publications.stanford.edu/HPSG/2015

Abstract

We show how linguistic grammars of two different yet related languages can be developed and implemented in parallel, with language-independent fragments serving as shared resources, and language-specific ones defined separately for each language. The two grammars in the focus of this paper are of Modern Hebrew and Modern Standard Arabic, and the basic infrastructure, or core, of the grammars is based on "standard" HPSG. We identify four types of relations that exist between the grammars of two languages and demonstrate how the different types of relations can be implemented in parallel grammars with maximally shared resources. The examples pertain to the grammars of Modern Hebrew and Modern Standard Arabic, yet similar issues and considerations are applicable to other pairs of languages that have some degree of similarity.

1 Introduction

Our goal in this paper is to develop deep linguistic grammars of two different yet related languages. We show that such grammars can be developed and implemented in parallel, with language-independent fragments serving as shared resources, and language-specific ones defined separately for each language. The desirability of reusable grammars is twofold. From an engineering perspective, reuse of code is clearly parsimonious. From a theoretical perspective, aiming to maximize the common core of different grammars enables better identification and investigation of language-specific and cross-linguistic phenomena (see Müller, 2015, for further discussion of the motivation for parallel development of grammars).

A number of projects have adopted the notion of parallel development of different HPSG grammars with a common core. In the CoreGram project (Müller, 2015), grammars of ten different languages belonging to diverse language families are being implemented in parallel, using the TRALE system (Meurers et al., 2002).[1] Within the DELPH-IN consortium[2], two projects target languages of the same language family. The ZHONG [|] project (Fan et al., 2015a,b) models grammars of Chinese languages with a common core. It currently includes grammars of Mandarine Chinese and Cantonese. SlaviCore (Avgustinova & Zhang, 2009) is a resource that contains basic analyses known to occur cross-linguistically within the Slavic language family. SlaviClimb (Fokkens & Avgustinova, 2013), an extension of SlaviCore, is a dynamic engineering component, similar to the LinGO Grammar Matrix customization system (Bender et al., 2002), which supports the development of grammars for Slavic languages.

The two grammars in the focus of this paper are of Modern Hebrew (MH) and Modern Standard Arabic (MSA), two related languages, belonging to the Semitic

[†] This research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 505/11).

[1] The set of languages includes: German, Danish, Persian, Maltese, Mandarin Chinese, Yiddish, English, Hindi, Spanish, and French.

[2] http://www.delph-in.net/

language family. HeGram, the MH grammar, is based on a starter grammar created with the Grammar Matrix customization system, but involves some major revisions to the "standard" grammar, mostly related to its novel argument-structure representation approach (see section 2.3). AraGram, the MSA grammar, is based on the infrastructure developed for HeGram.

Similarly to the CoreGram Project (Müller, 2015), the development process of the two grammars is "bottom-up". Namely, we examine linguistic phenomena in MH and MSA and identify generalizations which capture both grammars, on the one hand, and on the other, identify distinctions between the grammars. In some cases, to account for phenomena in one language we use a "*bottom-up with cheating*" approach (Müller, 2015); we reuse analyses that have been developed for one language to account for phenomena in the other language, as long as there is no contradicting evidence.

More generally, the parallel development of the two grammars revealed four types of relations that exist between the grammars of two languages:

(i) The two languages share some construction or syntactic phenomenon.

(ii) Some phenomenon is present in one language but is absent from the other.

(iii) The two languages share some construction, but impose different constraints on its realization.

(iv) Some phenomenon seems similar in the two languages, but is in fact a realization of different constructions.

While the challenge is to maximize the common parts of the grammars, it is important to be cautious with seemingly similar phenomena across the two languages. In some cases, as we will show, the solution is to define a shared construction with different language-specific constraints. Conversely, other cases are best accounted for by the definition of distinct constructions.

This paper demonstrates how the different types of relations can be implemented in parallel grammars with maximally shared resources. The examples pertain to the MH and MSA grammars, yet similar issues and considerations are applicable to other pairs of languages that have some degree of similarity.

2 Reusable grammars of Modern Hebrew and Modern Standard Arabic

2.1 Modern Hebrew and Modern Standard Arabic

Modern Hebrew is one of the official languages of Israel (along with Modern Standard Arabic). MH is a continuation of Biblical Hebrew (attested from 10th century BCE) and Mishnaic Hebrew (1st century CE). It was revived in Europe and Palestine toward the end of the 19th century and into the 20th century, influenced by

Yiddish, as well as Polish, Russian, German, English, Ladino and Arabic. It has had native speakers for about four generations.

Modern Standard Arabic is the literary standard of the Arab world. It is based on Classical Arabic (attested from the 6th century), which originated from Proto-Arabic or Old Arabic (attested from 7th century BCE). The modern period of Arabic dates approximately from the end of the eighteenth century with the spread of literacy, the concept of universal education, and journalism. MSA is the language of written Arabic media, e.g., newspapers, books, journals etc., and it is also the language of public speaking and news broadcasts on radio and television. However, MSA does not have native speakers, as Arabs are fluent in at least one dialect of spoken Arabic, which is their mother tongue, and only become literate in MSA in school (Ryding, 2005).

As MH and MSA are related, they exhibit a number of shared phenomena which can be attributed to their Semitic roots (see Figure 1). Nevertheless, since the languages diverged several millennia ago, the end grammars are quite different and do require language-specific accounts.

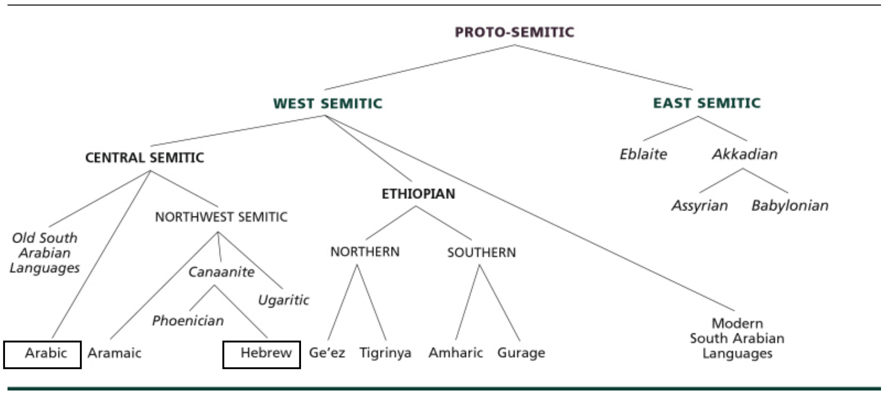


Figure 1: Semitic languages

2.2 Parallel Grammar Development

Our starting point is HeGram, a deep linguistic processing grammar of Modern Hebrew (Herzig Sheinfux et al., 2015). HeGram is grounded in the theoretical framework of HPSG and is implemented in the LKB (Copestake, 2002) and ACE systems. AraGram, the MSA grammar, utilizes the types defined in HeGram, as long as they are relevant for Arabic. In cases where the two languages diverge with respect to particular phenomena, language-specific types are defined in separate language-specific modules. More technically, the two grammars make extensive use of the ":+" operator provided by the LKB in order to define a type in a shared file, and to add language-specific constraints to its definition in distinct files (see (9)-(10) and (13)-(14) below).

The parallel development of the two grammars with their shared resources requires careful examination of the common and distinct properties of the two languages. Types, features, values and constraints can only be added or modified in a way that does not negatively affect the grammar of the other language. In order to guarantee that the changes introduced by the grammar of one language do not damage the grammar of the other we developed test suites of grammatical and ungrammatical sentences for both Arabic (160 sentences, 41 ungrammatical) and Hebrew (432 sentences, 106 ungrammatical) and test the grammar rigorously with [incr tsdb()] (Oepen, 2001). The test suites are continuously extended as analyses of more phenomena are introduced.

In the following sections we focus on a number of phenomena which illustrate different types of relations between the two languages and their implementation. We begin with a discussion of the way subcategorization is handled by the two grammars. We show that while semantic selection is found to be language-independent, the syntactic realization of arguments may be subject to language-specific constraints. Next, we describe the way the nominals of the two languages are represented in the lexical type hierarchy. In this case, the MH hierarchy is found to be a sub-hierarchy of the MSA one. Finally, we move on to clause structure. We discuss one case where two seemingly similar constructions are found to be licensed by distinct mechanisms, and another where the two languages share the same basic construction, yet impose different constraints on its realization.

2.3 Maximally shared resources: subcategorization

The architecture of HeGram embodies significant changes to the way argument structure is standardly viewed in HPSG. The main one is that it distinguishes between semantic selection and syntactic selection, and provides a way of stating constraints regarding each level separately. Moreover, one lexical entry can account for multiple subcategorization frames, including argument optionality and the realization of arguments with different syntactic phrase types (e.g., *want food* vs. *want to eat*). This involves the distribution of valence features across ten categories.[3] Each valence category is characterized in terms of its semantic role, as well as the types of syntactic phrases which can realize it (referred to as *syntactic realization classes*). Consequently, the semantic relations denoted by predicates consist of coherent argument roles, which are consistent across all predicates in the language.

Table 1 presents the ten valence categories used in HeGram, along with the corresponding semantic roles and syntactic realization phrases.[4] For example, Arg2 corresponds to the *Theme* semantic role, and can be realized in MH as an NP, an infinitive VP, a CP or a PP. The association between semantic roles and syntactic phrases is based on corpus investigation of MH which included at least 100 ran-

[3] Our restructuring of the VALENCE complex is inspired by Haugereid's packed argument frames (Haugereid, 2012).

[4] This architecture is similar in spirit to work done on Polish by Przepiórkowski et al. (2014).

domly selected examples of sentences containing each of the 50 most frequent verb lemmas in the 60-million token WaCky corpus of Modern Hebrew (Baroni et al., 2009).

Label	**Semantic Selection**	**Syntactic Realization**
Arg1	Actor, Perceiver, Causer	NP, PP
Arg2	Theme	NP, VP$_{inf}$, CP, PP
Arg3	Affectee, Benefactive, Malfactive , Recipient	NP, PP
Arg4	Attribute	AdjP, AdvP, PP, NP, VP$_{beinoni}$
Arg5	Source	PP
Arg6	Goal	PP
Arg7	Location	PP, AdvP
Arg8	Topic of Communication	PP
Arg9	Instrument	PP
Arg10	Comitative	PP

Table 1: Semantic roles and realization classes in HeGram

Each predicative lexical type in our grammars inherits from types which specify the possible semantic roles of its dependents and their possible syntactic realizations. As an example, consider the lexical type which licenses the MH verb *higiʕa* (‘*came*’).

(1) MH *higiʕa* (‘*came*’)

```
arg1-15-16-156_p_p := arg1_n & arg5_p & arg6_p &
[ SYNSEM.LOCAL.CAT.VAL.R-FRAME arg1-15-16-156 ].
```

The verb semantically selects three arguments: an *Actor* (arg1), a *Source* (arg5), and a *Goal* (arg6). Moreover, it requires that its *Actor* role be syntactically realized, yet allows for the omission of the latter two roles. This is captured by the value of its lexical type’s R(EALIZATION)-FRAME feature, *arg1-15-16-156*, which lists the different realization frames in which the verb can appear, separated by dashes. For example, *arg1* is an intransitive syntactic frame and *arg156* represents the realization of all three semantic arguments.

The syntactic realization of the semantic arguments is defined via inheritance. The lexical type in (1) inherits from three subtypes, each pertaining to one of its semantic arguments, and each determining the syntactic category of the phrases which realize that semantic role (noun, preposition, and preposition, respectively). The name of this type (i.e., *arg1-15-16-156_p_p*) reflects the different realization frames, as well as the syntactic category of its dependents (since Arg1 is always realized as an NP, its syntactic realization is omitted from the name of the type).

The MSA counterpart of *higiʕa* (‘*came*’) is *ʒaːʔa* (‘*came*’). The lexical type with which it is associated is illustrated in (2).

(2) MSA *ʒaːʔa* (‘*came*’)

```
arg1-15-16-156_p_np := arg1_n & arg5_p & arg6_np &
[ SYNSEM.LOCAL.CAT.VAL.R-FRAME arg1-15-16-156 ].
```

The only difference between the two types is in the realization of arg6. In MSA, *Goal* arguments can be realized by either NPs or PPs. This is captured in the type definition by the supertype *arg6_np*, which unlike its MH counterpart, *arg6_p*, also includes nouns as possible syntactic realizers. Consequently, the name of the type reflects this disjunctive value in its suffix (*np* instead of *p*). (3) and (4) demonstrate the realization of the *Goal* argument as a PP in MH and as an NP or a PP in MSA, respectively.

(3) *ha-qcinim higiʕu el ha-ʃagrirut ha-micrit*
*the-officers came.*3PM *to the-embassy the-Egyptian*
‘The officers came to the Egyptian Embassy.’

(4) *ʒaːʔuː dˤ-dˤubaːtˤ-u s-sifaːrat-a l-misˤriyyat-a / ʔilaː*
*came.*3PM *the-officers*-NOM *the-embassy*-ACC *the-Egyptian*-ACC */ to*
s-sifaːrat-i l-misˤriyyat-i
the-embassy-GEN *the-Egyptian*-GEN
‘The officers came to the Egyptian Embassy.’

The difference between the two languages with respect to the realization of *Goal* arguments required a slight modification of the MH schema shown in Table 1 to account for the MSA data. An additional modification involved the realization class of Arg2, since MSA uses the subjunctive in environments in which MH uses infinitives. Other than these slight language-specific details regarding syntactic realization, corpus investigations of the corresponding 50 MSA verbs using the 115-million token *arTenTen* corpus of Arabic (Arts et al., 2014) showed that they share the semantic frames identified for their MH counterparts, and consequently no changes were required in the overall argument representation scheme.

The non-standard argument structure representation of HeGram was found to be instrumental for distinguishing between general and language-specific properties of the grammar. In sum, the realization classes associated with different semantic roles are found to vary to some extent between languages while the semantic roles themselves appear to be more general.

2.4 Similarities between the languages: nominals in the lexical type hierarchy

MH and MSA are languages with rich, productive morphologies. Nouns in the two languages have natural or grammatical gender, and are marked for number. Adjectives decline according to a number-gender inflectional paradigm. Both categories are also morphologically marked for definiteness. Consequently, the grammars of the two languages require an elaborate nominal type hierarchy, where types are cross-classified according to the three dimensions: NUMBER, GENDER and DEFINITENESS.[5]

The nominal type hierarchy described above is sufficient for MH, while MSA requires an extension of the hierarchy in order to account for two additional properties: *dual number* and *Case*. A sketch of the basic shared hierarchy, along with the MSA extensions (in the boxes) is given in Figure 2 . All MH nominals (i.e., nouns and adjectives) are instances of types which realize all the cross-classification combinations of the three MH-relevant dimensions (e.g., *sm-def-nom-lex*).

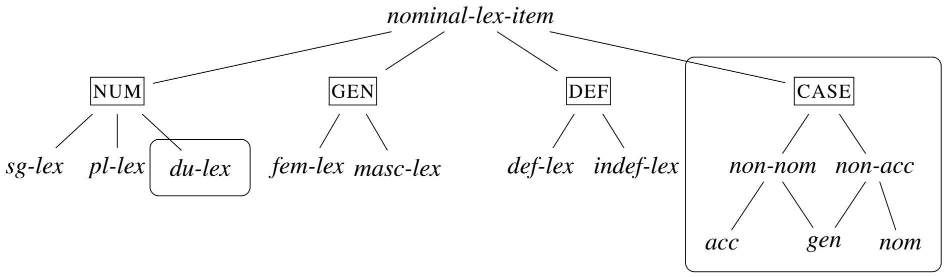


Figure 2: The nominal type hierarchy

Case in MSA is morphologically marked on all nominals by word-final vowels. Thus, in principle, all lexemes are cross-classified according to four dimensions: NUMBER, GENDER, DEFINITENESS, and CASE. The MH lexical entry for ‘boy’ (5) is an instance of a lexical type cross-classified according to three dimensions, whereas its MSA counterpart in (6) is an instance of a lexical type which is additionally classified as accusative (marked in a box).[6]

(5) MH *yeled* (‘*boy*’)

```
ild := indef-cmn-3sm-noun-lex &
       [ STEM < "ild" >,
       SYNSEM.LKEYS.KEYREL.PRED _boy_n_rel ].
```

(6) MSA *walad-an* (‘*boy*’)

```
wlda := indef-cmn-acc-3sm-lex &
       [ STEM < "wlda" >,
       SYNSEM.LKEYS.KEYREL.PRED _boy_n_rel ].
```

[5] Since the PERSON dimension is only relevant to nouns, not to adjectives, it is not presented here as part of the nominal type hierarchy.

[6] In our grammars we use 1:1 transliteration schemes for both MH and MSA. These schemes lack vowel representations as vowels are not represented in MH and MSA scripts. In glossed examples, however, we use phonemic transcription that includes vowels.

Note that the hierarchy below Case is structured to represent two different disjunctive groupings: non-nominative and non-accusative. As some MSA nominals are orthographically underspecified for Case, this intermediate level of the hierarchy was added as an engineering choice, in order to avoid repetition in the lexicon.

2.5 Deep and superficial similarities: clause structure

MH and MSA have different unmarked clause structures. In MH, SVO is the canonical word order, while in MSA it is VSO. Nevertheless, the unmarked clause order of MH is a marked structure in MSA, and vice versa. In addition, a notable property of MSA clauses is that subject-verb agreement depends on the subject position; verbs in SVO clauses exhibit full person-number-gender agreement with the subject, while in VSO clauses number agreement is suppressed and the verb is invariably singular. This is not the case in MH, where the verb fully agrees with the subject regardless of its position.[7]

2.5.1 Superficial similarities, different constructions: SVO

The SVO clauses of the two languages are remarkably similar; the finite verb exhibits full person-number-gender agreement with the subject which precedes it. As examples, consider the following SVO clauses in MH (7) and MSA (8).

(7) *ha-yeladim axlu et ha-leħem*
the-boys *ate*.3PM ACC *the-bread*
'The boys ate the bread.'

(8) *ʔl-ʔawlaːd-u ʔakaluː l-xubz-a*
the-boys-NOM *ate*.3PM *the-bread*-ACC
'The boys ate the bread.'

While superficially almost identical, the SVO clauses of the two languages are given distinct analyses in our grammars. The unmarked MH SVO clause is licensed by a *subject-head-phrase* phrase type. The syntactic tree pertaining to example (7) is shown in Figure 3.

The syntactic structure of VSO and SVO Arabic clauses has been thoroughly discussed in the literature (Fassi Fehri, 1993; Mohammad, 2000; Aoun et al., 2010; Alotaibi & Borsley, 2013, among others). The main challenge is the agreement asymmetries between SVO and VSO clauses. The analysis put forth by Aoun et al. (2010) and elaborated and cast in HPSG by Alotaibi & Borsley (2013) proposes that clause structure in MH is invariantly VSO, where number agreement is supressed. Full agreement on the verb is found only in SVO structures and in cases of *pro*-drop. In both constructions, they claim, the manifestation of full agreement is triggered by the existence of a post-verbal *pro* subject. In SVO structures this

[7] Exceptions to this generalization are colloquial verb-initial constructions (e.g., Melnik, 2006).

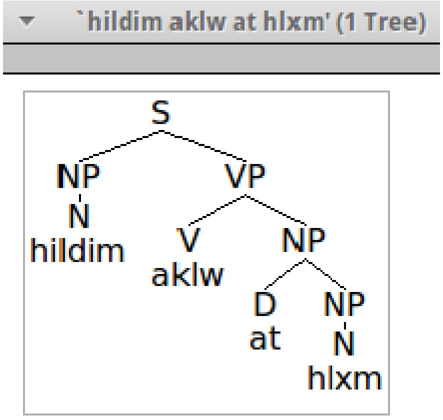


Figure 3: SVO in Modern Hebrew

pro subject is a resumptive pronoun which is associated with what looks like a preverbal subject, but is in fact a topic. The fact that subject arguments in SVO clauses are required to be definite supports this analysis.

We adopt the topic analysis of SVO clauses for MSA, and model such clauses as instances of a *filler-head-phrase* type. The syntactic tree of example (8) is given in Figure 4. Consequently, the *subject-head-phrase* type is defined only in the MH grammar.

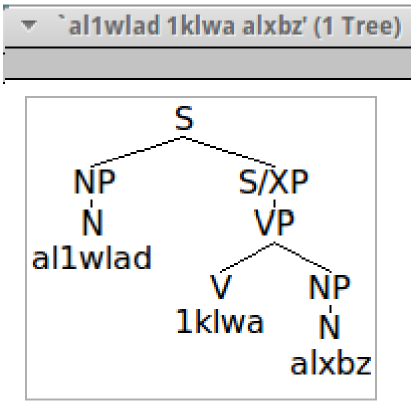


Figure 4: SVO in Modern Standard Arabic

The types dedicated to long-distance dependency constructions are shared by the two languages. Nevertheless, the MH grammar is more restrictive with regard to topicalization; it confines the phenomenon only to non-subjects in order to avoid vacuous structural ambiguity with SVO clauses, and restricts subject extraction only to *wh*-questions. MSA, on the other hand, allows all dependents to be topicalized, but restricts subject extraction to definite subjects. This disparity is implemented by using *extracted-subject-phrase* as a shared resource, and adding language-specific constraints in each grammar. This is easily done in the LKB by using the ":+" operator.

(9) MH: Subject extraction only occurs with questions

```
extracted-subj-phrase :+
[ SYNSEM.LOCAL.CONT.HOOK.INDEX.SF ques ].
```

(10) MSA: Extracted subjects must be definite

```
extracted-subj-phrase :+
[ SYNSEM.NON-LOCAL.SLASH.LIST.FIRST.CAT.HEAD.DEF + ].
```

The use of a shared type reflects the generalization that both languages have subject extraction and allows maximal reusability of the type hierarchy below the shared *extracted-subject-phrase* type.

2.5.2 Different constraints on the same construction: VSO

VSO constructions in both MH (11) and MSA (12) have a *head-subj-comp-phrase* phrase type, and thus its type definition is shared.[8]

(11) *et* *ha-leħem* *axlu* *ha-yeladim*
ACC *the-bread* *ate*.3PM *the-boys*
'The bread, the boys ate it.'

(12) *ʔakala* *l-ʔawlaːd-u* *l-xubz-a*
ate.3SM *the-boys*-NOM *the-bread*-ACC
'The boys ate the bread.'

There are, however, additional language-specific constraints which further restrict this clause type. In Hebrew, VSO constructions are only licensed in a V2 configuration, where some clause-initial material precedes the verb, e.g., *et ha-leħem* ('*ACC the-bread*') in (11). An additional Hebrew-specific constraint restricts this phrase type only to cases where the verb has undergone extraction (13). The MSA grammar, on the other hand, imposes its own language-specific constraint: the verb is invariably singular (14).

(13) MH Head Subject Complement constraint

```
VS-basic-head-subj-phrase :+
[ HEAD-DTR.SYNSEM.NON-LOCAL.SLASH 1-dlist ].
```

[8] Since only unary and binary branches are employed in the grammar, the *head-subj-comp-phrase* phrase type is implemented with two types: *head-subject* and *head-comp* (with a realized subject).

(14) MSA Head Subject Complement constraint

```
VS-basic-head-subj-phrase :+
[ HEAD-DTR.SYNSEM.LOCAL.CAT.HEAD.CNCRD png-s ].
```

This mechanism, where two languages share a construction and each language adds a different constraint to it without damaging the rest of the hierarchy, is an excellent utilization of HPSG type hierarchies, allowing maximal reusability in developing and implementing two grammars with a common core.

3 Current status and future prospects

We have adapted HeGram (Herzig Sheinfux et al., 2015) to Arabic along the lines discussed above. AraGram currently covers a plethora of syntactic phenomena, including Case marking, subject-verb and noun-adjective agreement, SVO and VSO word order, relatively free complement order, multiple subcategorization frames, selectional restrictions of verbs on their PP complements, topicalization, passive and unaccusative verbs. Many of these phenomena required only minor adaptations to the Hebrew grammar. Therefore, the development of AraGram took only several weeks (excluding corpus investigation and literature review). For comparison, the development of HeGram to its stage when we started developing AraGram took about a year. AraGram currently shares 95.5% of its types with HeGram, while HeGram currently shares 99.2% of its types with AraGram.

The development of AraGram is ongoing. In the near future, we will focus on additional constructions, including wh-questions, control, raising, the copular construction, and multi-word expressions. We also intend to work on automatic translation between the languages using semantic MRS transfer and generation.

References

Alotaibi, Mansour & Robert D. Borsley. 2013. Gaps and resumptive pronouns in Modern Standard Arabic. In Stefan Müller (ed.), *Proceedings of the 20th international conference on hpsg*, 6–26. Stanford: CSLI Publications.

Aoun, Joseph E., Elabbas Benmamoun & Lina Choueiri. 2010. *The syntax of Arabic*. Cambridge University Press.

Arts, Tressy, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff & Vit Suchomel. 2014. arTenTen: Arabic corpus and word sketches. *Journal of King Saud University-Computer and Information Sciences* 26(4). 357–371.

Avgustinova, Tania & Yi Zhang. 2009. Parallel grammar engineering for Slavic languages. Workshop on Grammar Engineering Across Frameworks at the 2009 ACL/IJCNLP conference, Singapore.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources And Evaluation* 43(3). 209–226.

Bender, Emily M., Dan Flickinger & Stephan Oepen. 2002. The grammar matrix: an open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Coling-02 workshop on grammar engineering and evaluation*, 1–7. Morristown, NJ, USA: Association for Computational Linguistics. doi:http://dx.doi.org/10.3115/1118783.1118785.

Copestake, Ann. 2002. *Implementing typed feature structure grammars*. Stanford: CSLI Publications.

Fan, Zhenzhen, Sanghoun Song & Francis Bond. 2015a. Building Zhong [|], a Chinese HPSG shared-grammar. In Stefan Müller (ed.), *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar, Singapore*, 97–110. Stanford, CA: CSLI Publications.

Fan, Zhenzhen, Sanghoun Song & Francis Bond. 2015b. An HPSG-based shared-grammar for the Chinese languages: Zhong [|]. In *Proceedings of the grammar engineering across frameworks (GEAF) 2015 workshop*, 17–24.

Fassi Fehri, Abdelkader. 1993. *Issues in the structure of Arabic clauses and words*. Dordrecht: Kluwer.

Fokkens, Antske & Tania Avgustinova. 2013. SlaviCLIMB: Combining expertise for Slavic grammar development using a metagrammar. In *Workshop on high-level methodologies for grammar engineering ESSLLI 2013*, 87.

Haugereid, Petter. 2012. A grammar design accommodating packed argument frame information on verbs. *International Journal of Asian Language Processing* 22(3). 87–106.

Herzig Sheinfux, Livnat, Nurit Melnik & Shuly Wintner. 2015. Representing argument structure in computational grammars. Submitted.

Melnik, Nurit. 2006. A constructional approach to verb-initial constructions in Modern Hebrew. *Cognitive Linguistics* 17(2). 153–198.

Meurers, W. Detmar, Gerald Penn & Frank Richter. 2002. A web-based instructional platform for constraint-based grammar formalisms and parsing. In *Proceedings of the acl workshop on effective tools and methodologies for teaching NLP and CL*, 18–25.

Mohammad, Mohammad A. 2000. *Word order, agreement, and pronominalization in Standard and Palestinian Arabic*. Amsterdam: John Benjamins.

Müller, Stefan. 2015. The CoreGram project: Theoretical linguistics, theory development and verification. *Journal of Language Modelling* 3(1). 21–86. http://hpsg.fu-berlin.de/~stefan/Pub/coregram.html.

Oepen, Stephan. 2001. [incr tsdb()] — competence and performance laboratory. User manual. Technical report Computational Linguistics, Saarland University Saarbrücken, Germany.

Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski & Marek Świdzib́ski. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international Conference on Language Resources and Evaluation, LREC 2014*, 2785–2792. Reykjavik, Iceland: ELRA. http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Ryding, Karin C. 2005. *A reference grammar of Modern Standard Arabic*. Cambridge university press.

A Constructional Analysis for the Skeptical

Doug Arnold
University of Essex

Andrew Spencer
University of Essex

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 41–60

http://csli-publications.stanford.edu/HPSG/2015

Abstract

This paper addresses the issue of phonologically null elements in HPSG by providing an analysis of the construction exemplified by NPs such as 'the rich', 'the beautiful', 'the unemployed', which lack an overt noun. The properties of this construction are explored in detail, and a number of approaches described: in particular approaches which posit a phonologically empty noun, and constructional approaches. It is shown that a constructional approach is empirically superior. This is interesting, theoretically, because empirical differences between such approaches have proved elusive hitherto.

1 Introduction

This paper addresses the issue of phonologically null elements in HPSG by providing an analysis of the construction exemplified by *the merely skeptical* in (1), and the examples in (2).

(1) This will not convince a cynic but may persuade [the merely skeptical].

(2) the (unconventionally) beautiful, the (recently) unemployed, the (severely) disabled, the (chronically) sick, the (truly) lazy, the (merely) incompetent, the (wretchedly) poor, the (obscenely) rich, the (unalterably) pious, the (unbearably) pompous, the old, the young. . .

The construction appears to involve an NP which lacks a nominal head, but which is otherwise normal. Our focus will be on discussing the empirical problems faced by analyses involving a null head, and on providing a constructional analysis that improves on existing accounts (though we will briefly discuss another approach which involves 'sharing'). Theoretically, this is an interesting result, because it is in general difficult to find empirical differences between constructional analyses and analyses involving a null head, and the choice has often been seen as a matter of simplicity, taste, or convenience.[1]

Apart from the presence of an adjective (potentially modified by an adverb), and absence of a head noun, the most obvious features of the NP in this construction are a definiteness requirement (so (3a) and (3b) are unacceptable), plurality (so in (3c) a plural verb form is required), and the interpretation (*the skeptical* means roughly 'individuals who are skeptical'):

[†] As well as HPSG 2015 in Singapore, versions of this paper have been presented at the Second European Workshop on HPSG (November 2014, in Paris), and the 2014 meeting of the Linguistics Association of Great Britain. We are grateful to participants at those meetings, to our colleagues at Essex, and anonymous referees for HPSG 2015, as well as Emily Bender, Olivier Bonami, Bob Borsley, Dan Flickinger, Daniele Godard, Taka Maekawa, Nurit Melnik, David Oshima, Frank Van Eynde, and Gert Webelhuth. Remaining deficiencies are entirely our responsibility: in particular, as example (1) makes clear, the title of this paper involves a weak pun, for which we apologize.

[1] See, for example, the different analyses of null-copula constructions in Bender (2001), and the discussion in Müller (2014, 102ff); other discussions bearing on the existence of phonologically null elements in HPSG include Nerbonne and Mullen (2000); Henri and Abeillé (2007); Laurens (2008); and Arnold and Borsley (2014).

(3) a. *This will convince even a skeptical. (cf. a skeptical individual)
b. *We met several skeptical. (cf. several skeptical individuals)
c. [The merely skeptical] are/*is easier to convince.

Following Fillmore et al. (2012), we will refer to this construction as the 'ANH' construction (Fillmore et al.'s `Adjective-as-Nominal.human` construction). These properties distinguish it from a number of superficially similar, but actually rather different, constructions which we will not discuss here (cf e.g. Huddleston and Pullum, 2002, 410ff).

First, there are normal NPs that are headed by nouns which happen to be homophonous with adjectives – nouns presumably derived from adjectives by a morphological conversion process. For example, *intellectual* as in (4b), behaves like a normal noun in accepting adjectival (rather than adverbial) modifiers, inflecting as a normal nouns (e.g. for plural, (4c)), and taking a full range of determiners (again, see (4c)).

(4) a. She is *an intellectual*.
b. She is *an (alleged/*allegedly) intellectual*.
c. *Some/All/Most intellectuals* accept these ideas.

There are also constructions which genuinely lack a nominal head, but which should also be distinguished from the construction we are concerned with. For example, superlative and definite comparative adjectives can appear without nominal heads, as in (6), but unlike the kind of NP we are interested in, such NPs can be singular, as shown in (5) and refer to inanimates, as in (6):

(5) [The most/more interesting] of his ideas has been ignored.
(6) [The older/oldest] of the books is also [the cheaper/cheapest].

There is also an elliptical construction, exemplified in (7): the *the merely geographical* in (7a) is interpreted as 'the merely geographical sense'; in (7b) *the abstract* means 'the abstract word'; in (7c) *a second* means 'a second child'. As example (7c) makes clear, the elliptical construction is not required to be either definite or plural:

(7) a. It is a distinct entity, in other senses than [the merely geographical].
b. Prefer the concrete word to [the abstract].
c. After having a first child, they decided they wanted [a second].

Finally, one should distinguish NPs which denote more or less abstract objects or qualities, as in (8), which are also singular, presumably because they denote uncountables:

(8) [The merely implausible] is often mistaken for [the completely impossible].

Of course, in the absence of a formal analysis any classification is at best tentative. However, we believe the construction we are concerned with here is sufficiently distinctive and productive to merit individual attention, and potentially provides a basis for a wider investigation of these other constructions.[2]

[2] Many examples will be ambiguous, e.g. *[The immortal] can seem beyond our understanding* might

The remainder of the paper is structured as follows. Section 2 will describe the key features of this construction in more detail, including some features that seem to have been overlooked. Section 3 will review some relevant literature and existing proposals, and develop explicit analyses: in Section 3.1 we briefly consider an approach based on multi-dominance or 'sharing', then in Section 3.2 we outline an analysis involving a phonologically empty noun, and discuss the problems it faces. In Section 3.3 we present a constructional analysis. Section 4 provides a summary.

For the sake of concreteness we assume the framework of Ginzburg and Sag (2001) (G&S).

2 Phenomenon

Typical examples of the ANH construction have been shown in (1) and (2). It is often thought of as a rather marginal construction, but as will be clear from attested examples like those in (9), it is highly productive (*pace, e.g.* Huddleston and Pullum, 2002, p417):

(9) a. Back in The Smoke (i.e. London) amongst [the habitually abusive] and [the floridly psychotic].
b. That mostly means the [habitually abusive] or [uncivil], or those who go out of their way to shill for a particular perspective. . .
c. When they don't find him (i.e. the ideal man), they . . . settle for [the sociable but unattractive], [the attractive but unsociable], and, as a last resort, for [the merely available].
d. Yet it's another Monday 4:30 am, in the land of [the barely awake].

Externally, ANH NPs behave like normal definite plurals – they allow e.g. possessive marking (10), post-modification by PP (11), restrictive relative clauses (12), non-restrictive relatives (13), and coordination with normal (i.e. lexically headed) NPs (14):

(10) the very poor's main problem. . . (possessive marking)
(11) the very poor in the country. . . (PP postmodification)
(12) the very poor who live in rural areas. . . (restrictive relative)
(13) the very poor, who are barely mentioned here,. . . (non-restrictive)
(14) [the very poor] and [some inhabitants of slum areas]. . . (coordination)

These NPs are plural, triggering plural agreement, and taking plural reflexives:

(15) [The very poor] are/*is present in every area of the city.
(16) [The very poor] find/*finds themselves/*herself without defence in these

be interpreted as involving an instance of the ANH construction ('those whose reputation does not die', perhaps) or an abstract object ('the phenomenon of immortality'), or as an ellipsis in a context like 'When you think simultaneously about his few immortal compositions and his massive commercial output, [the immortal] stand out more clearly.'

conditions.

The interpretation is 'generic' (loosely speaking), and primarily human — roughly 'individuals (people or perhaps beings including people) who are Adj', 'the kind of individual (person/being) who is Adj'. Thus, adjectives that are not plausibly applied to humans are hard to accept (so, e.g. *the aflame* is hard to accept, but *the aflame with enthusiasm* is acceptable), and in (17) *the immortal* means 'those who are immortal', but its primary interpretation relates (however implausibly) to humans, and does not include (e.g.) divinities.[3]

(17) [The immortal] do not truly appreciate the gift of immortality.

Moreover, only adjectives that can be applied to human individuals are permitted, adjectives that can only apply to collections or groups seem to be excluded. So, for example, while a group of people can be widespread, we cannot talk about **the widespread*.

As regards internal structure, ANH NPs have no nominal head, instead there is an adjective – in fact an AP – which can have complements as in (20a), and can be coordinated as in (19):

(18) These proposals will not help [the extremely poor].
(19) [The lazy, ignorant, and stupid] are harder to deal with than the merely stupid.

The adjective can be pre-modified by adverbs relatively freely, as in (20a), but one significant restriction is that the degree modifiers *how* and *however* are impossible – cf. (20b) and (20c) (this seems not to have been previously noted):

(20) a. the very rich, the nearly famous, the merely skeptical, the compulsively addicted to chocolate, the excessively fond of self-analysis, . . .
b. *[The however rich] do not care about taxation.
c. *[The how rich] do not care about taxation.

Most 'normal' adjectives are possible, so long as they are compatible with a 'generic' interpretation in relation to individual 'people' – in (21) there is an adjective (*awake*) that is normally postnominal, and (22) features an adjective with its complement which can only appear post-nominally. The examples in (23) involve what one would normally think of as 'stage level' predicates (which have been coerced to be 'characteristic' by adverbial pre-modifiers).

(21) the barely awake (**the barely awake individuals* vs *individuals barely awake*)
(22) the compulsively addicted to chocolate (**compulsively addicted to chocolate individuals*)
(23) the permanently upset, the congenitally unavailable, the merely available ('stage level')

However, though most 'normal' adjectives are possible, there are several classes

[3] This restriction to humans is shared by other 'null-nominal' constructions, e.g. those involving noun-less determiners, as in *All (welcome)*, *Some (came running)*, *Many (are called), but few (are chosen)*, and also by many nouns, for example *inhabitants* is prototypically taken to mean *'people* who live in a place' (excluding animals).

of adjective that are not possible in this construction.

'Process oriented' adjectives like *strong* in (24) (where it is interpreted as modifying *swim*, specifying the manner of swimming, rather than indicating a general attribute of strength) are excluded. So in (25), from a Robert Frost poem, *the strong* is interpreted as those who are strong in general (not in relation to some activity or process):

(24) Sam is a strong swimmer.
(25) [The strong] are saying nothing until they see.

Conceivably there is a semantic basis for this (for example, a process reading in (25) might be excluded because there is no process for *strong* to modify), but other restrictions are harder to explain. For example 'modal' adjectives like *alleged*, and *former* are excluded despite the theoretical possibility of interpreting them as denoting something like 'alleged people' (i.e. individuals who are alleged to be people) or 'former people' (individuals who used to be people):

(26) *[The alleged] have no opportunities here.
(27) *[The former] have no opportunities here.

Similarly, 'emotive' uses of adjectives are excluded. For example as a noun modifier *poor* can either be used descriptively to mean 'financially disadvantaged' or emotively to express the speaker's sympathy. Thus, (28a) is ambiguous. This ambiguity is absent in the ANH construction (28b). Similarly, an adjective like *frigging*, which has only an emotive use, is impossible in this construction (**the frigging*), though there is nothing wrong with *the frigging people* interpreted as 'the people' with an negative implication.

(28) a. The poor people need our help.
b. [The poor] need our help.

(29) a. The frigging people need our help.
b. *[The frigging] need our help.

These restrictions are often expressed in terms of only a subset of attributive adjectives being allowed (e.g. Huddleston and Pullum, 2002, pp 529,553): specifically, only attributive adjectives that can also be used predicatively. While it is clear that there is some kind of 'predicative' restriction at work here (for example, emotives, modals, and process oriented uses of adjectives are impossible predicatively), it is not clear to us that this is the right characterisation, because we find examples of this construction with adjectives which cannot be used attributively. For example, the adjectives *sorry*, *glad*, and *content* can all be used in this construction, in their predicative senses, senses which are excluded when they are used attributively:[4]

[4] The paraphrases given in italics are intended to clarify that these examples involve predicative senses. *Sorry* has an attributive use, meaning 'pathetic' (rather than regretful), as in a *sorry sight*, which is not involved here; attributively *glad* means 'causing happiness' (as in *glad tidings*), predicatively it means 'feeling happy', which is clearly the sense involved in (30b) (from a headline *The Guardian* newspaper); *content* does not appear attributively, instead we get *contented* (as in *a contented person*), thus, **a content person* is ruled out.

(30) a. This page is only for [the genuinely sorry]. (=*those who are genuinely sorry)*
b. the good, [the glad] and the celebrities (=*those who are glad*)
c. None but [the content] are truly happy. (=*those who are content*)

In fact, it seems to us that a better characterisation is that this construction excludes attributive adjectives, and is restricted to *predicative* adjectives.[5]

It seems to be generally assumed (e.g. in Huddleston and Pullum (2002) and Fillmore et al. (2012)) that the ANH construction requires the definite article (*the*). This is incorrect. One can find examples of other kinds of definite, as in attested examples like (33) and (34). However, indefinites are impossible, as are quantifiers.[6]

(31) [The very poor] are to be found everywhere.
(32) As a group, [America's poor] are far from being chronically undernourished.
(33) Most of [Asia's newly rich] are simply the first winners in a rush to own markets.
(34) . . . it must be appreciated that [those poor who were included in these surveys] were those who were deemed to be in need. . . [7]
(35) *[All/most/some/no very poor] have the same problems.

Finally, and interestingly, though as we have seen above, internal modification by adverbs like *merely* is possible, internal modification by adjectives (like *worried*, *lazy*, *well-educated*, and *deserving*) is also possible (and as (37) shows, both can appear at once):

(36) the worried well, the lazy rich, the well-educated young, the undernourished and deserving poor
(37) Asia's well-educated newly rich

There is a straightforward semantic contrast between adjectival and adverbial modification:

(38) the unconventionally beautiful
(39) the unconventional beautiful

The *unconventionally beautiful* are those who are beautiful in an unconventional way – whose beauty is unconventional. The *unconventional beautiful* are 'the

[5] We take 'predicative' to involve the semantic type $\langle e,t\rangle$. Predicative adjectives are those that can appear as complement to verbs like *be*, *become*, *seem* and *consider*. We take 'attributive' to mean noun-modifying – i.e. having semantic type $\langle\langle e,t\rangle\langle e,t\rangle\rangle$ – hence including both pre- and post-nominal adjectives.

[6] It is worth noting that the notion of 'definite' involved here is that involved in the 'downstairs' nominal in partitives (compare *two of the boxes* vs. **two of some boxes*). In particular, NPs involving just the quantifier *all* count as indefinite by this test (cf. **two of all boxes*), though they count as definite in other ways (e.g. by being unable to appear with existential *there* – cf. **There are all boxes in the corridor*). Thanks to Dan Flickinger for discussion of this point.

[7] Notice, however, that with *those*, as in (34), a relative clauses is needed: **[Those poor] are discussed below* vs *Those poor who were included are discussed below*. We have no account of this, but rather than being an issue with this construction, it may reflect a property of the demonstrative, because one sees the same behaviour in the contrast between **those came* vs. *those who were called came*.

beautiful' who are unconventional (i.e. as individuals) — this is exactly parallel to the interpretations with overt nouns (*unconventionally beautiful people* vs *unconventional beautiful people*).

We will return to all these properties below.

3 Analyses

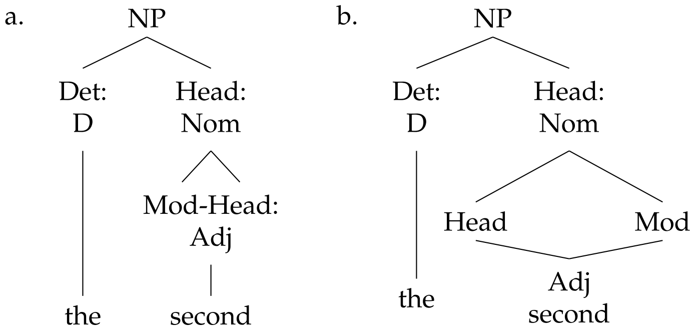
Descriptive discussion of this construction goes back at least to Jespersen (1987, 80-1), and Pullum (1975) for more formal discussion, but fully worked out formal analyses are thin on the ground. Hence, rather than attempting a full literature review, we will concentrate on three styles of approach that seem potentially feasible: one based on multi-dominance or 'sharing', one based on the existence of a phonologically null head, and a constructional approach.

3.1 Multi-dominance and Sharing Analyses

A form of multi-dominance or 'sharing' is presented in H&P, and a sharing analysis for similar constructions is proposed in Wescoat (2002), which develops an analysis of 'pronominal' determiners like *this* and *those* which can constitute an NP in the absence of a head noun (e.g. *This is a good idea, but those are better*).

H&P have relatively little to say about the ANH construction *per se* (it is just one of several instances of constructions involving '(fused) modifier-head with special interpretations' which are exemplified (Huddleston and Pullum, 2002, p417)), and no explicit representation is provided. However, H&P's approach to constructions of this kind assumes that two functions (for example the head and modifier functions) are 'fused' – that is realised simultaneously by one element – and they provide a representation for an example involving an ordinal adjective (*the second*) as in (40a).[8]

(40)

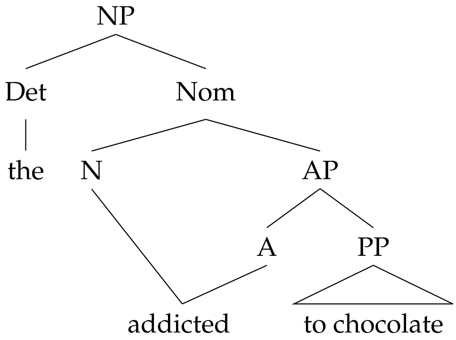


[8] See Huddleston and Pullum (2002, p412). For us, as for H&P, this is a distinct construction from the ANH construction, as it is not restricted in the same way: the construction involving an ordinal adjective can be singular, and indefinite, and is not restricted to humans, as in an example like *Having had one drink, I decided I wanted [a second].*

Of course, this is a descriptive, not a formal analysis, and H&P do not discuss how the various restrictions we have observed above might be captured. Nor is it immediately clear what a proper formal implementation should be. In particular, it might appear from (40a) that what H&P have in mind is essentially a constructional view, but here the representation is misleading because (as is evident from the analysis of other fused-head constructions) what H&P really have in mind involves multi-domination or 'sharing': the adjective in (40) fullfils two functions, and could be thought of as having two mothers, so an alternative representation might be as in (40b) (see for example Huddleston and Pullum (2002, p412), and the representation of fused (i.e. headless) relatives on p1073, where it is clear that this is that this is what H&P have in mind).

While this is not a formal analysis, a similar fully formalised analysis for similar constructions is proposed in Wescoat (2002). This involves *lexical* sharing, an idea which has sometimes been proposed in the HPSG literature (e.g. Kim et al., 2008). Applied to the ANH construction, it might give representations along the lines of (41).

(41)



Here the idea is that the single item *addicted* fullfils both the role of nominal head (of Nom), and adjectival head (of AP).

While this is an intriguing idea, it is still not an analysis (as well as accounting for the empirical restrictions described above, one would need to explain the precise combination of nominal and adjectival properties that one sees), but we will not pursue this here, because as Kim et al. (2008) point out, it is a theorem of Wescoat's axiomatisation of lexical sharing that a single word cannot be the exponent of multiple atoms unless those atoms are adjacent. So a prediction of this approach would be that nothing can intervene between the nominal and adjectival positions in this construction. This prediction is simply disconfirmed by examples where the adjective is pre-modified. For example in (42), the adverb *compulsively* intervenes between the nominal and adjectival positions (cf also many examples in Sections 1 and 2):

(42) a. the __ compulsively addicted to chocolate
b. the people compulsively addicted to chocolate

Accordingly, we will not pursue this analysis here.[9]

[9] It should be pointed out that H&P do not assume *lexical* sharing, so this might not be a problem for a formalisation of their approach.

3.2 Empty Noun Analyses

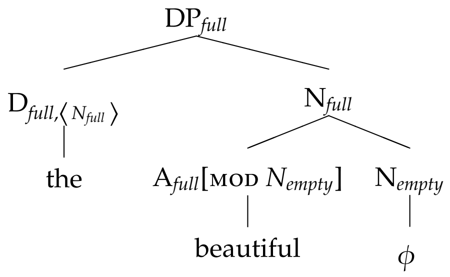
In this section, we will outline an approach to the ANH construction that involves an empty noun.[10]

Nerbonne and Mullen (2000) (N&M) give an analysis of some 'empty N' phenomena for German and English, which one could imagine extending, giving representations along the lines of (46). The idea of their analysis is that determiners should be classified with respect to whether they allow phonologically empty nominals, where whether a nominal is empty or not is determined by a feature LEFT-PERIPHERY that percolates up its left edge.

(43) every car/*ϕ left. (*every* requires a 'full' nominal)
(44) none *car/ϕ left. (*none* requires an 'empty' nominal)
(45) many cars/ϕ left. (*many* allows 'full' or 'empty' nominals)

One could extend this to a treatment of the construction we are concerned with by allowing adjectives to select the nominals they modify (via the usual MOD, or SELECT apparatus), giving representations like (46):[11]

(46)


DP_{full}
$D_{full,\langle N_{full} \rangle}$
the
N_{full}
A_{full}[MOD N_{empty}]
beautiful
N_{empty}
ϕ

Here *beautiful* has been given a MOD feature that allows it to modify the empty nominal. N&M claim that their analysis requires a 'DP' analysis, which takes the determiner to be the head of what G&S, in common with most other work in HPSG, call NPs (as can be seen in (46)). This in itself might be an objection to the analysis, but we doubt it is necessary. So far as we can see, the analysis could be re-cast straightforwardly using the SELECT apparatus introduced by Van Eynde (2007) (or indeed with the earlier MOD and SPEC features).

The general shape of a G&S-style analysis involving a phonologically empty noun is fairly easy to imagine: it would involve a lexical entry along the lines of (47), and give rise to structures like those in (48).

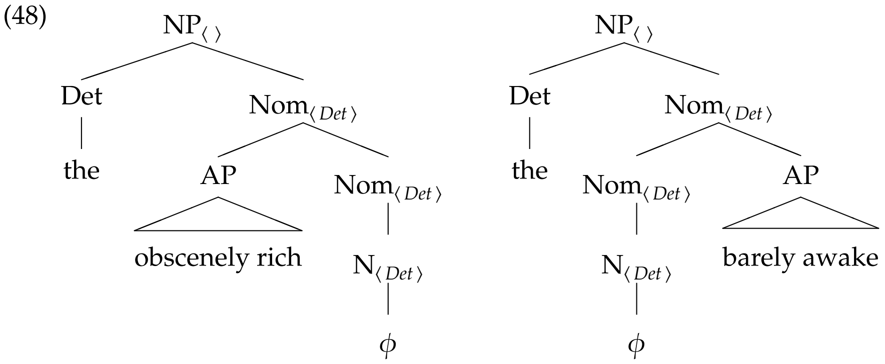
[10] See Borer and Roy (2010) for a recent empty noun analyses of closely related, but not identical, constructions in French and Hebrew in a generative framework.

[11] Note, however, that N&M themselves are quite tentative about the feasibility or desirability of such an extension (cf Nerbonne and Mullen, 2000, p156).

(47)
$$\begin{bmatrix} \textit{word} & \\ \text{PHON} & \langle\,\rangle \\ \text{SYNSEM} & \begin{bmatrix} \text{LOCAL} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{SPR} & \langle [\,] \rangle \\ \text{HEAD} & \begin{bmatrix} \textit{noun} & \\ \text{DEF} & + \\ \text{AGR} & \begin{bmatrix} \text{NUM} & \textit{pl} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \text{INDEX} & \boxed{1}\ \textit{pl} \\ \text{RESTR} & \{\textit{people}_{gen}\} \end{bmatrix} \end{bmatrix} \\ \text{EDGE}\,|\,\text{LEFT} & \textit{empty} \end{bmatrix} \end{bmatrix}$$

This is a phonologically empty *noun* which is plural and definite, with a non-empty SPR value (i.e. lacking a specifier) whose semantics is 'generic people', and which carries an *empty* EDGE feature which is intended to percolate up the left-branch of structures. Normal nouns will be *non-empty* for this feature.[12]

(48)



Clearly, this can provide an account of most of the phenomena described in Section 2: in particular, the plurality, definiteness, and the special interpretation all follow directly from the lexical entry.

The impossibility of examples like those in (49) poses a potential problem for empty-nominal analyses:

(49) a. *[the ϕ] (with an interpretation roughly 'people' in general)
b. *[The ϕ] can always surprise you. (intended: 'people in general')

However, these can be avoided here if we assume that *the* SELECTS *non-empty* heads.[13]

[12] Here *people*$_{gen}$ is intended as shorthand for however the semantics of plural generic reference should be represented. Plurality is also here expressed via both the AGR feature, and the *index* value. Both are necessary (because the noun is plural, both in terms of its agreement properties and in terms of the kind of entity it denotes), but one or other is probably redundant. We assume there is a HEAD feature [DEF *boolean*], so that a noun specified as [DEF +] will only appear with definite determiners. The use of an EDGE feature essentially re-implements the N&M proposal. The idea derives from Miller (1992), see e.g. Tseng (2003).

[13] However, notice that this requires EDGE to be a *synsem* feature, which is not standard. It is normally taken to be a feature at the level of *sign*s (see, e.g. Tseng, 2003). It is not clear if this is problematic.

The price of such an analysis is (i) an additional lexical entry, and (ii) some extra feature apparatus (the EDGE feature apparatus – which might be independently motivated).[14] It is thus an attractive approach.

However, it faces two empirical difficulties.

The first relates to the impossibility of adjective phrases with non-empty WH values, like *how rich* and *however rich* in this construction (cf. **the how rich*, and **the however rich*, see (20), above). The way WH-percolation works in the G&S framework is that head words amalgamate the WH values of their arguments, with the result being percolated by the Generalized Head Feature Principle (see Ginzburg and Sag, 2001, p189). *Wh*-expressions like *how* and *however* have non-empty WH values, and since G&S treat degree words as arguments of the head adjective, adjectives like *rich* will have non-empty WH values when accompanied by *how* or *however*, and so will APs like *how rich* and *however rich*. The problem is that the G&S framework (or indeed any other version of HPSG) provides no mechanism for heads to select adjuncts. Thus, there is no mechanism for the null noun to exclude expressions like *how rich* or *however rich* as adjuncts, and no way to avoid producing **the how rich ϕ*, and **the however rich ϕ*.[15]

The second, and we think fatal, difficulty is that the approach provides no account of why attributive adjectives – in particular, modal and emotive adjectives – are excluded from the construction, cf. the discussion of example in (28) and (29), and why predicative only adjectives are allowed, as witness examples (30). Why should this empty noun (uniquely among English nouns) reject normal attributive modifiers and accept predicative ones?

We consider these to be convincing reasons for rejecting this approach. Accordingly, in the following section we will develop a constructional analysis.

3.3 Constructional Analyses

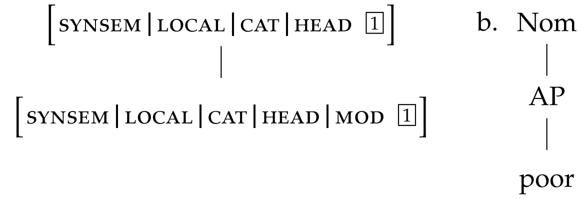
Branco and Costa (2006) provide an interesting analysis of elliptical NPs (e.g. *those two*, and examples like those in (7)), and while they mention examples of the ANH construction only in passing, it might be extended to deal with the ANH construction. In outline, what they propose is to exploit the fact that 'functor' daughters (i.e. daughters which are neither heads nor complements) select their head daughters, and use a unary rule in which a single functor

[14] Notice, however, that it really is an *additional* lexical item – its syntactic and semantic idiosyncracies mean it will not be possible to collapse it with the entries for any other empty nominals (this is why N&M were skeptical about extending their analysis, as noted in footnote 11).

[15] A potential response to this might be to argue the APs here are not adjuncts, but complements (or perhaps, following Bouma et al. (2001), adjuncts *and* complements), which might provide a way round this problem. However, note that it is not common to argue that *all* adjuncts should be treated in this way. For example, while Bouma et al. (2001) treat post-verbal adjuncts as complements they do not assume this for pre-verbal ones. But to deal with the facts here, one would need to treat both pre- and post-nominal adjectives as complements, since **the ϕ however awake* is just as bad as **the however rich ϕ*. While something along these lines might be technically possible, we think it would entirely eliminate the theoretical appeal of a null head analysis.

daughter projects a phrase whose HEAD properties are those of the (absent) head the functor would normally require, roughly as in (50a). Since an adjective like *poor* is specified as modifying a nominal, this would give rise to structures like (50b), which might combine with a determiner to produce an expression like *the poor*.

(50)


a.
[SYNSEM | LOCAL | CAT | HEAD 1]
[SYNSEM | LOCAL | CAT | HEAD | MOD 1]
b. Nom
AP
poor

Though it is an interesting approach to nominal ellipsis, we will not pursue it here, because while it does not posit an empty nominal, it suffers from the same flaw as approaches that do. Notice in particular, that only adjectives that have a *mod* feature will be able to participate in (2a), which is to say only *attributive* adjectives. This is entirely wrong, as we have seen.

Instead, we will take as our starting point the Sign-based Construction Grammar analysis of this construction provided in Fillmore et al. (2012, p350) (the '`Adjective-as-Nominal.human`' construction). The construction is specified as in (51).

(51)

$$\begin{array}{c}
\left[\begin{array}{ll}
\text{FORM} & \langle \textit{the}, X \rangle \\
\text{SYN} & \left[\begin{array}{ll} \text{CAT} & \left[\text{NUM}\ \textit{pl}\right] \\ \text{MRKG} & \textit{det} \end{array}\right] \\
\text{SEM} & \left[\begin{array}{ll} \text{INDEX} & i \\ \text{FRAMES} & \left\langle \left[\begin{array}{l} \textit{generic-fr} \\ \text{GENERIC-OBJ}\ \ i \end{array}\right], \left[\begin{array}{l} \textit{human-fr} \\ \text{ENTITY-OBJ}\ \ i \end{array}\right] \right\rangle \oplus L \end{array}\right]
\end{array}\right] \\
| \\
\left[\begin{array}{ll}
\text{FORM} & \langle X \rangle \\
\text{SYN} & \left[\begin{array}{ll} \text{CAT} & \textit{adj} \\ \text{VAL} & \langle\,\rangle \end{array}\right] \\
\text{SEM} & \left[\begin{array}{ll} \text{FRAMES} & L\textit{:list}\left(\left[\begin{array}{l} \textit{property-fr} \\ \text{ENTITY}\ \ i \end{array}\right]\right) \end{array}\right]
\end{array}\right]
\end{array}$$

Here a plural NP containing ⟨ *the* ⟩ as part of its FORM directly dominates a valence saturated adjective (i.e. an AP). The semantics given (in the FRAMES attribute) combines the semantics of the adjective with 'genericitity' and 'humanness' specifications by appending the FRAMES of the adjective to these specifications in the construction.

This clearly captures the main features of the construction – definiteness, plurality, and the special interpretation.

It is not clear whether it could avoid allowing examples like **the however rich*, or deal with the restriction to predicative adjectives, and we will not speculate here, because there are other problems with the formulation.

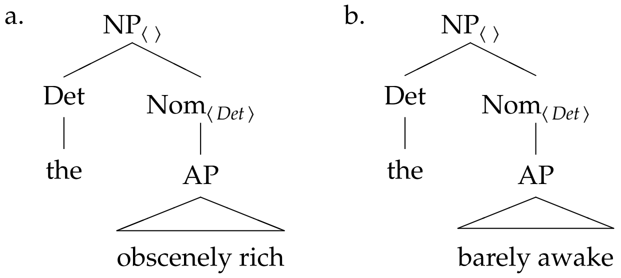
First, notice that the structure of an expression like *the poor* is simply an NP containing an AP – there is no internal N′ or Nom, hence no scope for adjectival modification following the determiner as in examples like those in (36) (*the worried well*, *the lazy rich* etc.)

Second, notice that the construction requires the presence of *the*. However, we have seen there are examples of this construction with other specifiers (cf. (33), *Asia's newly rich*, etc.)

Moreover, notice that the presence of *the* is simply stipulated as part of the FORM – the actual definite article is not part of the construction, which in fact lacks a determiner: *the* makes no semantic contribution, and its presence is unrelated to the definiteness of the construction, or general principles of English grammar (e.g. that only *indefinite* plural NPs lack determiners).[16]

However, we can improve on this straightforwardly. What we want is a construction that will build a Nominal (Nom, in X-bar terms an N′) out of an AP, to give structures along the lines of (52).

(52)



We can produce these with a construction which we will call *nominal-adj-ph*, a sub-sort of *non-headed-phrase*.

In outline, what this construction must do is take a predicative AP, and produce a nominal, where the *index* of the nominal has the semantic role associated with the subject of the adjective. That is, something like (53):

(53) $\text{Nom}_{\boxed{1}} \rightarrow \text{AP}_{\left\langle NP_{\boxed{1}} \right\rangle}$

Though the outline of this analysis is straightforward, getting the details of it right in the G&S framework involves a slight complication as regards the semantics (i.e. the CONTENT). G&S assume that a predicative adjective like *beautiful* projects a phrase like (54) (and similarly for phrases like *obscenely rich*,

[16] While it is true that the semantic contribution of *the* does not involve the kind of familiarity requirement that one normally expects, it *does* reflect some notion of uniqueness, e.g. *the rich* denotes the totality of rich individuals (which is unique, of course).

barely awake, and *compulsively addicted to chocolate*, with suitable changes to the PHON and CONT values).

(54)
$$\left[\begin{array}{ll} \textit{phrase} & \\ \textsc{phon} & \left\langle \textit{beautiful} \right\rangle \\ \textsc{ss}\,|\,\textsc{loc} & \left[\begin{array}{ll} \textsc{cat} & \left[\begin{array}{ll} \textsc{subj} & \left\langle \mathrm{NP}_{\boxed{1}} \right\rangle \\ \textsc{head} & \left[\begin{array}{ll} \textit{adj} & \\ \textsc{pred} & + \end{array}\right] \end{array}\right] \\ \textsc{cont} & \textit{beautiful-rel}(\boxed{1}) \end{array}\right] \end{array}\right]$$

This is a predicative phrase which is looking for an NP subject, whose *index* is identified with the 'instance' of the semantic relation *beautiful-rel*.[17]

What we need is a construction that will take such a structure as its daughter and produce a nominal. The complication here is that we need to convert the 'predicative' semantics of the adjective to the 'attributive' semantics of a nominal. The content of a predicative adjective is assumed to be a *'state-of-affairs'* (*soa*), just like that of a verb. The CONTENT (CONT) value of a nominal is a *scope-object*, that is, and *index* and a set of *restrictions*, as in (55). We need to embed the CONTENT of the adjective as the *soa* in such a structure:[18]

(55)
$$\left[\begin{array}{ll} \textsc{index} & \textit{index} \\ \textsc{restr} & \left[\begin{array}{ll} \textit{fact} & \\ \textsc{prop} & \left[\begin{array}{ll} \textit{proposition} & \\ \textsc{soa} & \textit{soa} \end{array}\right] \end{array}\right] \end{array}\right]$$

Taking this into account, the construction can be formulated as in (56).

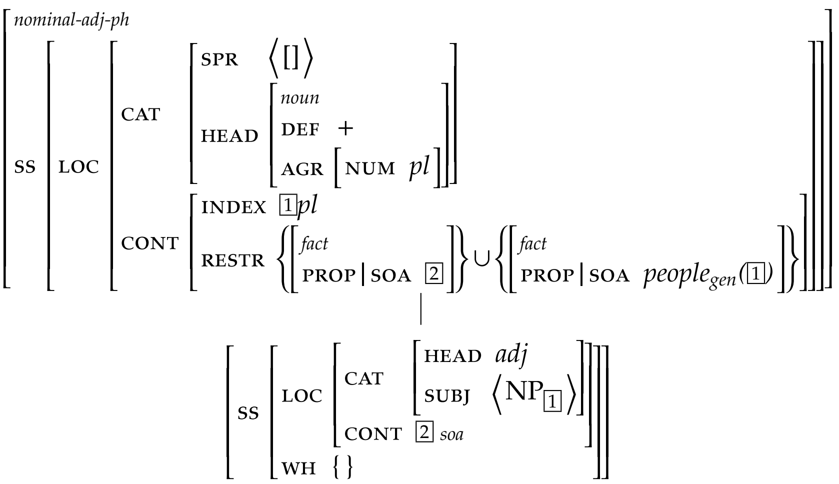
[17] *beautiful-rel*($\boxed{1}$)is an abbreviation for:

$$\left[\begin{array}{ll} \textit{soa} & \\ \textsc{nuc} & \left[\begin{array}{ll} \textit{beautiful-rel} & \\ \textsc{inst} & \boxed{1} \end{array}\right] \end{array}\right]$$

The PRED+ specification in (54) may be redundant, depending on one's view of whether attributive adjectives have SUBJ values – an issue we avoid here.

[18] This complication is a consequence of the G&S view of adjectival and nominal contents. It could be avoided if a more traditional semantics is assumed, where nominals and predictive adjectives are both of the semantic type $\langle e, t \rangle$. The semantics of the mother could just be given as the function in (i), applied to the content of the adjectival daughter.

(i) $\lambda P.\lambda x.P(x) \wedge people_{gen}(x)$

(56)


This construction takes an adjectival daughter, which is predicative (because lacking a SUBJ), and produces a nominal mother, which is definite, and plural. It combines the content of the adjective with a specification of the semantics of the nominal (*viz* that it is restricted to 'people$_{gen}$'), and identifies the *instance* of the adjective (i.e. the object the adjective is predicated of) with the index of the nominal. The intuitive effect is that an AP such as *beautiful* can be interpreted as denoting a plurality of beautiful individuals, as one would hope.

This nominal will combine with a determiner to produce structures like those in (52).

Let us now spell out how this construction accounts for the phenomena described in Section 2.

The definiteness restriction follows from the DEF+ specification on the mother nominal – only determiners that can combine with such a nominal will be permitted. Hence the contrast in (57):

(57) a. *This will convince even the skeptical.
b. *This will convince even a skeptical.

However, the construction places no constraints on its Specifier (the SPR value is required to be non-empty, but that is all). Thus it is predicted that any such determiner is possible, and we allow examples like (58) (= (33)):

(58) Most of [Asia's newly rich] are simply the first winners in a rush to own markets.

The mother nominal is specified as plural, hence the contrast in (59):[19]

[19] As with the lexical entry for the phonologically empty noun in (47), this specification is expressed both in the AGR value, and in the INDEX. Whether both need to be specified depends on how the association between these values is expressed. If it is expressed as a type constraint on phrases, then one or the other can be omitted here. However, if it is a lexical constraint, then both would be necessary.

(59) [The merely skeptical] are/*is easier to convince.

Since the AP is specified as having an unsaturated SUBJ, predicative adjectives are permitted, allowing (60) (and other examples from (30) above), and *only* predicative adjectives are permitted, excluding attributive only adjectives (process oriented, modal and emotive adjectives, as in (61):

(60) None but [the content] are truly happy.

(61) a. *[The alleged] have no opportunities here.
b. *[The frigging] need our help.

Notice that the daughter AP in (56) is specified as having an empty WH, as a result, while it will be able to include normal degree words, it will not be able to include expressions with non-empty WH values, such as *how* or *however*, thus accounting for the contrast in (62):[20]

(62) a. the very/obscenely/newly rich
b. *the how rich
c. *the however rich

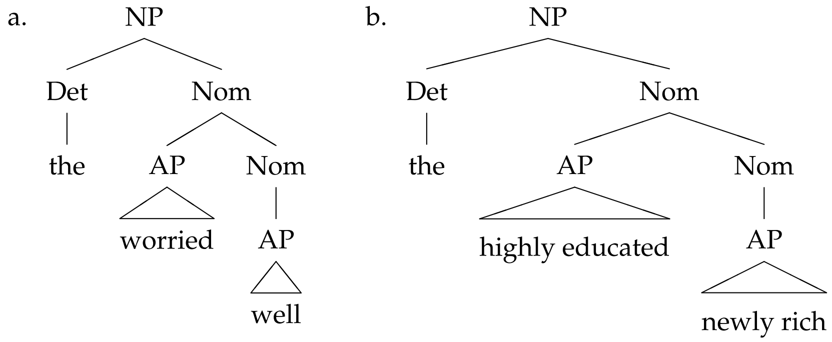
Other than the predicative WH restrictions, and the semantic constraint that it must be possible to use the adjective in relation to 'people$_{gen}$', there are no other restrictions on the adjectival expression. Thus, it can contain an adjective that normally appears pre-nominally or post-nominally, when used attributively, that can have complements, and be pre-modified:

(63) the barely awake (=(21))
(64) the compulsively addicted to chocolate (=(22))
(65) the extremely poor, the merely available, etc.

Adverbial modifiers appear within the AP, but since the construction produces a Nom, rather than an NP, there is no problem with adjectival modification — examples like *the worried well*, *the highly educated newly rich* will receive a representation along the lines of (66).

[20] As noted above, in the framework of G&S, WH-amalgamation and the Generalized Head Feature Principle ensure a phrase has a non-empty WH set if any of its constituents do. An empty WH value thus requires all sub-constituents to be similarly empty. *How* is a normal interrogative *wh*-word, so the non-empty WH specification is uncontroversial. *However* appears in 'exhaustive conditionals' like *However rich she becomes (I will not marry her)*, which are also interrogative (see e.g. Arnold and Borsley (2014) and references there), so it too should have a non-empty WH value.

(66)



This analysis thus avoids the empirical problems we have discussed in relation to the approaches which involve an empty noun, and existing constructional analyses.

4 Conclusion

The theoretical contribution of this paper has been to show that, contrary to what one might expect, it is possible to find clear empirical evidence that bears on the choice between a constructional analysis and one involving phonologically empty elements. In this case, the evidence favours a constructional account. In demonstrating this, we have given a detailed description of the characteristics of the ANH construction, a critique of some existing proposals, and provided explicit formalisations of both constructional and null-head analyses. Our constructional analysis in particular is empirically superior to existing accounts.

But of course, this is just one, rather idiosyncratic, construction in one language. It is the beginning, rather than the end, of the interesting questions.

The most immediate question it raises is where the *nominal-adj-phrase* specified in (56) fits into a general typology of non-headed constructions, in particular, the other English constructions mentioned in the Introduction. Equally interesting is the question of how this relates to similar constructions in other languages, where the facts are different – e.g. many languages allow *nominal-adj-phrase*-like constructions to be indefinite and singular, see *inter alia* Spencer (2002), Borer and Roy (2010).

References

Arnold, Doug and Borsley, Robert D. 2014. On the Analysis of English Exhaustive Conditionals. In Stefan Müller (ed.), *Proceedings of the 21th International Conference on Head-Driven Phrase Structure Grammar, University at Buffalo*, pages 27–47.

Bender, Emily M. 2001. *Syntactic Variation and Linguistic Competence: The Case of AAVE Copula Absence*. PhD thesis, Stanford University.

Borer, Hagit and Roy, Isabelle. 2010. The name of the adjective. In Patricia Cabredo Hofherr and Ora Matushansky (eds.), *Adjectives: Formal Analyses in Syntax and Semantics*, pages 85–113, Amsterdam: John Benjamins Publishing Co.

Bouma, Gosse, Malouf, Rob and Sag, Ivan A. 2001. Satisfying Constraints on Extraction and Adjunction. *Natural Language and Linguistic Theory* 1(19), 1–65.

Branco, Antnio and Costa, Francisco. 2006. Noun Ellipsis without Empty Categories. In Stefan Müller (ed.), *The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, pages 81–101, Stanford: CSLI Publications.

Fillmore, Charles J., Lee-Goldman, Russell R. and Rhomieux, Russell. 2012. The FrameNet Constructicon. In Hans C. Boas and Ivan A. Sag (eds.), *Sign-based Construction Grammar*, CSLI Lecture Notes, No. 193, pages 309–372, Stanford: CSLI Publications.

Ginzburg, Jonathan and Sag, Ivan A. 2001. *Interrogative Investigations: the form, meaning, and use of English Interrogatives*. Stanford, California: CSLI Publications.

Henri, Fabiola and Abeillé, Anne. 2007. The Syntax of Copular Construction in Mauritian. In Stefan Müller (ed.), *The Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*, pages 130–149, Stanford: CSLI Publications.

Huddleston, Rodney and Pullum, Geoffrey K (eds.). 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Jespersen, Otto. 1987. *Essentials of English Grammar: 25th impression*. London: Routledge.

Kim, Jong-Bok, Sells, Peter and Wescoat, Michael T. 2008. Korean copular constructions: A lexical sharing approach. In P. Clancy, M. E. Hudson, S.-A. Jun and P. Sells (eds.), *Japanese/Korean Linguistics*, volume 13, CSLI Publications.

Laurens, Frédéric. 2008. French Predicative Verbless Utterances. In Stefan Müller (ed.), *The Proceedings of the 15th International Conference on Head-Driven Phrase Structure Grammar*, pages 152–172, Stanford: CSLI Publications.

Miller, Philip H. 1992. *Clitics and Constituents in Phrase Structure Grammar*. New York: Garland.

Müller, Stefan. 2014. Elliptical Constructions, Multiple Frontings, and Surface-Based Syntax. In Paola Monachesi, Gerhard Jäger, Gerald Penn and Shuly Wintner (eds.), *Proceedings of Formal Grammar 2004, Nancy*, pages 91–109, Stanford, CA: CSLI Publications.

Nerbonne, John and Mullen, Tony. 2000. Null-headed nominals in German and English. In Frank Van Eynde, Ineke Schuurman and Ness Schelkens (eds.), *Proc. of Computational Linguistics in the Netherlands 1998*, pages 143–164, CLIN, Leuven.

Pullum, Geoffrey K. 1975. People Deletion in English. In ML Geis, SG Geoghegan and Arnold M Zwicky (eds.), *OSU Working Papers in Linguistics*, volume 14, pages 95–101, Columbus, OH: The Ohio State University.

Spencer, Andrew. 2002. Gender as an inflectional category. *Journal of Linguistics* 38, 279–312.

Tseng, Jesse L. 2003. Edge Features and French Liaison. In Jong-Bok Kim and Stephen Wechsler (eds.), *The Proceedings of the 9th International Conference on Head-Driven Phrase Structure Grammar*, pages 313–333, Stanford: CSLI Publications.

Van Eynde, Frank. 2007. The Big Mess Construction. In Stefan Müller (ed.), *The Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*, pages 415–433, Stanford: CSLI Publications.

Wescoat, Michael Thomas. 2002. *On lexical sharing*. PhD thesis, Stanford University.

Feeling our way to an analysis of English possessed idioms

Francis Bond
Nanyang Technological University

Jia Qian Ho
NanyanUniversity

Dan Flickinger
Stanford University

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 61–74

http://csli-publications.stanford.edu/HPSG/2015

Bond, Francis, Ho, Jia Qian, & Flickinger, Dan. 2015. Feeling our way to an analysis of English possessed idioms. In Müller, Stefan (Ed.), *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, 61–74. Stanford, CA: CSLI Publications.

Abstract

This paper describes an analysis for possessive idioms in English (e.g. *I twiddle my thumbs'* "I am idle"). The analysis relies on matching at the semantic level, to allow for syntactic variation. It has been implemented in the English Resource Grammar, and tested by parsing a subset of the British National Corpus. In addition to the syntactic analysis, we have linked the idioms to entries in the Princeton Wordnet, to allow for further lexical semantic analysis.

1 Introduction

Idiomatic constructions are very common in language, both at a type and token level. Despite considerable effort in categorizing and analyzing them (Nunberg et al., 1994; Moon, 1998; Sag et al., 2002) they are still not adequately represented in lexical resources, neither in lexicons such as wordnet (Fellbaum, 1998) or grammars such as the English Resource Grammar (Flickinger, 2000).

In this paper we focus on possessive idiomatic constructions: prototypically those in which one constituent is modified by a possessive pronoun co-indexed with a different constituent (typically the subject). A typical example is ***wrack one's brains*** "think hard", where the possessor of the brains must be the subject: *I wrack my brains*; *You wrack your brains*; *Kim wracks their brains*. These are interesting theoretically because of the interaction between syntax and semantics and are also of practical interest in translation (Bond, 2005). Most languages, even with similar idioms, do not include this possessive expression. For example, the equivalent phrase in Japanese is ***chie-wo shiboru*** "think hard: lit., squeeze knowledge". In this case it is a verb phrase with a fixed object, but there is no possessive.

The immediate motivation for this research was for machine translation: when translating out of English, typically the idiomatic possessive pronoun should be omitted. Going the other way, the possessive pronoun must be generated, and it must agree with the subject. Shallow statistical systems often get this wrong. A complete list of these idioms may also be useful for computer-assisted language learning. For example, an English learner can engage with the materials developed on corpora to understand figurative language, which is a more difficult aspect of language to learn, and to understand how pronouns operate in both literal and figurative English.

For example *Kim racks her brains* "Kim thinks hard" is given the unlikely literal translation by the statistical machine translations systems used by Google and Bing translate (1: translated on 2015-10-16).

(1) *Kim racks her brains*

a. キムは、 彼女の 脳を ラック
Kimu wa, kanojo no nou o rakku
Kim-TOP her-'s brain-ACC rack

Kim [dish] rack her brain (Google Translate)

a. キムは、 彼女の 脳を ラックします
Kimu wa, kanojo no nou o rakku shimasu
Kim-TOP her-'s brain-ACC rack do

Kim racks her brain (puts her brain in a [dish] rack) (Bing Translate)

In the following sections we present our idiom database, our analysis, and some corpus results,

2 The Idiom Database

In order to study their behavior we collected idioms from that included possession from a variety of sources, including WordNet (Fellbaum, 1998) and on-line lexicons such as Dictionary.com (2012). We ended up with 514 idioms:[1] very similar idioms have been merged into one entry (*to rack/wrack one's brains*) and idioms with two interpretations are treated as separate entries. These were categorized into different classes based on their syntactic and semantic structure. In addition, we attempted to give more literal paraphrases: ***wrack one's brains*** ~ ***think hard***. Because of the variance in the possessive pronoun, it is hard to extract these automatically even using sophisticated methods (Zhang et al., 2006). For this reason, we are trying to cover as many as possible manually.

These idioms were categorized into co-indexed and separate possessive idioms and further grouped syntactically. We list the most common types of co-indexed idioms in Table 1. X_{NP}, Y_{NP} and Z_{NP} denote variable noun phrases, N for invariable noun, V for verb, A for adjective, R for adverb, D for determiner, aux for auxiliary and neg for negation. Square brackets [ ] denote prepositional phrases (PP). Within these brackets, *P* denotes a preposition; elsewhere, *P* represents a particle.

We give two examples of individual **idiom entries** in (2) and (3).

Definitions were written based on online dictionaries. Individual open-class words were linked to senses wordnet (by intuition, no deep etymological search was made).

[1] Available from `http://compling.hss.ntu.edu.sg/idioms/possessed`.

Table 1: Types of Co-indexed Possessive Idioms

Structure	**Example**	**Frequency**
X_{NP} V_1 X's N_1	lose one's mind	137
X_{NP} V_1 [P_1 X's N_1]	fly off one's handle	40
X_{NP} V_1 X's N_1 [P_1 Y_{NP}]	cast one's lot [with someone/thing]	39
X_{NP} V_1 X's N_1 [P_1 D_1 N_2]	have one's head [in the clouds]	27
X_{NP} V_1 X's N_1 P_1	cry one's eyes out	22
X_{NP} V_1 X's own N_1	blow one's own horn	18
X_{NP} V_1+P_1 X's N_1	pull up one's socks	17
X_{NP} be [P_1 X's N_1]	off one's rocker	13
X_{NP} V_1 X's N_1 [P_1 X's N_2]	scratch one's ear [with one's elbow]	13
X_{NP} V_1 D_1 N_1 [P_1 X's N_2]	a dose [of one's medicine]	10
X_{NP} V_1 X's N_1 A_1	get one's hands dirty	10
X_{NP} V_1 Y_{NP} [P_1 X's N_1]	wind someone [around one's finger]	10
X_{NP} V_1 X's N_1(est)	do one's best	8
X_{NP} V_1 [P_1 X's N_1 [P_2 Y_{NP}]]	pour out one's heart [to someone]	7
X_{NP} aux+neg V_1 X's N_1	not mince one's words	5
X_{NP} V_1 Y_{NP} D_1 N_1 [P_1 X's N_2]	give someone a piece [of one's mind]	4
X_{NP} V_1 R_1 A_1 [P_1 X's N_1]	too big [for one's boots]	3
X_{NP} V_1 [P_1 D_1 N_1 P_2 X's N_2]	by the skin of one's teeth	2
X_{NP} V_1 N_1 [P_1 X's N_2]	have egg [on one's face]	2
X_{NP} V_1 X's N_1 [P_1 X]	have one's wits [about one]	2
X_{NP} V_1 X's N_1 and V_2 N_2	have one's cake and eat it	2
Remainder	let grass grow under one's feet	30
Total		**421**

This table lists the co-indexed possessive idioms, arranged in order of type frequency, with the exception of the last group, *remainder*

If the idiom is decomposable, a synonym or metaphorical extension for each component was chosen (marked with *) as in (2) and also linked to synsets in WordNet. Idiom decomposability is shown in @type.

Idiom decomposability was determined by **semantic substitution**: whether a lexical component can be replaced by appropriate word without altering its syntactic structure. In (2), *eat* is metaphorically extended to mean "withdraw" (*V_1) while *words* with "statement" (*N_1), to give "withdraw one's statement". This is the idiomatic meaning of the expressions, it is thus deccomposable. In contrast in (3), *twiddle* and *thumb* cannot be replaced with suitable synonyms nor metaphorical extensions, without altering the syntactic structure. The figurative meaning is "to be idle". Consequently, *twiddle one's thumb* is nondecomposable.

(2)

Idiom entry — fully projected	
Index form	eat one's words
Template	X_{NP} V_1 X's N_1
Example	Kim eats her words
Example	Kim is going to have to eat her words
Definition	to retract one's statement, especially with humility
V_1	(v) eat (take in solid food)
N_1	(n) words (the words that are spoken)
$^*V_1$	(v) swallow, take back, unsay, withdraw (take back what one has said)
$^*N_1$	(n) statement (a message that is stated or declared; a communication (oral or written) setting forth particulars or facts etc)
@type	decomposable

All non-decomposable idioms were given paraphrases, also linked to WordNet, marked with @ in their idiom entries. Decomposable idioms are paraphrasable with the extensions, so there is no need to list a separate paraphrase. In this case, the idiomatic meaning of the head (*V) will be the hypernym of the idiom. For non-decomposable examples, the head will also be the hypernym. However, where the paraphrase involves a copula and adjective, as in (3), the adjective paraphrase (@A) will be the hypernym of the idiom. This paraphrase captures the basic essence of each idiom and illustrates its hyponymy relation to lexical entries already listed in WordNet.

(3)

Idiom entry — non-projected	
Index form	twiddle one's thumbs
Template	X_{NP} V_1 X's N_1
Example	Kim twiddles her thumbs
Definition	to do nothing
V_1	(v) twiddle, fiddle with (manipulate, as in a nervous or unconscious manner)
N_1	(n) thumb, pollex (the thick short innermost digit of the forelimb)
@type	Nondecomposable
Paraphrase	X is idle
@template	X BE A
@A	(adj) idle (not in action or at work))

3 Analysis

The syntactic analysis uses idiom machinery inspired by Copestake (1994) and extended in Riehemann (2001); Copestake et al. (2002); Sag et al. (2002). It is implemented in the latest version of the English Resource Grammar (ERG: Flickinger, 2000, 2011). The relationship between the words in the idiom is captured using a fundamentally semantic mechanism, in our case encoded using Minimal Recursion Semantics (MRS: Copestake et al., 2005). Special lexical items introduce idiomatic predicates (marked as such in the lexicon). Idioms are treated as bags of predicates, with relations between them partially specified. If the semantics of a sentence can match this, then it has the idiomatic reading. This allows for considerable syntactic flexibility. During parsing, if a word has an idiom in it, a final check is made by the grammar when it enforces the root condition. Each idiomatic predicate must be licensed by at least one rule, otherwise the idiomatic interpretation is rejected.

Miyazaki et al. (1993) suggest that for some idioms we should allow nodes in a semantic hierarchy (so any noun with compatible semantics is allowed). We have linked the predicates in the idiom to their literal meanings (5) and the predicates in their paraphrases to the intended meaning using Wordnet synsets (6), but this is not used during parsing. Minor variations can easily be captured in the lexicon. For example, there are two alternative spellings of ***wrack***: *wrack* and *rack*. If we treat them as having no difference in meaning at all, then we represent them as two lexical items with different orthography, but the same predicate.

The interesting thing about the possessive idioms is that they also include an identity relation *id* to enforce the co-indexation. This is introduced by a special idiomatic verb-type, but could conceivable come from some kind of co-reference resolution. We give the bag of idioms that licenses ***wrack one's brains*** in (7).

(4) I$_i$ rack my$_i$ brains. [X Vs Y's Z; X=Y]

(5) Literal: ***rack***$_{v:9}$ "stretch on a rack"; ***brains***$_{n:1}$ "encephalon"

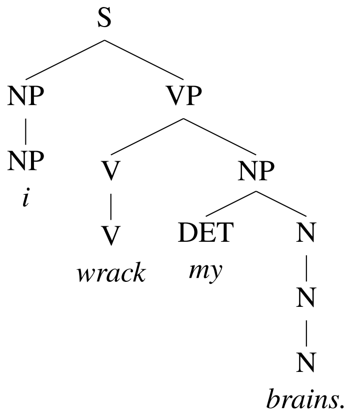
(6) Paraphrase: ***think***$_{v:1}$ "cogitate"; ***hard***$_{r:1}$ "with effort"

(7)
$$\left[\begin{array}{ll}
\textit{mrs} & \\
\text{LTOP} & \boxed{h1}\ h \\
\text{INDEX} & \boxed{e3}\ e \\
\text{RELS} & \left\langle \begin{array}{l}
\left[\begin{array}{ll} \textbf{\textit{\_rack\_v\_i}} & \\ \text{LBL} & \boxed{hv} \\ \text{ARG0} & \boxed{v} \\ \text{ARG1} & \boxed{x} \\ \text{ARG2} & \boxed{z} \end{array}\right],
\left[\begin{array}{ll} \textit{id} & \\ \text{LBL} & \boxed{hv} \\ \text{ARG0} & \boxed{id} \\ \text{ARG1} & \boxed{x} \\ \text{ARG2} & \boxed{y} \end{array}\right], \\
\left[\begin{array}{ll} \textit{poss} & \\ \text{LBL} & \boxed{hz} \\ \text{ARG0} & \boxed{ps} \\ \text{ARG1} & \boxed{z} \\ \text{ARG2} & \boxed{y} \end{array}\right],
\left[\begin{array}{ll} \textbf{\textit{\_brain\_n\_i}} & \\ \text{LBL} & \boxed{hz} \\ \text{ARG0} & \boxed{z} \end{array}\right]
\end{array}\right\rangle
\end{array}\right]$$

The idiomatic ***wrack one's brains*** thus has three elements in the grammar: a lexical entry that introduces ***_brains_n_i***, a lexical entry that introduces ***_rack_v_i*** and *id* and links them appropriately; and an idiomatic rule that makes sure all the relevant elements are there: the above three predicates, and the possessive relation. The linking is crucial: the identity rel is linked to the external argument (XARG) of the verb (the subject) and to the external argument of the first element of the COMPS list (the determiner of the object). This links the subject to the possessor of the object. The idiom allows for variation in number: both *I rack my brain* and *I wrack my brains* are attested. In this case, we underspecify number in the construction, and allow both.

We give the parse tree and full MRS for (4) in (8) and (9), respectively.

(8)



(9)
$$\begin{bmatrix}
\textit{mrs} & \\
\text{TOP} & \boxed{0}\,h \\
\text{INDEX} & \boxed{2}\,e \\
\text{RELS} & \left\langle \begin{array}{l}
\begin{bmatrix} \textit{pron\_rel} & \\ \text{LBL} & \boxed{4}\,h \\ \text{ARG0} & \boxed{3}\,x \end{bmatrix},
\begin{bmatrix} \textit{pronoun\_q\_rel} & \\ \text{LBL} & \boxed{5}\,h \\ \text{ARG0} & \boxed{3}\,x \\ \text{RSTR} & \boxed{6}\,h \\ \text{BODY} & \boxed{7}\,h \end{bmatrix}, \\
\begin{bmatrix} \textit{\_wrack\_v\_i\_rel} & \\ \text{LBL} & \boxed{1}\,h \\ \text{ARG0} & \boxed{2}\,e \\ \text{ARG1} & \boxed{3}\,x \\ \text{ARG2} & \boxed{8}\,x \end{bmatrix},
\begin{bmatrix} \textit{id\_rel} & \\ \text{LBL} & \boxed{1}\,h \\ \text{ARG0} & \boxed{9}\,i \\ \text{ARG1} & \boxed{3}\,x \\ \text{ARG2} & \boxed{10}\,x \end{bmatrix}, \\
\begin{bmatrix} \textit{def\_explicit\_q\_rel} & \\ \text{LBL} & \boxed{11}\,h \\ \text{ARG0} & \boxed{8}\,x \\ \text{RSTR} & \boxed{12}\,h \\ \text{BODY} & \boxed{13}\,h \end{bmatrix},
\begin{bmatrix} \textit{poss\_rel} & \\ \text{LBL} & \boxed{14}\,h \\ \text{ARG0} & \boxed{15}\,e \\ \text{ARG1} & \boxed{8}\,x \\ \text{ARG2} & \boxed{10}\,x \end{bmatrix}, \\
\begin{bmatrix} \textit{pronoun\_q\_rel} & \\ \text{LBL} & \boxed{16}\,h \\ \text{ARG0} & \boxed{10}\,x \\ \text{RSTR} & \boxed{17}\,h \\ \text{BODY} & \boxed{18}\,h \end{bmatrix},
\begin{bmatrix} \textit{pron\_rel} & \\ \text{LBL} & \boxed{19}\,h \\ \text{ARG0} & \boxed{10}\,x \end{bmatrix},
\begin{bmatrix} \textit{\_brain\_n\_1\_rel} & \\ \text{LBL} & \boxed{14}\,h \\ \text{ARG0} & \boxed{8}\,x \end{bmatrix}
\end{array} \right\rangle \\
\text{HCONS} & \text{(omitted for simplicity)}
\end{bmatrix}$$

The other idiom types are implemented in a similar way: the main predicate (verb or preposition) adds and links the identity relation. For some cases, such as ***keep one's cards close to one's chest*** (e.g. in *You$_i$ keep your$_i$ cards close to your$_i$ chest.*), it has to add two identity predicates. The idiomatic licensing rule for this is given in (11). A different kind of idiom observed was the double co-index idiom. The syntactic shape of such idioms is N$_1$ V N$_1$'S N$_2$ (PP) (CONJ) (V) N$_1$'s N$_3$. These instances belong to the less frequently observed extended structure idioms, which were essentially basic in shape but modified through post-insertions or by embedding. In this case, a second possessive noun phrase was added to an idiom that would otherwise be of basic shape but lack the idiom's meaning. A double co-indexing idiom looks similar to a basic idiom with the exception of the embedded possessive noun phrase. One instance of such an idiom is shown below:

(10) You$_i$ keep your$_i$ cards close to your$_i$ chest. [X keeps Y'S cards close to Z's chest; X=Y=Z]

(11)
$$\begin{bmatrix} \textit{mrs} & \\ \text{LTOP} & \boxed{h1}\ h \\ \text{INDEX} & \boxed{e3}\ e \\ \text{RELS} & \left\langle \begin{array}{l} \begin{bmatrix} \textit{\_keep\_v\_i\_rel} & \\ \text{LBL} & \boxed{h2}\ h \\ \text{ARG0} & \boxed{e3} \\ \text{ARG1} & \boxed{x} \\ \text{ARG2} & \boxed{card} \\ \text{ARG3} & \boxed{h9} \end{bmatrix}, \begin{bmatrix} \textit{id\_rel} & \\ \text{LBL} & \boxed{h2} \\ \text{ARG0} & \boxed{e3}\ i \\ \text{ARG1} & \boxed{x} \\ \text{ARG2} & \boxed{y} \end{bmatrix}, \begin{bmatrix} \textit{poss\_rel} & \\ \text{LBL} & \boxed{h13}\ h \\ \text{ARG0} & \boxed{e15}\ e \\ \text{ARG1} & \boxed{card} \\ \text{ARG2} & \boxed{y} \end{bmatrix}, \\ \begin{bmatrix} \textit{\_card\_n\_i\_rel} & \\ \text{LBL} & \boxed{h14} \\ \text{ARG0} & \boxed{card} \end{bmatrix}, \begin{bmatrix} \textit{\_close\_a\_to} & \\ \text{LBL} & \boxed{h21}\ h \\ \text{ARG0} & \boxed{e22}\ e \\ \text{ARG1} & \boxed{card} \\ \text{ARG2} & \boxed{chest} \end{bmatrix}, \begin{bmatrix} \textit{id\_rel} & \\ \text{LBL} & \boxed{h2} \\ \text{ARG0} & \boxed{e4}\ i \\ \text{ARG1} & \boxed{x} \\ \text{ARG2} & \boxed{z} \end{bmatrix}, \\ \begin{bmatrix} \textit{poss\_rel} & \\ \text{LBL} & \boxed{h27}\ h \\ \text{ARG0} & \boxed{e29}\ e \\ \text{ARG1} & \boxed{chest} \\ \text{ARG2} & \boxed{z} \end{bmatrix}, \begin{bmatrix} \textit{\_chest\_n} & \\ \text{LBL} & \boxed{h27}\ h \\ \text{ARG0} & \boxed{chest} \end{bmatrix} \end{array} \right\rangle, \\ \text{HCONS} & \text{(omitted for simplicity)} \\ \text{ICONS} & \langle\ \rangle \end{bmatrix}$$

There is a long tail of rare types: as Richter & Sailer (2009) point out, some of these idioms can even go across clause boundaries, for example: ***look as though butter wouldn't melt in one's mouth*** "appear innocent". Currently we have created idiom types for the most common classes of idiom (all those with a type frequency of greater than eight) and instantiated them with idiom rules for each of the entries in the database. In future work, we will keep working our way down the long tail.

While the two-place *id* predication appearing in the RELS lists of the above examples (7,11) was implemented and used for most of the empirical work reported here, we have also been developing an alternative representation of the identification of the possessor in our idioms with the external argument of the verb. Building on the notion of sets of constraints on pairs of individuals proposed for information structure by Song (2015), we can express the relevant identity in our idioms not as a predication but as an ICONS (”individual constraint”) pair. While binding constraints on intrasentential anaphors in general are still under development for the ERG, these ICONS pairs seem well-suited for expressing both coreference and noncoreference constraints imposed by the syntax, and that promise leads us to express these idiom-specific identities with the same formal mechanism. One advantage of removing the *id* predication from the RELS list is that we no longer have to engineer the assignment of the LBL for that predication; note that in our example above, that label value is identified with the label of *_rack_v_i*, but this is both awkward to ensure compositionally, and lacking in independent motivation. By using the ICONS representation instead, we clearly distingish coreference constraints between pairs of individuals from the contentful semantic predications that comprise the RELS list and are subject to scopal operators including quantifiers, modals, negation, and the like.

Sheinfux et al. (2015) also propose a method to handle idioms of this type in Hebrew. In their analysis, the verb selects for a special kind of argument, and the agreement properties are passed up using the XARG. This does not require our (independently motivated) idiom processing, but does require special lexical entries not just for the verb, but also the noun, the possessor and any prepositions involved in the idiom.

In future work, we will think further as to how to mark the idioms in the output semantic representation. Currently, the individual elements are marked as idiomatic. During processing we know which idiom was licensed (as we know which idiom rule applies), but this information is not part of the final MRS. Further, the possessive pronouns are not marked in any way, even though intuitively they are less meaningful than real referential pronouns. Both these issues are also relevant to the separate possessive idioms. One approach is to keep decomposed idioms as they are (but specify their predicates to have the idiomatic meanings) and paraphrase the non-decomposable ones, thus doing away with the non-referential pronouns altogether.

4 Testing on a corpus (the BNC)

We ran the extended ERG over the British National Corpus (Burnard, 2000) to identify actual examples of these idioms. We attempted to parse the first 3,494,381 sentences.[2] We were able to successfully parse 3,011,023 of the sentences (86%)

[2] This took 44 days on 20 CPUs, after which we had to stop to apply a security patch to the server. We are currently looking for a bigger server cluster.

and found 5,577 sentences with possible idioms (0.18%). We are the first to identify these idioms in the BNC. Up until now it has been hard to find these kinds of idioms, due to the complicated structure. With idioms implemented in a flexible grammar, they can be identified automatically.

A manual check of the first 319 idiom instances showed that 76.7% were being used idiomatically. The relatively high percentage shows that these complex idioms are typically used idiomatically. To distinguish between idiomatic and non-idiomatic uses we need to retrain the the parse ranking model with idiomatic examples and/or learn a special model to distinguish idiomatic from non-idiomatic uses (such as, Hashimoto & Kawahara, 2009).

The ten most common idiom types are shown in Table 2. The idiom ***shake one's head*** was the most common. In many cases, it was clearly referring both to the physical act of shaking one's head, and to the idiomatic meaning of "indicate disagreement". ***bite one's lip*** "forcibly prevent oneself from speaking" was similar: often both the literal and idiomatic meanings were applicable at the same time.

Table 2: Most common possessive idioms found in the British National Corpus

Idiom	**Frequency**	**Comment**
shake ones' head	2,055	
make one's way	359	often both idiomatic and literal
open one's eye	344	mainly non-idiomatic
find ones' way	205	
bite one's lip	145	
get one's way	131	
have one's way	139	
raise one's eyebrows	124	
shrug one's shoulders	118	
lose one's temper	113	

Current dictionaries rarely list idiom frequencies, this corpus-based study offers not just useful information for lexicographers, but also for improving translation systems by informing programmers which idioms to focus on. Future work can thus continue from this preliminary study and work on the other syntactic templates identified in section 2.

Finally, the BNC findings showed some interesting examples of syntactic flexibility, including modification, relativization and long distance dependencies, as shown (12). All of these were successfully identified by the ERG, although would be very hard to identify successfully using shallow chunk based systems. There were many more examples of modifications using adjectives such as *cannot believe my own bloody eyes*, *make one's unsteady way* and *have one's humorous moment*. This is an area we will continue to investigate by running a larger idiom sample through the corpus.

(12) a. *The butcher had lined his pockets too thickly in the past at their expense, and Faith's will had been a warning, a pointer to their future.*

b. *Now do thy speedy utmost, Meg,*

c. *Even if she is an overpaid brat in danger of losing her marbles, at least she provokes a reaction, and is 500 times more controversial than Madonna.*

d. *And if everybody starts getting very large discounts and the vendor loses control of the market, not only do the buyers lose all their advantage, but the vendor loses its corporate shirt.*

e. *Nor is it the case that the Federal Republic is using the issue of democratic accountability to drag its feet on EMU.*

f. *Mr Waddington, a former immigration minister and rightwinger, seems to have gritted his teeth at yesterday's meeting and stood by the compromise hammered out at Mrs Thatcher's insistence in a cabinet committee.*

g. *I'm starting to lose my bearings a bit—and my ball-bearings as well, come to that.*

With more data we can examine more reliably other aspects of syntactic flexibility, such as modification, quantification and topicalization, allowing us to test the claims of Nunberg et al. (1994). They distinguish idiomatically combining expressions (ICEs: our decompositional) and idiomatic phrases (IPs: our non-decompositional) with five tests: modification, quantification, topicalization, ellipsis, and anaphora.

5 Conclusions

We have implemented an analysis of co-indexed possessive idioms in HPSG, suitable for use in a computational grammar. We have tested an implementation of the major types of idiom in the English Resource Grammar and linked the predicates to wordnet. We are currently experimenting with expanding our variants and identifying corpus examples. As well as implementing in the ERG, the full idiom lexicon, including definitions, examples and links to wordnets is freely available under an open licence (CC-BY) at: `http://compling.hss.ntu.edu.sg/idioms/possessed/`.

Acknowledgments

We would like to thank the reviewers and participants of HPSG 2015 for their helpful comments and discussion. This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Bond, Francis. 2005. *Translating the untranslatable: A solution to the problem of generating English determiners* CSLI Studies in Computational Linguistics. CSLI Publications.

Burnard, Lou. 2000. *The British National Corpus users reference guide*. Oxford University Computing Services.

Copestake, Ann. 1994. Representing idioms. Presentation at the HPSG Conference, Copenhagen.

Copestake, Ann, Dan Flickinger, Ivan A. Sag & Carl Pollard. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation* 3(2). 281–332.

Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan Sag & Dan Flickinger. 2002. Multiword expressions: Linguistic precision and reusability. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 1941–7. Las Palmas, Canary Islands.

Dictionary.com. 2012. Free online English dictionary. `http://dictionary.reference.com/`.

Fellbaum, Christine (ed.). 1998. *WordNet: An electronic lexical database*. MIT Press.

Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6(1). 15–28. (Special Issue on Efficient Processing with HPSG).

Flickinger, Dan. 2011. Accuracy vs. robustness in grammar engineering. In Emily M. Bender & Jennifer E. Arnold (eds.), *Language from a cognitive perspective: Grammar, usage, and processing*, 31–50. Stanford: CSLI.

Hashimoto, Chikara & Daisuke Kawahara. 2009. Compilation of an idiom example database for supervised idiom identification. *Language Resources and Evaluation* 43(4). 355–384.

Miyazaki, Masahiro, Satoru Ikehara & Akio Yokoo. 1993. Combined word retrieval for bilingual dictionary based on the analysis of compound word. *Transactions of the Information Processing Society of Japan* 34(4). 743–754. (in Japanese).

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.

Nunberg, Geoffrey, Ivan A. Sag & Tom Wasow. 1994. Idioms. *Language* 70. 491–538.

Richter, Frank & Manfred Sailer. 2009. Phraseological clauses in constructional HPSG. In Stefan Müller (ed.), *Proceedings of the 16th international conference on Head-Driven Phrase Structure Grammar, university of Göttingen, germany*, 297–317. Stanford: CSLI Publications. `http://cslipublications.stanford.edu/HPSG/2009/`.

Riehemann, Susanne Z. 2001. *A constructional approach to idioms and word formation*: Stanford dissertation.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbuk (ed.), *Computational linguistics and intelligent text processing: Third international conference: Cicling-2002*, 1–15. Hiedelberg/Berlin: Springer-Verlag.

Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik & Shuly Wintner. 2015. Hebrew verbal multi-word expressions. In *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, 123–136.

Song, Sanghoun. 2015. Representing honorifics via individual constraints. In *ACL 2015 workshop on grammar engineering across frameworks (GEAF 2015)*, .

Zhang, Yi, Valia Kordoni, Aline Villavicencio & Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the workshop on multiword expressions: Identifying and exploiting underlying properties*, 36–44. Sydney, Australia: Association for Computational Linguistics. `http://www.aclweb.org/anthology/W/W06/W06-1206`.

Lacking Integrity: HPSG as a Morphosyntactic Theory

Guy Emerson
University of Cambridge

Ann Copestake
University of Cambridge

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 75–95

http://csli-publications.stanford.edu/HPSG/2015

Emerson, Guy, & Copestake, Ann. 2015. Lacking Integrity: HPSG as a Morphosyntactic Theory. In Müller, Stefan (Ed.), *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, 75–95. Stanford, CA: CSLI Publications.

Abstract

Standard accounts of HPSG assume a distinction between morphology and syntax. However, despite decades of research, no cross-linguistically valid definition of 'word' has emerged (Haspelmath, 2011), suggesting that no sharp distinction is justified. Under such a view, the basic units are morphemes, rather than words, but it has been argued this raises problems when analysing phenomena such as zero inflection, syncretism, stem alternations, and extended exponence. We argue that with existing HPSG machinery, a morpheme-based approach can in fact deal with such issues. To illustrate this, we consider Slovene nominal declension and Georgian verb agreement, which have both been used to argue against constructive morpheme-based approaches. We overcome these concerns through use of a type hierarchy, and give a morpheme-based analysis which is simpler than the alternatives. Furthermore, we can recast notions from Word-and-Paradigm morphology, such as 'rule of referral' and 'stem space', in our framework. We conclude that using HPSG as a unified morphosyntactic theory is not only feasible, but also yields fruitful insights.

1 Word Segmentation and Lexical Integrity

The Lexical Integrity Principle holds that syntactic rules do not have access to internal parts of words (Bresnan & Mchombo, 1995; Asudeh et al., 2013). Although this principle is often not explicitly stated in HPSG, it is usually implicitly assumed that there there is some notion of 'word', with a corresponding division of labour between lexical and phrasal rules. For example, Sag et al. (2003, p.228ff.) describe the use of lexemes as abstract structures from which we can derive families of wordforms differing only by inflection, where this process is carried out using lexical rules. However, while they take it for granted that we can identify words, the difficulties in defining the term 'word' have been known for some time:

> "Many forms lie on the border-line between bound forms and words, or between words and phrases; it is impossible to make a rigid distinction" — (Bloomfield, 1933)
>
> "What we call 'words' in one language may be units of a different kind from the 'words' in another language" — (Lyons, 1968)
>
> "There may be clear criteria for wordhood in individual languages, but we have no clear-cut set of criteria that can be applied to the totality of the world's languages" — (Spencer, 2006)

More recently, Haspelmath (2011) identifies ten possible criteria for defining words: potential pauses, free occurrence, mobility, uninterruptibility, non-selectivity, non-coordinability, anaphoric islandhood, non-extractability, phonological idiosyncrasies, and non-biuniqueness. They argue that none of these criteria, nor any combinations of them, coincide with linguistic or orthographic practice. Furthermore, we cannot retreat by saying that words are simply a language-specific concept. If we are forced to define words separately for each language, then we can quite easily define several word-like levels in any particular language, and we have no reason to give special status to one particular level. Moreover, given a language like Mandarin Chinese, where linguists cannot agree on what to call the Mandarin Word (Mair, 1990; Packard, 2000; Sun, 2006; Tang, 2010), this is not an obscure thought experiment, but a fundamental issue affecting the most widely spoken language on the planet. Haspelmath concludes:

> "Linguists have no good basis for identifying words across languages, and hence no good basis for a general distinction between syntax and morphology" — (Haspelmath, 2011)

Under this view, the distinction between morphology and syntax vanishes, leaving us with a single domain of morphosyntax, with abstract morphemes as the basic units. This pushes us towards an Item-and-Arrangement view of morphological phenomena, rather than Item-and-Process or Word-and-Paradigm (WP) views, since the latter approaches require a notion of 'word'. In Stump (2001)'s terms, we are pushed towards a lexical and incremental theory, rather than an inferential or realizational one. In Blevins (2006)'s terms, we are pushed towards a constructive theory, rather than an abstractive one. However, this is not to say that we must abandon progress made in these other frameworks. Far from it – many generalizations stated in word-based accounts can be re-expressed in morphemic terms, and we will discuss several in this paper.[1] Doing so allows us to frame them in a theory that is cross-linguistically more consistent, and where the analyses can mesh seamlessly with syntax above the 'word' level.

If we accept Haspelmath's conclusion, we are prompted to consider whether we can reformulate HPSG in terms of morphemes. In the following section, we argue not only that this is possible, but further that the use of type hierarchies makes HPSG particularly appealing as a morphosyntactic theory, as it can sidestep many problems attributed to morphemic approaches. In sections 3 and 4, we apply this framework to Slovene stem alternations and Georgian verb agreement, which have been claimed to pose problems for a morphemic approach. Along the way, we show how insights drawn from WP morphology can be recast in our framework.

2 Morphosyntactic HPSG

Recasting HPSG as a morphosyntactic theory can be done without fundamental changes to the architecture. HPSG is usually regarded as a lexicalist theory, but while the term 'lexicalism' has often been associated with lexical integrity, particularly as the term is used by transformational grammarians, we only require a relatively minor change to Sag et al.'s definition of 'strong lexicalism'. This states

[1] Roark & Sproat (2007) also demonstrate that lexical-incremental theories and inferential-realizational theories are computationally equivalent, since both can be implemented in the same model, using an FST.

firstly that the locus of grammatical and semantic information is the lexicon, and secondly that lexical entries correspond directly to the words present in a sentence.[2] We must only state instead that lexical entries correspond to morphemes, not words:

$$morpheme \rightarrow LE_1 \vee \cdots \vee LE_n$$

These lexical entries must be minimal, rather than derived by lexical rules. To formalize this idea, we propose the following (meta)principle:

> **The Morphemic Principle:** Phonological material may only be stipulated in lexical entries, not in syntactic or lexical rules.

This implies that the only way to combine phonological material is by combining lexical entries through non-unary syntactic rules, i.e. by combining morphemes. Furthermore, phonological material is not split between lexical rules and lexical entries – all morphemes are stored directly in the lexicon. This would remain true no matter what the orthographic conventions are, so adhering to such a principle would make grammars more consistent cross-linguistically.

A second reference to words lies in the Head-Complement Schema, which builds a phrase out of a word and its complements (Pollard & Sag, 1994; Sag et al., 2003). Without a notion of 'word', this instead becomes a process of building one type of phrase out of a second type of phrase and its complements. What this means is that the Head-Complement Schema must be restated in terms of pairs of types (t_1, t_2):

$$\begin{bmatrix} t_1 & \\ \text{...HEAD} & \boxed{1} \\ \text{...COMPS} & \left\langle \right\rangle \end{bmatrix} \rightarrow \begin{bmatrix} t_2 & \\ \text{...HEAD} & \boxed{1} \\ \text{...COMPS} & \left\langle \boxed{2} \right\rangle \end{bmatrix}, \boxed{2}$$

Allowing phrases to be the head daughter of a head-complement construction has in fact been motivated independently. Instead of a flat structure where the head combines with all complements at once, we can use a binary-branching structure where the head combines with one complement at a time, which allows adjuncts or subjects to intervene between the head and its complements. For example, such an approach is used in the English Resource Grammar (Flickinger et al., 2000), in the Grammar Matrix (Bender et al., 2002), to analyse the German Mittelfeld (Crysmann, 2003), and to analyse partial-VP fronting (Müller, 2015).

In conclusion, neither of the above changes are inherently problematic. However, after removing lexical rules from the theory, and assuming morphemes to be the basic units, we need to justify that it is still possible to capture phenomena traditionally regarded as morphological. In section 2.1, we clarify what we mean by 'morpheme'; in section 2.2, we review the difficulties attributed to a morphemic view; and in section 2.3, we show how the criticisms made on morphosyntactic grounds do not apply when using feature structures and a type hierarchy.

[2] The second half of this statement is also known as the Word Principle.

2.1 What is a Morpheme?

In order to shift to a morpheme-based view of morphosyntax, we have to ask whether morphemes can be more easily identified than words. However, a number of different definitions of 'morpheme' have been proposed in the literature, with some more problematic than others.

We follow Bender & Good (2005)'s notion of an 'abstract morpheme'. Under this view, we assume that a language can be split between the *morphophonology*, which establishes a correspondence between surface forms and sequences of abstract morphemes, and the *morphosyntax*, which establishes a correspondence between sequences of abstract morphemes and syntactic/semantic representations of utterances.

In this way, an abstract morpheme is a Saussurean sign, because it contains both semantic and phonological information. Furthermore, it is a minimal sign, because it is the smallest unit with both kinds of information.

While this definition may be 'weaker' than some, it is a substantive claim to say that we can analyse language in terms of abstract morphemes, and this view makes two assumptions explicit. Firstly, language is discrete,[3] in the sense that we can represent an utterance in terms of a finite number of elements from a discrete set. Secondly, morphophonology and morphosyntax are largely independent.

Where we differ from Bender and Good is to assume that the morphophonology acts not on an individual 'word', but on the whole utterance. This allows us to deal with mismatches between phonological and syntactic structure, for example Kwak'wala [kwk] definiteness and case markers, which are phonological suffixes but syntactic prefixes (Boas et al., 1947).

Assuming this overall architecture, the questions we need to ask are: can we systematically map between surface forms and abstract morpheme sequences? Can we systematically assign suitable structures to individual morphemes? And can we systematically build the semantics of the whole from the semantics of the parts? In the following sections, we discuss the challenges these questions raise, although the focus of this paper is on the second and third questions.

2.2 Challenges for Morphemes

Many objections have been raised against analysing language in terms of morphemes (Anderson, 1992; Matthews, 1991; Bochner, 1993), and they can be broadly split between considerations of phonological, semantic, and syntactic phenomena. The focus of this paper is on the latter, but we briefly discuss the first two issues now.

Various phonological phenomena resist segmentation, including metathesis, subtraction, discontinuous elements, infixation, reduplication, suprasegmental features, and apophony. However, a correspondence between surface forms and ab-

[3] An acoustic signal varies continuously in both time and amplitude, but it is nonetheless perceived categorically (Goldstone & Hendrickson, 2010)

stract morphemes does not need to explicitly involve segmentation; the correspondence is with the whole sequence of abstract morphemes, which may not be individually tied to parts of the input. This kind of analysis can be represented using a finite state transducer (FST), a simple and efficient formalism described in detail by Beesley & Karttunen (2003). Finite state techniques can express many phonological/morphological theories (such as autosegmental phonology (Kay, 1987), context-sensitive rewrite rules (Kaplan & Kay, 1994), and Paradigm Function Morphology (Karttunen, 2003), among others) and have been used to describe a variety of 'morphologically rich' languages (such as Finnish (Koskenniemi, 1983) and Turkish (Oflazer, 1994), among others). We believe that the above phenomena can be described using abstract morphemes and finite state techniques, although details are beyond the scope of this paper. What is important to note is that where we use PHON in the rest of the paper, we are not referring to the surface form, but to the representation of the abstract morpheme used by an FST.

Semantic idiosyncrasies, such as 'cranberry' morphemes and Latinate prefixes (*re-ceive, per-ceive*), have been proposed as posing difficulties for morphemic approaches. However, such phenomena are not limited to sub-word combinations, and idiosyncratic multi-word expressions are widespread (Sag et al., 2002). If the semantic objections to morphemes are valid, then we must also object to any constituent within a multiword expression. We view this conclusion as absurd, and we believe techniques used to analyse multiwords, such as those discussed by Sag et al., can also be applied to morphemes.

We now turn to syntactic objections, which can be reduced to the following:

1. Extended exponence (multiple overt morphemes expressing a feature)

2. 'Zero' inflection (no overt morphemes expressing a feature)

3. Syncretism (alternative feature values associated with the same morpheme)

4. Stem alternations (alternative morphemes associated with the same features)

Extended exponence can be dealt with using unification. Each exponent of a feature has that specified in its feature structure, and when multiple exponents occur, the features are unified, analogously to agreement.

Syncretism can be modelled using underspecified types. In some cases, this will involve a single type hierarchy for multiple featural dimensions, a technique which has been successfully used to analyse various languages, for example by Flickinger (2000) for person and number in English, and by Crysmann (2005) for number, gender, and case in German. Indeed, Krieger & Nerbonne (1993) argue that 'matrix-based' descriptions of paradigms can always be given a 'form-based' analysis, where each form is underspecified for a set of agreement values.

Although we could try to model zero inflection using morphemes without phonological material (since this is expressible using an FST), this would lead to rampant homophony between such morphemes. Instead, we first note that it only makes sense to postulate a zero element if it can be identified via overt elements competing for the same slot (Sanders, 1988). When an overt morpheme fills a slot, the type of the mother (the whole phrase) and the type of the other daughter (the rest of the phrase) will in general be different. We can therefore replace 'zero morphemes' by unary syntactic rules, with appropriate types for the mother and daughter, and which stipulate the features associated with the 'zero'.

It has been claimed that contextually-determined stem alternations and similar kinds of allomorphy constitute a problem, because multiple morphemes are associated with a set of features, but only one morpheme is used in a given context. However, in such cases, we can associate each stem or morpheme with the contexts in which it appears. The typed feature structure corresponding to the set of contexts may be highly underspecified, but this does not present a challenge to the theory. This is also true for 'morphomic stems' (Aronoff, 1994), where many features may play a role, and where values of these features may depend on one another – we will see such an example in the Slovene data below. In more extreme cases, some elements are called 'empty' morphemes, because they are allegedly associated with no features at all. However, we reject such a view, since such morphemes will only appear in some contexts but not others, and we can therefore associate the morpheme with the relevant features for those contexts. In a sense, because we can represent morphomic stems and empty morphemes with underspecified forms, we can see this as a special case of syncretism.

In short, none of the above objections represent an obstacle to a type-driven morphemic approach. However, it should also be noted that the same cannot be said of all morphemic theories. For example, our arguments do not apply to the influential framework of Distributed Morphology (Halle & Marantz, 1993), because that theory lacks the notions of underspecification and unification. Instead, they are forced to introduce other devices, such as competition between morphemes, which we will argue against in our analysis of Georgian. Of all the mechanisms that have so far been proposed, underspecification and unification seem to us to be the only straightforward way of capturing many-to-many mappings between morphemes and features.

2.3 Modelling Morphological Paradigms

Having described the general approach, we now describe the mechanical details. We focus on inflection in this paper, but we note that our approach could be extended to include derivational morphology. Indeed, Lieber (2004) and Booij (2005) argue that derivation can be handled in an Item-and-Arrangement theory, which fits neatly with our morpheme-driven framework.

Inflectional paradigms can often be represented in terms of a root and a number of affixes, falling into discrete position classes, or slots.[4] To model the affixation, we must decide whether the root or the affixes should act as heads.

[4] As noted by Crysmann & Bonami (2015), morpheme positions can vary. While we do not deal with morphotactics in detail here, we note that variable morpheme orders can in principle be dealt with in the same way as variable constituent orders in syntax.

If the root should act as head, we can introduce an MCOMPS list, with one item for each slot in the paradigm. This list should intuitively be separate from the COMPS and SPR lists, because inflection is separate from argument structure. Affixation would then be represented using a Head-MComp Schema:

$$\left[\begin{array}{ll} t_1 & \\ \text{...HEAD} & \boxed{1} \\ \text{...SUBJ} & \boxed{2} \\ \text{...COMPS} & \boxed{3} \\ \text{...MCOMPS} & \left\langle \right\rangle \\ \text{...ARG-ST} & \boxed{2} \oplus \boxed{3} \end{array}\right] \rightarrow \left[\begin{array}{ll} t_2 & \\ \text{...HEAD} & \boxed{1} \\ \text{...SUBJ} & \boxed{2} \\ \text{...COMPS} & \boxed{3} \\ \text{...MCOMPS} & \left\langle \boxed{4} \right\rangle \\ \text{...ARG-ST} & \boxed{2} \oplus \boxed{3} \end{array}\right], \boxed{4}$$

For an affix to share its features with the whole expression, we can introduce a re-entrancy between the head features of the root and the affix, as shown below. In the case of zero inflection, we can use a unary rule which removes an element from the MCOMPS list and unifies the appropriate head features with the root.

$$\left[\begin{array}{ll} \textit{root} & \\ \text{...HEAD} & \boxed{1} \\ \text{...MCOMPS} & \left\langle \left[\begin{array}{ll} \textit{affix} & \\ \text{...HEAD} & \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \left[\begin{array}{ll} \textit{affix} & \\ \text{...HEAD} & \left[\begin{array}{ll} \text{AGR} & \textit{agr} \end{array}\right] \end{array}\right]$$

If the affix should act as head, we can avoid introducing an MCOMPS list, and instead take the root or stem to be the specifier of the affix:

$$\left[\begin{array}{ll} \textit{affix} & \\ \text{SYNSEM|LOC|CAT} & \left[\begin{array}{ll} \text{HEAD} & \left[\begin{array}{ll} \text{AGR} & \textit{agr} \end{array}\right] \\ \text{SPR} & \left[\textit{stem}\right] \end{array}\right] \end{array}\right]$$

As above, we introduce re-entrancies between the head features of the root, affix, and whole expression, which we can do in the phrasal type:

$$\left[\begin{array}{ll} \textit{affixed-stem} & \\ \text{SYNSEM} & \boxed{3} \left[\begin{array}{ll} \text{LOC|CAT|HEAD} & \boxed{1} \end{array}\right] \\ \text{HEAD-DTR} & \left[\begin{array}{ll} \textit{affix} & \\ \text{SYNSEM|LOC|CAT} & \left[\begin{array}{ll} \text{HEAD} & \boxed{1} \\ \text{SPR} & \boxed{2} \end{array}\right] \end{array}\right] \\ \text{SPR-DTR} & \boxed{2} \left[\begin{array}{ll} \textit{stem} & \\ \text{SYNSEM} & \boxed{3} \end{array}\right] \end{array}\right]$$

For zero inflection, we stipulate the information in a unary rule with the same pair of types as used in the above head-specifier construction:

$$\left[\begin{array}{ll} \textit{affixed-stem} & \\ \text{SYNSEM} & \boxed{3}\left[\begin{array}{ll}\text{LOC|CAT|HEAD} & \left[\begin{array}{ll}\text{AGR} & \textit{agr}\end{array}\right]\end{array}\right] \\ \text{HEAD-DTR} & \left[\begin{array}{ll}\textit{stem} & \\ \text{SYNSEM} & \boxed{3}\end{array}\right]\end{array}\right]$$

In the following sections, we will use the affix-as-head analysis. This creates a natural similarity between auxiliaries and affixes, which is lost in the MCOMPS analysis. However, we want to stress that the general claim of this paper is not affected by the choice of mechanism: in either case, the claimed problems with morphemes can be overcome using a type-driven approach.

3 Slovene Stem Alternations

Here we consider a situation where the choice of a noun's stem is sensitive to number and case features. This situation exhibits all four of the issues mentioned above, and we will show how the use of a type hierarchy can overcome each of them. We will further show how notions developed in WP approaches, such as 'stem space' and 'rule of referral', can not only be re-expressed in our type-driven morphemic approach, but can in fact be expressed more robustly.

Slovene nouns inflect for three numbers (singular, dual, plural) and six cases. An example of the simplest kind of declension is shown in table 1, involving a single stem, and a slot for one case/number suffix. Some suffixes are syncretic for either case or number, such as *-oma* (dative or instrumental) and *-ih* (dual or plural). This can be modelled by organizing number and case in type hierarchies, with an underspecified type for each observed syncretism, as shown in figure 1.

	SINGULAR	DUAL	PLURAL
NOMINATIVE	*mést-o*	*mést-i*	*mést-a*
ACCUSATIVE	*mést-o*	*mést-i*	*mést-a*
GENITIVE	*mést-a*	*mést*	*mést*
DATIVE	*mést-u*	*mést-oma*	*mést-om*
INSTRUMENTAL	*mést-om*	*mést-oma*	*mést-i*
LOCATIVE	*mést-u*	*mést-ih*	*mést-ih*

Table 1: Declension with a single stem

Taking the suffix to be the head, and the noun stem to be its specifier, we get phrasal types and lexical entries as shown in figure 2. Where there is a 'zero' suffix, we introduce a unary rule.

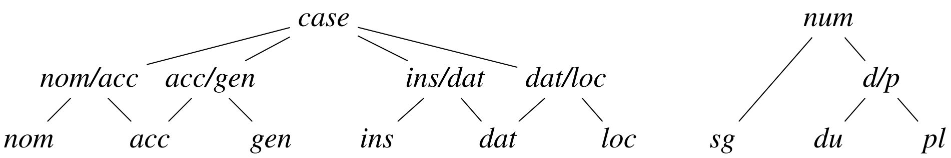


Figure 1: Case and number type hierarchies for Slovene

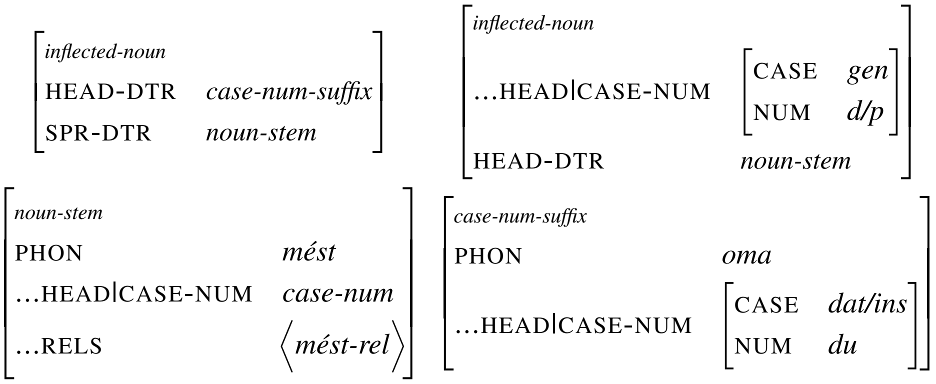


Figure 2: Phrasal types and lexical entries for case-number suffixes

A more complicated declension is shown in table 2, where an additional infixing element is present for all dual and plural forms, appearing between the noun root and the case suffix. This is an example of extended exponence, since each of the suffixes already indicates dual/plural number, and the *-ôv-* infix redundantly specifies it again. We can model this declension using the phrase structure shown in figure 3. The infix takes a noun root as its specifier, to yield a noun stem, which can then combine with a case-number suffix as before.[5] Phrasal types and lexical entries for this declension are shown in figure 4.

	SINGULAR	DUAL	PLURAL
NOMINATIVE	*grád*	*grad-ôv-a*	*grad-ôv-i*
ACCUSATIVE	*grád*	*grad-ôv-a*	*grad-ôv-e*
GENITIVE	*grad-ú*	*grad-ôv*	*grad-ôv*
DATIVE	*grád-u*	*grad-ôv-oma*	*grad-ôv-om*
INSTRUMENTAL	*grád-om*	*grad-ôv-oma*	*grad-ôv-i*
LOCATIVE	*grád-u*	*grad-ôv-ih*	*grad-ôv-ih*

Table 2: Declension with a distinct dual/plural stem

[5] There are differences in endings between the declensions for *grád* and *mést*, which exemplify two of the many declensions in Slovene. To model inflectional classes, each noun should also have a feature indicating its class, and each suffix should impose a constraint on the class of its specifier. If some suffixes appear in multiple classes (as is the case for Slovene), the classes can be organized in a hierarchy, and each suffix selects for an underspecified class.

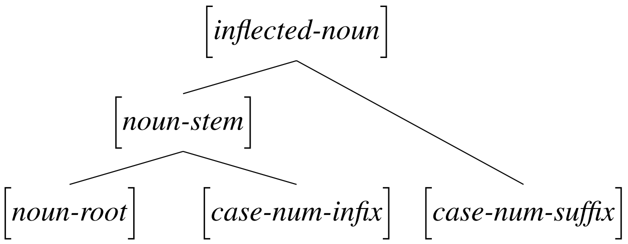


Figure 3: Phrase structure of an inflected noun

$$\begin{bmatrix} \textit{noun-stem} & \\ \text{HEAD-DTR} & \textit{case-num-infix} \\ \text{SPR-DTR} & \textit{noun-root} \end{bmatrix} \quad \begin{bmatrix} \textit{noun-stem} & \\ \text{...HEAD|CASE-NUM} & \begin{bmatrix} \text{CASE} & \textit{case} \\ \text{NUM} & \textit{sg} \end{bmatrix} \\ \text{HEAD-DTR} & \textit{noun-root} \end{bmatrix}$$

$$\begin{bmatrix} \textit{noun-root} & \\ \text{PHON} & \textit{grád} \\ \text{...HEAD|CASE-NUM} & \textit{case-num} \\ \text{...RELS} & \left\langle \textit{grád-rel} \right\rangle \end{bmatrix} \quad \begin{bmatrix} \textit{case-num-infix} & \\ \text{PHON} & \textit{ôv} \\ \text{...HEAD|CASE-NUM} & \begin{bmatrix} \text{CASE} & \textit{case} \\ \text{NUM} & \textit{d/p} \end{bmatrix} \end{bmatrix}$$

Figure 4: Phrasal types and lexical entries for case-number infixes

A number of Slovene nouns change stem, but with a pattern that involves both number and case. For example, *nágelj* (‘carnation’) has the stem *nágelj-n* for all forms other than nominative and accusative singular. We can deal with this in the same way as for *grád*, but unlike the *-ôv-* infix, which could be described using a pure number feature, the *-n-* infix requires a combined case-number feature.

The unique pair of stems *člôvek* and *ljud* (‘man/men’), exhibits an unusual pattern of suppletion, where *člôvek* is used for the singular, *ljud* is used for the plural, and they are split in the dual, as shown in table 3 (Priestly, 1993). Furthermore, the cases where the plural stem *ljud* is used for the dual are precisely those which display syncretism in the suffixes, suggesting a deeper generalization is to be found.

	SINGULAR	DUAL	PLURAL
NOMINATIVE	*člôvek*	*člověk-a*	*ljud-jé*
ACCUSATIVE	*člověk-a*	*člověk-a*	*ljud-í*
GENITIVE	*člověk-a*	*ljud-í*	*ljud-í*
DATIVE	*člověk-u*	*člověk-oma*	*ljud-ém*
INSTRUMENTAL	*člověk-om*	*člověk-oma*	*ljud-mí*
LOCATIVE	*člověk-u*	*ljud-éh*	*ljud-éh*

Table 3: Declension with suppletive stems

Corbett (2015) analyses this at the level of a paradigm, within the framework of Network Morphology, introducing ‘generalized referral’ rules that stipulate that the forms for the genitive and locative dual should be identical to the plural forms. Under such an analysis, however, we cannot immediately infer that using the wrong stem for the genitive dual is ungrammatical, as we need to compare it to other parts of the paradigm.

Instead, we give an analysis where the ungrammatical forms are directly ruled out by unification failure. By combining number and case into a single hierarchy, it is possible to introduce types so that each stem can only appear in the appropriate combinations of number and case. The fact that the two stems are part of the same paradigm is captured by the semantic predicate being the same for both.

$$\begin{bmatrix} \textit{\scriptsize gen.sg} & \\ \text{CASE} & \textit{gen} \\ \text{NUM} & \textit{sg} \end{bmatrix} \qquad \begin{bmatrix} \textit{\scriptsize d/p-cn} & \\ \text{CASE} & \textit{case} \\ \text{NUM} & \textit{d/p} \end{bmatrix} \qquad \begin{bmatrix} \textit{\scriptsize ljud-cn} & \\ \text{CASE} & \textit{case} \\ \text{NUM} & \textit{d/p} \end{bmatrix}$$

Figure 5: Combined number and case types

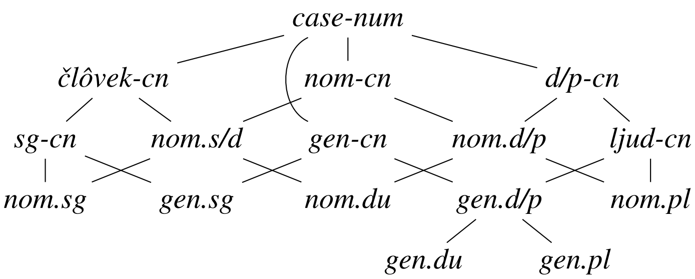


Figure 6: Case-number type hierarchy for Slovene

$$\begin{bmatrix} \textit{\scriptsize noun-stem} & \\ \text{PHON} & \textit{člôvek} \\ \text{...HEAD|CASE-NUM} & \textit{člôvek-cn} \\ \text{...RELS} & \left\langle \textit{člôvek-ljud-rel} \right\rangle \end{bmatrix}$$

$$\begin{bmatrix} \textit{\scriptsize noun-stem} & \\ \text{PHON} & \textit{ljud} \\ \text{...HEAD|CASE-NUM} & \textit{ljud-cn} \\ \text{...RELS} & \left\langle \textit{člôvek-ljud-rel} \right\rangle \end{bmatrix}$$

Figure 7: Lexical entries for *člôvek* and *ljud*

Another WP approach to modelling this would be to use the notion of a 'stem space' (Pirrelli & Battista, 2000; Bonami & Boyé, 2003). Under such an analysis, we divide the the paradigm into 'spaces' of cells, where each space uses the same stem. The underspecified types which we propose directly correspond to such spaces. However, by organizing these types in a hierarchy, we can efficiently refer to types at varying levels of granularity. In the present case, the types for *člôvek* and *ljud* are not relevant for *grád*, and vice versa; furthermore, none of these types are relevant for *mést*. For each of these nouns, we do not want to redundantly specify the same stem for multiple spaces. For a more complex paradigm, such as Italian verbal conjugation, as discussed by Montermini & Bonami (2013), this is a serious concern, as the number of spaces increases dramatically with the irregularity of the lexemes considered. By using a type hierarchy, we can simultaneously analyse a paradigm with varying numbers of stem spaces, thereby reducing redundancy in the lexicon: each lexical entry uses types at the relevant level of granularity.

In figure 6, we give a partial type hierarchy, with only nominative and genitive cases, which are sufficient to demonstrate the split in the dual for *člôvek* and *ljud*. The analysis follows similarly for the other cases.

So that we can still refer to case and number individually (which is important to get the correct semantics), each of these types has features for case and number, with examples given in figure 5. To distinguish types in the combined hierarchy from those in the separate hierarchies, we write *-cn* in the type name. For types with 'irregular' (non-rectangular) spaces in the paradigm, such as *ljud-cn*, the values for these features will the be the most specific ones that cover all relevant cells – the irregularity of the stem space is handled by the type's position in the hierarchy.

The generalization that the use of *ljud* in the dual matches the suffix syncretism is captured by *gen.d/p* being the only type immediately dominating *gen.du* and *gen.pl*. In fact, it would be impossible to maintain this property if we introduced a single underspecified type for singular and dual. Not only does this allow us to reproduce a 'rule of referral', but this is done without a need for directionality in the rule, which is known to be problematic to determine. Furthermore, the data is captured more directly, in the sense that each form can be described in terms of its parts, without referring to other cells in the paradigm.

In summary, our analysis of Slovene nominal declensions illustrates how all four of the problems discussed in section 2.2 can be overcome. Furthermore, we have seen how the WP notions of 'stem space' and 'rule of referral' can be robustly re-interpreted in a morphemic approach.

4 Georgian Verb Agreement

Georgian verbs present a situation involving multiple affixes which jointly determine the value of multiple features. The full verbal paradigm is notoriously complex, and Hewitt (1995, p.117) lists 11 different slots. We consider the two agreement affixes (one prefix and one suffix), which jointly agree with both subject and object, as shown in examples (1)-(3). The order of the nouns does not affect the argument structure, and we will not discuss case marking here.

The full agreement paradigm in the present tense is given in table 4, adapted from Harris (1981). Note that reflexives are marked separately in Georgian, so it is not possible for the subject and object to both be first person, or both be second person. For ease of exposition, a few distracting details are suppressed for now, and will be discussed at the end.

(1) მე ვაქებ ექიმს
me v-akeb ekim-s
I praise.1SG.3SG doctor-DAT
'I praise the doctor'

(2) მე გაქებ შენ
me g-akeb ʃen
I praise.1SG.2SG you
'I praise you'

(3) მე მაქებს ექიმი
me m-akeb-s ekim-i
I praise.3SG.1SG doctor-NOM
'the doctor praises me'

	Object					
Subject	1SG	1PL	2SG	2PL	3SG	3PL
1SG	—	—	*g*—∅	*g*—*t*	*v*—∅	*v*—∅
1PL	—	—	*g*—*t*	*g*—*t*	*v*—*t*	*v*—*t*
2SG	*m*—∅	*gv*—∅	—	—	∅—∅	∅—∅
2PL	*m*—*t*	*gv*—*t*	—	—	∅—*t*	∅—*t*
3SG	*m*—*s*	*gv*—*s*	*g*—*s*	*g*—*t*	∅—*s*	∅—*s*
3PL	*m*—*en*	*gv*—*en*	*g*—*en*	*g*—*en*	∅—*en*	∅—*en*

Table 4: Agreement in Georgian present tense verbs

This data has been traditionally analysed by noting certain weak correlations between affixes and agreement features, such as *v*- denoting a first person subject, and *g*- a second person object. Morphemes based on these weak correlations would overgenerate, leading many to invoke some other mechanism to prevent overgeneration. Harris (1981) uses deletion rules, where all morphemes are generated, but, for instance, *v*- is deleted in the presence of *g*-. Several other authors, working in a variety of frameworks, impose some ordering on applying lexical rules or inserting lexical items, so that one rule or item blocks the others (Anderson, 1986; Halle & Marantz, 1993; Carmack, 1997; Stump, 2001). The deletion analysis is implausible phonologically (since Georgian allows long consonant clusters), requires prediction of possible deleted elements when processing language, and makes it appear a coincidence that a Georgian verb can have at most one agreement suffix and one agreement prefix, since deletion rules would not guarantee this in general. Indeed, Harris neglects to state that the *-en* and *-t* suffixes cannot co-occur (the *-t* should 'delete'), although others do note this. The blocking analyses, however, hugely increase the complexity of the grammar, since we have to consider many alternative derivations in order to interpret a given form, or even determine if it is grammatical. Furthermore, as Blevins (2015) notes, competition between these rules cannot be regulated by a constraint which prioritizes more specific rules (such as 'Pāṇini's Principle' (Stump, 2001)), since we cannot say that a subject feature or an object feature is more specific than the other.

Here we present an alternative analysis, with a sign for each overt affix, and a unary rule for each 'zero', where each structure has both subject and object features. For example, *v-* indicates not only a first person subject, but also a third person object. For almost all the affixes, the paradigm cells form rectangular blocks, meaning that we can specify the subject and object features independently.

The exception is the suffix *-t*, which can appear with any subject except second singular and third plural, and with any object at all – but specifying these independently would lead to overgeneration. Instead, we can analyse this paradigm as having two homophonous *-t* suffixes, each with a rectangular shape. One specifies a first or second person plural subject, and any object. The other specifies a second person plural object, and a first or third person singular subject.[6] Indeed, traditional grammars often refer to these two distinct uses of *-t* separately.

PHON	Subj	Obj
v	*1*	*3*
g	*1/3*	*2*
m	*2/3*	*1sg*
gv	*2/3*	*1pl*
∅	*2/3*	*3*

PHON	Subj	Obj
t	*1/2pl*	*per-num*
t	*1/3sg*	*2pl*
s	*3sg*	*¬2pl*
en	*3pl*	*per-num*
∅	*1/2sg*	*¬2pl*

Table 5: Abbreviated lexical entries (left, prefixes; right, suffixes)

A summary of the agreement features of the full set of affixes and unary rules is given in table 5. The corresponding feature structures are shown in figure 10 for the unary rules, and in figure 11 for the overt affixes (just two are shown, for brevity). The corresponding person-number type hierarchy is given in figure 12. The phrasal types and the resulting phrase structure are shown in figures 8 and 9.

[6] We could also specify the subject as being anything but third plural, which would yield the same paradigm. However, doing so introduces a spurious ambiguity for *g–t*, in the case of a first plural subject and second plural object, since either homophone of *-t* could be used. For this reason, we prefer this more restrictive subject feature.

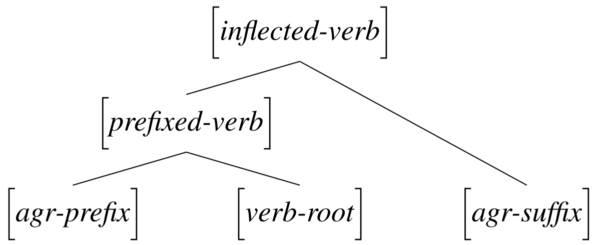


Figure 8: Phrase structure of an inflected verb

$$\begin{bmatrix} \textit{inflected-verb} & \\ \text{HEAD-DTR} & \textit{agr-suffix} \\ \text{SPR-DTR} & \textit{prefixed-verb} \end{bmatrix} \qquad \begin{bmatrix} \textit{prefixed-verb} & \\ \text{HEAD-DTR} & \textit{agr-prefix} \\ \text{SPR-DTR} & \textit{verb-root} \end{bmatrix}$$

Figure 9: Phrasal types

$$\left[\begin{array}{ll} \textit{inflected-verb} & \\ \text{HEAD-DTR} & \left[\begin{array}{ll} \textit{prefixed-verb} & \\ \text{...SUBJ} & \left[\text{...PER-NUM} \quad \textit{1/2s}\right] \\ \text{...COMPS} & \left\langle\left[\text{...PER-NUM} \quad \neg\textit{2p}\right]\right\rangle \end{array}\right] \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{prefixed-verb} & \\ \text{HEAD-DTR} & \left[\begin{array}{ll} \textit{verb-root} & \\ \text{...SUBJ} & \left[\text{...PER-NUM} \quad \textit{2/3}\right] \\ \text{...COMPS} & \left\langle\left[\text{...PER-NUM} \quad \textit{3}\right]\right\rangle \end{array}\right] \end{array}\right]$$

Figure 10: Unary rules

$$\left[\begin{array}{ll} \textit{agr-prefix} & \\ \text{PHON} & \textit{gv} \\ \text{...SPR} & \left[\begin{array}{ll} \text{...SUBJ} & \left[\text{...PER-NUM} \quad \textit{2/3}\right] \\ \text{...COMPS} & \left\langle\left[\text{...PER-NUM} \quad \textit{1pl}\right]\right\rangle \end{array}\right] \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{agr-suffix} & \\ \text{PHON} & \textit{s} \\ \text{...SPR} & \left[\begin{array}{ll} \text{...SUBJ} & \left[\text{...PER-NUM} \quad \textit{3sg}\right] \\ \text{...COMPS} & \left\langle\left[\text{...PER-NUM} \quad \neg\textit{2pl}\right]\right\rangle \end{array}\right] \end{array}\right]$$

Figure 11: Examples of expanded lexical entries

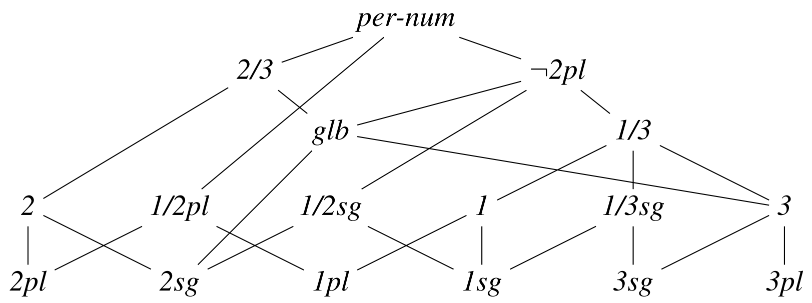


Figure 12: Person-number type hierarchy for Georgian. We introduce the type *glb* (greatest lower bound) so that the hierarchy forms a semilattice, but this type is not used in any well-formed structure.

After unification, this grammar generates all and only the forms in table 4, including leaving the gaps in the table for reflexives, without any spurious ambiguity, and without any additional ordering constraints or competition. This refutes previous claims in the literature that Georgian verb agreement cannot be modelled in a morphemic approach. Gurevich (2006) explicitly argues against the use of morphemes, but we have dealt with each of their objections (cumulative expression, zero morphs, empty morphs, and extended exponence), as explained in section 2.3. Similarly, Blevins (2015) claims that "a dynamic system of contrasts cannot be modelled by a set of static independent associations", but we have shown that this is indeed possible if the associations are with typed feature structures.

Although Blevins sets up a dichotomy between 'associative' and 'discriminative' approaches, the system of morphemes we propose can be viewed in both ways: each morpheme is associated with a feature structure, but the relevant feature values are organized in a type hierarchy so that they discriminate the appropriate meanings. For example, the *v*- prefix can be seen as being associated with a third person object, or conversely as discriminating against a second person object, since it is not unifiable with it. By organizing information using a rich type hierarchy, we can set up associations between morphemes and feature structures in a way that is perfectly compatible with a discriminative view. Indeed, the more underspecified a type is, the more it appears discriminative, rather than associative.

Some complexities of the system are evident in our analysis, such as the need for a *¬2pl* type, but this is in fact motivated twice. Moreover, a similar type is used by Flickinger (2000) to account for present tense verb agreement in English, since zero inflection indicates the subject can be anything except third person singular.

We avoid the need for blocking or competition by the use of more specific values for the person-number feature, and unlike the previously mentioned analyses, the grammaticality and interpretation of a form can be decided without reference to the rest of the paradigm.

In summary, our analysis of Georgian verb agreement illustrates how a type-driven morphemic approach can deal with many-to-many mappings between morphemes and features, contrary to previous claims.

4.1 Further Details

The agreement affixes are inverted between subject and object for a small class of verbs, and for one series of tense-aspect-mood combinations (called 'screeves' in traditional Georgian grammars). These require no change to the above analysis, and can be captured by switching how ARG-ST is linked to COMPS and SUBJ.

The third person subject suffixes are not always *-s* and *-en*, but depend on the tense-aspect-mood of the verb. To model this, we can stipulate several lexical entries as in figure 11, but each with a feature for tense-aspect-mood.

Verbs of motion require an additional agreement marker which effectively fills a separate slot – for example, *mi-v-di-var-t* 'we go', where *mi* is a directional prefix, and *var* indicates a first person subject. To model this, we can give the root *di* a distinct type from other verbs, which the affixes like *var* take as a specifier.

Agreement of intransitive verbs looks like the final column in table 4. To use the same lexical entries for agreement in both intransitives and transitives, we can define a unary rule for affixes whose mother has an empty list in ...SPR...COMPS, and whose daughter's ...SPR...COMPS...PER-NUM must be unifiable with third person. Although 'constructive', this analysis has much in common with the 'abstractive' approach to polyfunctionality described by Ackerman & Bonami (2015). In the general case, we can define a single 'abstract' lexical entry with all necessary information, and a set of unary rules modifying the morpheme for each function.

In the prestige dialect, agreement in ditransitive verbs is with the indirect object, and the direct object must be third person (first and second person objects are marked like reflexives in so-called 'object camouflage'). We can use the same lexical entries for affixes if the indirect object is the first element in the COMPS list. Some speakers have additional markers for third person indirect objects, although Harris notes that their use "is not consistent". The additional indirect object markers can be modelled by imposing an additional constraint on the verb, requiring that it is ditransitive. We neglect other dialectal variations for space reasons.

Third person plural subject agreement (with *-en*) is only triggered by animate nouns. To model this, we can extend the type hierarchy with additional subtypes of *3*, indicating both animacy and number. This does not affect the rest of the hierarchy (only the bottom right corner of figure 12), demonstrating the modular nature of our analysis.

5 Conclusion

In the light of work suggesting 'words' are not well-defined cross-linguistically, we have argued in favour of reformulating HPSG as a unified morphosyntactic theory. We have proposed the Morphemic Principle as a formalization of this approach, and shown how the use of underspecification and unification can avoid various objections to morphemic approaches. We have illustrated our framework by analysing Slovene stem alternations and Georgian verb agreement, giving simpler analyses than competing approaches, but while maintaining the same generalizations.

Acknowledgements

We would like to thank the three anonymous reviewers for their comments, Emily Bender, Dan Flickinger, and Farrell Ackerman for helpful discussion, and Khatuna Gelashvili for providing help with Georgian.

References

Ackerman, Farrell & Olivier Bonami. 2015. Systemic polyfunctionality and morphology-syntax interdependencies. In Andrew Hippisley & Nikolas Gisborne (eds.), *Defaults in morphological theory*, Oxford University Press.

Anderson, Stephen R. 1986. Disjunctive ordering in inflectional morphology. *Natural Language and Linguistic Theory* 4. 1–32.

Anderson, Stephen R. 1992. *A-morphous morphology*. Cambridge University Press.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes* (Linguistic Inquiry 22). MIT Press.

Asudeh, Ash, Mary Dalrymple & Ida Toivonen. 2013. Constructions with lexical integrity. *Journal of Language Modelling* 1(1). 1–54.

Beesley, Kenneth R & Lauri Karttunen. 2003. *Finite state morphology* (Studies in Computational Linguistics). CSLI Publications.

Bender, Emily & Jeff Good. 2005. Implementation for discovery: A bipartite lexicon to support morphological and syntactic analysis. In *Proceedings of the Annual Meeting of the Chicago Linguistic Society*, vol. 41 2, 1–16. Chicago Linguistic Society.

Bender, Emily M, Dan Flickinger & Stephan Oepen. 2002. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the 2002 workshop on grammar engineering and evaluation*, vol. 15, 1–7. Association for Computational Linguistics.

Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42(03). 531–573.

Blevins, James P. 2015. The minimal sign. In *The Cambridge handbook of morphology*, Cambridge University Press.

Bloomfield, Leonard. 1933. *Language*. Henry Holt.

Boas, Franz, Helene Boas Yampolsky & Zellig S Harris. 1947. Kwakiutl grammar with a glossary of the suffixes. *Transactions of the American Philosophical Society* 37(3). 203–377.

Bochner, Harry. 1993. *Simplicity in generative morphology* (Publications in Language Sciences 37). Walter de Gruyter.

Bonami, Olivier & Gilles Boyé. 2003. Supplétion et classes flexionnelles. *Langages* 102–126.

Booij, Geert. 2005. Compounding and derivation. *Morphology and its demarcations* 109–132.

Bresnan, Joan & Sam A Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory* 13(2). 181–254.

Carmack, Stanford. 1997. Blocking in Georgian verb morphology. *Language* 73(2). 314–338.

Corbett, Greville G. 2015. Morphosyntactic complexity: A typology of lexical splits. *Language* .

Crysmann, Berthold. 2003. On the efficient implementation of German verb placement in HPSG. In *Proceedings of RANLP 2003*, 112–116.

Crysmann, Berthold. 2005. Syncretism in German: a unified approach to underspecification, indeterminacy, and likeness of case. In *Proceedings of the 12th international conference on Head-Driven Phrase Structure Grammar*, 91–107.

Crysmann, Berthold & Olivier Bonami. 2015. Variable morphotactics in information-based morphology. *Journal of Linguistics* .

Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6(01). 15–28.

Flickinger, Dan, Ann Copestake & Ivan A Sag. 2000. HPSG analysis of English. In *Verbmobil: Foundations of speech-to-speech translation*, 254–263. Springer.

Goldstone, Robert L & Andrew T Hendrickson. 2010. Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(1). 69–78.

Gurevich, Olga I. 2006. *Constructional morphology: The Georgian version*: University of California, Berkeley dissertation.

Halle, Morris & Alec Marantz. 1993. Distributed Morphology and the pieces of inflection. In Kenneth Hale & Samuel Jay Keyser (eds.), *The view from Building 20*, 111–176. MIT Press.

Harris, Alice C. 1981. *Georgian syntax: A study in Relational Grammar* (Cambridge Studies in Linguistics). Cambridge University Press.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.

Hewitt, Brian George. 1995. *Georgian: A structural reference grammar*, vol. 2. John Benjamins Publishing.

Kaplan, Ronald M & Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20(3). 331–378.

Karttunen, Lauri. 2003. Computing with realizational morphology. In Alexander Gelbukh (ed.), *Computational linguistics and intelligent text processing* (Lecture Notes in Computer Science 2588), 203–214. Springer.

Kay, Martin. 1987. Nonconcatenative finite-state morphology. In *Proceedings of the 3rd conference of the European chapter of the Association for Computational Linguistics*, 2–10. Association for Computational Linguistics.

Koskenniemi, Kimmo. 1983. *Two-level morphology: A general computational model for word-form recognition and production*: University of Helsinki dissertation.

Krieger, Hans-Ulrich & John Nerbonne. 1993. Feature-based inheritance networks for computational lexicons. In Ted Briscoe, Valeria de Paiva & Ann Copestake (eds.), *Inheritance, defaults, and the lexicon* (Studies in Natural Language Processing), Cambridge University Press.

Lieber, Rochelle. 2004. *Morphology and lexical semantics* (Cambridge Studies in Linguistics). Cambridge University Press.

Lyons, John. 1968. *An introduction to theoretical linguistics*. Cambridge University Press.

Mair, Victor H. 1990. Implications of the Soviet Dungan script for Chinese language reform. *Sino-Platonic Papers* (18).

Matthews, Peter H. 1991. *Morphology* (Cambridge Textbooks in Linguistics). Cambridge University Press 2nd edn.

Montermini, Fabio & Olivier Bonami. 2013. Stem spaces and predictability in verbal inflection. *Lingue e linguaggio* 12(2). 171–190.

Müller, Stefan. 2015. *German sentence structure: An analysis with special consideration of so-called multiple fronting* (Empirically Oriented Theoretical Morphology and Syntax). Language Science Press.

Oflazer, Kemal. 1994. Two-level description of Turkish morphology. *Literary and linguistic computing* 9(2). 137–148.

Packard, Jerome L. 2000. *The morphology of Chinese*. Cambridge University Press.

Pirrelli, Vito & Marco Battista. 2000. The paradigmatic dimension of stem allomorphy in Italian verb inflection. *Italian Journal of Linguistics* 12. 307–379.

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

Priestly, T M S. 1993. Slovene. In Bernard Comrie & Greville G Corbett (eds.), *The Slavonic languages*, 388–451. Routledge.

Roark, Brian & Richard William Sproat. 2007. *Computational approaches to morphology and syntax*. Oxford University Press.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the third international conference on computational linguistics and intelligent text processing* (Lecture Notes in Computer Science 2276), 1–15. Springer.

Sag, Ivan A, Thomas Wasow & Emily M Bender. 2003. *Syntactic theory: A formal introduction*. CSLI Publications 2nd edn.

Sanders, Gerald. 1988. Zero derivation and the overt analogue criterion. *Theoretical Morphology* 155–175.

Spencer, Andrew. 2006. Morphological universals. In Ricardo Mairal & Juana Gil (eds.), *Lingustic universals*, 101–129. Cambridge University Press.

Stump, Gregory T. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge University Press.

Sun, Chao-Fen. 2006. *Chinese: A linguistic introduction*. Cambridge University Press.

Tang, Sze-Wing. 2010. *Formal Chinese syntax*. Shanghai Educational Publishing House.

Building Zhong, a Chinese HPSG Shared-Grammar

Zhenzhen Fan
National University of Singapore

Sanghoun Song
Incheon National University

Francis Bond
Nanyang Technological University

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 96–109

http://csli-publications.stanford.edu/HPSG/2015

Abstract

This paper describes some of our attempts in extending Zhong, a Chinese HPSG shared-grammar. New analyses for two Chinese specific phenomena, reduplication and the SUO-DE structure, are introduced. The analysis of reduplication uses lexical rules to capture both the syntactic and semantic properties (amplification in adjectives and diminishing in verbs). Words showing non-productive reduplication are entered in the lexicon, and the semantic relations will be captured in an external resource (the Chinese Open Wordnet). The SUO-DE structure constrains the meanings of relative clauses to a gapped-object interpretation.

1 Introduction

We are developing a Chinese HPSG shared-grammar named Zhong (Fan et al., 2015), that covers multiple varieties of Chinese. It is based on the existing work on Mandarin Chinese from the HPSG community. Our objective is to build a broad-coverage computational resource grammar that can be used for applications such as machine translation and computer aided language learning. We take a corpus-driven approach to improving its coverage through grammar rule enhancement and lexicon expansion.

Head-Driven Phrase Structure Grammar (HPSG: Pollard & Sag, 1994) is a lexicalized generative grammar theory developed by Carl Pollard and Ivan Sag at Stanford University. An HPSG-based grammar includes constraint-based grammar rules and a lexicon containing syntactic and semantic information about words, which makes it very useful as a grammar framework in natural language processing for deep linguistic analysis of human language aiming at content level understanding.

Computational linguists from different research centers worldwide have been collaborating to develop broad coverage HPSG grammars of different languages in a consortium called Deep Linguistic Processing with HPSG (DELPH-IN, http://www.delph-in.net). Broad coverage HPSGs for English (LinGO English Resource Grammar, ERG: Flickinger, 2000), German (GG: Müller & Kasper, 2000; Crysmann, 2005), Japanese (Jacy: Siegel & Bender, 2002), Korean (KRG: Kim et al., 2011), Spanish (SRG: Marimon, 2012), Norwegian (NorSource: Hellan, 2005), and several other languages have been developed and used in various applications.

In this paper we focus especially on two Chinese phenomena: reduplicated adjectives and verbs, and SUO-DE structure, and show how we implement them in our grammar.

2 Previous Works on Chinese HPSG

Since 1990s, linguistic analysis of specific Chinese phenomena in HPSG framework started to appear (Xue et al., 1994; Gao, 1994; Xue & McFetridge, 1995,?; Ng, 1997). Subsequently, two PhD theses (Gao, 2000; Li, 2001) documented the efforts towards a more comprehensive analysis of Chinese, covering major phenomena such as topic sentences, valence alternations (including BA, ZAI, and other constructions), as well as separable verbs and Chinese derivation and affixes.

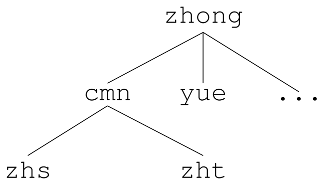
More recent works accompany linguistic analysis with computational implementation, leading to several independently developed HPSG grammars on Mandarin Chinese: MCG (Zhang et al., 2011), ManGO (Yang, 2007), and ChinGram (Müller & Lipenkova, 2013), all adopting Minimal Recursion Semantics (MRS) (Copestake et al., 2005) as the semantic representation format. These grammars focus on a variety of linguistic phenomena in Chinese, but typically only cover the words appearing in their testsuites.

3 Zhong

There are many varieties of Chinese, historically related but now separate languages. Zhong aims to model the common parts and the linguistic diversity across these varieties in a single hierarchy, inspired by the existing works on grammar sharing, such as the LinGO Grammar Matrix system (Bender et al., 2010), CoreGram (Müller, 2013), CLIMB (Fokkens et al., 2012), SLaviCore (Avgustinova & Zhang, 2009) and SlaviCLIMB (Fokkens & Avgustinova, 2013). The different Chinese grammars in Zhong share some elements, such as basic word order, and have other elements distinct, such as lexemes and specific grammar rules (e.g., classifier constructions).

Taking the original implementation of ManGO, we restructured it as follows:

(1)



All grammars build upon the common constraints and inherit from `zhong-lextypes.tdl`, `zhong.tdl`, and `zhong-letypes.tdl`. The differences between Mandarin and Cantonese, such as NP structures, are reflected in `cmn.tdl` and `yue.tdl`, respectively. The Mandarin Chinese grammars are further divided into `zhs` and `zht` depending on whether simplified characters or traditional characters are used. Further distinction between the two are modeled in `zhs.tdl` and `zht.tdl`, respectively.

The official webpage of Zhong, with demo and test results, is http://wiki.delph-in.net/moin/ZhongTop. And the entire data set can be freely downloaded from https://github.com/delph-in/zhong.

4 Chinese-specific Phenomena

As part of the efforts to enhance the grammar's coverage, we have analysed and implemented several Chinese-specific phenomena such as VV resultative compounds, A-NOT-A questions (Wang et al., 2015), NP structure (Sio & Song, 2015), sentence end particles, interjections and fragments. Here we present how we handled another two new phenomena, reduplicated adjectives and verbs, and the SUO-DE structure.

4.1 Reduplicated Adjectives and Verbs

According to Li & Thompson (1989), reduplication is a morphological process of repeating a morpheme to form a new word, which mainly applies to verbs and adjectives in Chinese. When a monosyllabic adjective or verb is reduplicated, the character is repeated (A → AA), as shown in (2) and (3).

(2) 红红
hónghóng
red-red

"very red"

(3) 看看
kànkàn
look-look

"take a look"

When reduplication is applied to disyllabic words, the two characters are repeated differently for adjectives (AB → AABB) and verbs (AB → ABAB), as illustrated in (4) and (5).

(4) 干干净净
gāngānjìngjìng
AABB-clean

"very clean"

(5) 休息休息
xiūxixiūxi
rest-rest

"have a rest"

Syntactically, the reduplicated adjectives can not be modified by degree adverbs (e.g. 很*hen* "very", 非常 *feichang* "extremely", 特别 *tebie* "specially", 极 *ji* "extremely", 十分 *shifen* "very much", 更 *geng* "more", 最 *zui* "most", 较 *jiao* "more", 比较 *bijiao* "more", etc.), as illustrated in (6).

(6) *很 干干净净
hěn gāngānjìngjìng
very AABB-clean

"very clean"

Reduplicated verbs, on the other hand, do not accept aspect markers like 了 *le*, 着 *zhe*, and 过 *guo*, as shown in (7).

(7) *看看 着
kànkàn zhe
look-look ASP

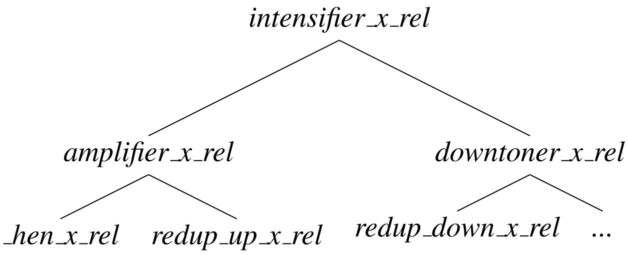
"take a look"

The meaning of the reduplicated adjectives (AA or AABB) is more vivid or intensified than its original form (A or AB) (Li & Thompson, 1989). For verbs, reduplication adds a tentative aspect (Chen et al., 1992), or signals a delimitative aspect (doing something "a little bit") (Li & Thompson, 1989).

Based on our position that sentences with similar meaning should have similar semantic representations, we model the semantic representation of reduplicated verbs or adjectives as the predicate of the original word (A or AB) and a predicate that acts as an intensifier. Depending on the semantic function of the intensifier, it can be either an **amplifier** (making the meaning more intensified) or a **downtoner** (scaling it down), following the analysis of Quirk et al. (1985, p589 onwards).

Two predicates are therefore defined, *amplifier_x_rel* and *downtoner_x_rel*, both inheriting from a common parent *intensifier_x_rel*. *redup_up_x_rel* (representing amplification using reduplication) and *redup_down_x_rel* (representing scaling-down using reduplication) inherit from *amplifier_x_rel* and *downtoner_x_rel* respectively, as illustrated in (8). Predicate for the most common intensifier, the degree adverb 很 (*hen*,"very"), is also added into this structure, but more detailed differentiation of degree scales is left to the Chinese Open Wordnet (Wang & Bond, 2013).

(8)



We use lexical rules to produce the reduplicated forms from the original form. The super type of the rules, *redup-type*, introduces the predicate *intensifier_x_rel*, as shown in (9).

(9)

$$\begin{bmatrix} \textit{redup-type} & \\ \text{CAT.HEAD} & \boxed{1} \\ \text{VAL} & \boxed{2} \\ \text{CONT} & \boxed{3}\ \text{HOOK} \begin{bmatrix}\text{LTOP} & \boxed{4}\\ \text{INDEX} & \boxed{5}\end{bmatrix} \\ \text{C-CONT} & \left\langle \begin{bmatrix} \textit{event-rel} & \\ \text{PRED} & \textit{intensifier\_x\_rel} \\ \text{LBL} & \boxed{4} \\ \text{ARG1} & \boxed{5}\end{bmatrix}\right\rangle \end{bmatrix} \rightarrow \begin{bmatrix}\text{CAT.HEAD} & \boxed{1}\\ \text{VAL} & \boxed{2} \\ \text{CONT} & \boxed{3}\end{bmatrix}$$

Two lexical rules, *redup-a-lr* and *redup-v-lr*, inherit from *redup-type*. *redup-a-lr* (10), which is for adjective reduplication (AA and AABB), requires an adjective, and defines that the predicate introduced is the amplifier *redup_up_x_rel*. It also adds the syntactic constraint that the specifier of the word is empty, preventing it from accepting degree adverbs. The rule for the reduplication of verbs (AA and ABAB), *redup-v-lr* (11), requires a verb, defines the predicate *redup_down_x_rel*, and states that the verb doesn't accept aspect markers.

(10)
$$\begin{bmatrix} \textit{redup-a-lr} \subset \textit{redup-type} & \\ \text{CAT.HEAD} & \text{+a (adjective)} \\ \text{VAL} & \begin{bmatrix}\text{SPR}\,\langle\rangle\end{bmatrix} \\ \text{C-CONT} & \left\langle \begin{bmatrix}\text{PRED} & \textit{redup\_up\_x\_rel}\end{bmatrix} \right\rangle \end{bmatrix}$$

ORTHOGRAPHY: A → AA (irregular AB → AABB)

(11)
$$\begin{bmatrix} \textit{redup-v-lr} \subset \textit{redup-type} & \\ \text{CAT.HEAD} & \text{+v (verb)} \\ \text{CONT.HOOK} & \begin{bmatrix}\text{ASPECT}\ \textit{non-aspect}\end{bmatrix} \\ \text{C-CONT} & \left\langle \begin{bmatrix}\text{PRED} & \textit{redup\_down\_x\_rel}\end{bmatrix} \right\rangle \end{bmatrix}$$

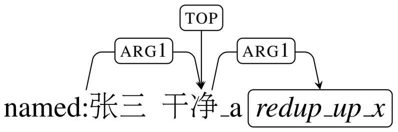
ORTHOGRAPHY: A → AA; A → A一A; (irregular AB → ABAB)

With the above definitions, for a sentence like (12), the dependency graph representing its MRS structure is provided in (13), which basically neans "Something called "张三" is *redup_up* clean".

(12) 张三 干干净净
zhāngsān gāngānjìngjìng
Zhangsan AABB-clean

"Zhangsan is very clean"

(13)



If we generate from an MRS representation "Something called "张三" is *amplifier* clean", we can get two possible surface forms:

(14) a. 张三 很 干净
zhāngsān hěn gānjìng
Zhangsan very clean

“Zhangsan is very clean”

b. 张三 干干净净
zhāngsān gāngānjìngjìng
Zhangsan redup_up-clean

“Zhangsan is very clean”

The above two lexical rules handle the A → AA reduplication for both verbs and adjectives. With pre-processing using regular expressions, another variation of the reduplication pattern of monosyllabic verbs, A → A 一 (*yi* “one”)A, can also be handled by (11). An example of this pattern is given below in (15).

(15) 看一看
kànyīkàn
look-one-look

“take a look/look a little”

Since AABB reduplication of AB adjectives and ABAB reduplication of AB verbs are not very productive in Chinese (i.e., there are many AB adjectives or verbs that can not be reduplicated this way), we list them as irregular derivation forms in `irregs.tab`. We have collected 92 entries for the AABB adjectives, and 74 entries for the ABAB verbs so far.

Another AB verb reduplication pattern is AB → AAB in (16), repeating the first character of some AB verbs. There is a similar pattern for some verbs with three characters. These verbs (so far 76) are also defined in `irregs.tab` to be handled in a similar manner.

(16) 说说话
shuōshuōhuà
AAB-talk

“have a talk/talk a little”

Other forms of AB verb reduplication, such as A了(*le*, “asp-marker”)A, and AA看(*kàn* “see”), will be added in future work.

ABB, shown in (17) and (18), is another commonly mentioned adjective reduplication pattern. Like other reduplicated words, it can’t be modified by degree adverbs. However, semantically it can’t be reduced down to an A or AB predicate and a general reduplication predicate *redup_up_x_rel*. Either the AB form of the word doesn’t exist, or its A form exists but the different reduplication BB adds different meaning to the same A form. These adjectives are directly added into the lexicon (103 entries) with a lexical type defined with the required syntactic constraint.

(17) 绿油油
lǜyóuyóu
green-oil-oil

“bright green”

(18) 绿茸茸
lǜróngróng
green-downy-downy

“mossy green”

The semantic connection between (17) and (18), that they are more specific but slightly different kinds of green (“bright green” and “mossy green”), will be captured in the Chinese Open Wordnet.

4.2 SUO-DE structure

In Mandarin Chinese, 所 *sǔo* is a particle used before a transitive verb to nominalize the structure “SUO+V” into a noun phrase (L ǚ , 1999). According to Lu & Ma (1985), in modern Chinese, SUO is used most commonly in the structure “(NP$_1$+)SUO+V+DE”, either to modify a noun following it (NP$_2$) or to act as a noun phrase itself. These variations are listed below in (19a-d). The last variation (19e) is used directly as an noun phrase in formal text.

(19) a. “$NP_1 + SUO + V + DE + NP_2$”
b. “$SUO + V + DE + NP_2$”
c. “$NP_1 + SUO + V + DE$” as NP
d. “$SUO + V + DE$” as NP
e. “$SUO + V$” as NP

One usage of SUO, for structure (19a) “NP$_1$+SUO+V+DE+NP$_2$”, is shown in example (20).

(20) 他 所 写 的 书
tā suǒ xiě de shū
he SUO write DE book

“the book he wrote”

We take the view of Deng (2009) that in structures where both SUO and DE appear (19a-d), DE plays the key role of nominalizing the phrase “(NP$_1$+)SUO+V+DE”, so that it can either be a noun phrase itself, or be a prenominal adjunct (relative clause) to NP$_2$. The role of SUO in the construction is to indicate that the missing argument of the verb is its patient or direct object.

Specifically, for structures in (19a & b), the lexical entry for the relativizing DE is presented in (21). The feature SPR of DE selects a preceding verbal clause containing a gap of one missing argument. DE heads the resulting relative clause, the missing argument of which is coreferential with the noun it modifies. The GAP value of DE's selected clause is defined to be identical to the NP in DE's MOD. DE's non-empty STOP-GAP feature ensures that it performs the gap-filling required.

DE also shares its HEAD feature with that of the selected clause. Semantically, DE does not introduce any information, so its RESTR list is empty, and its INDEX is the same as that of its selected clause.

(21)

$$\left\langle \text{的}, \begin{bmatrix} \text{SYN} & \begin{bmatrix} \text{HEAD}\ \boxed{2} & \\ \text{VAL} & \begin{bmatrix} \text{SPR} & \left\langle \text{V}\begin{bmatrix} \text{SYN} & \begin{bmatrix}\text{HEAD}\ \boxed{2}\\ \text{GAP}\left\langle\boxed{1}\right\rangle\end{bmatrix} \\ \text{SEM} \mid \text{INDEX}\ s & \end{bmatrix}\right\rangle \\ \text{COMPS} & \langle\rangle \\ \text{MOD} & \left\langle \boxed{1}\text{NP}\right\rangle \end{bmatrix} \\ \text{STOP-GAP}\left\langle\boxed{1}\right\rangle & \end{bmatrix} \\ \text{SEM} & \begin{bmatrix}\text{INDEX} & s \\ \text{RESTR} & \langle\rangle\end{bmatrix}\end{bmatrix}\right\rangle$$

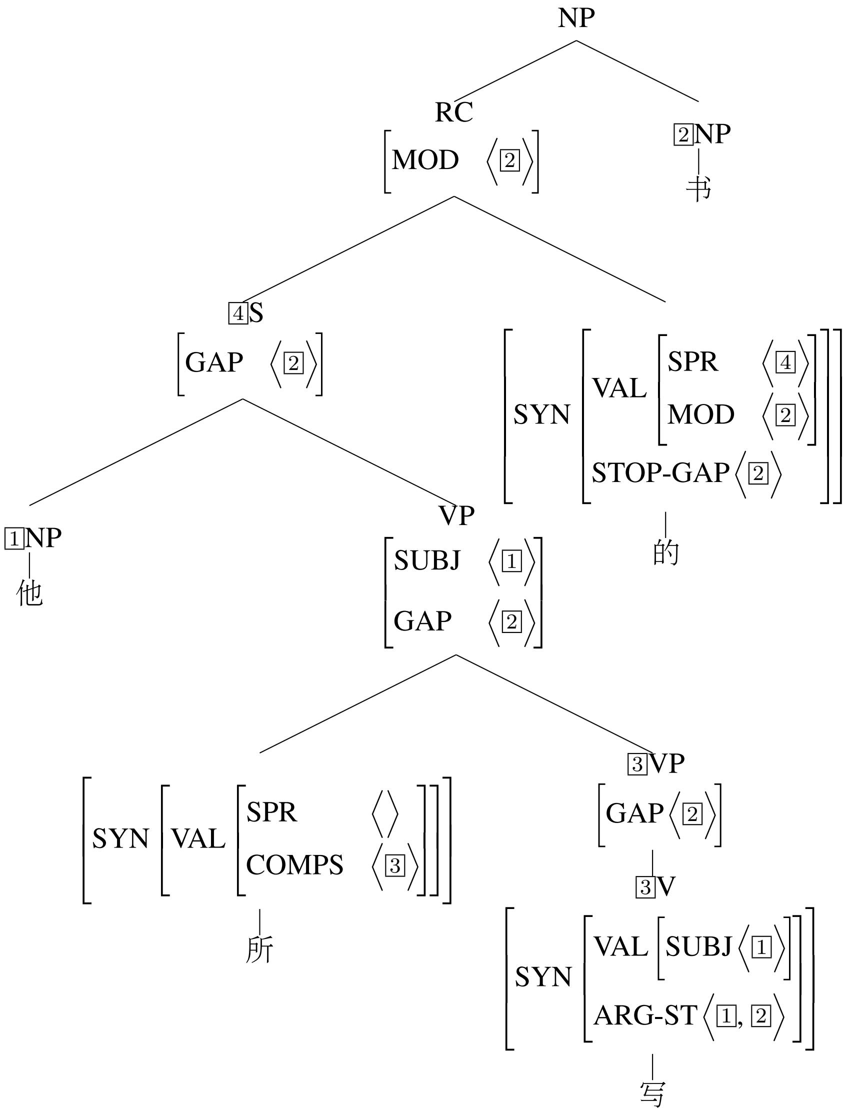
The lexical entry for SUO is shown in (22). SUO selects a transitive verb which has an unrealized subject and a GAP value referring to its direct object (2nd item on ARG-ST list). As a non-head marker marking the missing object, SUO has nothing to add on semantically. It's worth noting that SUO is redundant when NP_1 is present. When NP_1 is not present, SUO helps to restrict the reading of the gap.

(22)

$$\left\langle \text{所}, \begin{bmatrix} \text{SYN} & \begin{bmatrix} \text{HEAD}\ \textit{marker} & \\ \text{VAL} & \begin{bmatrix} \text{SPR} & \langle\rangle \\ \text{COMPS} & \left\langle \text{V}\begin{bmatrix} \text{SYN} & \begin{bmatrix}\text{HEAD}\ \boxed{3} & \\ \text{VAL} & \begin{bmatrix}\text{SUBJ} & \left\langle\boxed{1}\right\rangle \\ \text{COMPS} & \langle\rangle\end{bmatrix} \\ \text{GAP}\left\langle\boxed{2}\right\rangle & \end{bmatrix} \\ \text{ARG-ST}\left\langle\boxed{1}, \boxed{2}, \ldots\right\rangle & \\ \text{SEM} \mid \text{INDEX}\ s & \end{bmatrix}\right\rangle \end{bmatrix} \end{bmatrix} \\ \text{SEM} & \begin{bmatrix}\text{INDEX} & s \\ \text{RESTR} & \langle\rangle\end{bmatrix}\end{bmatrix}\right\rangle$$

(21) and (22) interact to produce the noun phrase structure for (20) in (23). In the tree, SUO constrains the missing argument of the verb to be the direct object. This information, contained in feature GAP, is passed up the tree, until the S or VP combines with DE to form a relative clause.

(23)



We have implemented SUO and the relativizing DE into our grammar for SUO-DE structures in (19a & b). The MRS representation for (20) is presented in (24), where the ARG2 of the predicate *_写_v_1_rel* “write” links to the predicate *_书_n_1_rel* “book”. The implementation for (19c & d) is currently in progress.

(24)
$$\left[\begin{array}{ll}
\textit{mrs} & \\
\text{TOP} & \boxed{0}\,h \\
\text{INDEX} & \boxed{2}\,e \\
\text{RELS} & \left\langle \begin{array}{l}
\left[\begin{array}{ll} \textit{pron\_rel} & \\ \text{LBL} & \boxed{9}\,h \\ \text{ARG0} & \boxed{10}\,x \end{array}\right],
\left[\begin{array}{ll} \textit{pronoun\_q\_rel} & \\ \text{LBL} & \boxed{11}\,h \\ \text{ARG0} & \boxed{10}\,x \\ \text{RSTR} & \boxed{12}\,h \\ \text{BODY} & \boxed{13}\,h \end{array}\right],
\left[\begin{array}{ll} \textit{\_写\_v\_1\_rel} & \\ \text{LBL} & \boxed{14}\,h \\ \text{ARG0} & \boxed{15}\,e \\ \text{ARG1} & \boxed{10}\,x \\ \text{ARG2} & \boxed{8}\,x \end{array}\right], \\
\left[\begin{array}{ll} \textit{\_书\_n\_1\_rel} & \\ \text{LBL} & \boxed{16}\,h \\ \text{ARG0} & \boxed{8}\,x \end{array}\right],
\left[\begin{array}{ll} \textit{exist\_q\_rel} & \\ \text{LBL} & \boxed{17}\,h \\ \text{ARG0} & \boxed{8}\,x \\ \text{RSTR} & \boxed{18}\,h \\ \text{BODY} & \boxed{19}\,h \end{array}\right]
\end{array}\right\rangle \\
\text{HCONS} & \left\langle
\left[\begin{array}{ll} \textit{qeq} & \\ \text{HARG} & \boxed{0}\,h \\ \text{LARG} & \boxed{16}\,h \end{array}\right],
\left[\begin{array}{ll} \textit{qeq} & \\ \text{HARG} & \boxed{12}\,h \\ \text{LARG} & \boxed{9}\,h \end{array}\right],
\left[\begin{array}{ll} \textit{qeq} & \\ \text{HARG} & \boxed{18}\,h \\ \text{LARG} & \boxed{16}\,h \end{array}\right]
\right\rangle \\
\text{ICONS} & \left\langle
\left[\begin{array}{ll} \textit{focus-or-topic} & \\ \text{IARG1} & \boxed{15}\,e \\ \text{IARG2} & \boxed{8}\,x \end{array}\right]
\right\rangle
\end{array}\right]$$

5 Conclusion

We have extended our grammar of Chinese with new analyses for reduplication and the SUO-DE structure. The analysis of reduplication uses lexical rules to capture both the syntactic and semantic properties (amplification in adjectives and diminishing in verbs). Words showing non-productive reduplication are entered in the lexicon, and the semantic relations will be captured in an external resource (the Chinese Open Wordnet). Classifier reduplication is left until we have a fuller analysis of classifiers. The SUO-DE structure constrains the meanings of relative clauses to a gapped-object interpretation.

Treebanking using the current version of Zhong has revealed many gaps, especially in dealing with longer sentences found in real text, where different phenomena tend to interact to make constraint specification challenging . We plan to focus our subsequent efforts on phenomena that would help parse such longer sentences. Some of the tasks on the immediate agenda are: relative clauses, variations of nominalisation, serial verb constructions, conjunctions, other forms of VV compounds, etc. Lexical acquisition for Mandarin Chinese using traditional characters, `zht`, and Cantonese, `yue`, will also be performed to expand their lexical coverage.

Acknowledgments

We would like to thank the reviewers, the participants of the Linguistic Analysis Design (LAD) sessions at the DELPH-IN 2015 summit, attendees of the HPSG 2015 conference as well as all others who have given their useful input.

This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Avgustinova, Tania & Yi Zhang. 2009. Parallel Grammar Engineering for Slavic Languages. In *Workshop on grammar engineering across frameworks at the ACL/IJCNLPx*, .

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson & Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation* 8(1). 23–72. http://dx.doi.org/10.1007/s11168-010-9070-1. 10.1007/s11168-010-9070-1.

Chen, Feng-yi, Ruo-ping Jean Mo, Chu-Ren Huang & Keh-Jiann Chen. 1992. Reduplication in Mandarin Chinese: Their formation rules, syntactic behavior and ICG representation. In *Proceedings of rocling v computational linguistics conference v*, 217–233. The Association for Computational Linguistics and Chinese Language Processing.

Copestake, Ann, Dan Flickinger, Ivan A. Sag & Carl Pollard. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation* 3(2). 281–332.

Crysmann, Berthold. 2005. Relative clause extraposition in German: An efficient and portable implementation. *Research on Language and Computation* 3(1). 61–82.

Deng, Dun. 2009. 《现代汉语“所”及“所”字结构的重新审视与定性》 xiàndài hànyǔ suǒ jí suǒ zì jiégòu de chóngxīn shěnshì yǔ dìngxìng [a new analysis and definition of suo and the suo construction in modern Chinese]. 《汉语学习》 *Hànyǔ xuéxí [Chinese Language Learning]* 2. 106–112.

Fan, Zhenzhen, Sanghoun Song & Francis Bond. 2015. Building Zhong [|], a Chinese HPSG meta-grammar. In *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, 97–110.

Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6(1). 15–28. (Special Issue on Efficient Processing with HPSG).

Fokkens, Antske & Tania Avgustinova. 2013. SlaviCLIMB: Combining expertise for Slavic grammar development using a metagrammar. In *Workshop on high-level methodologies for grammar engineering*, 87–92.

Fokkens, Antske, Tania Avgustinova & Yi Zhang. 2012. CLIMB Grammars: Three Projects using Metagrammar Engineering. In *Proceedings of the eight international conference on language resources and evaluation*, 1672–1679. Istanbul.

Gao, Qian. 1994. Chinese NP Structure. *Linguistics* 32. 475–510.

Gao, Qian. 2000. *Argument Structure, HPSG, and Chinese Grammar*: Ohio State University dissertation.

Hellan, Lars. 2005. Implementing Norwegian reflexives in an HPSG grammar. In *Proceedings of the 12th international conference on head-driven phrase structure grammar (HPSG)*, 519–539. Stanford: CSLI Publications.

Kim, Jong-Bok, Jaehyung Yang, Sanghoun Song & Francis Bond. 2011. Processing of Korean and the development of the Korean resource grammar. *Linguistic Research* 28(3). 635–672.

Li, Charles N. & Sandra A. Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. University of California Press.

Li, Wei. 2001. *The morpho-syntactic interface in a Chinese phrase structure*: Simon Fraser University dissertation.

Lu, Jianming & Zhen Ma. 1985. 《"的"字结构和"所"字结构》 dé zì jiégòu hé suǒ zì jiégòu [de construction and suo construction]. In 《现代汉语虚词散论》 *xiàndài hànyǔ xūcí sǎnlùn [a collection of articles on functional words in modern Chinese]*, 231–248. Beijing: Beijing University Press.

L ǚ , Shuxiang (ed.). 1999. 《现代汉语八百词》*xiàndài hànyǔ bābǎi cí [eight hundred words of modern Chinese]*. Beijing: The Commercial Press Ltd.

Marimon, Montserrat. 2012. The Spanish DELPH-IN grammar. *Language Resources and Evaluation* 47(2). 371–397.

Müller, Stefan. 2013. The CoreGram project: Theoretical linguistics, theory development and verification. Ms. Freie Universität Berlin. http://hpsg.fu-berlin.de/~stefan/Pub/coregram.html.

Müller, Stefan & Walter Kasper. 2000. HPSG analysis of German. In Wolfgang Wahlster (ed.), *Verbmobil: Foundations of speech-to-speech translation*, 238–253. Berlin, Germany: Springer.

Müller, Stefan & Janna Lipenkova. 2013. Chingram: A TRALE implementation of an HPSG fragment of Mandarin Chinese. *Sponsors: National Science Council, Executive Yuan, ROC Institute of Linguistics, Academia Sinica NCCU Office of Research and Development* 240.

Ng, Say Kiat. 1997. *A double specifier account of Chinese NPs using head-driven phrase structure grammar*. University of Edinburgh Department of Linguistics Msc thesis.

Pollard, Carl & Ivan A. Sag. 1994. *Head driven phrase structure grammar*. Chicago: University of Chicago Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Siegel, Melanie & Emily M. Bender. 2002. Efficient deep processing of Japanese. In *Proceedings of the 3rd workshop on Asian language resources and international standardization at the 19th international conference on computational linguistics*, 1–8. Taipei.

Sio, Joanna Ut-Seong & Sanghoun Song. 2015. Divergence in expressing definiteness between Mandarin and Cantonese. In *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, 178–195.

Wang, Shan & Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th workshop on Asian language resources, a workshop at IJCNLP-2013*, 10–18. Nagoya.

Wang, Wenjie, Sanghoun Song & Francis Bond. 2015. A constraint-based analysis of A-NOT-A questions in Mandarin Chinese. In *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, 196–215.

Xue, Ping & Paul McFetridge. 1995. DP structure, HPSG and the Chinese NP. In *Proceedings of the 14th annual conference of the Canadian linguistics association*, .

Xue, Ping, Carl Pollard & Ivan A Sag. 1994. A new perspective on Chinese ziji. In *Proceedings of the thirteenth west coast conference on formal linguistics*, .

Yang, Chunlei. 2007. *Expert systems for pragmatic interpretations of ziji and quantified noun phrases in HPSG*: Shanghai International Studies University dissertation.

Zhang, Yi, Rui Wang & Yu Chen. 2011. Engineering a deep HPSG for Mandarin Chinese. In *Proceedings of the 9th workshop on Asian language resources*, .

Unique lexical entries in a subconstructional grammar

Petter Haugereid
Bergen University College

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 110–121

http://csli-publications.stanford.edu/HPSG/2015

Abstract

Function words like prepositions, adverbs, particles, and complementizers may be assigned more than one category due to the different functions they can have. In this paper I present an approach that assumes unique lexical entries for words that are assigned more than one category. I will focus on prepositions and how they may function as heads of modifying PPs, selected prepositions, or as particles.

1 Introduction

The Norwegian LFG grammar NorGram (Dyvik, 2000) has a long list of lexical entries where one form is assigned more than one category. Table 1 shows for each pair of a selected set of categories, the number of word forms that are assigned both categories. There are 43 adjectives (A) that also can be degree adverbs (ADVdeg). One of them, *merkelig*, is illustrated in (1) as an adjective (1a)) and as a degree adverb ((1b)).

(1) a. Det var en merkelig følelse.
it was a strange feeling
It was a strange feeling.

b. Rommet blir merkelig stille.
room-DEF becomes oddly quiet
The room becomes oddly quiet.

As the table shows, many prepositions also can be adverbs (66), particles (PRT) (38) and selected prepositions (Psel) (53). One of them, *unna* ('away'), is exemplified in (2) where it is an adverb ((2a)), a preposition ((2b)), a particle ((2c)), and a selected preposition ((2d)).

(2) a. Han kjørte unna.
he drove away
He drove out of the way.

b. De gikk unna flammene.
they walked away flames-DEF
They walked away from the flames.

c. Han smatt unna.
he escaped away
He escaped.

d. Han sluntret unna pliktene sine.
he idled away duties his
He shirked his duties.

[†] I would like to thank four anonymous reviewers, the INESS group in Bergen and the audience at the HPSG 2015 conference in Singapore for very useful comments and suggestions.

	A	ADV	ADVdeg	ADVs	Cadv	P	PRT	Psel
Psel	0	38	1	0	4	53	31	-
PRT	5	39	2	3	3	38	-	
P	5	66	1	3	9	-		
Cadv	4	8	4	7	-			
ADVs	6	15	31	-				
ADVdeg	43	15	-					
ADV	13	-						
A	-							

Table 1: Pairing of categories and the number of words assigned to both categories in NorGram.

The most obvious way to treat these words in the lexicon, is to create separate lexical items for each category assigned to it. This is not entirely satisfying, given the the intuition that most of them share a meaning. The aim of this paper is to show that these forms can be assigned unique lexical items that will be compatible with the functions that are required from them.

2 Multiple lexical items

There are several reasons for assuming several lexical entries for one form, specially within a framework like HPSG where there are no derivations and no information gets lost. In particular, this holds for semantic relations. Once a semantic relation is entered on the RELS list by a lexical item, a lexical rule or a syntactic rule, the compositional nature of HPSG requires that this relation also is a part of the semantic representation of the phrase that the lexical item, lexical rule or rule is a part of. So if the noun *tabs* introduces a relation *_tab_n_rel* and the preposition *on* introduces a relation *_on_p_rel*, these relations have to appear in the resulting semantic representation. This is a little problematic in the case of idioms like *He kept tabs on the competition*. The composition of semantic relations requires the *_tab_n_rel* and the *_on_p_rel* to be a part of the resulting representation, even though the idiomatic meaning is to *observe*.

Sag et al. (2003, 347–355) solves this problem by assuming a special lexical entry for the idiomatic version of *keep* that has three items on the SUBCAT list; (i) the NP subject, (ii) an idiomatic noun *tabs*, and (iii) a constituent marked by the preposition *on*. (See (3).) The relation of the idiomatic version of *keep* is *observe*, and the idiomatic noun *tabs* and the selected preposition *on* are both assumed to be semantically empty. This gives the intended *oberve*-relation between the OBSERVER (*he*) and the OBSERVED (*the competition*).

(3)
$$\begin{bmatrix} \textit{ptv-lxm} \\ \text{STEM} \left\langle \text{keep} \right\rangle \\ \text{ARG-ST} \left\langle \text{NP}_i \text{ , } \begin{bmatrix}\text{FORM tabs}\end{bmatrix}, \begin{bmatrix}\text{FORM on} \\ \text{INDEX } j\end{bmatrix} \right\rangle \\ \text{SEM} \begin{bmatrix} \text{INDEX } s \\ \text{RESTR} \left\langle \begin{bmatrix} \text{RELN} & \textbf{observe} \\ \text{SIT} & s \\ \text{OBSERVER} & i \\ \text{OBSERVED} & j \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

The problem with this approach is that in addition to an idiomatic and non-idiomatic version of the verb *keep*, it also presupposes an empty preposition (in addition to the standard preposition with an *_on_p_rel*) and an idiomatic noun *tabs* in addition to the regular word *tabs* with the relation *_tab_n_rel*.

There is a whole range of linguistic phenomena that one way or another forces the use of multiple lexical entries for the same form in Norwegian:

- Verbs, nouns or adjectives can have several argument frames, and the standard way to account for that in lexicalist frameworks like HPSG and LFG is to assume multiple lexical entries, alternatively deriving lexemes from lexemes by lexical rules. An example of a verb with many frames is the verb *få* ('get') in NorGram which has 38 frames, each of which is expanded into a lexical entry during parsing. Verbs, nouns and adjectives also can appear in idioms, in which case they do not retain their original meaning, and separate lexical items are assumed.

- Adjectives also can be degree adverbs (see (1)).

- Adverbs and prepositions also can be complementizers.

- Prepositions also may have other roles, as head of a modifying PP, as a selected preposition, as an adverbial or as a particle (see (2)).

- Certain pronouns can function as personal pronouns or reflexives, or possessive pronouns or possissive reflexives.

- Certain determiners can function as numerals or articles, as pronouns or as definite determiners.

3 Incremental parsing with left-branching structures

Instead of assuming that lexical entries are specific to the extent that multiple lexical entries are needed for the same form (where the basic meaning is the same), I suggest an approach where lexical items are allowed to be underspecified with

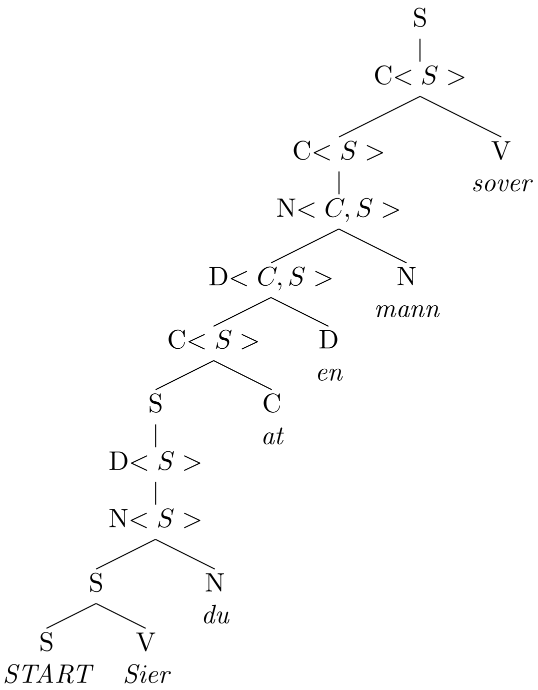


Figure 1: Parse tree

regard to what function they fill. This approach depends on three factors; (i) underspecification, (ii) multiple inheritance, and (iii) category specific phrase structure rules that access the words in question. While the the first two factors are common practice in HPSG, the third factor is an innovation. It can be achieved by means of incremental parsing with left-branching structures.

In my approach I assume that parse trees are distinct from constituent trees, and that the parse trees are completely left-branching (Haugereid & Morey, 2012). The strategy is that of a shift reduce parser, namely to use a stack to store information about constituents that are not completed. This gives us parse trees without center-embeddings, and allows for incremental processing of sentences.

There are mainly three types of rules: (i) *embedding rules*, that initiate a constituent, (ii) *attaching rules*, that add words to an already initiated constituent, and (iii) *popping rules*, that mark the completion of a constituent.

The syntactic structure is built incrementally, word by word, as shown in Figure 1. The analysis starts with a START sign in the bottom left. The START sign is combined with the first word of the sentence with a binary rule, in this case the rule for attaching the verb *Sier* (an attaching rule). The structure that now consists of the start sign and the first word (represented by the node S) is then combined with the next word *du* with a rule that initiates nominal constituents (an embedding rule) (N<S>). The features of the *S* are then put on a stack. The next rule is a unary rule

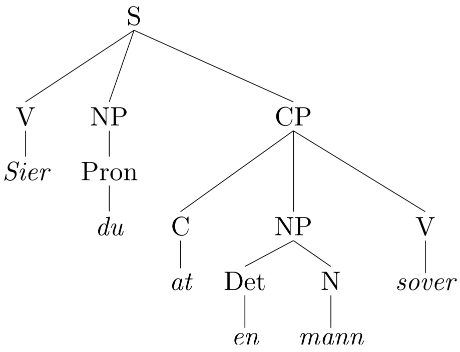


Figure 2: Constituent tree

that adds a quantifier relation (D<S>), and the following rule is a rule that pops the features of the start symbol from the stack, and the category goes back to S. Similar embedding, attaching and popping rules apply for the rest of the clause. The constituent tree is formed simply by adding a left bracket when there is an embedding rule and a right bracket when there is a popping rule. The constituent tree corresponding to the parse tree in Figure 1 is shown in Figure 2.

This left-branching design opens for subconstructions that attach single words, and not full constituents, and it gives us the possibility to tailor subconstructions for every category of words, and the words attached by the subconstructions are allowed to be more or less specific.

4 Analysis of prepositions as unique lexical entries

In this section I will focus on prepositions and show how a preposition can be attributed one lexical entry that accounts for all its functions. It is allowed by a combination of the constructionalist approach sketched in Section 3, underspecification, and the exploitation of types. The analysis is implemented in an HPSG-like grammar of Norwegian within the LKB system (Copestake, 2001).

A preposition like *on* can be both a particle (*I logged on*) and a selected preposition (*He relied on the kindness of strangers*/*We kept tabs on our checking account*). In addition, it can also be a regular preposition as in *He sleeps on the floor.*

My approach to prepositions is inspired by the treatment of particles and selected prepositions in the English Resource Grammar (ERG) (Flickinger, 2000), where the lexical entry for *on* as a particle or selected preposition is shown in (4).

(4)
$$\left[\begin{array}{l} \text{ORTH}\left\langle \text{”on”} \right\rangle \\ \text{CAT} \left[\begin{array}{l} \text{HEAD} \left[\begin{array}{l} \textit{prep} \\ \text{MOD}\,\langle\rangle \end{array}\right] \\ \text{VAL}|\text{COMPS} \left\langle \left[\begin{array}{l} \textit{synsem} \\ \text{CAT}|\text{HEAD } \textit{nom} \\ \text{CONT}|\text{HOOK } \boxed{1} \end{array}\right] \right\rangle \end{array}\right] \\ \text{CONT} \left[\begin{array}{l} \text{HOOK } \boxed{1} \\ \text{RELS} \left\langle !! \right\rangle \end{array}\right] \\ \text{KEYREL} \left[\begin{array}{l} \textit{basic\_arg12\_relation} \\ \text{PRED } \textit{\_on\_p\_sel\_rel} \end{array}\right] \end{array}\right]$$

The ERG lexical entry for selected prepositions/particles has an empty RELS list, which means it is semantically empty. Still, it has specified a KEYREL with a PRED value (*_on_p_sel_rel*) that will be required by the verb that selects it. But this relation does not end up on the RELS list.

My approach is similar in that I assume a lexical entry with an empty RELS list. (See the lexical entry for *på* ('on') in (5).) It also has a relation as value of KEYREL, but the PRED value is an underspecified type, *på_prd*, which allows it to function as a normal preposition, as a selected preposition, and as a particle.

(5)
$$\left[\begin{array}{ll} \textit{prep-word} & \\ \text{ORTH} & \left\langle \text{”på”} \right\rangle \\ \text{CAT} & \left[\text{HEAD } \textit{prep}\right] \\ \text{CONT} & \left[\text{RELS} \left\langle !! \right\rangle\right] \\ \text{KEYREL} & \left[\text{PRED } \textit{\_på\_prd}\right] \end{array}\right]$$

I can do this, firstly, because the PRED value is underspecified, which means that it is compatible with different relations as *_på_p_rel* (regular preposition relation) and all predicates that include *på* as a part of a complex predicate, like *_fokusere*på_14_rel* ('focus on') and *_logge-på_1_rel* ('log on'). Secondly, I use phrasal subconstructions, which makes it possible to decompose argument frames and predicates and let each sign of the grammar, be it a lexical item, an inflectional rule, or a syntactic rule, only contribute that piece of information that positively can be attributed to it, even if it is underspecified information. When the signs are put together, the pieces of information contributed by each sign about the argument frame and the predicate are unified, and the predicate is determined. The simplified type hierarchy in Figure 3 shows how the type *på_prd* is compatible with the predicates *_logge-på_1_rel*, *_fokusere*på_14_rel*, and *_på_p_rel*.[1]

[1] The predicate names also indicate the number of arguments as well as their function. This is discussed in Haugereid (2014).

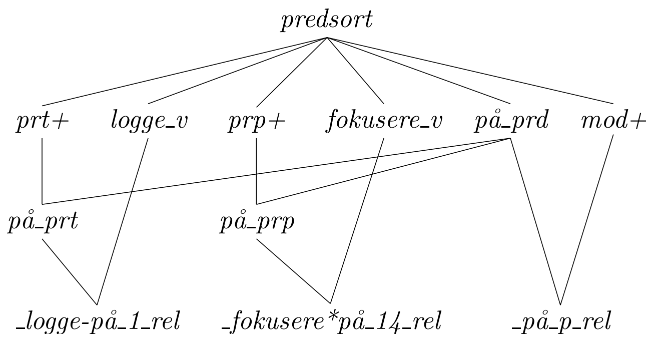


Figure 3: Type hierarchy of pred values of *på* ('on')

It is the function *på* has in the clause that determines which predicate it will end up with. If it functions as a particle of *logge* ('log'), *på_prd* will be unified with the PRED value of *logge* (*logge_v*), and the resulting relation will be *_logge-på_1_rel*. If it functions as a selected preposition of *fokusere* ('focus'), *på_prd* will be unified with *fokusere_v*, yielding the predicate *_fokusere*på_14_rel*. And if it functions as a modifier, *på_prd* will be unified with the type *mod+*, which gives the predicate *_på_p_rel*.

The subconstruction rule that attaches particles is given in Figure 4. It unifies the KEYREL value of the structure built so far (the first daughter) with that of the particle, and also the mother. It marks the PART value of the first daughter as *prt+*, and this value is unified with that of KEYREL|PRED. This ensures that *på* is interpreted as a particle.

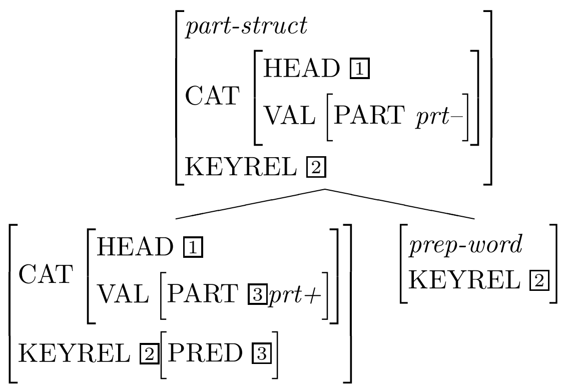


Figure 4: Rule for attaching particles

Similar to this rule attaching particles, the grammar also has a rule *marker-struct* that attaches selected prepositions.

The subconstruction rule for attaching verbs (*vbl-struct*) is shown in Figure 5. It selects the verb via the VBL feature, and the VBL requirement of the verb is transferred to the mother. Like the subconstruction rules for particles and prepositions, this rule unifies the KEYREL value of the structure built so far (the first daughter) with that of the attached word (the verb), and the mother.

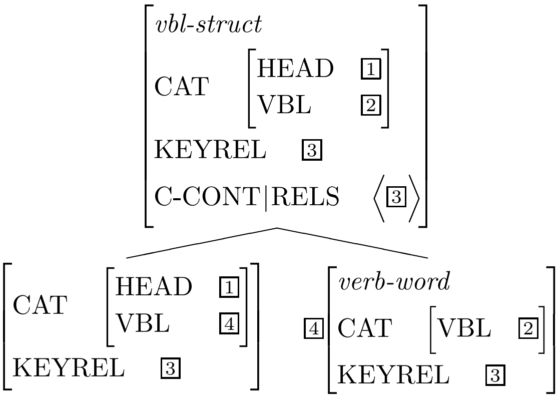


Figure 5: Rule for attaching verbs

The unification of KEYREL values in *part-struct* and *vbl-struct* ensures that when they apply in the same clause, the PRED values of the verb and the particle have to unify. Only the combinations of verb predicate and preposition/particle predicate that are defined in the type hierarchy are licenced by the grammar.

5 Implementation

The approach has been tested with a large computational lexicon, the NorKompLeks (NKL) (Nordgård, 1996), which is a lexicon with about 75,000 lexical entries, of which 7,400 are verbs. The verbs are listed with one or more argument frames. In all, there are 13,330 argument frames, on average about 2 per verb. The lexicon has 1,322 lexical items that may function as prepositions, adverbs or particles.

I have created a table where I match the argument frame codes in NKL with subconstruction types in Norsyg. An intransitive verb like *abdisere* ('abdicate') is in NKL given the argument frame code `intrans1`. This code is matched with the subconstruction types *1np*, *arg2–*, *arg3–*, *arg4–*, and *prt–*, which means a frame with an (external) NP subject (*1np*) and no other arguments or particles. The argument frame type associated with the lexical entry for *abdisere* gets the following definition:

_abdisere_1_rel := abdisere_v & prt– & 1np & arg2– & arg3– & arg4–.

Here, the type *abdisere_v* is the type that is specified on the verb.[2] The lexical entry for *abdisere* is given in (6). Note that, as with prepositions, the RELS list of the verb is empty. It is rather the subconstruction rule for adding verbs, *vbl-struct*,

[2] Since the verb only has one frame associated with it, it could also have been specified with its only subtype, *_abdisere_1_rel*.

that enters the KEYREL value of the verb onto the RELS list. (See Figure 5.) In this way we are not committing ourselves to the existence of a specific verbal relation if a verb appears in a sentence. The verb may for example be a part of an idiom or function as a light verb in a serial verb construction.

(6)
$$\begin{bmatrix} \textit{verb-word} & \\ \text{ORTH} & \left\langle \text{"abdisere"} \right\rangle \\ \text{CAT} & \begin{bmatrix}\text{HEAD } \textit{verb}\end{bmatrix} \\ \text{CONT} & \begin{bmatrix}\text{RELS} \left\langle !! \right\rangle\end{bmatrix} \\ \text{KEYREL} & \begin{bmatrix}\text{PRED } \textit{abdisere\_v}\end{bmatrix} \end{bmatrix}$$

The verb *få* ('get'), which in NKL is listed with 22 frames,[3] is given the lexical entry in (7). It is specified with the same information as the intransitive verb *abdisere* ('abdicate'). Only the ORTH and KEYREL values are different.

(7)
$$\begin{bmatrix} \textit{verb-word} & \\ \text{ORTH} & \left\langle \text{"få"} \right\rangle \\ \text{CAT} & \begin{bmatrix}\text{HEAD } \textit{verb}\end{bmatrix} \\ \text{CONT} & \begin{bmatrix}\text{RELS} \left\langle !! \right\rangle\end{bmatrix} \\ \text{KEYREL} & \begin{bmatrix}\text{PRED } \textit{få\_v}\end{bmatrix} \end{bmatrix}$$

This illustrates the shift of the burden of valence alternations from the lexicon to the hierarchy of subconstruction types. The KEYREL|PRED type *få_v* is given 22 subtypes. Three of them are shown below:

_få_12_rel := få_v & prt- & 1np & 2np & arg3- & arg4-.

_få-bort_12_rel := få_v & bort_prt & 1np & 2np & arg3- & arg4-.

*_få*med-refl_124_rel := få_v & prt- & 1np & 2np & arg3- & med_prp & 4refl.*

The subtype *_få_12_rel* allows *få* to be realized as a regular transitive verb with an NP subject (*1np*) and an NP object (*2np*).

The subtype *_få-bort_12_rel* is a transitive frame for the particle verb *få bort* 'remove'. As with *_logge-på_1_rel* in Figure 3, the KEYREL|PRED value of the verb *få_v* is unified with the KEYREL|PRED of the particle (*bort* 'away').

The subtype *_få*med-refl_124_rel* is a frame for the verb *få* with the selected preposition *med* and a reflexive pronoun as object of the preposition; *få med seg (noe)* 'manage to bring/understand (something)'.

The crossclassification of the verb predicates (7,400), function word predicates (1,322), and about 30 other subconstruction types gives 13,330 argument frame types of which 1,781 involve particles, 5,536 involve selected prepositions, and 84 frames involve both selected prepositions and particles. The hierarchy

[3] As mentioned in Section 2, the NorGram lexicon, which is more developed, lists på with 38 frames.

takes 1 hour and 43 minutes to compile with ACE (`http://sweaglesw.org/linguistics/ace/`). However, once the grammar is compiled, the size of the hierarchy of subconstruction types does not seem to have a serious effect on the efficiency of the parser. The parsing time of a sentence parsed when a small lexicon with 2,000 lexical entries is loaded is 0.01534 seconds, and the parsing time for the same sentence when the full lexicon (75,000 lexical entries) is loaded is 0.01778 seconds. Whether the increase is due to the size of the lexicon or the size of the hierarchy of subconstruction types is unknown.

6 Future work

The modifier rule is given in Figure 6. It is an embedding rule, which means that the key features of the structure built so far (here, the CAT and the KEYREL of the first daughter) are put on a STACK in the mother, and the HEAD and the KEYREL features of the word initiating the modifying constituent are unified with those of the mother. The KEYREL of the modifier is entered onto the C-CONT|RELS list. In addition, its PRED value is unified with the *mod+* type, which means that if the word initiating the modifying constituent is the preposition *på*, its PRED value *_på_prd* will be unified with the type *mod+*, yielding the PRED value *_på_p_rel*, which appears in the semantic representation of the sentence.

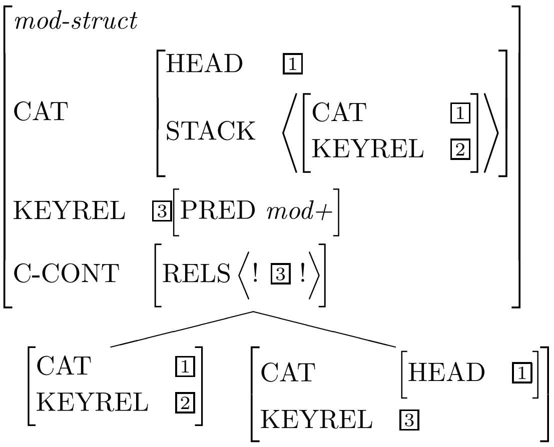


Figure 6: Embedding rule for attaching modifiers

Also other categories are treated in the same fashion. Nouns are not specified with a relation on the RELS list. Like the prepositions, their relation is specified as value of KEYREL, and the relation is entered on the RELS list when the words are added by their respective rules. This allows us to have special subconstructions for idiom nouns, like *tabs* in *keep tabs on*, that rather than treating the relation of the noun as a separate relation by entering it on the RELS list, unifies its predicate with the predicate of the verb (*keep*) and the preposition (*on*), resulting in a single idiom predicate.

The aim is to extend this analysis also to other categories, like adjectives that can be degree adverbs (see (1)), and complementizers that can be prepositions or adverbs. I want to develop a grammar that ultimately has unique lexical entries for all the words in the lexicon, regardless of whether they are content words or function words.

References

Copestake, Ann. 2001. *Implementing typed feature structure grammars* CSLI Lecture Notes. Stanford: Center for the Study of Language and Information. `http://cslipublications.stanford.edu/site/1575862603.html`.

Dyvik, Helge. 2000. Nødvendige noder i norsk: Grunntrekk i en leksikalsk-funksjonell beskrivelse av norsk syntaks. In Øivin Andersen, Kjersti Fløttum & Torodd Kinn (eds.), *Menneske, språk og felleskap*, Novus forlag.

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

Haugereid, Petter. 2014. VP idioms in Norwegian: A subconstructional approach. In Stefan Müller (ed.), *Proceedings of the 21st international conference on head-driven phrase structure grammar, university at buffalo*, 83–102. Stanford, CA: CSLI Publications. `http://cslipublications.stanford.edu/HPSG/2014/haugereid.pdf`.

Haugereid, Petter & Mathieu Morey. 2012. A left-branching grammar design for incremental parsing. In Stefan Müller (ed.), *Proceedings of the 19th international conference on head-driven phrase structure grammar, chungnam national university daejeon*, 181–194. `http://cslipublications.stanford.edu/HPSG/2012/haugereid-morey.pdf`.

Nordgård, Torbjœrn. 1996. Norkompleks: Some linguistic specifications and applications. In *Allc-ach '96*, 214–216. Bergen.

Sag, Ivan A., Thomas Wasow & Emily M. Bender. 2003. *Syntactic theory: A formal introduction*. Stanford: CSLI Publications 2nd edn. `http://cslipublications.stanford.edu/site/1575864002.html`.

Hebrew Verbal Multi-Word Expressions

Livnat Herzig Sheinfux
University of Haifa

Tali Arad Greshler
University of Haifa

Nurit Melnik
The Open University of Israel

Shuly Wintner
University of Haifa

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 122–135

http://csli-publications.stanford.edu/HPSG/2015

Herzig Sheinfux, Livnat, Arad Greshler, Tali, Melnik, Nurit, & Wintner, Shuly. 2015. Hebrew Verbal Multi-Word Expressions. In Müller, Stefan (Ed.), *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, 122–135. Stanford, CA: CSLI Publications.

Abstract

Multi-word expressions (MWEs) are challenging for grammatical theories and grammar development since they blur the traditional distinction between the lexicon and the grammar, and vary in the degree of idiosyncrasy with respect to their semantic, syntactic, and morphological behavior. Nevertheless, the need to incorporate MWEs into grammars is unquestionable, especially in light of estimates claiming that MWEs account for approximately half of the entries in the lexicon. In this study we focus on verbal MWEs in Modern Hebrew: we consider different types of this class of MWEs, and propose an analysis in the framework of HPSG. Moreover, we incorporate this analysis into HeGram, a deep linguistic processing grammar of Modern Hebrew.

1 Introduction

Multi-word expressions (MWEs) in Modern Hebrew (MH), as in other languages, are not simple to characterize, since they vary in the degree of idiosyncrasy with respect to their semantic, syntactic, and morphological behavior. In this study we focus on verbal MWEs: we consider different types of this class of MWEs, and propose an analysis in the framework of HPSG (Pollard & Sag, 1994). Moreover, we incorporate this analysis in HeGram (Herzig Sheinfux et al., 2015), a deep linguistic processing grammar of Modern Hebrew.

Our motivation is twofold. First, the need to incorporate MWEs into the grammar is unquestionable, especially in light of estimates claiming that MWEs account for approximately half of the entries in the lexicon (Sag et al., 2002). Second, we view MWEs as a challenging test case for the innovative architecture implemented in HeGram.

2 Multi-word expressions

MWEs are lexical units that consist of more than one word. They tend to be semantically idiosyncratic. Consider, for example, (1) and (2), in which the idiomatic reading cannot be derived from the idioms' literal parts. One would only understand the meaning if the MWE was already known to him.

(1) *dan yaca me-ha-kelim*
Dan came.out from-the-tools
Literal: 'Dan came out of the tools.'
Idiomatic: 'Dan lost his temper.'

(2) *dan higdil roʃ*
Dan made.grow head
'Dan took.on responsibility.'

[†] This research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 505/11).

In addition, MWEs are characterized by having constrained syntactic behavior. Namely, MWEs can't necessarily be passivized, or undergo wh-questions about the idiomatic arguments ((3) and (4), respectively). However, wh-questions about the literal arguments can occur (5).

(3) dan huca me-ha-kelim
*Dan was.taken.out from-the-tools
'Dan was taken out of the tools' (only odd literal)

(4) mi-ma dan yaca
*from-what Dan came.out
'What did Dan come out of?' (only literal)

(5) mi yaca me-ha-kelim
who came.out from-the-tools
'Who lost his temper?'

MWEs are challenging for grammatical theories and grammar development, but as they account for approximately half of the entries in the lexicon (Sag et al., 2002), incorporating them into grammars is important. Moreover, identifying MWEs is important for natural language processing applications – if MWEs are not identified as such, that will probably cause problems further down the processing pipeline.

3 Verbal MWEs in Hebrew

3.1 The Patterns

Hebrew verbal MWEs vary with respect to the specificity of the arguments they take and the relations that hold among them. We identify the following patterns:

Idiomatic NP & PP complements

MWEs can be headed by verbs which lexically select for a particular NP complement (2) or for a PP headed by a particular preposition and complemented by a particular NP (6).

(6) dan yarad me-haʕec
Dan went.down from-the.tree
'Dan conceded.'

Possessive idioms

Some MWEs are headed by verbs which select for possessive NPs, either as complements of the verb (7) or as complements in the PP complement of the verb (8), and impose agreement between the possessor and one of the verb's dependents:

(7) *dan$_i$ ṭaman yad-o$_i$ ba-calaħat*
*Dan buried.*3SM *hand-his in.the-plate*
'Dan refrained from acting.'

(8) *dan$_i$ yaca mi-kelav$_i$*
Dan came.out from-tools.his
'Dan lost his temper.'

Idioms with "empty slots"

MWEs can include "empty slots", filled by non-idiomatic and unrestricted complements (e.g., *Dana* in (9)).

(9) *dan heʕemid et dana$_i$ ʕal ṭaʕut-a$_i$*
Dan made.stand ACC Dana on mistake-her
'Dan proved Dana wrong.'

3.2 The Challenges

The occurrence of verbal MWEs poses a number of challenges to any linguistic theory. Following are a number of challenges which we observe in our data and which we account for in our grammar.

The verbs which head most verbal MWEs play a dual function in language as both literal and idiomatic expressions. One challenge is to capture the commonalities of the different instantiations, while accounting for their differences. As an example, consider the following sentences illustrating a literal and an idiomatic *hoci* ('*take.out*').

(10) a. *dan hoci et ha-sefer (me-ha-argaz)*
Dan took.out ACC the-book (from-the-box)
'Dan took the book out (of the box).'

b. *dan hoci et dana$_i$ me-ha-kelim / mi-keleiha$_i$*
Dan took.out ACC Dana from-the-tools / from-tools.her
'Dan made Dana lose her temper.'

Most of the characteristics of the literal and idiomatic instantiations of the verb *hoci* ('*take.out*') are shared. The verb semantically selects two complements, *Theme* and *Source*, which are realized as NP and PP, respectively, with the PP headed by the preposition *me-* ('*from*'). Moreover, the syntactic structure of the two instantiations is identical.

The two senses diverge in a number of ways. As expected, the idiomatic sense is more restrictive in terms of its selectional restrictions. The *Source* argument can only be realized by an NP headed by the idiomatic plural definite noun *ha-kelim* ('*the tools*'). Moreover, the *Source* NP can optionally appear with a possessor suffix, provided that it is co-indexed with the *Theme* argument of the verb. Another difference is that the *Source* argument is obligatory in the idiomatic sense, and optional in the literal one. Any divergence from these restrictions eliminates the idiomatic reading.

While MWEs are quite specific with respect to their lexical selection, in some cases, they do allow for some flexibility. Consider, for example, the plural subject counterpart of (7):

(11) a. *ha-anaʃim*$_i$ *ṭamnu yad-am*$_i$ *ba-calaħat*
*the-people buried.*3P *hand.**S**-their in.the-plate.**S***

b. *ha-anaʃim*$_i$ *ṭamnu yadei-hem*$_i$ *ba-calaħat*
*the-people buried.*3P *hand.**P**-their in.the-plate.**S***
'The people refrained from acting.'

(12) *ha-anaʃim*$_i$ *ṭamnu yadei-hem*$_i$ *ba-calaħot*
*the-people buried.*3P *hand.**P**-their in.the-plate.**P***
'The people buried their hands in the plates.' (only odd literal)

With the MWE *ṭaman yad-o ba-calaħat* ('*buried his hand in the plate*'), plural subjects can either bury their singular *hand* (11a) or plural *hands* (11b) in the (singular) *plate*. Neverthless, once *plate* becomes plural (12), the idiomatic reading is lost. These constraints, of course, are expression-specific and need to be specified in the lexicon.

A different case of constrained flexibility involves internal modification. Internal modifiers can be adverbs, which, in MH, can intervene between the verb and its complement (e.g., (13)).

(13) *dan yarad* ***ba-sof*** *me-haʕec*
Dan went.down at.the-end from-the.tree
'Finally Dan conceded.'

Alternatively, internal modifiers can be adjectives which syntactically modify one of the complements, as in (14) and (15).

(14) *ha-cibur nafal ba-paħ* ***ha-pirsumi***
the-public fell in.the-bin the-advertising
'The public was tricked by advertisement.'

(15) *ha-irgunim ha-lahaṭabim mehadqim et ha-ħagora*
the-organizations the-LGBT are tightening ACC
ha-vruda *ʃelahem*
the-belt the-pink their
'The LGBT organizations are tightening their pink belt.'[1]

[1] This is an attested MH counterpart to Manfred Sailer's (p.c.) example: *They had to tighten their Gucci belts*.

Note that in all three cases the modifier is optional. Nevertheless, its occurrence inside an idiomatic verb phrase rules out the possibility of analyzing idioms under a 'word with spaces' account.

A final challenge is posed by non-local selection phenomana, of which there are two types: In the case of PP complements, such as that in (10b), there is a chain of lexical selection, where a verb selects for a PP with a particular prepositional head, which in turn selects for an NP with a particular nominal head; Additional non-local constraints are imposed in the case of possessive idioms, which require the obligatory co-indexation between possessors and arguments. For example, in (10b) the possessor of the NP complement in the *Source* PP *mi-keleiha* ('*from-her.tools*') must be co-indexed with the *Theme* NP *Dana*. Consequently, in order for this relation to hold, the index of a possessor within an NP must be "visible" at the level of the PP of which the NP is a complement.

4 The incorporation of MWEs into the grammar

4.1 HeGram

Our proposed analysis is cast in the context of HeGram (Herzig Sheinfux et al., 2015), a deep linguistic processing grammar of Modern Hebrew, which is based on a starter grammar created with the Lingo Grammar Matrix customization system (Bender et al., 2002) and implemented in the LKB (Copestake, 2002) and ACE systems. Morphology is handled outside the grammar, as the lexicon is comprised of automatically analyzed forms.

HeGram currently covers a variety of phenomena, including case marking, subject-verb and noun-adjective agreement, SVO and V2 word order, relatively free complement order, multiple subcategorization frames, selectional restrictions of verbs on their PP complements, topicalization, wh-questions, passive and unaccusative verbs, control verbs, raising verbs, and the copular construction (including zero copula). HeGram is developed in parallel with AraGram (see Arad Greshler et al., 2015), a grammar of Modern Standard Arabic.

The architecture of HeGram embodies significant changes to the way argument structure is standardly viewed in HPSG. The main one is that it distinguishes between semantic selection and syntactic selection, and provides a way of stating constraints regarding each level separately. Moreover, one lexical entry can account for multiple subcategorization frames, including argument optionality and the realization of arguments with different syntactic phrase types (e.g., *want food* vs. *want to eat*). This involves the distribution of valence features across ten categories.[2] Each valence category is characterized in terms of its semantic role, as well as the types of syntactic phrases which can realize it (referred to as *syntactic realization classes*). Consequently, the semantic relations denoted by predicates

[2] Our restructuring of the VALENCE complex is inspired by Haugereid's packed argument frames (Haugereid, 2012).

consist of coherent argument roles, which are consistent across all predicates in the language.

Table 1 presents the ten valence categories used in HeGram, along with the corresponding semantic roles and syntactic realization phrases.[3] For example, Arg2 corresponds to the *Theme* semantic role, and can be realized in MH as an NP, an infinitive VP, a CP or a PP.

Label	**Semantic Selection**	**Syntactic Realization**
Arg1	Actor, Perceiver, Causer	NP, PP
Arg2	Theme	NP, VP$_{inf}$, CP, PP
Arg3	Affectee, Benefactive, Malfactive , Recipient	NP, PP
Arg4	Attribute	AdjP, AdvP, PP, NP, VP$_{beinoni}$
Arg5	Source	PP
Arg6	Goal	PP
Arg7	Location	PP, AdvP
Arg8	Topic of Communication	PP
Arg9	Instrument	PP
Arg10	Comitative	PP

Table 1: Semantic roles and realization classes in HeGram

Each predicative lexical type in our grammars inherits from types which specify the possible semantic roles of its dependents and their possible syntactic realizations. As an example, consider the lexical type which licenses the (literal) MH verb *hoci* (‘*took out*’).

(16) MH *hoci* (‘*took out*’):

```
arg12-125_n_p := arg1_n & arg2_n & arg5_p &
       [ SYNSEM.LOCAL.CAT.VAL.R-FRAME arg12-125 ].
```

The verb semantically selects three arguments: an *Actor* (arg1), a *Theme* (arg2), and a *Source* (arg5). Moreover, it requires that its *Actor* and *Theme* roles be syntactically realized, yet allows for the omission of the *Source*. This is captured by the value of its lexical type’s R(EALIZATION)-FRAME feature, *arg12-125*, which lists the different realization frames in which the verb can appear, separated by dashes; *arg12* is a transitive syntactic frame and *arg125* represents the realization of all three semantic arguments.

The syntactic realization of the semantic arguments is defined via inheritance. The lexical type in (16) inherits from three subtypes, each pertaining to one of its semantic arguments, and each determining the syntactic category of the phrases which realize that semantic role (noun, noun, and preposition, respectively). The

[3] This architecture is similar in spirit to work done on Polish by Przepiórkowski et al. (2014).

name of this type (i.e., *arg12-125_n_p*) reflects the different realization frames, as well as the syntactic category of its dependents (since Arg1 is always realized as an NP, its syntactic realization is omitted from the name of the type).

The association between semantic roles and syntactic phrases is based on corpus investigation of MH, which included at least 100 randomly selected examples of sentences containing each of the 50 most frequent verb lemmas in the 60-million token WaCky corpus of Modern Hebrew (Baroni et al., 2009). Whereas the semantic classes are expected to be more or less universal, some language-specific differences are expected in the syntactic realizations. Corpus investigations on Modern Standard Arabic in the context of the development of AraGram confirmed these expectations (for more elaboration, see Arad Greshler et al., 2015).

4.2 Verbal MWEs in HeGram

The example sentence in (10b) repeated here as (17), poses most of the challenges described above.

(17) *dan hoci et dana*$_i$ *mi-keleiha*$_i$
Dan took.out ACC Dana from-tools.her
'Dan made Dana lose her temper.'

It is an "empty slot" MWE, with an idiomatic PP complement with a possessed NP whose possessor is obligatorily co-indexed with the literal NP complement filling the "slot". In what follows we use this example to illustrate our approach to the analysis of verbal MWEs.

4.2.1 Verbs with dual instantiations and their selectional restrictions

Verbs which can head VP MWEs can also occur in "standard" VP constructions. The degree of overlap between the behavior of the verb in its standard guise and in its idiomatic role is mostly verb-specific. Nevertheless, regardless of the degree, our lexical inheritance hierarchy enables us to distinguish between shared properties and those which differ in the two instantiations.

The subcategorization properties of the literal instantiation of *hoci* ('*take.out*') are expressed in its VALENCE (see Figure 1), which includes the three relevant arguments: DEP1 (*Actor*), DEP2 (*Theme*) and DEP5 (*Source*) (the rest are suppressed for space reasons). Moreover, the value of its R(EALIZATION)-FRAME is *arg12-125*, indicating that while the *Actor* and *Theme* arguments are obligatory, the *Source* argument is optional. These characteristics are all a result of the fact that the literal instantiation is an instance of the type *arg12-125_n_p_past_le* (for further elaboration, see (16) in section 4.1).

The idiomatic instantiation of *hoci* ('*take.out*') is an instance of a distinct, yet very similar type, *arg125_n_pi_xarg25_past_le*. Its syntactic selection properties are identical to its literal counterpart. However, in contrast to the literal *hoci* ('*take.out*'), the idiomatic one has a different R(EALIZATION)-FRAME value,

$$\left[\begin{array}{ll}
\multicolumn{2}{l}{\textit{arg12-125\_n\_p\_past\_le}}\\
\multicolumn{2}{l}{\text{STEM}\left\langle \text{“}\textit{hoci}\text{”}\right\rangle\ \textbf{‘took out’}}\\
\text{..CAT} \mid \text{VAL} & \left[\begin{array}{ll}
\multicolumn{2}{l}{\text{R-FRAME } \textit{arg12-125}}\\
\text{DEP1..} & \left[\begin{array}{ll}
\multicolumn{2}{l}{\text{CAT} \mid \text{HEAD } \textit{noun}}\\
\text{CONT} \mid \text{HOOK} & \left[\begin{array}{l}\text{INDEX } \boxed{1}\\ \text{TOPREL} \mid \text{PRED } \textit{l-rel}\end{array}\right]
\end{array}\right]\\
\text{DEP2..} & \left[\begin{array}{ll}
\multicolumn{2}{l}{\text{CAT} \mid \text{HEAD } \textit{noun}}\\
\text{CONT} \mid \text{HOOK} & \left[\begin{array}{l}\text{INDEX } \boxed{2}\\ \text{TOPREL} \mid \text{PRED } \textit{l-rel}\end{array}\right]
\end{array}\right]\\
\text{DEP5..} & \left[\begin{array}{ll}
\multicolumn{2}{l}{\text{CAT} \mid \text{HEAD } \textit{adp}}\\
\text{CONT} \mid \text{HOOK} & \left[\begin{array}{l}\text{INDEX } \boxed{5}\\ \text{TOPREL} \mid \text{PRED } \textit{l-rel}\end{array}\right]
\end{array}\right]\\
\multicolumn{2}{l}{\text{PPSORT} \mid \text{DEP5-P } \textit{\_from\_p\_rel}}
\end{array}\right]\\
\text{..CONT} \mid \text{HOOK} \mid \text{TOPREL} & \left[\begin{array}{l}\textit{\_take-out\_v\_rel}\\ \text{ARG1 } \boxed{1}\\ \text{ARG2 } \boxed{2}\\ \text{ARG5 } \boxed{5}\end{array}\right]
\end{array}\right]$$

Figure 1: The literal *hoci* (‘*take.out*’)

arg125, indicating that all arguments are obligatory (an abbreviated description is shown in Figure 5).

The main distinction between the two variants is in their semantic content, and semantic selection. In order to distinguish between literal and idiomatic words, and to control their distribution, semantic relations are divided into *l(iteral)-rel*s and *i(diomatic)-rel*s (Copestake, 1994; Sag et al., 2002; Kay & Sag, 2012, among others). Consequently, the semantic relation denoted by the literal verb, *_take-out_v_rel*, is a subtype of *l-rel*, and the one denoted by the idiomatic verb, *_i-take_out-cause_lose_v_rel*, is a subtype of *i-rel*.[4] The TOPREL feature is a pointer to the main semantic relation (in RELS) denoted by a lexeme (for more about TOPREL, see the following section).

Selectional restrictions of verbs in HeGram are specified in the respective DEP feature. The literal verb requires the *Source* (Arg5) PP to be headed by the specific preposition *me* (or *mi*). This requirement is defined in the PPSORT feature

[4] Please note that our analysis does not distinguish decomposable from non-decomposable idioms, as we only have a relatively superficial semantic representation of MWEs. All the idiomatic components of an MWE have separate idiomatic entries in the lexicon, which include an approximated paraphrase of their idiomatic meaning.

complex, under DEP5-P, whose value is set to *_from_p_rel*. Naturally, its idiomatic counterpart is more selective. It requires that its *Source* PP be headed by a specific idiomatic *_from_tools_ip_rel* relation. This selective preposition *me* (or *mi*), in turn, selects for an NP with an idiomatic *_i-tools-temper_n_rel* relation. This notwithstanding, the *Theme* argument of the idiomatic *hoci* (‘*take.out*’) is an “open slot” and can be filled by any NP complement, provided that it is not idiomatic (i.e., has an *l-rel*), and the same applies to the NP *Actor* in subject position.

4.2.2 A chain of lexical selection

Indirect non-local lexical selection such as the one described above, where a verb selects for a preposition which selects for a noun, forms a type of a chain, where heads of phrases select heads of other phrases. This mechanism is supported by the TOPREL feature, an independently motivated feature in HeGram, which identifies the main semantic relation denoted by a lexeme. Idiomatic selectors target this feature, which percolates from head daughter to the “mother” phrase.[5]

The AVM in Figure 2 illustrates the selection chain which characterizes the idiomatic form of the preposition *mi*, which is selected by the idiomatic *hoci* (‘*take.out*’). The co-indexation of XARG will become relevant in the next section.

$$\begin{bmatrix} \textit{poss-raise-adposition-lex-np-i} \\ \text{STEM}\left\langle \textit{“mi”} \right\rangle \\ \ldots \begin{bmatrix} \text{CONT}\begin{bmatrix}\text{TOPREL } \textit{\_from\_tools\_ip\_rel} \\ \text{HOOK} \mid \text{XARG } \boxed{1}\end{bmatrix} \\ \text{CAT} \mid \text{VAL} \mid \text{DEP2} \mid \ldots \begin{bmatrix}\text{XARG } \boxed{1} \\ \text{TOPREL} \mid \text{PRED } \textit{\_i-tools-temper}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

Figure 2: The idiomatic form of the preposition *mi*

Admittedly, using a selection chain to ensure that idiomatic verbs that select specific PPs only combine with the correct complements introduces some redundancy to the lexicon. However, this solution does solve the non-local selection problem.[6]

Semantic selection via the TOPREL of dependents is instrumental in accounting for cases of internal modification (e.g., (14) and (15)). The TOPREL of a phrase is identical to the main relation of the head, regardless of whether it is modified or not. This is illustrated in Figure 3.

[5] Kay & Sag (2012) suggest a similar feature, LEXICAL-ID (LID).

[6] Although there is no independent evidence for the existence of an idiomatic form of prepositions, usage patterns diverge: the one used in an MWE selects for a specific complement, whereas the standard preposition does not.

$$\left[\begin{array}{l} \textit{mrs} \\ \text{HOOK} \mid \text{TOPREL } \boxed{1} \\ \text{RELS} \left\langle \begin{bmatrix} \text{PRED} & \textit{\_pink\_j\_rel} \\ \text{LBL} & \boxed{3} \\ \text{ARG1} & \boxed{2} \end{bmatrix}, \boxed{1} \begin{bmatrix} \text{PRED} & \textit{\_i-belt-expenses\_n\_rel} \\ \text{LBL} & \boxed{3} \\ \text{ARG0} & \boxed{2} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \textit{\_def\_q\_rel} \\ \text{ARG0} & \boxed{2} \end{bmatrix} \right\rangle \end{array}\right]$$

Figure 3: The MRS of the idiomatic *the pink belt*

4.2.3 Possessive idioms

Possessive idioms present a second type of non-local selection. In such idioms the possessor of NP dependents, or NP complements of PP dependents, must be co-indexed with the verb's subject or complement (depending on the MWE). This requires that the index of the possessor be "visible" at the NP, and even PP level. The feature which projects the lower possessor to this higher level is the XARG feature (Kay & Sag, 2012; Bond et al., 2015).

The account of possessive idioms builds on our analysis of possessive nouns. Consider as an example the (literal) noun *keleiha* ('*her.tools*'), shown in Figure 4. The agreement property of this particular noun is 3rd-person-plural-masculine, and this is defined in its PNG feature, which is structure-shared with CNCRD (tagged $\boxed{2}$). Its possessor is realized by the 3rd-person-single-feminine pronominal clitic *ha*. This information is represented in the semantic XARG feature. Finally, the semantic relations denoted by the NP include *tool-rel*, which is the main relation (structure-shared with TOPREL), and *poss-rel*, which identifies the possessor ($\boxed{1}$) and possessed ($\boxed{3}$).

$$\left[\begin{array}{l} \textit{poss-cmn-3pm-3sf-noun-lex} \\ \text{STEM} \left\langle \textit{keleiha 'tools.her'} \right\rangle \\ \text{SYNSEM} \left[\text{LOCAL} \left[\begin{array}{ll} \text{CAT} \mid \text{HEAD} & \begin{bmatrix} \text{CNCRD } \boxed{2}\textit{png-3pm} \\ \text{CLT } \textit{poss-clt} \end{bmatrix} \\ \text{CONT} & \left[\begin{array}{ll} \text{HOOK} & \begin{bmatrix} \text{INDEX } \boxed{3} \begin{bmatrix}\text{PNG } \boxed{2}\end{bmatrix} \\ \text{TOPREL } \boxed{4} \\ \text{XARG } \boxed{1} \begin{bmatrix}\text{PNG } \textit{png-3sf}\end{bmatrix} \end{bmatrix} \\ \text{RELS} & \left\langle \boxed{4} \begin{bmatrix} \textit{tool-rel} \\ \text{ARG0 } \boxed{3} \end{bmatrix}, \begin{bmatrix} \textit{poss-rel} \\ \text{PSR } \boxed{1} \\ \text{PSD } \boxed{3} \end{bmatrix} \ldots \right\rangle \end{array}\right] \end{array}\right] \right] \end{array}\right]$$

Figure 4: A possessive noun

The XARG feature exposes the INDEX features of the “inner” possessor at the NP level, and thus makes it visible to an idiomatic selector. When a possessed NP is a complement of a preposition, the XARG features of its possessor percolate to the PP level. This is illustrated in the AVM describing the idiomatic form of the preposition *mi* in Figure 2.

Different idiomatic MWEs have different patterns of co-indexed possession, so the exact structure-sharing pattern is lexically specified per verb type.[7] In (7) the subject must be co-indexed with the possessor of the NP *Theme* complement (Arg2), while in (8) it must be co-indexed with the possessor of the NP complement inside the PP. In (10b) it is the NP complement which is co-indexed with the possessor of the NP complement inside the PP. Each one of these co-indexation relations between arguments is represented in the grammar by a lexical type, from which the relevant lexemes inherit. For example, the idiomatic *hoci* (‘*take.out*’) is an instance of a general lexical type *arg125_n_pi_xarg25_past_le*, which requires the co-indexation between the Arg2 complement and the possessor within the Arg5 argument.

The different components of the analysis of the MWE in the example sentence in (17) are shown together in Figure 5.

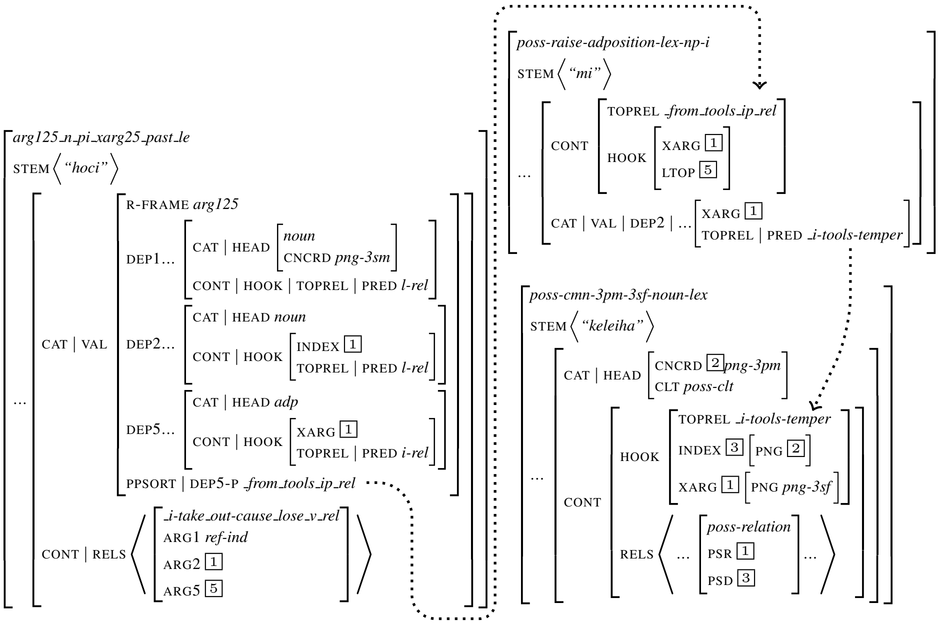


Figure 5: The selection chain in possessive idioms

[7] Bond et al. (2015) introduce an extra *identity* relation to the semantics of idiomatic verbs, which identifies the possessor and the index of the appropriate argument. This solution requires post-processing with MRS rewriting rules, which are not needed in our analysis.

5 Conclusion

We presented an account of Hebrew verbal MWEs in an existing HPSG grammar. The analysis covers a multitude of MWE types, including challenging phenomena such as (possessive) co-indexation and internal modification. Moreover, the grammar now produces two analyses for most MWEs, corresponding to their idiomatic and literal readings.

MWEs are challenging because they blur the traditional distinction between the lexicon and the grammar. In our analysis, support of MWEs required minimal changes to the grammar: most crucially, the division of *rels* to either *i-rels* or *l-rels*. All other changes involve the lexicon: we make extensive use of HPSG's type hierarchies in order to state generalizations over lexical types.

The main contribution of this work is of course the extension of the coverage of HeGram to verbal MWEs. To the best of our knowledge, this is the first account of Hebrew MWEs in a linguistically-motivated grammar. Moreover, the mechanisms that we advocate are fully applicable to other languages, and can be incorporated into existing HPSG grammars with minimal effort.

In the future we intend to explore syntactic constraints on MWEs and account for their full behavior. This includes phenomena such as topicalization, wh-questions, coordination, etc.

References

Arad Greshler, Tali, Livnat Herzig Sheinfux, Nurit Melnik & Shuly Wintner. 2015. Development of maximally reusable grammars: Parallel development of Hebrew and Arabic grammars. In Stefan Müller (ed.), *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar, Singapore*, 27–40. Stanford, CA: CSLI Publications.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources And Evaluation* 43(3). 209–226.

Bender, Emily M., Dan Flickinger & Stephan Oepen. 2002. The grammar matrix: an open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Coling-02 workshop on grammar engineering and evaluation*, 1–7. Morristown, NJ, USA: Association for Computational Linguistics. doi:http://dx.doi.org/10.3115/1118783.1118785.

Bond, Francis, Jia Qian Ho & Daniel Flickinger. 2015. Feeling our way to an analysis of English possessed idioms. In Stefan Müller (ed.), *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar, Singapore*, 61–75. Stanford, CA: CSLI Publications.

Copestake, Ann. 1994. Representing idioms. Paper presented at The 20th International Conference on HPSG, Copenhagen.

Copestake, Ann. 2002. *Implementing typed feature structure grammars*. Stanford: CSLI Publications.

Haugereid, Petter. 2012. A grammar design accommodating packed argument frame information on verbs. *International Journal of Asian Language Processing* 22(3). 87–106.

Herzig Sheinfux, Livnat, Nurit Melnik & Shuly Wintner. 2015. Representing argument structure in computational grammars. Submitted.

Kay, Paul & Ivan A. Sag. 2012. A lexical theory of phrasal idioms. Unpublished manuscript, Stanford University.

Pollard, Carl & Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press and CSLI Publications.

Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski & Marek Świdzib́ski. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international Conference on Language Resources and Evaluation, LREC 2014*, 2785–2792. Reykjavik, Iceland: ELRA. http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the third international conference on intelligent text processing and computational linguistics (cicling 2002)*, 1–15. Mexico City, Mexico.

‘Agreement mismatch’ between sort/kind/type and the determiner

Takafumi Maekawa
Ryukoku University

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 136–156

http://csli-publications.stanford.edu/HPSG/2015

Abstract

A singular countable noun in English normally needs a determiner and they should agree in number. However, there is a type of noun phrase, such as *these sort of skills*, which does not conform to this generalisation. As a singular countable common noun the noun *sort* requires a determiner, but there is an agreement mismatch here: *sort* is singular but the determiner is plural. Rather, the determiner agrees with the NP after the preposition *of*. There are several possible analyses that might be proposed, but the best analysis is the one in which *sort* and the preposition *of* are 'functors', non-heads selecting heads.

1 Introduction

A plural countable noun in English can stand on its own, without a determiner (1a).[1] A singular countable noun, however, normally needs a determiner in order to be grammatical. The noun *book* in (1b), which is a singular and countable common noun, requires a determiner to combine with, and the determiner *this* would satisfy this requirement.

(1) a. books
    b. *(this) book

Moreover, the determiner should agree in number with the head noun, as in (2).

(2) a. this book
    b. *these book

In (2b) the noun and the determiner do not agree in number. Thus, it might be possible to make a generalisation of the following sort.

(3) A singular countable noun in English requires a determiner and they should agree in number.

(1b) and (2b) do not conform to this generalisation.

Determiners are often assumed to be a specifier of a head noun in HPSG (Pollard & Sag 1994, Sag et al. 2003, Kim 2004, Kim & Sells 2008). In this

[†] I would like to thank the participants at HPSG 2015, especially Frank Van Eynde and Dan Flickinger, for their feedback and discussions. Thanks are also due to Bob Borsley for his valuable comments on the earlier version of this paper. I am also grateful to four anonymous reviewers for HPSG 2015 and the participants at the presentation given at the meeting of Rokko English Linguistics Circle held on 27th December 2014. Any shortcomings are my responsibility. This research was supported by the Japan Society for the Promotion of Science (Grant-in-Aid for Scientific Research (C) 26370466).

[1] Following Huddleston & Pullum (2002:355) we assume that the term 'determiner' refers to the following things: determinatives (*the tie*), determiner phrases (*almost every tie*), genitive NPs (*my tie*), plain NPs (*what colour tie*), PPs (*over thirty ties*).

assumption the partial lexical description for a singular countable noun is something like the following (cf. Sag et al. (2003:107), Kim (2004:1114), Kim & Sells (2008:108)).

(4)
$$\begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{noun} & \\ \text{AGR} & \boxed{1}\begin{bmatrix}\text{N} & \textit{sg}\end{bmatrix}\end{bmatrix} \\ \text{SPR} & \left\langle \begin{bmatrix}\text{AGR} & \boxed{1}\end{bmatrix} \right\rangle \end{bmatrix}$$

The value of the HEAD feature includes the AGR (AGREEMENT) feature. The value of the latter represents information about morpho-syntactic properties of the expression. The N (NUMBER) value represents the information about the grammatical number. (4) indicates that this word is morpho-syntactically singular. The SPR (SPECIFIER) feature shows that this expression has a specifier and indicates what kind of specifier it is. Thus, the determiner requirement of a countable singular noun is encoded as a matter of valency. The boxed tag $\boxed{1}$ in (4) means that the specifier has the same AGR value as the head noun, representing determiner-noun agreement. Overall, (4) states that a singular countable noun should have a specifier which agrees with it in number. Thus it can capture the generalisation stated in (3) and account for the unacceptability of (1b) *(*this*) *book* and (2b) **these book*: the former lacks a specifier and the latter does not show specifier-noun agreement.

Note that in (4) the determiner-noun agreement is represented on the basis of the SPR specifications of the head noun . This means that if the head noun is a singular countable noun not only the determiner requirement but also the determiner-noun agreement refers to the SPR specifications of the head noun.

However, there is a type of noun phrase in English which does not conform to this generalisation but is acceptable at least in an informal style.

(5) a. these *sort* of skills
b. those *kind* of pitch changes
c. these *type* of races (Keizer 2007:170)

These noun phrases contain a singular countable noun *sort*, *kind* and *type*, respectively. We will refer to them collectively as '*sort*-nouns'. In (5) the *sort*-noun is preceded by the plural determiner and followed by the preposition *of*, which in turn is followed by the plural noun. We will call these constructions in (5) as 'Plural Determiner plus *Sort*-Noun Construction (PDSNC)'.

The *sort*-noun in PDSNCs requires a determiner because it is a singular countable common noun. The only possible determiner that can satisfy this requirement is the one just before it (Hudson 2004:38). It should be noted that there is a sort of agreement mismatch here: the *sort*-noun is singular but the determiner is plural. Rather, the determiner agrees with the NP after the preposition *of*. It is clear that this is incompatible with the generalisation stated in (3) and described in (4).

The purpose of this paper is to investigate the syntactic properties of *sort*-nouns and PDSNCs, and consider how they might be analysed within the framework of Head-driven Phrase Structure Grammar (HPSG). We will argue that the *sort*-noun and the preposition *of* in PDSNCs are functors, non-heads selecting heads (Van Eynde 2006, Allegranza 1998).

The organisation of this paper is as follows. In section 2 we sketch some analyses which have been proposed for PDSNCs, and at the same time look at some data which are problematic for them. Sections 3 and 4 look at two possible analyses, both of which include important weaknesses. Section 5 presents the functor analysis and we look at how it is able to deal with the facts. In section 6 we also look at some further data which we argue is no problem to our approach. Section 7 is the conclusion.

2 Earlier Approaches

The PDSNCs have been discussed in many places, including studies from the viewpoint of meaning and function (Keizer 2007) and the diachronic development (Denison 2002, De Smedt et al. 2007, Davidse et al. 2008, Brems & Davidse 2010, Brems 2011). It seems that there are no fully worked out analyses of the synchronic syntactic properties of the constructions, but the above studies touch upon some of them.

Some suggest that the determiner, the *sort*-noun and *of* make a group, constituting a complex determiner (De Smedt et al. 2007, Davidse et al. 2008, Brems & Davidse 2010, Brems 2011). This is schematically represented as follows.

(6) [complex determiner: *these sort of*][head: *skills*]

However, there are at least two reasons for rejecting this view. First, it is possible to put an adjective before the *sort*-noun, as the following examples illustrate (see also Kim & Moon (2014:530)).

(7) a. these *steady-state* type of organisations (BYU-BNC[2]: CM0 W_commerce)
b. these *weird* sort of criticisms (COCA[3]: 2009 SPOK NPR_TellMore)
c. those *feminine* kind of things (COCA: 1991 FIC AntiochRev)
d. those *needy* sort of Americans (COCA: 1990 ACAD Raritan)

[2] Davies (2004–)

[3] Davies (2008–) The Corpus of Contemporary American English

The extra element between the determiner and the *sort*-noun makes the complex determiner analysis dubious.

Second, as pointed out by Denison (2002) and Keizer (2007), it is possible to delete the preposition *of* and the following NP.

(8) a. They won't last long, mate, **these type** never *do*.
(BYU-BNC; Keizer (2007:174))

b. But **these kind** *are* good for us. (COCA: 1995 NEWS Houston)

c. It was a game for the hardy, with talent and drive to spare, and **those sort** *were* precious few. (COCA: 2001 FIC Salmagundi)

These facts suggest that the preposition *of* does not make a complex with the determiner and the *sort*-noun. It seems, then, that the complex determiner approach is not satisfactory.

Others suppose that the *sort*-noun plays a role as a postdeterminer in PDSNCs (Denison 2002, Keizer 2007). Keizer (2007:175) provide the following structure for PDSNCs.

(9) [$_{\text{NP}}$ [$_{\text{Det}}$ those][[$_{\text{NomPostD}}$ sort][$_{\text{LE}}$ of][$_{\text{N}}$ things]]] (Keizer 2007:175)

Keizer (2007:175) assumes that a *sort*-noun is a nominal postdeterminer, which is NomPostD in (9), and preposition *of* is a linking element (LE), which is required when a postdeterminer is followed by another noun. [4] It is not difficult for this approach to accommodate the examples in (7) and (8): the *sort*-noun can have an adjectival modifier as in (7) because it is a nominal postdeterminer; and (8) is no problem because it is the case where the head noun is elided along with the linking element.

However, the postdeterminer approach is not without problems. The syntactic status of the postdeterminer position is not clear. For example, there is no consensus about what lexemes can occur in this position (Van de Velde 2011). For some, including Quirk et al. (1985:261), quantifiers and numerals are classified as postdeterminer, whereas for others adjectives like *other*, *same* or *usual* are postdeterminers (e.g. Sinclair (1990:70)). Moreover, there are some who do not assume a postdeterminer as an independent syntactic position (Huddleston & Pullum 2002), and others have explicitly argued against the idea of postdeterminers in the NP configuration (Van de Velde 2009).[5]

[4] Keizer (2007:175) states that the same linking element occurs in such expressions as *in front of*. The following examples illustrate that it cannot occur when it is not followed by an NP.

(i) I parked the car in front *(of) the building.

(ii) I parked the car in front (*of). (Keizer 2007:175)

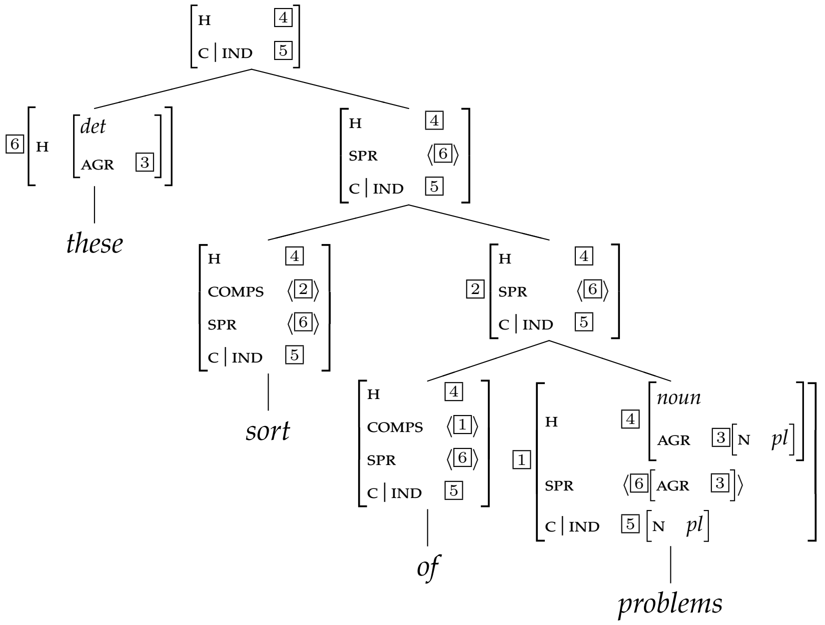
[5] Kim & Moon (2014:527ff) propose that the *sort*-noun and the preposition *of* make a complex word, which functions as a complex determiner. The examples in (8) are problematic to their analysis.

It seems, then, that both of the complex determiner approach and the postdeterminer approach contain some problems. In the rest of this article we will provide an analysis without such problems in the framework of HPSG. We will look at three possible HPSG analyses. Two of them appear to be unsatisfactory, but the third seems to give a satisfactory account of the facts.

3 Weak Head Analysis

We have argued that in PDSNCs the determiner agrees with NP after *of*. One might argue that this agreement pattern is possible if the *sort*-noun and the preposition *of* function as weak heads. A weak head is a lexical head which shares the HEAD (H) value and some other important properties with its complement (Tseng 2002, Abeillé et al. 2006). Both the *sort*-noun and the preposition *of* can be treated as weak heads. With this mechanism, we would have structures like (10).

(10)



As a weak head, the preposition *of* shares the SPR value with its complement, *problems*. It is propagated to the mother node, which is a complement of the *sort*-noun. The *sort*-noun then inherits the SPR value as a weak head. The value finally reaches the phrase *sort of problems*. This enables the combination of *these* and *sort of problems* because the latter inherits the SPR value from *problems*.

This analysis can handle the problems noted with the earlier approaches (Section 2). First, the determiner, the *sort*-noun and *of* do not make a complex determiner, so it is possible for an adjective to intervene between the determiner and the *sort*-noun, as in (7). Second, the examples in (8) can be accommodated if we assume that the complement of the *sort*-noun is optional. Finally, this analysis is free from the unclear notion of 'postdeterminer'.

It appears that the notion of weak head plays a role in explaining the pattern of agreement with the verb when a PDSNC is a subject. The following example shows that a PDSNC subject causes plural agreement with the verb.

(11) Well I'd actually expect that *those sort of courses* **are**/***is** very uh heavily subscribed uh, heavy just like *these sort of problems* **are**/***is** very hard to solve. (Keizer 2007: 175; adapted from ICE-GB)

In (11) subject-verb agreement is triggered by *courses* and *problems*, respectively. This means that the grammatical number of the full NP is determined by the grammatical number of the NP which is the complement of *of*. To capture this, let us assume that a weak head preserves the INDEX (IND) value of its complement on the mother node. In (10) the preposition *of* preserves the *plural* value of the IND feature of its complement on the mother node. That value is further preserved by another weak head, *sort*, on the full NP. The IND feature represents what the expression refers to in the real world, and its value determines the form of subject-verb agreement (Kathol 1999, Wechsler & Zlatić 2003, Kim 2004). In (10) the value of the IND feature that is propagated to the full NP is [N *pl*]], which indicates that the expression with this property is semantically singular.[6] The propagation of the IND value described above ensures that the *plural* value of the IND | N feature is propagated to the full NP node from the complement of *of*.

Thus, the weak head analysis outlined above appears to be able to deal with determiner-noun agreement and subject-verb agreement that PDSNCs show. However, there is an objection to this analysis. (10) shows that the IND | N value of the *sort*-noun is identical to that of the complement of *of*. This entails that the *sort*-noun and the complement of *of* are semantically plural. There is evidence against this view.

(12) a. [This kind of dog] is dangerous.
b. [These kind of dogs] are dangerous.
c. [These kinds of dogs] are dangerous.
(Huddleston & Pullum 2002:352)

Huddleston & Pullum (2002:353) states that '[t]he meaning of the bracketed NP in [(12b)] is like that of the one in [(12a)] in that we have a single kind of dog, not a plurality, as in [(12c)]'. Following this statement, we can assume that the *sort*-noun in PDSNCs has a singular interpretation. It is clear that

[6] The N (NUMBER) value represents the information about the grammatical number (Section 1).

the singular interpretation of the *sort*-noun is not compatible with the weak head analysis outlined above, which requires it to have a plural interpretation.

It seems, then, that the weak head analysis is unsatisfactory.

4 Transparent Head Analysis

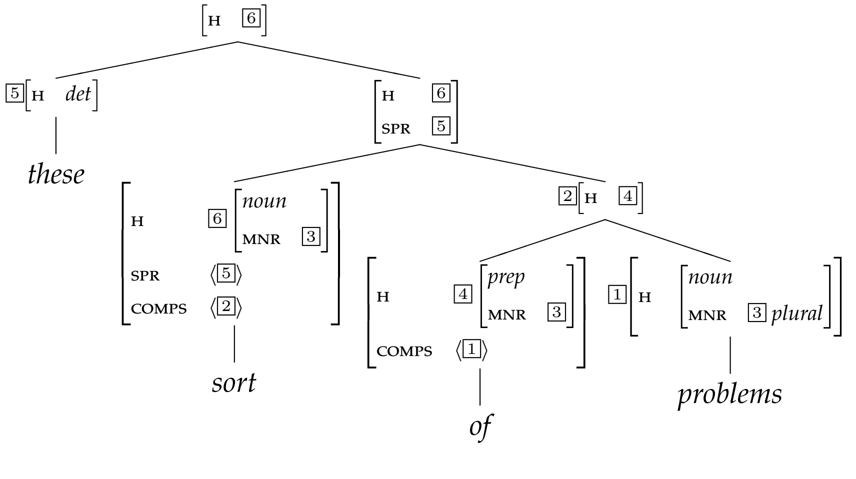
One might employ 'transparent heads' to allow the propagation of information from non-heads to phrases. Flickinger (2008) observes that in partitive NPs as in (13), where the partitive head *some* takes as its complement a PP headed by *of*, the grammatical number of the full NP is determined by the grammatical number of the complement NP of *of*.

(13) a. Some of the rice is ruined.
b. Some of the books are ruined.
c. *Some of the rice are ruined.
d. *Some of the books is ruined. (Flickinger 2008:90)

Flickinger (2008) introduces the MINOR feature as a HEAD feature so that a head selecting for a complement can preserve some properties of the complement on the phrase. The transparent head *of* in (13) identifies its MINOR value with that of their complement. The value is then propagated to the mother by the head feature principle. As another transparent head, the partitive head *some* also preserves the MINOR value of its complement and propagates it to the mother. If we assume that the number property is represented as a MINOR value, it can propagate up from the lower non-head and can be visible on the full partitive NP. The MINOR value of the full partitive NP then determines the form of subject-verb agreement.

A transparent head approach to PDSNCs would require that the *sort*-noun and *of* should identify their MINOR value with that of their respective complement. With this assumption, we will have structures like (14).

(14)


[H 6]
5[H det]
these
[H 6, SPR 5]
[H 6[noun, MNR 3], SPR ⟨5⟩, COMPS ⟨2⟩]
sort
2[H 4]
[H 4[prep, MNR 3], COMPS ⟨1⟩]
of
1[H [noun, MNR 3 plural]]
problems

As in partitive NPs, *of* in (14) identifies its MINOR value with that of their complement. The value is propagated to the mother by the head feature principle. The *sort*-noun then inherits that MINOR value from its complement and passes it up to the mother. That value is again propagated to the full NP by the head feature principle. The grammatical number of the full NP is thus determined by the MNR value propagated from the complement of *of*.

This analysis can avoid the problems noted with the earlier approaches (Section 2), as can the weak head approach outlined in the last section: it is possible for an adjective to intervene between the determiner and the *sort*-noun, as in (7); it is easy to make the complement of the *sort*-noun optional as in (8); and this analysis do not employ a postdeterminer as a syntactic position.

However, the objection that we raised against the weak head approach is also applicable here. (14) shows that the MNR value of the *sort*-noun is identical to that of the complement of *of*. This means that the grammatical number of the *sort*-noun is the same as that of the complement of *of*: they are both plural. This is incompatible with the fact that the *sort*-noun in PDSNCs has a singular interpretation (12).

It seems, then, that the transparent head analysis too is unsatisfactory.

5 Functor Analysis

We will turn now to an analysis which we think provides a satisfactory account of the data. This is an analysis in which the determiner, the *sort*-noun and the preposition *of* are functors: non-heads which select the head (Van Eynde 2006, Allegranza 1998).[7]

5.1 Functors

We assume that a singular determiner *this* has a partial lexical description like the following.

(15) *this*:

$$\begin{bmatrix} \textsc{head} & \begin{bmatrix} \textit{determiner} & \\ \textsc{agr} & \boxed{1}\begin{bmatrix}\textsc{n} & \textit{sg}\end{bmatrix} \\ \textsc{sel} & \begin{bmatrix}\textsc{head} & \begin{bmatrix}\textit{noun} & \\ \textsc{agr} & \boxed{1}\end{bmatrix}\end{bmatrix}\end{bmatrix} \\ \textsc{mrk} & \textit{marked}\end{bmatrix}$$

[7] The analysis provided in this section is partly based on the ideas given in Maekawa (2010).

Those non-heads that select the heads are called functors. The information about selection is indicated by the SEL (SELECT) feature of a non-head, and it represents the constraints which the non-head daughter imposes on the head daughter. The SEL value of (15) shows that *this* selects a singular noun. The AGR value $\boxed{1}$ shared between *this* and its head noun means determiner-noun agreement between them. MARKING (MKG) indicates whether the expression involves a determiner or a numeral, or whether it can stand alone without these elements (Van Eynde 2006). The *marked* value means that the expression contains a determiner or is a determiner itself.

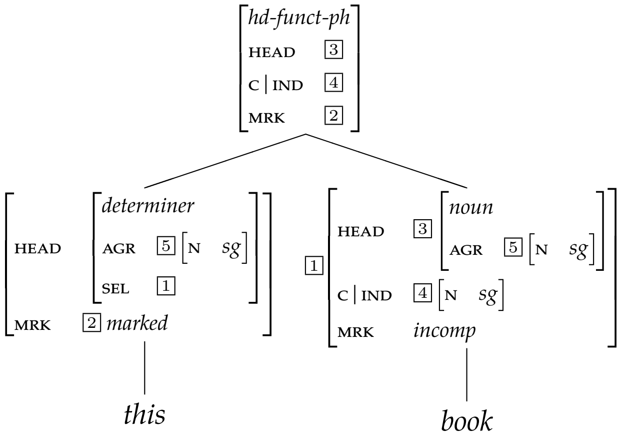
The combination of a determiner and a head nominal is an instance of a head-functor phrase, which is subject to the following constraint (Van Eynde 2006:164,166).

(16)
$$\textit{hd-funct-ph} \rightarrow \begin{bmatrix} \text{MRK} & \boxed{1} \\ \text{DTRS} & \left\langle \begin{bmatrix} \text{MRK} & \boxed{1} \\ \text{SEL} & \boxed{2} \end{bmatrix}, \boxed{3}\begin{bmatrix} \text{SYNSEM} & \boxed{2} \end{bmatrix} \right\rangle \\ \text{H-DTR} & \boxed{3} \end{bmatrix}$$

The constraint in (16) states that in a phrase of type *head-functor-phrase* (*hd-funct-ph*) the non-head daughter selects the head daughter, and the MRK value of the mother is token-identical to that of the non-head daughter.

Let us see how functor *this* combines with a singular countable noun.

(17)



The MKG feature of *book* has a value whose type is *incomplete* (*incomp*), which means that the word is incomplete on its own, requiring some sort of determiner. In (17) both the IND | N and the AGR | N values of *book* are *sg*, indicating that it is a singular nominal. The combination shown in (17) is an instance of a head-functor phrase. In (17) *this* selects the head noun and the MRK value *marked* is inherited to the mother node. We assume that the IND value, as well as the HEAD value, is propagated from the head daughter to the mother node (Sag et al. 2003:144).[8]

In this approach generalisation (3) is captured in terms of two separate specifications: the determiner requirement of a singular countable noun is represented by the *incomp* value of the MRK feature of the head nominal, whereas the determiner-noun agreement is represented by the shared value of the AGR | N feature between the determiner and the head noun. This is in clear contrast with the standard HPSG treatment given in (4), where the determiner requirement and the determiner-noun agreement both depend on the SPR specifications of the head noun.

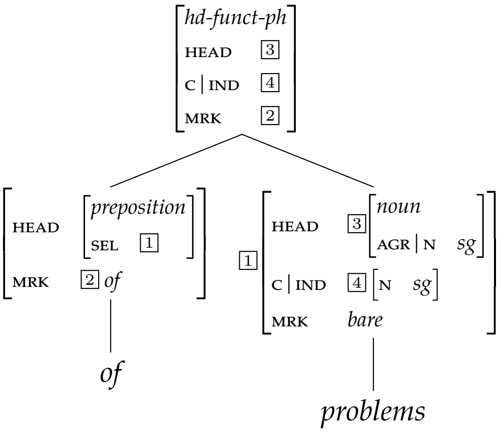
Finally, we assume that the preposition *of* is a functor (Van Eynde 2005) and has something like the following partial lexical description .

(18) *of* (functor):

$$\begin{bmatrix} \textsc{head} & \begin{bmatrix} \textit{preposition} & \\ \textsc{sel} & \begin{bmatrix}\textsc{head} & \textit{noun}\end{bmatrix} \end{bmatrix} \\ \textsc{mrk} & \textit{of} & \end{bmatrix}$$

The SEL value of (18) states that this preposition selects a head-daughter which is a nominal. Let us consider how functor *of* combines with the head nominal.

(19)



The combination of the preposition *of* and *problems* is an instance of a head-functor phrase, in which the functor *of* selects the head nominal.[9] The MRK

[8] The propagation of the HEAD and IND values is due to the constraint on phrases of type *headed-phrase* (*hd-ph*), which is a supertype of *hd-funct-ph*. This is also a supertype of *head-complement-phrase*, which we will see later.

[9] The resulting expression is an NP. A piece of evidence that the functor *of* and the head nominal make an NP comes from Dutch. Dutch has constructions similar to PDSNCs, but they are different from the English counterparts in lacking an intermediating preposition between the *sort*-noun and its complement.

(i) dit/dat soort auto/auto's
this/that kind car/cars

value of *of* is inherited to the mother node. See Van Eynde (2000, 2004, 2005) for analyses of some prepositions as functors.

5.2 PDSNCs

We will finally turn to the functor analysis of PDSNCs. We will first discuss what is the head of the PDSNCs. Let us consider (11), which is repeated in the following.

(20) Well I'd actually expect that *those sort of courses* **are**/***is** very uh heavily subscribed uh, heavy just like *these sort of problems* **are**/***is** very hard to solve. [= (11)]

Here, the PDSNC subjects *those sort of courses* and *these sort of problems* show plural agreement with the verb. The agreement triggers are the nouns following *of*: *courses* and *problems*, respectively. Let us assume, then, that the noun following *of* is the head of the whole structure of PDSNCs .

Given the above discussions about the headedness of the PDSNCs, we can say that the *sort*-noun does not function as the head. Instead, we can propose that the *sort*-noun in PDSNCs is a functor, selecting the *of*-marked NP head-daughter. The partial lexical description of a functor *sort*-noun will look like the following.

(21) *sort* (functor):

$$\begin{bmatrix} \textsc{head} & \begin{bmatrix} \textit{noun} & \\ \textsc{agr}\,|\,\textsc{n} & \textit{sg} \\ \textsc{sel} & \left\langle \begin{bmatrix}\textsc{mrk} & \textit{of}\end{bmatrix} \right\rangle \end{bmatrix} \\ \textsc{mrk} & \textit{incomp} \\ \textsc{c} & \begin{bmatrix}\textsc{ind}\,|\,\textsc{n} & \textit{sg}\end{bmatrix} \end{bmatrix}$$

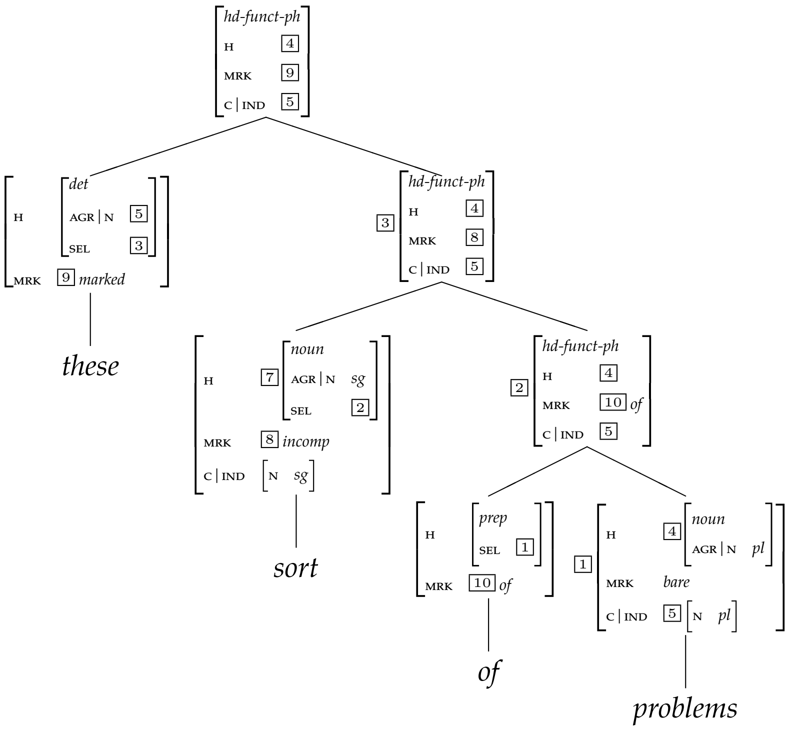
(21) states that the functor *sort*-noun selects an *of*-marked head-daughter. Note that the determiner requirement of a *sort*-noun as a singular countable noun is indicated by the *incomp* value of the MRK feature.

Our syntactic analysis of a PDSNC is given in (22).

'this/that kind of car/cars' (Broekhuis & den Dikken 2012:631)

Here, *soort* is a Dutch *sort*-noun, and it is directly followed by the bare nominal *auto/auto's* 'car/cars'. Thus, we can say that a *sort*-noun selects an NP in both English and Dutch.

(22)



We have already seen above how the *of*-phrase is constructed, so we will not discuss it here. The *sort*-noun in this construction is a functor with the property in (21). As a functor, it selects the *of*-marked phrase via the SEL value $\boxed{2}$. In this head-functor phrase the *sort*-noun is a non-head daughter, and the head-daughter is *of problems*. The HEAD and C | IND values of the mother node come from the head daughter. The *pl* value of AGR | N, which is propagated from *problems* via the HEAD feature, enables this phrase to combine with the plural determiner *these*. The combination of the determiner with the head nominal is an instance of a head-functor phrase, as discussed in section 5.1. Therefore, the MRK value *marked* is inherited from *these* to *these sort of problems*.

The AGR | N and IND values of the top node come from *sort of problems*. Because these values originally come from *problems*, the whole phrase is plural both morpho-syntactically and semantically. The semantic plurality accounts for the plural agreement with the verb, illustrated in (11). The morpho-syntactic plurality accounts for the plural agreement with the determiner.

It is important to note here that the determiner requirement from the *sort*-noun as a singular countable noun is fully satisfied in (22). It is the plural determiner that satisfies this requirement. Agreement mismatch does not occur here because the determiner and the *sort*-noun do not have a determiner-head relationship. The head of the whole structure is the plural noun *problems*, with which the determiner has an agreement relationship via the AGR | N feature. This analysis is possible because the determiner requirement and the determiner-noun agreement are represented separately in our approach.

This approach can capture the facts in (7) and (8), which, as discussed in section 2, are problematic to the earlier analyses of PDSNCs. The relevant parts of (7) and (8) are repeated in (23) and (24), respectively.

(23) a. these *steady-state* type of organisations
b. these *weird* sort of criticisms
c. those *feminine* kind of things
d. those *needy* sort of Americans

(24) a. (...), *these type* never do.
b. But *these kind* are good for us.
c. (...), and *those sort* were precious few.

First, the determiner, the *sort*-noun and *of* do not make a complex determiner in our approach, so it enables an adjective to intervene between the determiner and the *sort*-noun, as in (23). Second, the preposition *of* and the following noun make a constituent, which makes it easy to delete it, as in (24). Finally, our analysis is free from the unclear notion of 'postdeterminer'.

Moreover, our functor analysis is free from the problems involved in the other HPSG analyses which we discussed in the last two sections. The number mismatch between the *sort*-noun and the head noun do not occur in our analysis because the *sort*-noun do not preserve the grammatical number of the head noun.

It seems, then, that our functor analysis is superior to the other analyses which we discussed.

6 Other Variations

In this section we will look at constructions which look like PDSNCs but are actually not. The functor analysis of *sort*-noun can be applied to some of these constructions. We will first consider the variants in which the *sort*-noun works as a head of the whole construction.

6.1 *Sort*-Noun as a Head

PDSNCs are 'very informal and is considered incorrect by some people' (*OALD*).[10] According to Huddleston & Pullum (2002:353), however, they are 'very well established, and can certainly be regarded as acceptable in informal style' . They are in contrast with the less informal variants, which are often found in dictionaries. Some of them are illustrated in the following.

[10] http://www.oxfordlearnersdictionaries.com/definition/english/kind_1

(25) a. *This kind of question* often appears in the exam.

b. *These kinds of questions* often appear in the exam. (*OALD*: *ibid*)

These variants, like PDSNCs, include a determiner, a *sort*-noun and an *of*-phrase. However, the *sort*-noun in these constructions agrees in number with the preceding determiner, in contrast with PDSNCs where the determiner and the *sort*-noun do not show number agreement.

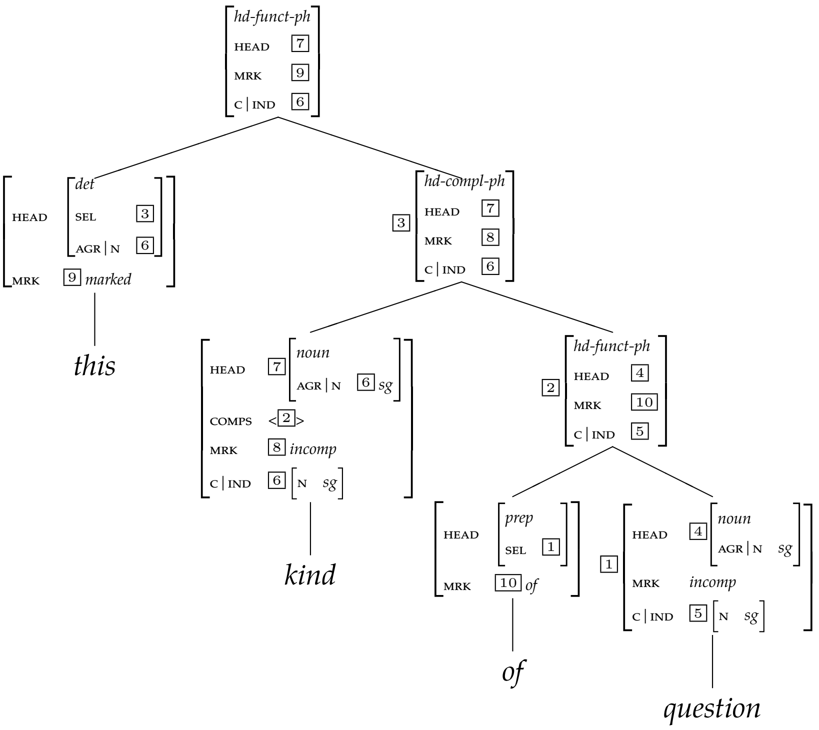
The following example show that when these constructions are subjects, number agreement with the verb is induced by the number of the *sort*-noun.

(26) *These sorts of behaviour* are not acceptable. (*OALD*: *ibid*)

In these examples the noun after *of* is an uncountable noun, which is always singular. (26), in which there is plural subject-verb agreement, shows that the *sort*-noun, not the noun after *of*, is the trigger of subject-verb agreement.

Now let us consider how these examples are analysed in HPSG. The structure for (25a) is given in (27).

(27)



The *sort*-noun *kind* in (27) is a head, not a functor. As a singular countable noun, the AGR | N and the IND | N values are *sg*. The MRK value is *incomplete* (abbreviated as *incomp* here) as it needs a determiner in order to occur in NP positions. The COMPS list of *sort*-noun in (27) indicates that it takes a complement marked with *of*. The combination of *kind* and *of question* is a structure of a head-complement phrase (which is of type *head-complement-phrase* (*hd-compl-ph*)). Because it is a subtype of *hd-ph*, the AGR | N value *sg* is inherited from *kind* to the mother node, which enables this phrase to combine with the singular determiner *this*. The IND value is also inherited from the head-daughter to the mother node, so the *sg* value reaches the top node. This makes the whole phrase semantically singular, which leads to the singular agreement with the verb when the phrase is in the subject position, as illustrated by (25a). Thus, the forms of determiner-noun agreement and subject-verb agreement are both determined by the properties of the head noun *kind*. Therefore, the form of *question* is irrelevant for the both types of agreement.

In (25b) and (26), the head of the whole structure is the plural nouns *kinds* and *sorts*, respectively. Their partial lexical description is something like the following.

(28) *sorts/kinds*:

$$\left[\begin{array}{ll} \text{HEAD} & \left[\begin{array}{ll} \multicolumn{2}{l}{\textit{noun}} \\ \text{AGR}\,|\,\text{N} & \textit{pl} \end{array}\right] \\ \text{COMPS} & \langle \left[\text{MRK} \;\; \textit{of}\right] \rangle \\ \text{MRK} & \textit{bare} \\ \text{C} & \left[\text{IND}\,|\,\text{N} \;\; \textit{pl}\right] \end{array}\right]$$

(28), which is a partial lexical description of the plural common noun *sorts*, is the same as that of a singular *sort*-noun, except for the AGR | N, IND | N and MRK values. The former two are *pl*. The MRK value is *bare*, which indicates that *sorts* does not have to have a determiner to be used in NP positions. The forms of determiner-noun agreement and the subject-verb agreement are determined by the AGR | N and the IND | N values of *kinds/sorts*, respectively. In this structure they are both *pl*, indicating that both types of agreement should be in plural, as shown by (25b) and (26). The form of *questions/behaviour* is irrelevant for the purpose of agreement.

6.2 Variants with Agreement Ambiguity

There is a variant in which the nominal after *of* is the only plural element in the phrase.

(29) a. this type of promoters (BYU-BNC: FTE W_ac_nat_science)
b. this kind of activities (COCA: 1992 SPOK NPR_Weekend)
c. this sort of things (COCA: 1999 MAG Money)

The structure in (27) also accommodates the variant in (29).[11,12] In these examples the right-most noun is plural, but as discussed above, it is irrele-

[11] The MRK value of the noun following *of* is *bare* in these cases.

[12] The following example is a supportive evidence for this claim.

vant for the both types of agreement because it is not the head. The head is the singular *sort*-noun, so it triggers singular agreement not just with the determiner but also with the verb. The following examples illustrate this.

(30) a. (...) **this type of promoters** is more frequent in B.subtilis than in E.coli (11). (BYU-BNC: FTE W_ac_nat_science)

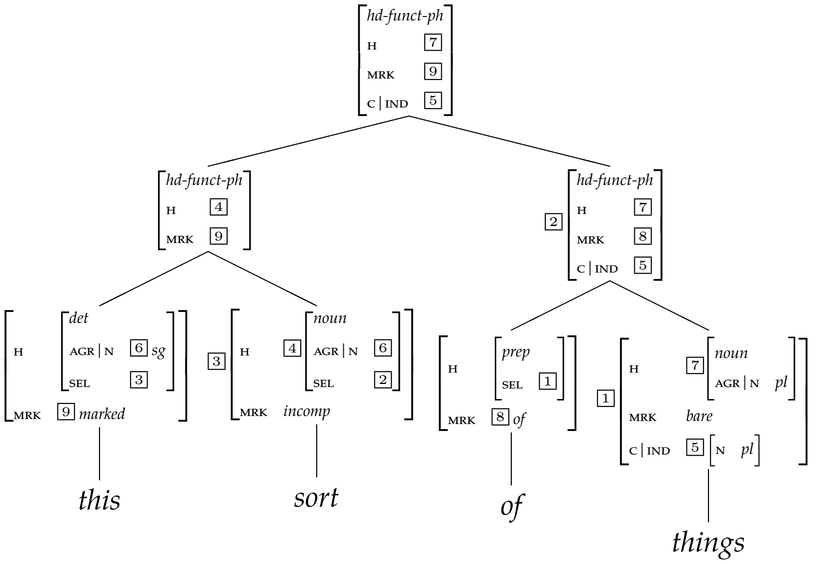
b. **this kind of activities** is one of the most important for our bank. (COCA: 1992 SPOK NPR_Weekend)

c. "**This sort of things** happens all the time," Bradley says, (...) (COCA: 1999 MAG Money)

In the examples in (30) the singular *sort*-noun triggers singular agreement with the determiner and the verb.

An interesting point about the functor analysis of *sort*-nouns given in (21) is that it also allows the following structure, in which the combination of the determiner and the *sort*-noun acts as a complex functor, selecting the *of* phrase.

(31)



In (31) the determiner selects *sort*. It should be singular because its head is [AGR | N *sg*]. The SEL value of *sort* is inherited to the mother node because it

(i) This kind of questions and sort of answers are/*is helpful.

In our approach this can be analysed as a case of N-bar coordination.

(ii) this [$_{N'}$ kind of questions] and [$_{N'}$ sort of answers]

Determiner-noun agreement and subject-verb agreement in (12) have exactly the same patterns as the clear case of N-bar coordination such as the following.

(iii) This boy and girl are/*is eating a pizza (King & Dalrymple 2004:70)

Thus, we can conclude that the NPs in (29) have structures like (27). I am grateful to Dan Flickinger for bringing this point to the my attention.

is a HEAD feature. Like PDSNCs, the head of the whole phrase is the head-daughter of the *of* phrase. If it is a plural NP, then the whole phrase is plural. This accounts for plural agreement with the verb.

(32) a. **This kind of rankings** have given ammunition to conservatives (...) (COCA: 2001 NEWS CSMonitor)

b. (...) **this type of women** like to be around rich and powerful men. (COCA: 2008 SPOK Fox_Gibson)

Now, note that this structure generates the same sequence as (29), i.e., singular D + singular *sort*-noun + of + plural N. The examples are repeated here.

(33) a. this type of promoters

b. this kind of activities

c. this sort of things [= (29)]

Recall that our analysis of (33) assumed that the singular *sort*-noun was the head of the whole phrase, and it was responsible for the singular agreement both with the determiner and the verb, as in (30). Thus, our dual treatment of a *sort*-noun, as a head and a functor, accounts for the fact that the variant in (33) triggers both singular agreement (30) and plural agreement (32) with the verb.

The dual patterns of subject-verb agreement can be seen in the following pair as well, where the determiner is *one*.

(34) a. My dear child, there is only ***one*** *kind of canals* that excites imagination. (COCA: 1999 FIC MassachRev)

b. ***One*** *kind of policies* are the missions (...) (http://middleburycampus.com/article/1-in-8700-glenn-lower/)

In (34a) the *sort*-noun is the head, triggering singular agreement with the underlined elements. In (34b) *kind* is a functor and the head is *policies*, which accounts for the plural agreement with the verb.

7 Conclusion

This study started with the observation about singular countable nouns, and we made a tentative generalisation in (3), which is repeated here.

(35) A singular countable noun in English requires a determiner and they should agree in number. [= (3)]

However, a *sort*-noun in PDSNCs does not seem to conform to this generalisation: it is a singular countable noun requiring a determiner, but the determiner satisfying this requirement is not in the agreement relation with it. The determiner agrees with the NP following *of*. We claimed that a *sort*-noun in PDSNCs is a functor, a non-head selecting a head. We argued that the functor treatment of *sort*-nouns can provide a satisfactory account of the PDSNC data. We also suggested that the dual patterns of subject-verb agreement which one of the variants shows (e.g. *this sort of things*), observed in (30) and (32), can be accounted for by assuming that a *sort*-noun is ambiguous: it can be either a head of a full NP or a functor (21).

In HPSG it has been assumed that a determiner is a specifier of a head noun and the determiner-noun agreement is based on the SPR specifications of the head noun. In our analysis, however, the determiner-noun agreement is not based on the SPR specifications: it is dissociated from the determiner requirement of a singular countable noun. This enables the plural determiner to satisfy the determiner requirement of a singular *sort*-noun while agreeing with the head of the whole structure.

References

Abeillé, Anne, Olivier Bonami, Danièle Godard & Jesse Tseng. 2006. The syntax of French *à* and *de*: An HPSG analysis. In Patrick Saint-Dizier (ed.), *Syntax and Semantics of Prepositions* (Text, Speech and Language Technology 29), 147–162. Springer Verlag.

Allegranza, Valerio. 1998. Determiners as functors: NP structure in Italian. In Sergio Balari & Luca Dini (eds.), *Romance in HPSG*, 55–108. Stanford: CSLI Publications.

Brems, Lieselotte. 2011. *Layering of Size and Type Noun Constructions in English*. Berlin: Mouton de Gruyter.

Brems, Lieselotte & Kristin Davidse. 2010. The grammaticalization of nominal type noun constructions with *kind/sort of*: chronology and paths of change. *English Studies* 91. 180–202.

Broekhuis, Hans & Marcel den Dikken. 2012. *Syntax of Dutch: Nouns and Noun Phrases*, vol. 2. Amsterdam: Amsterdam University Press.

Davidse, Kristin, Lieselotte Brems & Liesbeth De Smedt. 2008. Type noun uses in the English NP: a case of right to left layering. *International Journal of Corpus Linguistics* 13. 139–168.

Davies, Mark. 2004–. BYU-BNC (Based on the British National Corpus from Oxford University Press). `http://corpus.byu.edu/bnc/`.

Davies, Mark. 2008–. The Corpus of Contemporary American English: 450 Million Words, 1990-Present. `http://corpus.byu.edu/coca/`.

De Smedt, Liesbeth, Lieselotte Brems & Kristin Davidse. 2007. NP-internal functions and extended uses of the 'type' nouns *kind*, *sort* and *type*: towards a comprehensive, corpus-based description. In Roberta Facchinetti (ed.), *Corpus Linguistics 25 Years On*, 225–255. Amsterdam: Rodopi.

Denison, David. 2002. History of the *sort of* construction family. Paper presented at ICCG2: Second International Conference on Construction Grammar, Helsinki.

Flickinger, Dan. 2008. Transparent heads. In Stefan Müller (ed.), *The Proceedings of the 15th International Conference on Head-Driven Phrase Structure Grammar*, 87–94. Stanford: CSLI Publications. `http://cslipublications.stanford.edu/HPSG/9/`.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hudson, Richard. 2004. Are determiners head? *Functions of Language* 11(1). 7–24.

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

Keizer, Evelien. 2007. *The English Noun Phrase*. Cambridge: Cambridge University Press.

Kim, Jong-Bok. 2004. Hybrid agreement in English. *Linguistics* 42(6). 1105–1128.

Kim, Jong-Bok & Grace Ge-soon Moon. 2014. The SKT construction in English: A corpus-based perspective. *Linguistic Research* 31(3). 519–539.

Kim, Jong-Bok & Peter Sells. 2008. *English Syntax: An Introduction*. Stanford: CSLI Publications.

King, Tracy Holloway & Mary Dalrymple. 2004. Determiner agreement and noun conjunction. *Journal of Linguistics* 40. 69–104.

Maekawa, Takafumi. 2010. Dependency structure vs. phrase structure: two analyses of English determiners. *Hokusei Review, Junior College* 8. 51–62.

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

Quirk, Randolph, Sydney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

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

Sinclair, John (ed.). 1990. *The Collins COBUILD English Grammar*. London: Collins.

Tseng, Jesse L. 2002. Remarks on marking. In Frank Van Eynde, Lars Hellan & Dorothee Beermann (eds.), *Proceedings of the 8th International Conference on Head-driven Phrase Structure Grammar*, 267–283. Stanford: CSLI Publications. http://cslipublications.stanford.edu/HPSG/2/.

Van Eynde, Frank. 2000. On the notion 'minor preposition'. In A. Kathol & D. Flickinger (eds.), *Proceedings of the 7th International Conference on Head-driven Phrase Structure Grammar*, 81–99. Stanford University: CSLI Publications. http://web.stanford.edu/group/cslipublications/cslipublications/HPSG/2000/.

Van Eynde, Frank. 2004. Minor adpositions in Dutch. *The Journal of Comparative Germanic Linguistics* 7. 1–58.

Van Eynde, Frank. 2005. Minor prepositions in nominal projections. In Aline Villavicencio & Valia Kordoni (eds.), *Proceedings of the 2nd ACL-SIGSEM Workshop on Prepositions and their Use in Computational Linguistics Formalisms and Application*, 54–63. Colchester: Association for Computational Linguistics.

Van Eynde, Frank. 2006. NP-internal agreement and the structure of the noun phrase. *Journal of Linguistics* 42(1). 139–186.

Van de Velde, Freek. 2009. Do we need the category of postdeterminer in the NP? *Transactions of the Philological Society* 107(3). 293–321.

Van de Velde, Freek. 2011. Left-peripheral expansion of the English NP. *English Language and Linguistics* 15(2). 387–415.

Wechsler, Stephen & Larisa Zlatić. 2003. *The Many Faces of Agreement*. Stanford: CSLI Publications.

Ellipsis of SAY, THINK, and DO in Japanese subordinate clauses: A constructional analysis

David Y. Oshima
Nagoya University

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 157–176

http://csli-publications.stanford.edu/HPSG/2015

Abstract

This paper addresses some Japanese constructions where the predicate heading a subordinate clause – specifically, a suspensive form of IU ‘say’, OMOU ‘think’, or SURU ‘do’ – appears to be elided. I will discuss that these elliptic constructions are subject to certain syntactic and interpretative constraints which do not apply to their non-elliptic counterparts, and develop an SBCG-analysis that aims to model these constraints without postulating a covert element in the place of the missing verb.

1 Introduction

This paper discusses the Japanese constructions exemplified with (1a), (2a), and (3a), which appear to involve “omission” of the predicate heading a subordinate clause. The missing predicate can be “recovered” as a suspensive form (i.e., the gerund or infinitive form) of the lexemes: IU ‘say’, OMOU ‘think’, or SURU ‘do’, as in (1b,c), (2b,c), and (3b).[1]

(1) *SAY-ellipsis construction*
Ken-ga [“Ohayoo” to {a. ∅/b. itte/c. ii}] haitte kita.
K.-Nom good.morning Quot {∅/say.Ger/say.Inf} enter.Ger come.Pst
‘Ken came in, (saying) “Good morning”.’

(2) *THINK-ellipsis construction*
Ken-wa [“Masaka” to {a. ∅/b. omotte/c. omoi}] furikaetta.
K.-Top no.way Quot {∅/think.Ger/think.Inf} look.back.Pst
‘Ken looked back, (thinking to himself) “No way”.’

(3) *DO-ellipsis construction*
Ken-wa [akanboo-o se-ni {a. ∅/b. shite}] atari-o shibaraku
K.-Top baby-Acc back-Dat {∅/do.Ger} vicinity-Acc for.a.while
sansaku-shita.
stroll.Pst
‘Ken strolled around for a while, (carrying) the baby on his back.’

The existence of these constructions has long been acknowledged. Previous studies of the SAY- and THINK-ellipsis constructions, which I group as the QV-ellipsis construction (QV = quotative verb), include Fujita (2000), Oshima and Sano (2012), Oshima (2013), and Kim (2013). Previous studies of the DO-ellipsis construction include Muraki (1983), Teramura (1983), and Dubinsky and Hamano (2003).

[1] The abbreviations used in glosses are: Acc = accusative, Adv = adverb marker, Asp = aspectual auxiliary, Ben = benefactive auxiliary, Caus = causative, Dat = dative, DP = discourse particle, Gen = genitive, Ger = gerund, Inf = infinitive, Ipfv = imperfective auxiliary, Loc = location, Neg = negation, Nom = nominative, Plt = polite, Psv = passive, Prs = present, Pst = past, Quot = quotative particle, Top = topic.

The constructions in question do not involve the canonical kind of ellipsis, such as the English VP-ellipsis illustrated in (4), where (i) the missing element is semantically recovered with the aid of contextual cues, and (ii) the elliptic and non-elliptic versions are semantically equivalent.

(4) A: Has John left?
    B: No, he hasn't {left/∅}.

Rather, they are reminiscent of the English construction which Fillmore et al. (2012) refer to as the **`adjective-as-nominal.Human`** construction:

(5) **The rich** exploit **the poor**, and **the poor** exploit **the poorer**.

Even without contextual information, the "nounless" NPs in (5) can be interpreted as referring to humans. Furthermore, they are not semantically equivalent to their "headed" counterparts, in that they receive the generic interpretation; note that (5) is more properly paraphrased as "Rich people exploit poor people, ...", than as "*The* rich people exploit *the* poor people, ...".

My analysis to be proposed below is similar to the one proposed by Lyons (1991) for the nounless NP construction, which in spirit is "constructionalist", as well as to those proposed by Fillmore et al. (2012: 357–360) and Arnold and Spencer (2015 (this volume)), which are explicitly so. (6) illustrates the interpretative rule proposed by Lyons (1991).

(6) *Lyons' (1991) "Adjective Head Rule"* (with some adaptations)

a. The sequence of the form: [*the* + Adj.] may constitute a plural NP referring to humans.

b. If the adjective is [−nationality], then the NP obligatorily receives the generic interpretation. If the adjective is [+nationality], then the NP optionally receives the generic interpretation.

2 Background: Basic facts about the infinitive and gerund clause constructions

The suspensive clause construction (Susp-Cx), which subsumes the infinitive and gerund clause constructions (Inf-Cx and Ger-Cx), refers to a hypotactic structure where the subordinate clause is headed by a predicate in its infinitive form (*ren'yoo* form) or gerund form (*te*-form).

In the literature, the Susp-Cx has often been considered to semantically convey only the logical conjunction of the two component clauses, on a par with the English *and*-coordination structure (e.g., Fukushima 1999; Lee and Tonhauser 2010). This view, however, does not hold scrutiny; if the Inf-Cx and Ger-Cx merely represent logical conjunction, then (7b) is expected not to be pragmatically odd, like the English sentence provided to illustrate its intended interpretation.[2]

(7) a. Hiroshi-wa man'nenhitsu-o Ginza-no depaato-de
H.-Top fountain.pen-Acc G.-Gen department.store-Loc
{**kai/katte**}, sono man'nenhitsu-o chichioya-ni purezento-shita.
buy.Inf/buy.Ger that fountain.pen-Acc father-Dat present.Pst
'Hiroshi bought a fountain pen at a department store in Ginza, and he gave it to his father.'

b. #Hiroshi-wa chichioya-ni man'nenhitsu-o **purezento-shi**(**te**), sono
H.-Top father-Dat fountain.pen-Acc present.Inf(Ger) that
man'nenhitsu-o Ginza-no depaato-de katta.
fountain.pen-Acc G.-Gen department.store-Loc buy.Pst
(Hiroshi {gave/will give} his father a fountain pen, and he bought it at a department store in Ginza.)

Based on such observations, in Oshima (2012) I argued that the Inf-Cx and Ger-Cx have multiple meanings, all of which are more specific than mere logical conjunction, and accordingly postulated three constructs in the SBCG (Sign-Based Construction Grammar) sense.

The Inf-Cx and Ger-Cx may convey either (i) that the eventuality described in the subordinate clause (E_1) *temporally precedes or coincides with* the one described in the main clause (E_2), or (ii) that the propositions described by the two clauses stand in the rhetorical relation of *contrast*. Furthermore, the Ger-Cx, but not the Inf-Cx, has a third interpretation where the *resulting state* of E_1 temporally subsumes E_2; this interpretation is available only when the subordinate predicate belongs to a limited class of telic verbs that includes TATSU 'stand up', KIRU 'put on (clothes)', and MOTSU 'grab, take in one's hand'. The three interpretations are schematically illustrated in (8).

(8) (Eventuality E_1 and proposition P_1 correspond to the subordinate clause, and E_2 and P_2 to the main clause.)

i. *"non-subsequence" interpretation*: $\mathrm{E}_1 \leq \mathrm{E}_2$

ii. *"contrast" interpretation*: **Contrast**(P_1, P_2)

iii. *"resulting state" interpretation*: **ResultingState**(E_1) $\supseteq \mathrm{E}_2$ (available only with the Ger-Cx)

The "non-subsequence" interpretation is exemplified in (7a) above and (9) below.[3]

[2] In Oshima (2012), it is reported that out of the 22 native-speaker consultants, 15 evaluated (7b) as 'contradictory', two 'not sure', and five 'not contradictory'.

[3] The "non-subsequence" variety of the Susp-Cx can be used to describe a situation where E_1 is a state, E_2 is an event, and E_1 temporally subsumes E_2 ($\mathrm{E}_1 \supseteq \mathrm{E}_2$)

(i) a. (Kesa) niwa-ni risu-ga ite, sono koto-o kaisha-de
this.morning garden-Dat squirrel-Nom exist.Ger that matter-Acc company-Loc

(9) Kyuu-ni kion-ga **sagatte**, kaze-mo tsuyoku natta.
suddenly temperature-Nom fall.Ger wind-also strong.Inf become.Pst
'All of sudden, the temperature dropped and the wind became stronger, too.'

The "contrast" interpretation is illustrated in (10).

(10) Akira-wa kinoo **toochaku-shi(te)**, Hiroshi-wa
A.-Top yesterday arrive.Inf(Ger) H.-Top
ototoi toochaku-shita.
the.day.before.yesterday arrive.Pst
'Akira arrived yesterday, and (on the other hand) Hiroshi arrived the day before yesterday.'

(11) illustrates a sentence that allows both "non-subsequence" and "resulting state" readings. On the former reading, it implies that Ken's putting on a hat takes place within the topic time (in Klein's 1994 sense); on the latter, it does not. The former is not, and the latter is, compatible with a situation where Ken has an unusual habit

dooryoo-ni hanashita.
colleague-Dat tell.Pst
'There was a squirrel in the garden of my house (this morning), and I told my colleagues about it in the office.' ($E_1 \leq E_2$)
b. Magarikado-ni ookina iwa-ga atte, sore-ni jitensha-ga butsukatta.
corner-Dat big rock-Nom exist.Ger that-Dat bicycle-Nom hit.Pst
'There was a big rock on a street corner, and a bicycle ran into it.' ($E_1 \supseteq E_2$)

It cannot be used, on the other hand, to describe a situation where E_1 is an event, E_2 is a state, and E_2 temporally subsumes E_1.

(ii) a. Jooshi-kara idoo-no hanashi-o kiite, ie-ni
superior-from personnel.transfer-Gen speech-Acc hear.Ger home-Dat
kaette-kara-mo kibun-ga omokatta.
return.Ger-since-also feeling-Nom heavy.Pst
'Having heard from my superior that I will be transferred, I felt heavy-hearted even after coming home.' ($E_1 \leq E_2$)
b. #Kaichoo-ga toochaku-shite, subete-no yakuin-ga demukae-no tame
president-Nom arrive.Ger, all executive-Nom greeting-Gen for.purpose
ikkai robii-ni {ita/ atsumatte ita}.
first.floor lobby-Dat exist.Pst gather.Ger Ipfv.Pst
(The president arrived, and all the executives {were/were assembling} in the ground floor lobby to greet him.) ($E_1 \subseteq E_2$)
cf. Kaichoo-ga toochaku-shita toki, subete-no yakuin-ga demukae-no
president-Nom arrive.Pst when, all executive-Nom greeting-Gen
tame ikkai robii-ni {ita/ atsumatte ita}.
for.purpose first.floor lobby-Dat exist.Pst gather.Ger Ipfv.Pst
'When the president arrived, all the executives {were/were assembling} in the ground floor lobby to greet him.'

The analysis in (8-i) is not fully adequate in failing to account for this contrast. In this work, however, I adopt this simplifying analysis for convenience; as I only consider cases where both E_1 and E_2 are events, this simplification should not lead to any practical problem.

of wearing a hat all the time, and has not taken it off for years.

(11) Ken-wa booshi-o **kabutte** e-o kaita.
K.-Top hat-Acc put.on.Ger picture-Acc paint.Pst
i. 'Ken put on a hat and painted a picture.' (the "non-subsequence" interpretation)
ii. 'Ken painted a picture wearing a hat.' (the "resulting state" interpretation)

Logical representations of the two readings of (11) are given in (12), where **TT** stands for topic time and τ represents the temporal trace function (a function from eventualities to their temporal locations; Krifka 1998).

(12) a. ("non-subsequence" interpretation of (11))
$\exists e_2[\exists e_1[\textbf{put.on.hat}(e_1, \textbf{hiroshi}) \wedge \tau(e_1) \subseteq \textbf{TT} \wedge \tau(e_1) \leq \tau(e_2) \wedge \textbf{draw.picture}(e_2, \textbf{hiroshi}) \wedge \tau(e_2) \subseteq \textbf{TT} \wedge \tau(e_2) < \textbf{now}]]$
b. ("resulting state" interpretation of (11))
$\exists e_2[\exists e_1[\exists e_3[\textbf{put.on.hat}(e_1, \textbf{hiroshi}) \wedge \textbf{RS}(e_3, e_1) \wedge \tau(e_3) \supseteq \textbf{TT} \wedge \tau(e_3) \supseteq \tau(e_2) \wedge \textbf{draw.picture}(e_2, \textbf{hiroshi}) \wedge \tau(e_2) \subseteq \textbf{TT} \wedge \tau(e_2) < \textbf{now}]]]$

For the ease of exposition, in the following I will leave out reference to the topic time in semantic representations.

Below I will argue that the QV-ellipsis construction is a special subtype of the suspensive clause construction with the "non-subsequence" meaning, and that the DO-ellipsis construction is a special subtype of the gerund clause construction with the "resulting state" meaning.

3 Constraints on the QV-ellipsis construction

QV-ellipsis constructions generally can be paraphrased using the gerund or infinitive form of IU 'say' or OMOU 'think'. It is not always possible, however, to elide a form of IU/OMOU heading a suspensive clause. The possibility of ellipsis depends on both syntactic and semantic factors.

On the syntactic side, the subordinate clause in the QV-ellipsis construction must consist solely of the (direct or indirect) quotative phrase, and cannot contain any other (explicit) dependent.

(13) a. [**Oogoe-de** "Dareka imasen-ka?" to *(itte)] doa-o
loud.voce-by anybody exist.Plt.Neg-DP Quot say.Ger door-Acc
tataita.
knock.Pst
'He knocked on the door, saying "Is anybody here?" in a loud voice.'

b. [**Boku-ni** “Jaa-na” to *(itte)] dete itta.
I-Dat bye Quot say.Ger exit.Ger go.Pst
‘He left the room, saying “Bye” to me.’

The subject of the subordinate clause is not necessarily co-referential with the one of the main clause; however, conforming to the aforementioned constraint, it cannot be explicitly expressed (Fujita 2000).

(14) a. [(***Shujin-ga**) “Omachidoosama” to] soba-ga
manager-Nom sorry.to.have.kept.you.waiting Quot soba.noodle-Nom
okareta.
put.Psv.Pst
‘(The restaurant manager) said “Sorry to have kept you waiting”, and a bowl of *soba* noodles was put in front of me.’
b. [(**Shujin-ga**) “Omachidoosama” to itte]
manager-Nom sorry.for.having.you.wait Quot say.Ger
soba-ga okareta.
soba.noodle-Nom put.Psv.Pst
‘*idem*’

On the semantic side, the interpretation of the QV-ellipsis construction is more restricted than that of the “non-subsequence” variety of the suspensive clause construction (Oshima and Sano 2011).

As mentioned above, the suspensive clause construction on the “non-subsequence” interpretation entails that P_1 and P_2 both hold, and that E_1 is *not* temporally subsequent to E_2. Due to pragmatic enrichment, oftentimes it further conversationally implicates a more specific relation between P_1 and P_2 or E_1 and E_2, in a way similar to how the English *and*-coordination construction might implicate a causal relation, manner relation, etc. (e.g., “Hans pressed the spring and the drawer opened” may conversationally implicate that the drawer opened *because* Hans pressed the spring in order to open the drawer, that Hans pressed the spring *in order to* open the drawer, etc.; Levinson 2000).

(15) a. Ha-o **migaite**, hige-o sotta.
tooth-Acc brush.Ger beard-Acc shave.Pst
‘He brushed his teeth and (then) shaved.’ (temporal precedence)
b. Kyuu-ni kion-ga **sagatte**, kaze-mo tsuyoku natta.
suddenly temperature-Nom fall.Ger wind-also strong.Inf become.Pst
‘All of sudden, the temperature dropped and the wind became stronger.’ (temporal coincidence)
c. Basu-ni **notte**, kaisha-ni itta.
bus-Dat ride.Ger company-Dat go.Pst
‘He went to work, taking a bus.’ (manner relation)

d. Ishi-ni **tsumazuite**, koronda.
stone-Dat stumle.Ger fall.Pst
'He stumbled on a stone and fell.' (causal relation)

Interestingly, the SAY-ellipsis construction cannot be used to describe a situation where P_1 is (naturally inferred to be) the cause/reason of P_2; in other words, it entails that P_1 is *not* the reason of P_2.

(16) a. Hiroshi-wa ["Futorimashita-ne" to #(**itte**)] Yumi-o azen-to
H.-Top become.fat.Pst.Plt-DP Quot say.Ger Y.-Acc appalled-Adv
saseta.
do.Caus.Pst
'Hiroshi appalled Yumi, saying "You've gained some weight, haven't you?".' (causal relation present)

b. Hiroshi-wa ["Futorimashita-ne" to (**itte**)] Yumi-no hara-o
H.-Top become.fat.Pst.Plt-DP Quot say.Ger Y.-Ger belly.Acc
tsutsuita.
poke.Pst
'Hiroshi poked Yumi's belly, (saying) "You've gained some weight, haven't you?".' (causal relation absent)

The THINK-ellipsis construction, on the other hand, requires that either the causal relation hold between P_1 and P_2, as in (17a), or the manner relation hold between E_1 and E_2, as in (17b).

(17) a. ["Moo doose maniawanai" to (**omotte**)]
already anyway be.on.time.Neg.Prs Quot think.Ger
hashiru-no-o yameta.
run.Prs-Nominalizer-Acc stop.Pst
'He stopped running, (thinking) "I won't make it anyway".' (causal relation present)

b. ["Dare-ni-demo shippai-wa aru" to (**omotte**)] jibun-o
who-Dat-even mistake-Top exist.Prs Quot think.Ger self-Acc
nagusameta.
console.Pst
'He consoled himself, (thinking) "Anyone can make a mistake".' (manner relation present)

(18a) illustrates that, when neither the causal nor manner relation holds, the THINK-ellipsis construction cannot be felicitously used.

(18) ('I was watching a baseball game. The team I was supporting had a big lead, but at the ninth inning the opponent team closed to within two runs . . .')

a. ["Nandaka kumoyuki-ga ayashiku natte kita-na" to
somehow weather-Nom strange become.Ger Asp.Pst-DP Quot
#(**omotte**)] kansen-shite iru-to, kekkyoku surii-ran
think.Ger watch.game.Ger Ipfv.Prs-after eventually three-run
hoomuran-ga tobidashite gyakuten-make-o kisshite shimatta.
home.run-Nom pop.Ger reversal-loss-Acc receive.Ger end.up.Pst
'I was watching the game, thinking to myself "Darn, the tide is turning", and then a three-run home run of the opponent team turned around the game and we ended up losing.' (neither causal nor manner relation present)

b. ["Nandaka kumoyuki-ga ayashiku natte kita-na" to
somehow weather-Nom strange become.Ger Asp.Pst-DP Quot
(**omotte**)] yakimoki-shite iru-to, kekkyoku suriiran
think.Ger chafe.Ger Ipfv.Prs-after eventually three-run
hoomuran-ga tobidashite gyakuten-make-o kisshite shimatta.
home.run-Nom pop.Ger reversal-loss-Acc receive.Ger end.up.Pst
'I was being restless, thinking to myself "Darn, the tide is turning", and then a three-run home run of the opponent team turned around the game and we ended up losing.' (causal relation present)

To summarize the section:

(19) i. The SAY-ellipsis construction can be paraphrased with *itte* (gerund) or *ii* (infinitive); the THINK-ellipsis construction can be paraphrased with *omotte* (gerund) or *omoi* (infinitive).
ii. In both SAY- and THINK-ellipsis constructions, the subordinate clause must consist solely of the quotatitve phrase accompanied by *to*, and must not contain an explicit subject or an adverbial modifier.
iii. The SAY-ellipsis construction implies that there is no causal relation between P_1 and P_2.
iv. The THINK-ellipsis construction implies that there is a causal relation between P_1 and P_2, or a manner relation between E_1 and E_2.

4 Constraints on the DO-ellipsis construction

The DO-ellipsis construction can be classified into two major types (Teramura 1983), which I refer to as the HOLD-type and the "accompanying circumstance"-type. In the HOLD-type, elided *shite* can be regarded as a predicate of possession.

(20) *The HOLD-type*

a. Watashi-wa [saifu-o katate-ni {a. ∅/b. **shite**}] heya-o
I-Top wallet-Acc one.hand-Dat {∅/do.Ger} room-Acc
tobidashita.
dash.out.Pst
'I dashed out of the room, (holding) my wallet in my hand.'

b. Ken-wa [akanboo-o se-ni {a. ∅/b. **shite**}] atari-o shibaraku
K.-Top baby-Acc back-Dat {∅/do.Ger} vicinity-Acc for.a.while
sansaku-shita.
stroll.Pst
'Ken strolled around for a while, (carrying) the baby on his back.'

In the "accompanying circumstance"-type, on the other hand, the semantic contribution of *shite* is unclear and possibly absent.

(21) *The "accompanying circumstance"-type*

a. Sono senshu-wa [tairyoku-no otoroe-o riyuu-ni
that athlete-Top strength-Gen decline-Acc reason-Dat
{a. ∅/b. **shite**}] sakunen intai-shita.
{∅/do.Ger} last.year retire.Pst
'That athlete retired last year, the reason being the decline of his physical strength.'

b. Keisatsu-wa [hisseki-o tegakari-ni {a. ∅/b. **shite**}] memo-o
police-Top handwriting-Acc clue-Dat {∅/do.Ger} note-Acc
kaita jinbutsu-o tokutei-shita.
write.Pst person-Acc identify.Pst
'The police identified the person who wrote the note, using the traits of the handwriting as a clue.'

This work focuses on the HOLD-type, leaving the formal treatment of the "accompanying circumstance"-type to future research.

SURU as a verb of possession refers to a telic, punctual process (i.e., an achievement), rather than a state.

(22) Ken-wa kan-biiru-o te-ni shita.
K.-Top can-beer-Acc hand-Dat do.Pst
'Ken took a can of beer in his hand.'
NOT: 'Ken was holding a can of beer in his hand.'

The gerund clause headed by possessive *shite* is ambiguous between the "non-subsequence" and "resulting state" interpretations (or, between the "take" and "hold" interpretations); the infinitive clause headed by possessive *shi*, on the other hand, allows only the "non-subsequence" interpretation.

(23) Ken-wa [kan-biiru-o te-ni **shite**], uta-o utatta.
K.-Top can-beer-Acc hand-Dat do.Ger song-Acc sing.Pst
i. 'Ken took a can of beer in his hand, and sang a song.' (non-subsequence reading); OR
ii. 'Ken sang a song, holding a can of beer in his hand.' (resulting state reading)

(24) Ken-wa [kan-biiru-o te-ni **shi**], uta-o utatta.
K.-Top can-beer-Acc hand-Dat do.Inf song-Acc sing.Pst
'Ken took a can of beer in his hand, and sang a song.' (non-subsequence reading only)

The DO-ellipsis construction allows only the "resulting state" interpretation.

(25) Ken-wa [kan-biiru-o te-ni ∅], uta-o utatta.
K.-Top can-beer-Acc hand-Dat song-Acc sing.Pst
'Ken sang a song, holding a can of beer in his hand.' (resulting state reading only)

The subject of the subordinate clause of the DO-ellipsis construction must (i) not be explicitly expressed and (ii) be coreferential with the matrix subject. This property is shared by gerund clauses on the resulting state reading in general; to illustrate, (26), where the subjects of the subordinate and main clauses are referentially disjoint, does not allow the resulting state interpretation.

(26) Hiroshi-ga booshi-o kabutte, Yumi-ga sono sugata-o
H.-Nom hat-Acc put.on.Ger Y.-Nom that appearance-Acc
shashin-ni totta.
photograph-Dat take.Pst
'Hiroshi put on a hat, and Yumi took a picture of him wearing it.' (non-subsequence reading only)

As is the case with the QV-ellipsis construction, the subordinate clause of the DO-ellipsis construction appears to resist occurrence of an adverbial modifier.

(27) a. [Roopu-o te-ni (**shite**)] furiotosarenai yoo-ni
rope-Acc hand-Dat do.Ger shake.off.Psv.Neg.Prs in.purpose.to
funbatta.
stand.firm.Pst
'I stood firm holding a rope in my hand so as not to fall off.'
b. [Roopu-o **shikkari-to** te-ni ?(**shite**)] furiotosarenai
rope-Acc tightly hand-Dat do.Ger shake.off.Psv.Neg.Prs
yoo-ni funbatta.
in.purpose.to stand.firm.Pst
'I stood firm holding a rope tightly in my hand so as not to fall off.'

To summarize the section:

(28) i. The DO-ellipsis construction has two varieties: the HOLD-type and the "accompanying circumstance"-type.
ii. The subordinate clause of the DO-ellipsis construction consist solely of the dative and accusative NP's.
iii. The DO-ellipsis construction (or at least the HOLD-type thereof) can be paraphrased with *shite* (gerund), but not by *shi* (infinitive).
iv. In the HOLD-type, the subject of the subordinate clause must be coreferential with the matrix subject. This property is shared by – or is inherited from – the non-elliptic counterpart.

5 Evidence for the bi-clausal structure

One might be tempted to consider that the QV-ellipsis and DO-ellipsis constructions are mono-clausal (QuotP = quotative phrase).

(29) (= (1a))
a. Ken-ga [$_{\text{QuotP}}$ "Ohayoo" *to*] haitte-kita. (mono-clausal analysis)
b. Ken-ga [$_{\text{S}}$ [$_{\text{QuotP}}$ "Ohayoo" *to*]] haitte-kita. (bi-clausal analysis)

(30) (= (3a))
a. Ken-wa [$_{\text{AdvP}}$ akanboo-o se-ni] atari-o . . . (mono-clausal analysis)
b. Ken-wa [$_{\text{S}}$ akanboo-o se-ni] atari-o . . . (bi-clausal analysis)

One piece of evidence against the mono-clausal analysis comes from the scopal interaction between the putative subordinate clause and negation in the matrix clause. When the matrix predicate is negated, the putative subordinate clause of a QV- or DO-ellpsis construction does not necessarily fall under the scope of negation, patterning the same as the suspensive subordinate clause in general.

(31) ["Hara-wa hette masen" to (itte)] kuchi-o
stomach-Top lessen.Ger Ipfv.Prs.Plt Quote say.Ger mouth-Acc
tsukenakatta.
put.Neg.Pst
'He did not even have a bite, (saying) "I'm not hungry".'

(32) Ken-wa [yari-o te-ni (shite)] dare-mo toosanakatta.
K.-Top spear-Acc hand-Dat do.Ger anybody let.pass.Neg.Pst
'Ken did not let anyone in, (holding) a spear in his hand.'

Non-clausal adverbials, on the other hand, cannot escape from the scope of negation on the predicate (as in: *John did not sing* {***loudly/in the office***}), except for discourse-oriented ones (as in: ***Fortunately****, John did not sing*). It can thus be concluded that the quotative phrase in the QV-ellipsis constriction, and the "X-o Y-ni" phrase in the DO-ellipsis construction, are not non-clausal adverbiabls.

6 An SBCG analysis

This section provides a formal analysis of the SAY-, THINK-, and DO-ellipsis constructions in the framework of Sign-Based Construction Grammar (SBCG; Sag 2012). In the version of SBCG used in the current work, Montagovian semantics (rather than Frame Semantics or Minimal Recursion Semantics) is used as the primary means of semantic representation.

6.1 Background assumptions

I will assume the general construction (constraint) for Japanese clauses to be (33), and the one for the declarative clause to be (34).

(33) *clause-construct* ⇒

$$\left[\begin{array}{ll}
\text{MTR} & \left[\begin{array}{ll} \textit{clause} & \\ \text{SYN} & /\,\boxed{1}\,!\left[\text{VAL}\ \ \langle\,\rangle\right] \\ \text{SEM}|\text{LF} & /\downharpoonleft_{\omega}(\downharpoonleft_{\beta}(\ldots(\downharpoonleft_{\psi}(\downharpoonleft_{0}(\downharpoonleft_{\alpha})\ldots(\downharpoonleft_{1}))))) \end{array}\right] \\
\text{HD-DTR} & /\,\boxed{2}\left[\begin{array}{ll} \text{SYN} & \boxed{1}\left[\begin{array}{ll}\text{CAT} & \textit{predicate}\\ \text{VAL} & \boxed{\text{A}}\end{array}\right] \\ \text{SEM}|\text{LF} & \upharpoonleft_{0} \\ \text{ARG-ST} & \boxed{\text{B}}\ \langle \text{X}_1{:}[\text{LF}\upharpoonleft_{1}], \ldots, \text{X}_n{:}[\text{LF}\upharpoonleft_{\alpha}]\rangle \\ \text{DEPS} & \boxed{\text{B}}\oplus\langle \text{Y}_1{:}[\text{LF}\upharpoonleft_{\beta}],\ldots,\text{Y}_n{:}[\text{LF}\upharpoonleft_{\psi}]\rangle \end{array}\right]\\
\text{DTRS} & /\,\boxed{\text{A}}\oplus\langle\,\boxed{2}\,\rangle \\
\text{CX-CONT} & \upharpoonleft_{\omega}
\end{array}\right]$$

(34) **Declarative Clause Cx**
declarative-clause-construct ⇒

$$\left[\begin{array}{ll}
\text{HD-DTR} & \left[\text{SYN}|\text{CAT}|\text{PRDFORM}\ \ \textit{finite}\right] \\
\text{CX-CONT} & \lambda P_{\langle v,t\rangle}\exists e_0[P(e_0)]
\end{array}\right]$$

Some background assumptions and notational conventions are explained below:

(35) i. Type *sem-obj*, the value of SEM(ANTICS), has two attributes: INDEX and L(OGIAL)F(ORM). LF in function corresponds to Sag's (2012) FRAMES, and its value is an expression of lambda calculus.

ii. Subscripted arrow symbols are meta-variables over logical expressions. The direction of arrows (upward or downward) is just for expositional ease.

iii. The value of CX-CONT is the meaning component contributed by the construct itself (Copestake et al. 2005).

iv. "/" indicates that the constraint on the right is a default constraint. "!" indicates that the feature structure on the right is exempted from the domain of structural identity (Sag 2012: note 71).

v. Following Bouma et al. (2001), it is assumed that typically adverbials, including adverbial clauses, are dependents of a predicate, rather than adjuncts on a clause.

vi. It is assumed that Japanese clauses generally have a "flat" structure, where the subject appears on the same level as more oblique arguments and adverbials.

Declarative clauses are thus required to satisfy the constraints shown in (36), which incorporates the ones posed by *declarative-clause-construct* with the ones inherited from its supertype *clause-construct*.

(36)

$$\begin{bmatrix} \textit{declarative-clause-cxt} & \\ \text{MTR} & \begin{bmatrix} \text{SYN} & \boxed{1}\,!\begin{bmatrix}\text{VAL} & \langle\,\rangle\end{bmatrix} \\ \text{SEM}|\text{LF} & \downarrow_\omega(\downarrow_\beta(\ldots(\downarrow_\psi(\downarrow_0(\downarrow_\alpha)\ldots(\downarrow_1))))) \end{bmatrix} \\ \text{HD-DTR} & \boxed{2}\begin{bmatrix} \text{SYN} & \boxed{1}\begin{bmatrix} \text{CAT} & \begin{bmatrix} \textit{predicate} & \\ \text{PRDFORM} & \textit{finite} \end{bmatrix} \\ \text{VAL} & \boxed{\text{A}} \end{bmatrix} \\ \text{SEM}|\text{LF} & \uparrow_0 \\ \text{ARG-ST} & \boxed{\text{B}}\,\langle \text{X}_1\text{:[LF}\uparrow_1], \ldots, \text{X}_n\text{:[LF}\uparrow_\alpha]\rangle \\ \text{DEPS} & \boxed{\text{B}} \oplus \langle \text{Y}_1\text{:[LF}\uparrow_\beta], \ldots, \text{Y}_n\text{:[LF}\uparrow_\psi]\rangle \end{bmatrix} \\ \text{DTRS} & \boxed{\text{A}} \oplus \langle\,\boxed{2}\,\rangle \\ \text{CX-CONT} & \uparrow_\omega\text{: } \lambda P \exists e_0[P(e_0)] \end{bmatrix}$$

The meaning of a clause is generally calculated by the following steps: (i) the meaning of the heading predicate (corresponding to $\uparrow_0$/$\downarrow_0$ in (36)) is cyclically applied to those of the arguments, from the most oblique to the least oblique (i.e., the subject), (ii) if there are any adjuncts, their meanings are cyclically applied to the result of step (i), and (iii) the "constructional meaning" ($\uparrow_\omega$/$\downarrow_\omega$) is applied to the result of steps (i) and (ii). In the case of the declarative clause, step (iii) is existential closure of the eventuality variable. To illustrate with a specific example, the meaning of declarative clause (37a) is calculated as in (38), via the β-conversion shown in (39).[4]

(37) [$_{\text{S}}$[$_{\text{NP}}$ Hiroshi-ga] [$_{\text{NP}}$ Yumi-o] [$_{\text{AdvP}}$ Shinjuku-de] [$_{\text{V}}$ mita]].
H.-Nom Y.-Acc S.-Loc see.Pst
'Hiroshi saw Yumi in Shinjuku.'

[4] A box surrounding an AVM indicates that the AVM is a description of a specific linguistic entity, rather than a description of a grammatical entity (grammatical constraint, etc.); see Sag (2012).

(38)
$$\boxed{\left[\begin{array}{ll}
\multicolumn{2}{l}{\textit{declarative-clause-cxt}}\\
\text{MTR} & \left[\begin{array}{ll}\text{SYN} & \boxed{1}\,!\left[\text{VAL}\ \ \langle\,\rangle\right]\\ \text{SEM}|\text{LF} & \exists e_0[\mathbf{see}(e_0,\mathbf{h},\mathbf{y})\wedge\tau(e_0)<\mathbf{now}\wedge\mathbf{in}(e_0,\mathbf{s})]\end{array}\right]\\
\text{HD-DTR} & \left[\begin{array}{ll}\text{SYN} & \boxed{1}\left[\text{CAT}\ \left[\begin{array}{ll}\multicolumn{2}{l}{\textit{predicate}}\\ \text{PRDFORM} & \textit{finite}\end{array}\right]\right]\\ \text{SEM}|\text{LF} & \lambda y[\lambda x[\lambda e_1[\mathbf{see}(e_1,x,y)\wedge\tau(e_1)<\mathbf{now}]]]\\ \text{ARG-ST} & \boxed{\text{B}}\ \langle\text{NP:}[\text{LF }\mathbf{h}],\text{NP:}[\text{LF }\mathbf{y}]\rangle\\ \text{DEPS} & \boxed{\text{B}}\oplus\langle\text{AdvP:}[\text{LF }\lambda Q_{\langle v,t\rangle}[\lambda e_2[Q(e_2)\wedge\mathbf{in}(e_1,\mathbf{s})]]]\rangle\end{array}\right]\\
\text{CX-CONT} & \lambda P\exists e_0[P(e_0)]
\end{array}\right]}$$

(39) $\lambda P[\exists e_0[P(e_0)](\lambda Q_{\langle v,t\rangle}[\lambda e_2[Q(e_2)\wedge\mathbf{in}(e_1,\mathbf{s})]](\lambda y[\lambda x[\lambda e_1[\mathbf{see}(e_1,x,y)\wedge\tau(e_1)<\mathbf{now}]]](\mathbf{y})(\mathbf{h})))\Rightarrow_\beta\exists e_0[\mathbf{see}(e_0,\mathbf{h},\mathbf{y})\wedge\tau(e_0)<\mathbf{now}\wedge\mathbf{in}(e_0,\mathbf{s})]$

6.2 Regular suspensive clauses

Turning now to (regular, non-elliptic) suspensive clauses, I propose (40) as a construction that licenses the "non-subsequence" variety of the suspensive clause:

(40) **"Non-Subsequence" Suspensive Clause Cx**
temporal-suspensive-clause-construct ⇒
$$\left[\begin{array}{ll}
\text{HD-DTR} & /\left[\text{SYN}|\text{CAT}\ \left[\begin{array}{ll}\text{PRDFORM} & \textit{suspensive}\\ \text{SELECT} & \left[\text{SYN}|\text{CAT}\ \ \textit{predicate}\right]\end{array}\right]\right]\\
\text{CX-CONT} & /\ \lambda P[\lambda Q_{\langle v,t\rangle}[\lambda e_2[\exists e_1[P(e_1)\wedge Q(e_2)\wedge\tau(e_1)\leq\tau(e_2)]]]]
\end{array}\right]$$

This will assign meaning (12a) to (11) (except that reference to the topic time is omitted). Note that here suspensive clauses are, like other adverbials (see (35v)), treated as dependents of a predicate. In this regard I depart from Oshima (2012), where they are treated as adjuncts on a clause.

The construction that licenses the "resulting state" variety of the gerund clause is given in (41):

(41) **"Resulting State" Gerund Clause Cx**
resultingstate-gerund-clause-construct ⇒
$$\left[\begin{array}{ll}
\text{HD-DTR} & /\left[\begin{array}{ll}\text{SYN}|\text{CAT} & \left[\begin{array}{ll}\text{PRDFORM} & \textit{gerund}\\ \text{SELECT} & \left[\begin{array}{ll}\text{SYN}|\text{CAT} & \textit{predicate}\\ \text{ARG-ST} & \langle \mathrm{Z}_i,\ldots\rangle\end{array}\right]\end{array}\right]\\ \text{ARG-ST} & \langle \textit{pro}_i,\ldots\rangle\end{array}\right]\\
\text{CX-CONT} & /\left(\begin{array}{l}\lambda P[\lambda Q[\lambda e_2[\exists e_1[\exists e_3[P(e_1)\wedge Q(e_2)\wedge\\ \mathbf{RS}(e_3,e_1)\wedge\tau(e_3)\supseteq\tau(e_2)]]]]]\end{array}\right)
\end{array}\right]$$

This will assign (12b) to (11) (again, except that reference to the topic time is omitted).

6.3 Special suspensive clauses

I propose, finally, (42)–(44) as the constructions that license elliptic, headless suspensive clauses. Specifically, (42) and (43) respectively license the subordinate clause of the SAY-ellipsis construction and the THINK-ellipsis construction (which involve a direct quotative phrase); (44) licenses the subordinate clause the DO-ellipsis construction (of the HOLD-type). Their DTRS attributes are specified to be singleton and doubleton, which guarantees the absence of an explicit subject or an adverbial within it.

(42) **Special Suspensive Clause Cx (SAY, direct quote)**
elliptic-speech-temporal-suspensive-clause-construct ⇒

$$\begin{bmatrix} \text{MTR} & \begin{bmatrix} \text{SYN} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \textit{predicate} & \\ \text{PRDFORM} & \textit{suspensive} \\ \text{SELECT} & \begin{bmatrix} \text{SYN}|\text{CAT} & \textit{predicate} \end{bmatrix} \end{bmatrix} \\ \text{VAL} & \langle\,\rangle \end{bmatrix} \\ \text{SEM}|\text{LF} & \downarrow_3(\textbf{say}_{dir}(\downarrow_2)(\downarrow_1)) \\ \text{ARG-ST} & \langle \textit{pro}\text{:}[\text{LF}\ \uparrow_1], \boxed{1}\,\rangle \end{bmatrix} \\ \text{HD-DTR} & \textit{none} \\ \text{DTRS} & \langle\,\boxed{1}\ \text{QuotP:}[\text{MRKG}\ \textit{to}, \text{LF}\ \uparrow_2]\rangle \\ \text{CX-CONT} & \uparrow_3\text{:} \begin{pmatrix} \lambda P[\lambda Q[\lambda e_2[\exists e_1[P(e_1) \wedge Q(e_2) \wedge \tau(e_1) \leq \tau(e_2) \wedge \\ \neg\exists\langle t_1, t_2\rangle[\textbf{because}(\hat{}\exists e_3[P(e_3) \wedge \tau(e_3) = t_1], \\ \hat{}\exists e_4[Q(e_4) \wedge \tau(e_4) = t_2]) \wedge t_1 \leq t_2]]]]] \end{pmatrix} \end{bmatrix}$$

(43) **Special Suspensive Clause Cx (THINK, direct quote)**
elliptic-thought-temporal-suspensive-clause-construct ⇒

$$\begin{bmatrix}
\text{MTR} & \begin{bmatrix} \text{SYN} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \textit{predicate} & \\ \text{PRDFORM} & \textit{suspensive} \\ \text{SELECT} & \begin{bmatrix}\text{SYN}|\text{CAT} & \textit{predicate}\end{bmatrix} \end{bmatrix} \\ \text{VAL} & \langle\ \rangle \end{bmatrix} \\ \text{SEM}|\text{LF} & \downarrow_3(\textbf{think}_{dir}(\downarrow_2)(\downarrow_1)) \\ \text{ARG-ST} & \langle \textit{pro}\text{:[LF } \uparrow_1], \boxed{1} \rangle \end{bmatrix} \\
\text{HD-DTR} & \textit{none} \\
\text{DTRS} & \langle \boxed{1}\ \text{QuotP:[MRKG } \textit{to}\text{, LF } \uparrow_2]\rangle \\
\text{CX-CONT} & \uparrow_3\!: \begin{pmatrix} \lambda P[\lambda Q[\lambda e_2[\exists e_1[P(e_1) \wedge Q(e_2) \wedge \tau(e_1) \leq \tau(e_2) \wedge \\ [\textbf{by.means.of}(e_1, e_2) \vee \exists\langle t_1, t_2\rangle[\textbf{because}(\hat{}\exists e_3[P(e_3) \wedge \\ \tau(e_3) = t_1], \hat{}\exists e_4[Q(e_4) \wedge \tau(e_4) = t_2]) \wedge t_1 \leq t_2]]]]]] \end{pmatrix}
\end{bmatrix}$$

(44) **Special Gerund Clause Cx (DO, HOLD-type)**
elliptic-possession-resultingstate-gerund-clause-construct ⇒

$$\begin{bmatrix}
\text{MTR} & \begin{bmatrix} \text{SYN} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \textit{predicate} & \\ \text{PRDFORM} & \textit{gerund} \\ \text{SELECT} & \begin{bmatrix}\text{SYN}|\text{CAT} & \textit{predicate} \\ \text{ARG-ST} & \langle\, \text{X}_i, \ldots \rangle\end{bmatrix} \end{bmatrix} \\ \text{VAL} & \langle\ \rangle \end{bmatrix} \\ \text{SEM}|\text{LF} & \downarrow_4(\textbf{take.in}(\downarrow_3)(\downarrow_2)(\downarrow_1)) \\ \text{ARG-ST} & \langle \textit{pro}_i\text{:[LF } \uparrow_1], \boxed{1}, \boxed{2} \rangle \end{bmatrix} \\
\text{HD-DTR} & \textit{none} \\
\text{DTRS} & \langle \boxed{1}\ \text{NP:[CASE } \textit{acc}\text{, LF } \uparrow_2], \boxed{2}\ \text{NP:[CASE } \textit{dat}\text{, LF } \uparrow_3]\rangle \\
\text{CX-CONT} & \uparrow_4\!: \begin{pmatrix} \lambda P[\lambda Q[\lambda e_2[\exists e_1[\exists e_3[P(e_1) \wedge Q(e_2) \wedge \\ \textbf{RS}(e_3, e_1) \wedge \tau(e_3) \supseteq \tau(e_2)]]]]] \end{pmatrix}
\end{bmatrix}$$

In (42) and (43), $\textbf{say}_{dir}$ and $\textbf{think}_{dir}$ are logical predicates corresponding to IU 'say' and OMOU 'think' selecting a direct quotative phrase. To deal with QV-ellipsis constructions with an indirect quotative phrase, slightly different constructions will be required. In (44), **take.in** is a predicate that selects, besides the eventuality argument, (i) the possessor argument, (ii) the possessum argument, and (iii) the location argument.

The semantics of (42)–(44) are more specific than those of (40) and (41). In all of (42), (43), and (44), the meaning of the mother sign has one less "open slot", the place for the predicate meaning being filled by a constant. (42) and (43), furthermore, convey a more specific meaning than their non-elliptic counterpart which merely conveys the temporal relation of "E_1 precedes or coincides with E_2". These provide justification for treating the elliptic clauses as subtypes of the regular suspensive clauses. It should be noted that the absence of the causal relation encoded in the SAY-ellipsis construction, and the presence of the causal or manner ("by means of") relation encoded in the THINK-ellipsis construction, are presumably part of the "not-at-issue" (conventionally implicated) meaning, rather than the "at-issue" (proffered) meaning. To represent them in more precise terms, a more elaborate apparatus for semantic representation, where multiple levels/dimensions of meaning can be distinguished, will be required (see, e.g., Potts 2005; McCready 2010).

The mother sign of each of these constructs (i.e., a headless clause) is specified to have the ARG-ST attribute; this is required to constrain long-distance anaphoric binding into the headless subordinate clause, as in (45a,b), as well as to express the obligatory coreference between the subjects of the main and subordinate clause in the DO-ellipsis construction.[5]

(45) a. Hiroshi$_i$-wa [[kimi-ga jibun$_i$-o kizukatte kurenai] to
H.-Top you-Nom self-Acc be.concerned.Ger Ben.Neg.Prs Quot
(itte)] namida-o nagashite ita-yo.
say.Ger tear-Acc shed.Ger Ipfv.Pst-DP
'Hiroshi$_i$ was shedding tears, saying that you don't care about him$_i$ at all.'

b. Ken$_i$-ga [jibun$_i$-no yari-o te-ni (shite)] tachihadakatta.
K.-Nom self-Gen spear-Acc hand-Dat do.Ger block.way.Pst
'Ken$_i$ blocked the way, holding his$_i$ spear in his$_i$ hand.'

7 Summary

This paper discussed the syntactic and semantic properties of three "special" hypotactic constructions in Japanese, where the heading predicate of the subordnate clause is not explictly present. The subordinate clauses of the three constructions respectively involve "omission" of a suspensive form of IU 'say', OMOU 'think', and SURU 'do'. It was shown that the elliptic subordinate clauses have more specific meanings than the corresponding canonical subordinate clauses (headed by a suspensive form of a verb), and thus the former can be sensibly regarded as special subtypes of the latter. Using the framework of Sign-Based Construction Grammar, a formal analysis of the three constructions was presented, which does not postulate a phonologically null element serving as the head of a subordinate clause.

[5] See Przepiórkowski (2001) for justification for allowing phrasal (non-lexical) expressions to have ARG-ST (and its extension DEPS).

References

Arnold, Doug & Andrew Spencer. 2015. A constructional analysis for the skeptical. In Stefan Müller (ed.), *Proceedings of The 22nd International Conference of Head-Driven Phrase Structure Grammar*, Stanford: CSLI Publications.

Bouma, Gosse, Robert Malouf & Ivan A. Sag. 2001. Satisfying constraints on extraction and adjunction. *Natural language and linguistic theory* 19. 1–65.

Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation* 3. 281–332.

Dubinsky, Stanley & Shoko Hamano. 2003. Case checking by AspP. In William McClure (ed.), *Japanese/Korean linguistics*, vol. 12, 231–242. Stanford: CSLI Publications.

Fillmore, Charles J., Russel R. Lee-Goldman & Russel Rhomieux. 2012. The FrameNet construction. In Hans C. Boas & Ivan A. Sag (eds.), *Sign-Based Construction Grammar*, 309–372. Stanford: CSLI Publications.

Fujita, Yasuyuki. 2000. *Kokugo inyoo kooobun no kenkyuu [A study of Japanese quotative constructions].* Osaka: Izumi Shoin.

Fukushima, Kazuhiko. 1991. Bound morphemes, coordination and bracketing. *Journal of Linguistics* 35. 297–320.

Kim, Hyunah. 2013. Inyoo koobun ni okeru hatsuwa dooshi no senzai: Fukubun to shite no bunseki [Latency of speech-act verbs in quotative structures: Analysis of complex sentence]. *Nihongo bunpoo* 13(1). 52–67.

Klein, Wolfgang. 1994. *Time in language*. New York: Routledge.

Krifka, Manfred. 1998. The origins of telicity. In Susan Rothstein (ed.), *Events and grammar*, 197–235. Dordrecht: Kluwer.

Lee, Jungmee & Judith Tonhauser. 2010. Temporal interpretation without tense: Korean and Japanese coordination constructions. *Journal of Semantics* 27. 307–341.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge: The MIT Press.

Lyons, Christopher G. 1991. English nationality terms: Evidence for dual category membership. *Journal of Literary Semantics* 20. 97–116.

McCready, Eric. 2010. Varieties of conventional implicature. *Semantics and Pragmatics* 3(8). 1–57.

Muraki, Shinjiro. 1983. "Chizu o tayori ni, hito o tazuneru" toiu iikata [The expression of the form "Chizu o tayori ni, hito o tazuneru"]. In Minoru Watanabe (ed.), *Fukuyoogo no kenkyuu [Studies on modificational expressions]*, 267–292. Tokyo: Meiji Shoin.

Oshima, David Y. 2012. On the semantics of the Japanese infinitive/gerund-clause constructions: Polysemy and temporal constraints. In Stefan Müller (ed.), *Proceedings of The 19th International Conference of Head-Driven Phrase Structure Grammar*, 292–309. Stanford: CSLI Publications.

Oshima, David Y. 2013. Inyoo jutsugo no arawarenai hatsuwa/shikoo hookokubun: "Shooryaku" ka "koobun" ka [Japanese speech/attitude report sentences without a quotative predicate: Ellipsis or construction?]. *Research Bulletin of International Student Center, Ibaraki University* 11. 113–128.

Oshima, David Y. & Shin-ichiro Sano. 2012. On the characteristics of Japanese reported discourse: A study with special reference to elliptic quotation. In Isabelle Buchstaller & Ingrid van Alphen (eds.), *Quotatives: Cross-linguistic and cross-disciplinary perspectives*, 145–171. Amsterdam: John Benjamins.

Potts, Christopher. 2005. *The logic of conventional impricatures*. Oxford: Oxford University Press.

Przepiórkowski, Adam. 2001. ARG-ST on phrases. In Dan Flickinger & Andreas Kathol (eds.), *Proceedings of The 7th International Conference of Head-Driven Phrase Structure Grammar*, 267–284. Stanford: CSLI Publications.

Sag, Ivan A. 2012. Sign-Based Construction Grammar: An informal synopsis. In Hans C. Boas & Ivan A. Sag (eds.), *Sign-Based Construction Grammar*, 69–202. Stanford: CSLI Publications.

Teramura, Hideo. 1983. "Futai jookyoo" hyoogen no seiritsu no jooken: "X o Y ni . . . suru" toiu bunkei o megutte [Licensing conditions on expressions of "accompanying circumstance": On the construction of the form "X o Y ni . . . suru"]. *Nihongogaku* 2(10). 38–46. Reprinted in 1992 in *Teramura Hideo ronbunshuu I: Nihongo bunpoo hen [Collection of papers by Hideo Teramura I: Japanese grammar]*, pages 113–126, Tokyo: Kurosio Publishers.

Divergence in Expressing Definiteness between Mandarin and Cantonese

Joanna Ut-Seong Sio
Nanyang Technological University

Sanghoun Song
Incheon National University

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 177–194

http://csli-publications.stanford.edu/HPSG/2015

Sio, Joanna Ut-Seong, & Song, Sanghoun. 2015. Divergence in Expressing Definiteness between Mandarin and Cantonese. In Müller, Stefan (Ed.), *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, 177–194. Stanford, CA: CSLI Publications.

Abstract

In this paper, we model the dialectal variation in the expression of definiteness in Mandarin and Cantonese adopting the Head-Driven Phrase Structure Grammar (HPSG) framework (Pollard & Sag, 1994) and Minimal Recursion Semantics (MRS) (Copestake et al., 2005).

1 Introduction

Definiteness is a grammatical category that applies to noun phrases. A noun phrase is definite if there is sufficient information in the context for the hearer to identify the referent. Identifiability is a pragmatic notion relating to the assumptions made by the speaker on the cognitive status of a referent in the mind of the addressee in the context of utterance (Chen, 2004).

Unlike English, there are no articles (e.g. *a*, *the*) in Chinese indicating the definiteness value of a noun phrase. The referential interpretations of some Chinese noun phrases are flexible and thus ambiguous given appropriate contexts. In addition, dialects vary in terms of which surface forms are ambiguous. Amongst seven Chinese dialectal groups (viz., Northern, Wu, Xiang, Gan, Hakka, Yue and Min (Yuan, 1983), the present work focuses on Mandarin (abbreviated as ‘cmn’ in examples), which is a member of the Northern Group, and Cantonese (abbreviated as ‘yue’ in examples), which is a member of Yue.

2 Basic Data

2.1 Four Basic Types of Noun Phrases in Chinese

Table 1: Definiteness

type	example	Mandarin	Cantonese
DEM-CL-N	這隻狗	definite	
NUME-CL-N	三隻狗	indefinite	
CL-N	隻狗	indefinite	(in)definite
N	狗	(in)definite	indefinite

Noun phrases (NPs) in Chinese come in four basic forms: [DEM-CL-N], [NUME-CL-N], [CL-N] and [N]. [DEM-CL-N] phrases are always definite in Chinese while

[†] We are thankful to Francis Bond for his help and comments on this paper. We would like to thank participants of the Linguistic Analysis Design (LAD) sessions (DELPH-IN ’15 summit), attendees of the HPSG 2015 conference as well as all others who had given their useful input. This research was supported in part by the MOE Tier 2 grant *That’s what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13). The research reported here is also supported by the project ’Grammar Matrix Reloaded: Syntax and Semantics of Affectedness’ (MOE 2013-T2-1-016) funded by the Ministry of Education (MOE) in Singapore.

[NUME-CL-N] phrases are always indefinite. The definiteness interpretation of [CL-N] and [N] phrases vary depending on the dialect. Bare noun, [N], can always have a kind reading. In Mandarin (cmn) and Cantonese (yue), the definiteness interpretations of noun phrases are presented in Table 1 (Cheng & Sybesma, 1999; Sio, 2006).

The definiteness of a noun phrase can affect its distribution. Generally, only definite noun phrases can appear in the subject or topic position in Chinese (Chao, 1968; Lee, 1986; Li & Thompson, 1989, among others). Even though a [CL-N] phrase in Cantonese can be interpreted as either definite or indefinite, a [CL-N] phrase in the subject or topic position can only be interpreted as definite. This is illustrated in (1a) and (1b).[1] The same applies to Mandarin bare nouns, which are only interpreted as definite (or kind) in the subject or topic position as exemplified in (2a) and (2b).[2]

(1) a. 隻 狗 要 過 馬路。
zek3 gau2 jiu3 gwo3 ma5lou6
CL dog want cross road

'The dog wants to cross the road.' [yue]

b. 隻 狗 冇 人 要 呀。
zek3 gau2, mou5 jan4 jiu3 aa3
CL dog, no one want SFP

'That dog, no one wants it.' [yue]

(2) a. 狗 要 過 馬路。
gǒu yāo guò mǎlù
dog want cross road

'The dog wants to cross the road.'
NOT 'A dog wants to cross the road.' [cmn]

b. 狗，我 不 想 要 了。
gǒu, wǒ bù xiǎng yào le
dog I not want have SFP

'The dog, I don't want to have it (anymore).' [cmn]

For a noun phrase that cannot be interpreted as definite, putting it in the subject or topic position would lead to ungrammaticality. This applies to [CL-N] phrases in Mandarin, (3a), (3b) and bare nouns in Cantonese, (4a), (4b), with the exception of a kind reading, (5).

[1] CL: CLassifiers, SFP: Sentence Final Particle

[2] The examples presented in (1a) and (2a) are taken from Cheng & Sybesma (1999).

(3) a. *隻 狗 要 過 馬路。
zhī gǒu yāo guò mǎlù
CL dog want cross road [cmn]

b. *隻狗, 我 不 想 要 了。
zhī gǒu, wǒ bù xiǎng yàoo
CL dog I not want have

'The dog, I don't want to have it.' [cmn]

(4) a. *狗 要 過 馬路。
gau2 jiu3 gwo3 ma5lou6
dog want cross road [yue]

b. *狗, 冇 人 要 㗎。
gau2, mou5 jan4 jiu3 gaa3
dog, no one want SFP

'The dog, no one wants it.' [yue]

(5) 狗 鍾意 食 骨頭。
gau2 zung1ji3 sik6 gwat1tau4
dog like eat bone

'Dogs like to eat bones.' [yue]

[NUME-CL-N] phrases do not show distributive differences between Mandarin and Cantonese with respect to definiteness. They are indefinite in both dialects. However, the distribution of [NUME-CL-N] phrases regarding the subject/topic restriction is more intriguing than the other types of noun phrases. Li (1998) argues that [NUME-CL-N] phrases have two interpretations: quantity-denoting (concerning quantity) or individual-denoting (concerning the existence of certain individuals). A [NUME-CL-N] phrase cannot appear in the subject or topic position unless it has a quantity-denoting reading. We will illustrate the contrast with the subject position using Mandarin data. In (6a), the subject is a [NUME-CL-N] phrase, and it is ungrammatical unless *you* 'have' is added in the front, as in (6b).[3]

(6) a. *三 個 學生 在 學校 受傷 了。
sān gě xuéshēng zài xuéxiào shòushāng le
three CL student at school hurt SFP

'Three students were hurt at school.' [cmn]

b. 有 三 個 學生 在 學校 受傷 了。
yǒu sān gě xuéshēng zài xuéxiào shòushāng le
have three CL student at school hurt SFP

'There are three students hurt at school.' [cmn]

[3] (6a) and (6b) are taken from Li (1998).

(7), on the other hand, is grammatical. (7) is a non-episodic sentence and the [NUME-CL-N] phrase in (7) has what Li (1998) calls a quantity-denoting reading. It indicates the rice-eating capacity of (any)‘three people’ rather than the existence of three specific individuals.

(7) 三 個 人 可以 吃 得 完 一 桶 飯。
sān gè rén kěyǐ chī dè wá yī tǒng fàn
three CL person can eat to.the.extent finish one bucket rice

‘Three people can finish one bucket of rice.’ [cmn]

Adding *you* ‘have’ in (7) will make it ungrammatical as *you* ‘have’ asserts the existence of individuals and thus is only compatible with an individual-reading.

In addition to *you* ‘have’, it is also possible to save a sentence with a [NUME-CL-N] phrase as the subject by adding *dou* ‘all’ (Li, 1998). *Dou* ‘all’ ranges over an entire set of individuals and gives rise to a universal quantification reading. This is illustrated in (8). [4]

(8) 三 個 學生 都 來 這 裡 了。
sān gà xuéshēng dōu lái zhè lǐ le
three CL student all come this place SFP

‘Three students all came here.’ [cmn]

2.2 The Definite Article and Demonstratives

There are generally 6 situations where the English definite article is used (Lyons, 1977; Hawkins, 1978; Chen, 2004):

(9) a. Situational: Bring me **the hammer**.

b. Anaphoric: I saw a man pass by with a dog. **The dog** was very small and skinny, but **the man** was very large.

c. Shared knowledge: Be quiet. Do not wake up **the baby** (who is sleeping in the next room).

d. Uniqueness: Mary is **the smartest student in my class**.

e. Association: John went to a wedding last weekend. **The bride** was beautiful.

f. With an establishing relative clause: Do you know **the student who slapped the principal in the last Christmas party**?

[4] Example (8) is taken from Li (1998).

In the situational use in (9a), by using the definite article, the speaker indicates to the addressee that he will be able to identity the hammer in the context of the utterance. The use of the definite article in (9b) is anaphoric. The referents of 'the dog' and 'the man' are introduced into the universe of discourse by the previous sentence. In (9c), the definite article is used because the referent, 'the baby', is shared knowledge. In (9d), the definite article is used because the referent is unique (a superlative). (9e) illustrates a case of identifiability via association. The mention of 'a wedding' triggers the identifiability of all the things that are related to 'a wedding' (e.g. bride, cake, etc.) by association. In (9f), the identifibility of the student comes from the post-nominal relative clause. Hawkins (1978) calls the relative clause an 'establishing' relative clause. It establishes the identity of the referent.

The use of demonstratives fall into four major types (Himmelmann, 1996; Chen, 2004):

(10) a. Situational: Could you carry **this huge bag** for me?

b. Discourse Deictic: Your wife is not answering the phone. **This** is not good.

c. Anaphoric: There is a shopping mall about a block from here. You won't find anything interesting in **that mall** though.

d. Recognitional: It was filmed in California, **those dusty kind of hills that they have out here by Stockton and all.**[5]

Demonstratives in situational use is different from the situational use of definite article in that in the former, the subject in question must be visible to the addressee. Consider the following two sentences (Chen, 2004):

(11) a. Beware of the dog.

b. Beware of that dog.

(11b) is felicitous only if the dog is visible. In fact, the implication that there is a dog supposedly visible in the surrounding makes it a much scarier sign.

Demonstratives primarily encodes spatial notions (e.g. proximal vs. distal with respect to the speaker). They are most natural in a contrastive environment, explaining their incompatibility with unique objects:

(12) a. The sun is so bright.

b.??That sun is so bright.

[5] Example (10d) is taken are from Himmelmann (1996).

The anaphoric use of demonstratives involves the transference of spatial notions to the temporal dimensions (Lyons, 1977, p. 670). Deictic location is reinterpreted as location in the universe of discourse. The anaphoric use of the demonstratives is much less common in comparison with their deictic use. When the demonstratives are used anaphorically, it is often with a contrastive sense (Chen, 2004).

The recognitional use is when the speaker does not know with certainty whether a referent is identifiable enough for the addressee. In such situations, the speaker usually prefers a definite expression, which presume some familiarity on the part of the addressee with the referent, rather than using an indefinite expression which treats the referent as non-identifiable (Chen, 2004).

In Mandarin, there are two demonstratives, proximal and distal. Both demonstratives appear in two related forms:[6]

(13) a. proximal: zhè, zhèi

b. distal: nà, nèi

Both forms of the demonstratives can be added directly to a noun (Cheng & Sybesma, 2015):

(14) 這 孩子 真 頑皮。
zhè/zhèi háizi zhēn wánpí
this/that child really naughty

'This child is very naughty.' [cmn]

The only distributional difference is that *zhèi* and *nèi* cannot constitute a phrase. They cannot appear alone, neither as subjects, (17) nor as objects, (16). Unlike *zhè* and *nà*, which can be used alone as subjects, (15), though not as objects, (17) (Chao, 1968, p. 649).

(15) 這/那 也 不 要 緊。
zhè/Nà yě bù yàojìn
this/that also NEG matter

'This/That also doesn't matter.' [cmn]

(16) *我 要 這/那。
wǒ yāo zhè(zhèi)/nà(nèi)
1SG want this/that

Intended reading: 'I want this/that.' [cmn]

[6] It is generally believed that *zhèi* is historically *zhè* + *yī* 'one' and *nèi* is *nà* + *yī* 'one'

(17) 這/那 是 什麼？
zhèi/*nèi shì shénme
This/That BE what

Intended reading: 'What is this/that?' [cmn]

In Cantonese, the proximal demonstrative is *lei1* 'this' and the distal demonstrative is *go2* 'that'. Unlike Mandarin, in Cantonese, the demonstratives cannot stand alone and it cannot combine with the noun directly.

[DEM-CL-N] phrases are definite in both Mandarin and Cantonese, and they have similar grammatical properties. Chen (2004) claims that demonstratives in Mandarin have developed some functions which are typically served by definite articles in languages like English. He claims that the Mandarin demonstratives can be used anaphorically in a non-contrastive environment in (18), in situation of shared general knowledge in (19), association in (20) and with an establishing relative clause as in (21).[7]

(18) 有 一 個 獵人 養 著 一 隻 狗。
yǒu yī gè lièrén yǎn zhe yī zhī gǒu
have one CL hunter keep PROG one CL dog
這 隻 狗 很 懂事。
zhè zhī gǒu hěn dǒngshì
this CL dog very intelligent

'There was a hunter who had a dog. That dog was very intelligent.' [cmn]

(19) 這 天氣 真 怪， 十二 月 了，可 一點 不 冷。
zhè tiānqì zhēn guài shièryuè le kě yī diǎn bù lěng
this weather really strange December SFP but one little.bit not cold

'The weather is really strange. It is December now, but it is not cold at all.' [cmn]

(20) 他 買 了 一 輛 舊 車，那 輪胎 都 磨平 了。
tā mǎi le yī liàng jiù chē nà lúntāi dōu mópíng le
3SG buy PERF one CL old car that tire all rub.flat SFP

'He bought an old car. All the tires are worn out.' [cmn]

(21) 上 個 月 來 看 你 的 那 個 人，
shàng gè yuè lái kàn nǐ de nà gè rén
previous CL month come see you DE that CL person
我 今天 又 見 到 他 了。
wǒ jīntiān yòu jiàndào tā le
1SG today again see 3SG SFP

'The person who came to see you last month, I saw him again today.' [cmn]

[7] The Mandarin examples presented in here, (18)-(22), (25), (26), are taken from (Chen, 2004)

In (18), the [DEM-CL-N] phrase *zhè zhī gǒu* is used anaphorically in the absence of contrast. It is also possible to have a bare noun in place of the [DEM-CL-N] phrase, as in (22). In the Cantonese counterpart, either a [DEM-CL-N] phrase or a [CL-N] phrase would be appropriate.

(22) 有 一 個 獵人 養 著 一 隻 狗。
yǒu yī gè lièrén yǎn zhe yī zhī gǒu
have one CL hunter keep PROG one CL dog
狗 很 懂事。
gǒu hěn dǒngshì.
dog very intelligent

'There was a hunter who had a dog. The dog was very intelligent.' [cmn]

The same applies to the Mandarin examples in (19) and (20). It is possible to replace the [DEM-N] phrase in (19) and (20) with just a bare noun. [DEM-N] phrases are not grammatical in Cantonese. For Cantonese, a [CL-N] phrase or a [DEM-CL-N] phrase could be used for the equivalents of (19) and (20), as shown in (23) and (24) below. [8]

(23) (lei1) di1 tin1hei3 zan1 hai6 gwai3, dou1 sap6ji3jyut6 la1, zung6
this CL weather really be strange, already December SFP, still
m4 dung3
not cold

'The weather is really strange. It is December now, but it is not cold at all.' [yue]

(24) keoi5 maai5-zo2 bou6 gau6 ce1, (go2) di1 taai1 dou1
3SG buy-PERF CL old car, that CL tire all
mo4ping4-saai3 ga3 la3
polish.flat-completely SFP SFP

'He bought an old car. All the tires are worn out.' [yue]

Based on the above Mandarin examples, Chen (2004) concludes that the Chinese demonstratives serve some of the functions that are characteristic of the definite article like *the* in English. However, Chinese demonstratives are not yet full-fledged definite articles (Chen, 2004). First of all, they still respect the visibility requirement in situational use. Consider the contrast between (25) and (26) below:

(25) 安靜 點，別 把 那 孩子 吵醒 了。
ānjìng diǎn, bié bǎ nà háizi chǎo-xǐng le
quiet a.little.bit not make that baby wake-up SFP

'Be quiet. Don't wake up that baby.' [cmn]

[8] No Cantonese characters are given in these examples because some characters could not be displayed properly.

(26) 安靜 點 ，別 把 孩子 吵醒 了。
ānjìng diǎn , bié bǎ háizi chǎo-xǐng le
quiet a.little.bit , not make baby wake-up SFP

‘Be quiet. Don’t wake up the baby.’ [cmn]

(25) is infelicitous unless the addressee can see the baby. Furthermore, the demonstratives still require contrastiveness. (27) is unnatural and a bare noun should be used, as in (28).

(27) *那 個 太陽 出 來 了。
nà gè tàiyáng chūlái le
that CL sun come.out SFP

‘That sun came out.’ [cmn]

(28) 太陽 出 來 了。
tàiyáng chūlái le
sun come.out SFP

‘The sun came out.’ [cmn]

For Cantonese, a [CL-N] phrase will be appropriate for (26) and (28). In fact, for (28), a bare noun would also be appropriate. It could be because *tàiyáng*‘sun’ can be interpreted as a proper name and proper names can always appear bare in Chinese.

3 Analysis

The previous section can be summarized as follows. First, there are four basic types of NPs in Mandarin and Cantonese, viz. [DEM-CL-N], [NUME-CL-N], [CL-N], and [N]. [N] in Mandarin and [CL-N] in Cantonese are comparable to both [the x] or [a/an x] in English, except when they appear in the subject or topic position, then they can only mean [the x]. [NUME-CL-N] phrases are always indefinite. They can however still appear in the subject or topic position if they have a quantity-denoting rather than an individual-denoting reading in the sense of (Li, 1998). Chen (2004) shows that the demonstratives in Mandarin show characteristics of some of the functions of the definite articles in languages like English in allowing a non-contrastive anaphoric usage, situational usage, recognitional usage as well as can be used in contexts of shared general knowledge. Cantonese shows similar patterns. There are, however, at least two aspects showing that the Chinese demonstratives are not full-fledged definite articles. In the context of shared knowledge, the visibility requirement still applies. The demonstrative is only admissible if the referent is visible to the addressee. Furthermore, the demonstratives cannot be used with unique objects.

Adopting the framework of HPSG (Pollard & Sag, 1994) and MRS(Copestake et al., 2005), this section presents an analysis that models the different definiteness interpretations of the four types of NPs in Mandarin and Cantonese, as well as the requirement that Chinese subjects need to be definite. Not all the observations presented earlier on can be modeled at this stage. We will leave those further research.

3.1 Cognitive Status

Quite a few previous studies have dealt with definiteness and/or givenness using HPSG so far. The analysis proposed here is along the line of Borthen & Haugereid (2005) and Bender & Goss-Grubbs (2008). These studies address a property of referents within the HPSG formalism and propose *cog-st* (cognitive status), which specifies the relationship between referents and the common ground in discourse. This feature structure places a constraint on the availability of types of NPs in particular constructions.

The constraint has much to do with the morphosyntactic markers of expressing definiteness. Borthen & Haugereid (2005) and Bender & Goss-Grubbs (2008) argue that the binary distinction such as definite vs. indefinite is sometimes not precise enough to deal with the various types of definiteness in NPs. As exemplified in the previous section (and in many other human languages), NPs are often ambiguous, though a more specific meaning is provided up to the entire parse tree. Furthermore, language processing, as of now, normally does not go beyond a sentence (i.e. intrasentential). Contextual information can only be partially resolved in our language application. In other words, not all NP structures can be analyzed as two-fold (i.e., definite vs. indefinite) within the context of grammar engineering. Instead of the binary distinction, Borthen & Haugereid (2005) and Bender & Goss-Grubbs (2008) use the givenness hierarchy (Prince, 1981; Gundel et al., 1993). From right to left in Table 2, each type is exemplified in (29).

Table 2: Givenness hierarchy

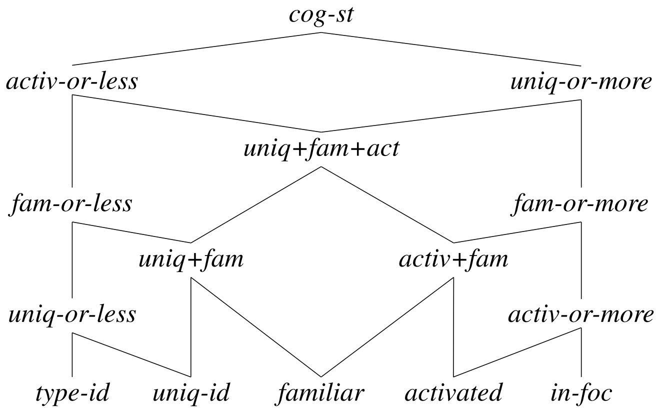
In focus	>Activated	>Familiar	>Uniq. id	>Referential	>Type id
it	*this*, *that* *this N*	*that N*	*the N*	indefinite *this N*	*a N*

(29) a. I couldn't sleep last night.

b. i. A dog (next door) kept me awake.
ii. This dog (next door) kept me awake.
iii. The dog (next door) kept me awake.
iv. That dog (next door) kept me awake.
v. That kept me awake.
vi. It kept me awake.
(Borthen & Haugereid, 2005, p. 230)

Along this line, Borthen & Haugereid (2005) provide an HPSG-based type hierarchy of cognitive status, which was then slightly refined by Bender & Goss-Grubbs (2008) as sketched out in (30).

(30)



This hierarchical approach to NP meanings enables us to represent partial information and thereby facilitates maintaining the phrase structure rules of forming NPs in a flexible way.

Building upon the type hierarchy provided in (30), Table 1 is now converted into Table 3.

Table 3: Cognitive status

type	example	Mandarin	Cantonese
DEM-CL-N	這 隻 狗	*uniq-or-more*	
NUME-CL-N	三 隻 狗	*type-id*	
CL-N	隻 狗	*type-id*	*cog-st*
N	狗	*cog-st*	*type-id*

First, if a particular construction conveys only definite meaning, the phrase places the *uniq-or-more* feature to the head noun as indicated in the second row in Table 3. Notice that in the *cog-st* hierarchy provided in (30) *uniq-or-more* excludes the leftmost item (i.e. *type-id*) that signals indefiniteness from its subtypes. In this way, *uniq-or-more* indicates that the NP can be evaluated as containing definiteness. Note also that ‘Activated’ and ‘Familiar’ in Table 2 are instantiated as NPs with demonstratives (i.e., *this N*, and *that N*). Since *uniq-or-more* includes these meanings, [DEM-CL-N] in the second row of Table 3 is not inconsistent with the constraint. Second, if a particular construction conveys only indefinite meaning, the phrase is constrained as *type-id*. Notice that the *type-id* node in the *cog-st* hierarchy is exclusive of any definite meaning. Finally, if a particular construction is ambiguous (i.e. (in)definite), the cognitive status of the phrase remains underspecified as *cog-st*. This means that an NP whose value of cognitive status is underspecified can be interpreted as either indefinite or definite.[9]

[9] We defer to the corpus-based findings provided in Gundel et al. (1993).

3.2 Phrase Structure Rules

In Table 3, note that Mandarin and Cantonese exhibit contrasting features in the fourth row and the fifth row whereas they share the same features in the second row and the third row. The constraints on such a divergence of expressing definiteness between Mandarin and Chinese are as follows.

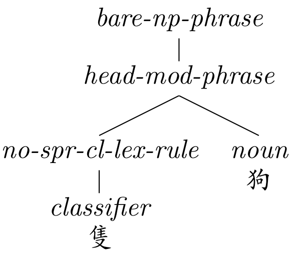
First of all, Mandarin and Cantonese share the following lexical type of classifiers, in which the element of MOD goes for the head noun, the element of SPR (i.e. specifier) goes for demonstratives and numerals. For example, in 這 隻 狗 'this CL dog', 這 and 狗 are constrained as SPR and MOD, respectively.

(31)
$$\begin{bmatrix} \textit{classifier} & \\ \text{LTOP} & \boxed{1} \\ \text{INDEX} & \boxed{2} \\ \text{ARG1} & \boxed{2} \\ \text{MOD} & \left\langle \begin{bmatrix} \textit{noun} & \\ \text{LTOP} & \boxed{1} \\ \text{INDEX} & \boxed{2} \\ \text{SPR} & \left\langle [\,] \right\rangle \\ \text{COG-ST} & \textit{cog-st} \end{bmatrix} \right\rangle \\ \text{SPR} & \left\langle \begin{bmatrix} \text{LTOP} & \boxed{1} \\ \text{INDEX} & \boxed{2} \end{bmatrix} \right\rangle \end{bmatrix}$$

Classifiers themselves impose no COG-ST constraint on the head noun, given that any types of COG-ST value can be assigned to the NPs with classifiers.

When classifiers are not specified by demonstratives and numerals (i.e. [CL-N]) in Mandarin, the NP involves an indefinite interpretation. This is constrained by a lexical rule, as presented in the AVM of (32). This rule makes the SPR list empty and places a constraint on the head noun's cognitive status as *type-id* responsible for indefinite. A sample derivation is given on the right side.

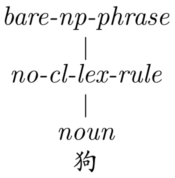
(32)
$$\begin{bmatrix} \textit{no-spr-cl-lex-rule} & \\ \text{MOD} & \left\langle \begin{bmatrix} \text{COG-ST} & \textit{type-id} \end{bmatrix} \right\rangle \\ \text{SPR} & \langle\, \rangle \\ \text{ARGS} & \left\langle \begin{bmatrix} \textit{classifier} \\ \text{SPR} \left\langle [\,] \right\rangle \end{bmatrix} \right\rangle \end{bmatrix}$$



Note that this constraint is Mandarin-specific. Since the definiteness of the [CL-N] form in Cantonese is ambiguous, this rule is not necessary for Cantonese.

Mandarin and Chinese also differ in how bare NPs are constrained. Cantonese, in which the [N] form is not ambiguous, employs the following lexical rule for nouns. This rule functions the same as the rule presented in (32), but it takes nouns as its daughter. The rule is Cantonese-specific.

(33)
$$\begin{bmatrix} \textit{no-cl-lex-rule} & \\ \text{MOD} & \left\langle \begin{bmatrix}\text{COG-ST} & \textit{type-id}\end{bmatrix} \right\rangle \\ \text{SPR} & \left\langle\ \right\rangle \\ \text{ARGS} & \left\langle \begin{bmatrix} \textit{noun} \\ \text{SPR}\left\langle \begin{bmatrix}\ \end{bmatrix} \right\rangle \end{bmatrix} \right\rangle \end{bmatrix}$$



Bare-np-phrase used in the parse trees of (32-33) is constrained as represented in the following AVM. This non-branching rule signals *cog-st* (i.e. underspecified) and introduces an existential quantifier (i.e. *exist_q_rel*) into the RELS list.

(34)
$$\begin{bmatrix} \textit{bare-np-phrase} & \\ \text{HD} & \begin{bmatrix} \textit{noun} & \\ \text{COG-ST} & \textit{cog-st} \\ \text{LTOP} & \boxed{1} \\ \text{INDEX} & \boxed{2} \end{bmatrix} \\ \text{C-CONT} & \begin{bmatrix} \text{RELS} & \left\langle !\ \begin{bmatrix} \text{PRED} & \textit{exist\_q\_rel} \\ \text{ARG0} & \boxed{2} \\ \text{RSTR} & \boxed{3} \end{bmatrix}\ ! \right\rangle \\ \text{HCONS} & \left\langle !\ \begin{bmatrix} \textit{qeq} & \\ \text{HARG} & \boxed{3} \\ \text{LARG} & \boxed{1} \end{bmatrix}\ ! \right\rangle \end{bmatrix} \end{bmatrix}$$

If the daughter of this phrase can have a more specific value of COG-ST, the value is unified. For instance, the daughters of *bare-np-phrase* in parse trees of (32-33) are constrained as [COG-ST *type-id*]. Because *type-id* is a specific subtype of *cog-st*, the COG-ST feature is unified as *type-id* (i.e. indefinite).

Finally, in order to disallow indefinite items to be used as subjects in Mandarin and Cantonese, the ordinary *subj-head-phrase* rule additionally includes one language-specific constraint as provided in (35).[10]

[10] Since proper names and clausal subjects are not indefinite, this constraint does not affect other types of subjects.

(35) $\begin{bmatrix} \textit{subj-head-phrase} \\ \text{NHD} \mid \text{COG-ST} \quad \textit{uniq-or-more} \end{bmatrix}$

Note that *uniq-or-more* is mutually exclusive with *type-id*, as represented in the type hierarchy (30). For instance, the structures provided in (32-33) cannot take the subject position because their COG-ST feature is inconsistent with the constraint on *subj-head-phrase*.

4 Sample Derivations

This section provides two sample derivations in Cantonese and Mandarin, respectively. The sentences are listed in (36) and (37).

(36) 隻 狗 走 啦。
zek3 gau2 zau2 la3
CL dog leave SFP

'The dog is leaving.' [yue]

(37) 狗 走 了。
gǒu zǒu le
dog leave SFP

'The dog is leaving.' [cmn]

The two sentences share almost the same meaning. The subjects are evaluated as conveying a definite interpretation, as only definite NPs can appear as subjects in Chinese.

Figure 1 representing (36) shows the derivation of a Cantonese sentence, an intransitive verb taking a [CL-N] phrase as the subject. Even though [CL-N] phrases can be interpreted either as definite or indefinite in Cantonese, when appearing in the subject position, it can only be interpreted as definite. The Mandarin counterpart of this sentence would be ungrammatical as [CL-N] phrases can only be indefinite in Mandarin. In the MRS structure on the right side, the COG-ST value of the subject 狗 'dog' is specified as *uniq-or-more* following the constraint presented in (35). Note that the NP 隻 狗 'CL-dog' itself is assigned *cog-st* as the value of COG-ST, as shown on the tree. The value becomes more hierarchically specific when the NP is used as the non-head daughter of *subj-head-phrase*: When the NP is combined with the verb 走 'leave' to form a *subj-head-phrase*, the subject is assigned [COG-ST *uniq-or-more*].

Figure 2 representing (37) shows the derivation of a Mandarin sentence, an intransitive verb taking an [N] phrase as subject. Even though [N] phrases can be interpreted either as definite or indefinite in Mandarin, when appearing in the

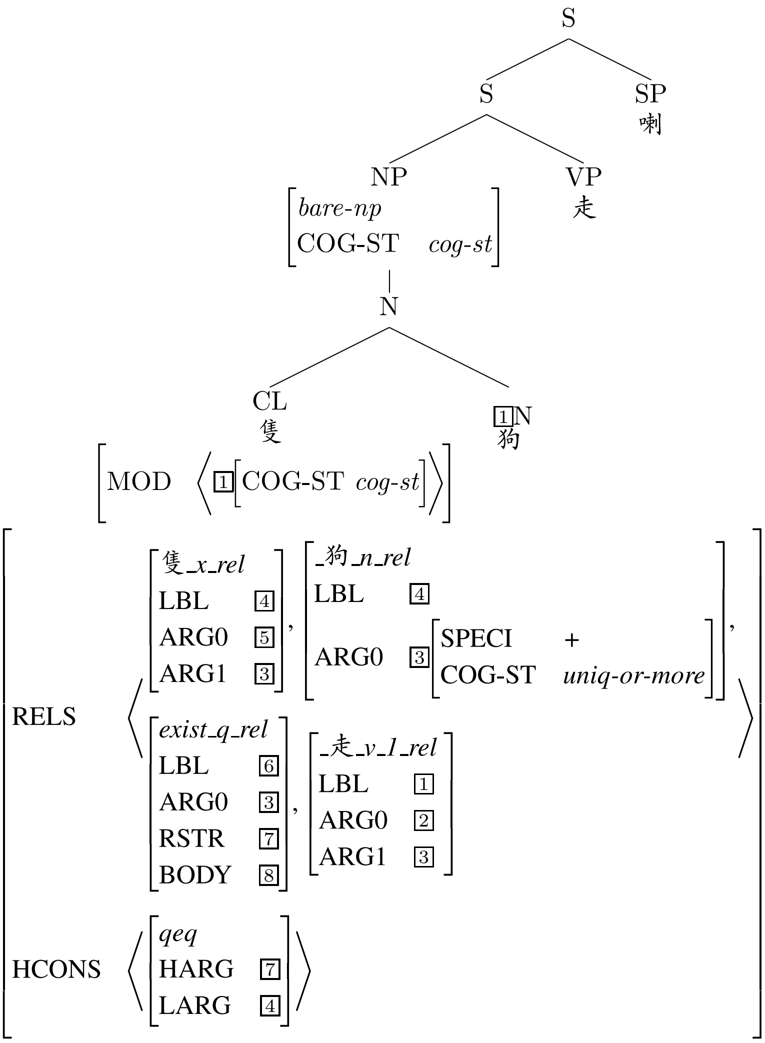


Figure 1: A sample derivation in Cantonese

subject position, it can only be interpreted as definite. The Cantonese counterpart of this sentence would be ungrammatical as [N] phrases can only be indefinite in Cantonese. The COG-ST of the subject 狗 'dog' in the MRS representation is specified as *uniq-or-more* in the same way as Figure 1.[11]

References

Bender, Emily M. & David Goss-Grubbs. 2008. Semantic Representations of Syntactically Marked Discourse Status in Crosslinguistic Perspective. In *Proceed-*

[11] Note that *cog-st* is hearer-oriented. The speaker-oriented status is represented as [SPECI *bool*] (i.e. specificity) (Borthen & Haugereid, 2005; Bender & Goss-Grubbs, 2008).

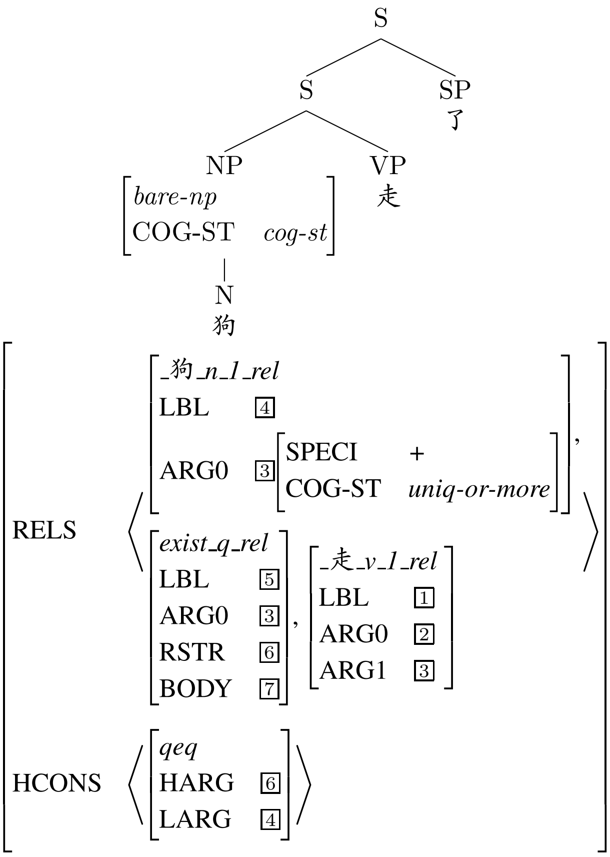


Figure 2: A sample derivation in Mandarin

ings of the 2008 conference on semantics in text processing, 17–29. Association for Computational Linguistics.

Borthen, Kaja & Petter Haugereid. 2005. Representing Referential Properties of Nominals. *Research on Language and Computation* 3(2-3). 221–246.

Chao, Yuan Ren. 1968. *Grammar of Modern Spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Chen, Ping. 2004. Identifiability and Definiteness in Chinese. *Linguistics* 42(6). 1129–1184.

Cheng, Lisa Lai-Shen & Ring Sybesma. 2015. Mandarin. In Tibor Kiss & Artemis Alexiadou (eds.), *Syntax – theory and analysis. an international handbook. handbooks of linguistics and communication science*, 42.1–3. Berlin: Mouton de Gruyter.

Cheng, Lisa Lai-Shen & Rint Sybesma. 1999. Bare and not-so-bare Nouns and the Structure of NP. *Linguistic Inquiry* 30(4). 509–542.

Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan A. Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language & Computation* 3(4). 281–332.

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

Hawkins, John A. 1978. *Definiteness and Indefiniteness: a Study in Reference and Grammaticality Prediction*. London: Croom Helm.

Himmelmann, Nikolaus P. 1996. Demonstratives in Narrative Discourse: A Taxonomy of Universal Uses. In Barbara Fox (ed.), *Studies in anaphora*, 205–254. Amsterdam: John Benjamins.

Lee, Hun-tak Thomas. 1986. *Studies on Quantification in Chinese*: University of California, Los Angeles dissertation.

Li, Charles N & Sandra A Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley and Los Angeles: University of California Press.

Li, Yen-hui Audrey. 1998. Argument Determiner Phrases and Number Phrases. *Linguistic Inquiry* 29(4). 693–702.

Lyons, John. 1977. *Semantics. Vols. I and II.* Cambridge: Cambridge University Press.

Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, IL: The University of Chicago Press.

Prince, Ellen F. 1981. Toward a Taxonomy of Given-New Information. In Peter Cole (ed.), *Radical pragmatics*, 223–256. New York: Academic Press.

Sio, Joanna Ut-Seong. 2006. *Reference and Modification in the Chinese Nominal*. the Netherlands: LOT Publication.

Yuan, Jiahua. 1983. *Hànyǔ fāngyán gàiyào [Outline of Chinese Dialects]*. Beijing: Wenzi Gaige.

A Constraint-based Analysis of A-NOT-A Questions in Mandarin Chinese

Wenjie Wang
Nanyang Technological University

Sanghoun Song
Incheon National University

Francis Bond
Nanyang Technological University

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 195–214

http://csli-publications.stanford.edu/HPSG/2015

Wang, Wenjie, Song, Sanghoun, & Bond, Francis. 2015. A Constraint-based Analysis of A-NOT-A Questions in Mandarin Chinese. In Müller, Stefan (Ed.), *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, 195–214. Stanford, CA: CSLI Publications.

Abstract

The A-NOT-A structure is one way to express polar questions in Mandarin Chinese. The present study provides a constraint-based analysis of A-NOT-A questions in Mandarin Chinese within the framework of HPSG (Pollard & Sag, 1994) and MRS (Copestake et al., 2005). We propose two possible approaches to analysing the A-NOT-A structure — a morphological/lexical approach as well as a syntactic approach — and illustrate their implementation, as well as their respective strengths and weaknesses.

1 Introduction

1.1 Basic Properties

The A-NOT-A structure is one way to express polar questions in Mandarin Chinese. The structure is so termed because it consists of an element (A) that is followed immediately by the same element but of negative polarity (NOT-A). For ease of reference, we shall refer to these elements as A_1 and A_2 respectively.

The A-NOT-A structure exists in various forms, which are exemplified below:

(1) a. Basic: A-NOT-A
张三 喜欢$_{A1}$ 不 喜欢$_{A2}$ 狗 ?
Zhāngsān xǐhuān bù xǐhuān gǒu ?
Zhangsan like NOT like dog PU

b. Contracted: A′-NOT-A
张三 喜 不 喜欢 狗 ?
Zhāngsān xǐ bù xǐhuān gǒu ?
Zhangsan like NOT like dog PU

c. Phrasal: AO-NOT-AO
张三 喜欢 狗 不 喜欢 狗 ?
Zhāngsān xǐhuān gǒu bù xǐhuān gǒu ?
Zhangsan like dog NOT like dog PU

d. Phrasal: AB-NOT-A
张三 喜欢 狗 不 喜欢 ?
Zhāngsān xǐhuān gǒu bù xǐhuān ?
Zhangsan like dog NOT like PU

All variations presented in (1) convey almost the same meaning: “Does Zhangsan like dogs?”

1.1.1 Reduplication

As shown in (1), A_1 and A_2 are reduplicates of each other. Reduplication for A-NOT-A can be performed *partially*: (1b) shows that A_1 can be reduplicated with just its first character/syllable, while (1d) shows that the verb can be reduplicated without its complement. Note that in both cases, A_2 must itself be *fully* reduplicated. As such, the following are ungrammatical:

(2) a. * 张三 喜欢 不 喜 狗 ?
* Zhāngsān xǐhuān bù xǐ gǒu ?
Zhangsan like NOT like dog PU

b. * 张三 喜欢 狗 不 喜 ?
* Zhāngsān xǐhuān gǒu bù xǐ ?
Zhangsan like dog NOT like PU

1.1.2 What can be A?

All lexical types capable of behaving as a syntactic head of predicates in Mandarin Chinese, such as verbs, adjectives, and prepositions, can participate in the A-NOT-A structure as A elements (Tseng, 2009). In the examples below, adjectives and prepositions (co-verbs) are shown playing the role of A elements:

(3) a. 张三 高 不 高 ?
Zhāngsān gāo bù gāo ?
Zhangsan tall NOT tall PU

‘Is Zhangsan tall (or not tall)?’

b. 张三 在 不 在 家 ?
Zhāngsān zài bù zài jiā ?
Zhangsan at NOT at home PU

‘Is Zhangsan at home (or not at home)?’

Adverbs are not allowed to be A elements, with the exception of frequency adverbs such as 常 *cháng* “often”:

(4) a. * 张三 很 不 很 高 ?
* Zhāngsān hěn bù hěn gāo ?
Zhangsan very NOT very tall PU

(Intended: ‘Is Zhangsan very tall?’)

b. 张三 常 不 常 迟到 ?
Zhāngsān cháng bù cháng chí-dào ?
Zhangsan often NOT often late-arrive PU

‘Is Zhangsan often late?’

A elements cannot be reduplicated elements themselves. As such, although the frequency adverb 常 *cháng* can be an A element, its reduplicated form 常常 *cháng cháng* cannot.

(5) a. * 张三 常常 不 常常 迟到 ?
* Zhāngsān cháng-chang bù cháng-chang chí-dào ?
Zhangsan often NOT often late-arrive PU

‘Is Zhangsan often late?’

1.1.3 What can be NOT?

Mandarin Chinese employs two negative operators (不 *bù* and 没 *méi*), the choice of which hinges on the aspectual property of the verbal item that they are attached to: 不 *bù* for statives and imperfectives, and 没 *méi* for bound events and perfectives. This is exemplified in (6).

Both of them can participate in the *A*-NOT-*A* structure as NOT, and likewise the aspect of the *A* element determines which is used. They also have slightly different co-occurrence constraints.

(6) a. 去 不 去 ?
qù bù qù ?
go NOT go PU

‘Are you going?’

b. 去 没 去 ?
qù méi qù ?
go NOT go PU

‘Have you gone (somewhere)?’

1.2 Basic Constraints

1.2.1 Modifiability of *A* elements

The *A* elements in *A*-NOT-*A* cannot take modifiers, such as degree adverbs, or aspectual markers:

(7) a. * 张三 很 高 不 很 高 ?
* Zhāngsān hěn gāo bù hěn gāo ?
Zhangsan very tall NOT very tall PU

‘Is Zhangsan very tall?’

b. * 张三 去 了 不 去 了 ?
* Zhāngsān qù le bù qù le ?
Zhangsan go LE NOT go LE PU

(Intended: ‘Zhangsan went?’)

The exception is the A-MEI-A sub-pattern, which can be post-modified by the experiential aspectual marker 过 *guò*.

(8) 张三 去 过 没 去 过 ?
Zhāngsān qù guò méi qù guò ?
Zhangsan go GUO NOT go GUO PU

'Has Zhangsan been there before?'

1.3 Co-occurrence Constraints

1.3.1 Sentence-final particles

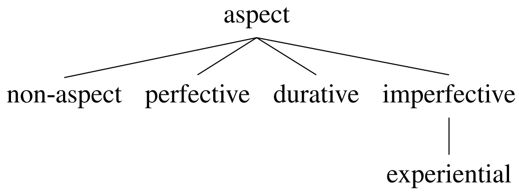
A-NOT-*A* questions are not permitted to occur with certain sentence-final particles. In the cases of 了 *lè*, 吗 *ma*, 吧 *ba*, 哦 *o* and 耶 *ye*, it is because only propositions can be used with these sentence-final particles, whereas *A*-NOT-*A* is a question.

Other sentence-final particles like the emphatic markers 嘛 *ma*, 呀 *ya* and 呢 *nē* do not, however, restrict themselves to only propositions and are therefore permitted to be used with *A*-NOT-*A*.

1.3.2 Aspectual markers

Chinese is an aspect-based language, in which aspect is linguistically and necessarily expressed, and plays an important role in syntax. The aspect hierarchy of Chinese (as implemented in ZHONG [|]) is roughly sketched out in (9):

(9)



Grammatical aspect in Chinese is largely expressed by verbal markers. There are three aspectual markers in Mandarin Chinese: 了 *lè*, 着 *zhè*, and 过 *guò*, which indicate the perfective, durative, and experiential aspects respectively. Since each verb lexically selects these markers, not all these three items can be necessarily attached to all verbs. For example, 去 *qù* 'go' does not canonically co-occur with *zhè*. These markers are collectively known as LE-ZHE-GUO or LZG, and they are hierarchically constrained as described in (10) in Type Definition Language.

(10)
```
+vjp :+ [ LZG lzg ].
lzg := avm.
le := lzg.
zhe := lzg.
guo := lzg.
```

```
no-lzg := lzg.
le+zhe := le & zhe.
le+guo := le & guo.
zhe+guo := zhe & guo.
le+zhe+guo := le & zhe & guo.
```

The LE-ZHE-GUO markers are also restricted in their co-occurrence with A-NOT-A, either with the entire A-NOT-A phrase, or with the individual A elements (See Section 1.2.1). The markers *lè* and *zhè* are not allowed to co-occur with A-NOT-A at all, while *guò* can only occur with A-NOT-A if the NOT element is 没 *méi*.

1.4 Versus MA-questions

The MA-question is another type of polar question, in which a sentence-final particle 吗 *mā* is used.

For example, (11) has a similar meaning to (1).

(11) 张三 喜欢 狗 吗 ?
Zhāngsān xǐhuān gǒu ma ?
Zhangsan like dog MA PU

'Does Zhangsan like dogs?'

On the surface, both (1) and (11) are translated as "Does Zhangsan like dogs?", and thus appear allo-structural and the semantic representation should be almost the same in order for one form to be paraphrased into the other form. However, there are at least three reasons for believing that they are not equivalent:

Firstly, they are pragmatically different. MA-questions are seen as being biased towards the overtly indicated proposition (*p*), whereas A-NOT-A questions are neutral as both propositions (*p* and ¬*p*) are indicated (Liing, 2014), barring the differences arising due to sequential order.

Secondly, they differ in terms of information structure. MA-questions can have focus on any of its constituents. For instance, in (11), either the subject 张三 *Zhāngsān*, the object 狗 *gǒu*, or the verb 喜欢 *xǐhuān* can be evaluated as containing focus. Should focus be required, the asker employs a specific prosodic clue and/or the focus marker 是 *shì*. (12) presents that different constituents in MA-questions can be freely clefted.

(12) a. 是 张三 喜欢 李四 吗 ?
shì Zhāngsān xǐhuān Lǐsì ma ?
SHI Zhangsan like Lisi MA PU

'Is it Zhangsan (and not anyone else) who likes Lisi?'

b. 张三 是 喜欢 李四 吗 ?
Zhāngsān shì xǐhuān Lǐsì ma ?
Zhangsan SHI like Lisi MA PU

'Is it that Zhangsan likes Lisi?'

c. 张三 喜欢 的 是 李四 吗 ?
Zhāngsān xǐhuān dè shì Lǐsì ma ?
Zhangsan like DE SHI Lisi MA PU

'Is it Lisi whom Zhangsan likes?'

This is because the scope of *mā* is not explicitly observable from the sentence itself. By contrast, *A*-NOT-*A* does not signal focus to any other elements but the structure itself (i.e., no ambiguity). The subject and the object in *A*-NOT-*A* questions cannot pass the cleft test exemplified in (12). In other words, *A*-NOT-*A* always bears focus (i.e., predicate focus).

Thirdly, they differ semantically. When a universal quantifier 都 *dōu* is used, a scope ambiguity happens with MA-questions but not with *A*-NOT-*A* questions, as shown in (13). (McCawley, 1994)

(13) a. 他们 都 喜欢 不 喜欢 开车 ?
tāmen dōu xǐhuān bù xǐhuān kāichē ?
they all like NOT like drive PU

'Do they all like to drive?'

b. 他们 都 喜欢 开车 吗 ?
tāmen dōu xǐhuān kāichē ma ?
they all like drive MA PU

'Do they all like to drive?' or
'Do all of them like to drive?'

2 HPSG Account

This section proposes two possible approaches to handling the *A*-NOT-*A* structure: 1) the morphological/lexical approach and 2) the syntactic approach.

2.1 Approach 1: Morphological/Lexical Approach

In this approach, the *A*-NOT-*A* structure is handled from the lexicon and thus its morphology. The *A*-NOT-*A* structure is dealt with as a single morphological word, and this allows us to treat the *A*-NOT-*A* element as a single predicate in the semantics. This approach aligns with the implementation chosen for reduplicated adjectives in ZHONG [|] (Fan et al., 2015). The treatment of the A-NOT-A structure as a “monolithic” morphological word also means that the modification of the A element is naturally prevented from happening. Constraints on the modification of the entire A-NOT-A structure are also much more easily implemented.

2.1.1 Parent/Super Lexical Rule

A super-type lexical rule (*a-not-a-lex-rule*) provides the general constraints and definition of the structure. This rule is responsible for the conversion of any lexical entry that can participate as A elements into A-NOT-A and thereafter provides the relevant information for the structure. Key sections of the *a-not-a-lex-rule* are illustrated below.

(14)
$$\begin{bmatrix} \textit{a-not-a-lex-rule} & \\ \text{SYNSEM} & \begin{bmatrix} \text{ASPECTED} & - \\ \text{SPART} & \textit{no-spart} \\ \text{LOCAL} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{MC} & \textit{luk} \\ \text{HEAD} & \boxed{1}\begin{bmatrix}\text{MODIFIABLE} & -\end{bmatrix} \\ \text{VAL} & \boxed{2}\end{bmatrix} \\ \text{CONT} & \begin{bmatrix}\text{INDEX} & \boxed{3} \\ \text{I-KEY} & \boxed{4} \\ \text{SF} & \textit{ques}\end{bmatrix}\end{bmatrix}\end{bmatrix} \\ \text{DTR} & \begin{bmatrix}\text{SYNSEM} & \begin{bmatrix}\text{BOUND} & - \\ \text{LOCAL} & \begin{bmatrix}\text{CAT} & \begin{bmatrix}\text{HEAD} & \boxed{1}\begin{bmatrix}\text{MODIFIABLE} & -\end{bmatrix} \\ \text{LENGTH} & \textit{one-or-two} \\ \text{VAL} & \boxed{2}\end{bmatrix} \\ \text{CONT} & \begin{bmatrix}\text{ASPECT} & \textit{non-aspect}\end{bmatrix}\end{bmatrix}\end{bmatrix}\end{bmatrix} \\ \text{C-CONT} & \begin{bmatrix}\text{ICONS} & \boxed{4}\left\langle\, ! \begin{bmatrix}\textit{focus} & \\ \text{IARG1} & \boxed{3} \\ \text{IARG2} & \boxed{3}\end{bmatrix} ! \,\right\rangle\end{bmatrix}\end{bmatrix}$$

The lexical rule indicates that the A-NOT-A structure bears the sentence force (SF) of *ques*. A feature type MODIFIABLE is used to state that the A-NOT-A structure cannot be modified. Focus is represented via ICONS (Individual CONstraints) (Song, 2014). The I-KEY feature points to ICONS, indicating that A-NOT-A is the focus of the sentence. IARG1 and IARG2 both point to the INDEX of the A-NOT-A structure itself.

Two lexical rules will inherit from this parent lexical rule. These two child lexical rules are for the A-不-A and A-没-A sub-patterns discussed earlier in Section 1.1.3.

2.1.2 Lexical Rule for A-不-A sub-pattern

This child lexical rule handles the A-不-A sub-pattern, and inherits from *a-not-a-lex-rule*. The additional constraints are indicated below:

(15)
$$\begin{bmatrix} \textit{a-not-a-bu-lex-rule} & \\ \text{SYNSEM} \mid \text{CONT} & \begin{bmatrix} \text{ASPECT} & \textit{non-aspect} \\ \text{LZG} & \textit{no-lzg} \end{bmatrix} \end{bmatrix}$$

The ASPECT has been indicated as *non-aspect*, which prevents sentence-final particles from modifying the structure. The LZG feature is given a value of *no-lzg*, which prevents the aspectual markers from modifying the structure.

2.1.3 Lexical Rule for A-没-A sub-pattern

This child lexical rule handles the A-没-A sub-pattern, and inherits from *a-not-a-lex-rule*. The additional constraints are indicated below:

(16)
$$\begin{bmatrix} \textit{a-not-a-mei-lex-rule} & \\ \text{SYNSEM} \mid \text{CONT} & \begin{bmatrix} \text{ASPECT} & \textit{imperfective} \\ \text{LZG} & \textit{guo} \end{bmatrix} \end{bmatrix}$$

The ASPECT feature is given the value of *imperfective*. As this sub-pattern allows the experiential aspectual marker 过 *guò* to modify the structure, we provide the LZG feature with a value of *guo*.

2.1.4 Handling A′-NOT-A

From (1b), duplicated here as (17), we see an example of the A′-NOT-A sub-pattern.

(17) Contracted: A′-NOT-A

张三	喜	不	喜欢	狗	?
Zhāngsān	xǐ	bù	xǐhuān	gǒu	?
Zhangsan	like	NOT	like	dog	PU

To recap, this pattern exhibits partial reduplication, where only the first syllable/character of A_1 is reduplicated. Nevertheless, apart from surface form, it is identical to its fully reduplicated counterpart. As such, this pattern is first transformed into the fully reduplicated pattern before being handled by the lexical rules. The mechanism for this is described in Section 2.1.6.

2.1.5 Sample Derivation

Using the sentence 张三 喜欢 不 喜欢 狗 ? 'Does Zhangsan like dogs?', we derive the MRS in (18):

(18)

$$\begin{bmatrix}
\text{INDEX} & \boxed{2}\begin{bmatrix}\text{SF} & \textit{ques}\\ \text{ASPECT} & \textit{non-aspect}\end{bmatrix} \\
\text{RELS} & \left\langle \begin{array}{l}
\begin{bmatrix}\textit{named\_rel} & \\ \text{LBL} & \boxed{4} \\ \text{CARG} & \text{‘张三’} \\ \text{ARG0} & \boxed{3}\end{bmatrix},
\begin{bmatrix}\textit{proper\_q\_rel} & \\ \text{LBL} & \boxed{6} \\ \text{ARG0} & \boxed{3} \\ \text{RSTR} & \boxed{7} \\ \text{BODY} & \boxed{8}\end{bmatrix}, \\
\begin{bmatrix}\textit{\_喜欢\_v\_1\_rel} & \\ \text{LBL} & \boxed{1} \\ \text{ARG0} & \boxed{2} \\ \text{ARG1} & \boxed{3} \\ \text{ARG2} & \boxed{9}\end{bmatrix},
\begin{bmatrix}\textit{\_狗\_n\_1\_rel} & \\ \text{LBL} & \boxed{10} \\ \text{ARG0} & \boxed{9}\end{bmatrix},
\begin{bmatrix}\textit{exist\_q\_rel} & \\ \text{LBL} & \boxed{11} \\ \text{ARG0} & \boxed{9} \\ \text{RSTR} & \boxed{12} \\ \text{BODY} & \boxed{13}\end{bmatrix}
\end{array}\right\rangle \\
\text{HCONS} & \left\langle \begin{bmatrix}\textit{qeq} & \\ \text{HARG} & \boxed{0} \\ \text{LARG} & \boxed{1}\end{bmatrix},
\begin{bmatrix}\textit{qeq} & \\ \text{HARG} & \boxed{7} \\ \text{LARG} & \boxed{4}\end{bmatrix},
\begin{bmatrix}\textit{qeq} & \\ \text{HARG} & \boxed{1}\textit{2} \\ \text{LARG} & \boxed{1}\textit{0}\end{bmatrix}, \right\rangle \\
\text{ICONS} & \left\langle \begin{bmatrix}\textit{focus} & \\ \text{IARG1} & \boxed{2} \\ \text{IARG2} & \boxed{2}\end{bmatrix}, \right\rangle
\end{bmatrix}$$

The semantic head $\boxed{2}$ has [SF *ques*], which indicates that the sentence is interrogative. Within the semantics, the *A*-NOT-*A* structure has only a single predicate *_喜欢_v_1_rel*. The element in ICONS is specified as *focus*, and both IARG1 and IARG2 are coindexed with the INDEX of the verb. This means that the *A*-NOT-*A* structure is the focus within the clause.

In most areas, the MRS for the *A*-NOT-*A* structure is close to that of MA-questions, with a number of key differences explained in an earlier section pertaining to the areas such as the focus and the aspect. Barring these, MA-questions can technically be generated from our implementation of the *A*-NOT-*A* structure, and can likewise be provided alongside *A*-NOT-*A* questions as suitable candidates during machine translation.

2.1.6 Implementation in Zhong [|]

In a nutshell, the input is first cleaned up by a regular expression preprocessor (REPP) and readied for parsing. The cleaning up includes removal of spaces left over from segmentation (Eg: 高 不 高 → 高不高), and replacing the reduplicated parts with the character 々 [1] (Eg: 高不高 → 高不々). The segment 不々 is treated by the parser as a suffix, which it can then remove and reduce the structure to just the *A* element, and subsequently match with its appropriate lexical entry (Eg: 高不々 → 高). This lexical item will then be passed through the *a-not-a-lex-rule* and the relevant child lexical rule, and will then be given the features and semantics of the *A*-NOT-*A* structure.

[1] This character is adopted from Japanese, which uses it to indicate the reduplication of the character that precedes it.

As explained earlier in Section 2.1.4, the A′-NOT-A sub-pattern (the contracted pattern) is identical — apart from surface form — to its fully reduplicated counterpart. As such, the REPP will pick up these contracted patterns in the input and transform them into their fully-reduplicated forms, removing any spaces along the way (Eg: 喜 不 喜欢 → 喜欢不喜欢), and also replace the reduplicated element with the character 々 (Eg: 喜欢不喜欢 → 喜欢不々). As with the above, the structure is then reduced to the A element, and then be passed through the *a-not-a-lex-rule*.

It should be noted that the implementation is done based on the functions and limitations of the system, and it does not reflect any assumptions on the actual parsing of the structure by a speaker.

2.1.7 Limitations

This method does not allow us to constrain the objects to be identical in the AO-NOT-AO pattern, as the O elements can be diverse and be too vast to feasibly implement. The O elements can also, potentially, be of an arbitrarily long length as long as it is a grammatically correct verb phrase. However, such long sentences are not necessarily accepted by speakers due to the cumbersome nature of it, even if they do not violate any grammatical rules.

Also, because of the treatment of the A-NOT-A structure as a single morphological word, the formation of the structure remains opaque to the grammatical system, which only sees the structure as a single lexical entry.

This approach does not allow us to cover A-MEI-A patterns where the A element is modified by *guò*, as illustrated in (19):

(19) a. (Unmodified)
吃 没 吃
chī méi chī
eat MEI eat

b. (With Experiential GUO)
吃 过 没 吃 过

Using this approach, it will require that this pattern be also generated in the morphology, or automatically detected when parsing, and then given additional rules that take into account the aspectual marker.

An initial issue that we had believed might arise from this approach was that the lexicon could become very large if each A element were to have a separate lexical entry for its respective A-NOT-A form(s). However, with the A-NOT-A structure now being automatically detected and pre-processed (such that separate lexical entries are no longer needed), this disadvantage is largely removed.

2.2 Approach 2: Syntactic Approach

The syntactic approach builds the A-NOT-A structure as three components: A_1, NOT and A_2.

2.2.1 Characters

The A elements in A-NOT-A are full or partial reduplicates of each other. One such form is that only the first character of A_1 is reduplicated. With this in mind, we introduce new feature types to the lexicon entries, as underlined in (20):

(20)
$$\begin{bmatrix} \textit{+vjp} & \\ \text{STEM} & \boxed{1} \\ \underline{\text{BOUND}} & \textit{luk} \\ \underline{\text{SPART}} & \textit{spart} \\ \text{HEAD} & \begin{bmatrix} \underline{\text{CHAR}} & \begin{bmatrix} \textit{char} & \\ \underline{\text{FCHAR}} & \textit{string} \\ \underline{\text{WCHAR}} & \boxed{1} \\ \underline{\text{LENGTH}} & \textit{length} \end{bmatrix} \\ \underline{\text{P-KEY}} & \boxed{2} \end{bmatrix} \\ \text{PRED} & \boxed{2} \end{bmatrix}$$

The feature types WCHAR and FCHAR specify all characters and the first character of a lexical entry, respectively. The feature WCHAR is identical to the STEM of the lexical entry. Next, the LENGTH specifies that an entry has *one* or *more-than-one* character. Finally, the *luk* feature BOUND specifies if an entry is a bound or non-bound form.[2] This is to ensure that one-character A_1 forms of a multi-character word are not used outside of A-NOT-A, as they are not independent morphemes. The P-KEY feature is identical to the PRED feature so as to block homographs from co-occurring as the A elements. An example of such a homograph is 撒 *sā / sǎ*, which can mean ‘let go’ and ‘scatter’, respectively. These two will have different PRED values: *_撒_v_1_rel* and *_撒_v_2_rel*. Finally, the SPART feature indicates the type of sentence-final particle that can co-occur with the structure.

To provide a clearer idea, the entries in (21) illustrate the bound and non-bound forms of 喜欢, respectively. As they are identical to each other apart from length, they take the same PRED value.

[2] The *luk* constraint consists of three components, such as $+$, $-$, and *na* (not-applicable).

(21) a.
$$\begin{bmatrix} \text{喜} & \\ \text{STEM} & \boxed{1}\left\langle \text{‘喜’} \right\rangle \\ \text{BOUND} & + \\ \text{CHAR} & \begin{bmatrix} \text{FCHAR} & \text{‘喜’} \\ \text{WCHAR} & \boxed{1} \\ \text{LENGTH} & \textit{one} \end{bmatrix} \\ \text{PRED} & \text{\_喜欢\_}\textit{v\_rel} \end{bmatrix}$$

b.
$$\begin{bmatrix} \text{喜欢} & \\ \text{STEM} & \boxed{1}\left\langle \text{‘喜欢’} \right\rangle \\ \text{CHAR} & \begin{bmatrix} \text{FCHAR} & \text{‘喜’} \\ \text{WCHAR} & \boxed{1} \\ \text{LENGTH} & \textit{more-than-one} \end{bmatrix} \\ \text{PRED} & \text{\_喜欢\_}\textit{v\_rel} \end{bmatrix}$$

The use of the features FCHAR and WCHAR to access the characters of a word is due to a limitation in the present system. It is expected that future iterations will store the characters as a list, and that the characters will be accessed via their indices.

2.2.2 Supertype

The present analysis uses the NOT element as the “origin” of the A-NOT-A structure, which will then select the A elements. A generic A-NOT-A lexical type A-NOT-A-ADV-LEX is defined for this NOT element. As shown in (22), the element of MOD goes for A_1, the element of COMPS goes for A_2, and both take +*vjp* (verb, adjective or preposition) as their head type. Both A elements are semantically identical, so they take the same SUBJ and COMPS, and share the same ASPECT and P-KEY values. A_1, being the head of the structure, bears the sentential force (SF) of *ques*.

(22)
$$\left[\begin{array}{ll}
\textit{a-not-a-adv-lex} & \\
\text{POSTHEAD} & + \\
\text{MOD} & \left\langle \left[\begin{array}{ll} \textit{+vjp} & \\ \text{SF} & \textit{ques} \\ \text{I-KEY} & \boxed{1} \\ \text{INDEX} & \boxed{2} \\ \text{P-KEY} & \boxed{3} \\ \text{ASPECT} & \boxed{4} \\ \text{SUBJ} & \boxed{5} \\ \text{COMPS} & \boxed{6} \\ \text{SPART} & \textit{no-spart} \end{array}\right] \right\rangle \\
\text{COMPS} & \left\langle \left[\begin{array}{ll} \textit{+vjp} & \\ \text{P-KEY} & \boxed{3} \\ \text{ASPECT} & \boxed{4} \\ \text{SUBJ} & \boxed{5} \\ \text{COMPS} & \boxed{6} \\ \text{SPART} & \textit{no-spart} \\ \text{BOUND} & - \end{array}\right] \right\rangle \\
\text{ICONS} & \left\langle\ !\ \boxed{1}\left[\begin{array}{ll} \textit{focus} & \\ \text{IARG2} & \boxed{2} \end{array}\right]\ !\ \right\rangle
\end{array}\right]$$

The focus meaning is represented via Individual CONStraint (Song, 2014). Its I-KEY feature points to the ICONS element, which indicates that the A-NOT-A structure is the focus of the sentence. Thus, IARG2 in ICONS is identical to INDEX of A_1. In addition, A_2 has the constraint [BOUND $-$], as bound forms cannot participate as A_2.

(23) a.
$$\left[\begin{array}{ll}
\textit{不\_polar\_basic} & \\
\text{STEM} & \left\langle \text{‘不’} \right\rangle \\
\text{COMPS} & \left\langle \left[\begin{array}{ll} \text{ASPECT} & \textit{non-aspect} \\ \text{LZG} & \textit{no-lzg} \end{array}\right] \right\rangle
\end{array}\right]$$

b.
$$\left[\begin{array}{ll}
\textit{没\_polar\_basic} & \\
\text{STEM} & \left\langle \text{‘没’} \right\rangle \\
\text{COMPS} & \left\langle \left[\begin{array}{ll} \text{ASPECT} & \textit{imperfective} \\ \text{LZG} & \textit{guo} \end{array}\right] \right\rangle
\end{array}\right]$$

As mentioned before, the NOT element can be either 不 *bù* or 没 *méi*, depending on the A elements' aspectual property. As we see in (23), the aspectual properties of their A elements are indicated in their respective COMPS' ASPECT constraints. When the NOT element is *bù*, the A elements cannot co-occur with any of the LE-ZHE-GUO markers (*no-lzg*), whereas when the NOT element is *méi*, it can co-occur with *guò*.

As we have seen in Section 1.1, there are a few patterns for the A-NOT-A structure. With the generic A-NOT-A lexical type we defined in (22), we create two sub-types for A-NOT-A and A'-NOT-A, as shown in Section 2.2.3 and Section 2.2.4.

2.2.3 Subtype: A-NOT-A

The sub-type for the basic form is as follows:

(24)
$$\begin{bmatrix} \textit{a-not-a-basic-adv-lex} & \\ \text{MOD} & \left\langle \begin{bmatrix} \text{LIGHT} & + \\ \text{WCHAR} & \boxed{1} \\ \text{BOUND} & - \end{bmatrix} \right\rangle \\ \text{COMPS} & \left\langle \begin{bmatrix} \text{LIGHT} & + \\ \text{WCHAR} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

The basic form of A-NOT-A contains two identical A elements, as shown in (25):

(25) 张三 喜欢 不 喜欢 狗 ?
Zhāngsān xǐhuān bù xǐhuān gǒu ?
Zhangsan like NOT like dog PU

As such, both MOD (A_1) and COMPS (A_2) have identical WCHAR values. The MOD is constrained to [BOUND −] to block it from parsing the contracted form. Lastly, both MOD and COMPS are constrained with [LIGHT +] such that the A-NOT-A structure will be treated as a single lexical item instead of as a phrase (cf. Abeillé & Godard (2001)). The constraints presented so far will account for the following ungrammatical sentences:

(26) a. *张三 讨厌 不 喜欢 狗 ?
Zhāngsān tǎoyàn bù xǐhuān gǒu ?
Zhangsan hate NOT like dog PU

b. *张三 喜 不 喜 狗 ?
Zhāngsān xǐ bù xǐ gǒu ?
Zhangsan like NOT like dog PU

2.2.4 Subtype: A′-NOT-A

The sub-type for the contracted form is as follows:

(27)
$$\begin{bmatrix} \textit{a-not-a-contracted-adv-lex} & \\ \text{MOD} & \left\langle \begin{bmatrix} \text{LIGHT} & + \\ \text{WCHAR} & \boxed{1} \\ \text{BOUND} & + \\ \text{LENGTH} & \textit{one} \end{bmatrix} \right\rangle \\ \text{COMPS} & \left\langle \begin{bmatrix} \text{LIGHT} & + \\ \text{FCHAR} & \boxed{1} \\ \text{LENGTH} & \textit{more-than-one} \end{bmatrix} \right\rangle \end{bmatrix}$$

In the A′-NOT-A variant, only the first character of A_1 is reduplicated, as shown in (28):

(28) 张三 喜 不 喜欢 狗 ?
Zhāngsān xǐ bù xǐhuān gǒu ?
Zhangsan like NOT like dog PU

As such, the LENGTH value of MOD (A_1) is constrained to *one*, while its WCHAR — being a single-character word — is identical to the FCHAR of COMPS (A_2). As it is a bound form, we constrained it to [BOUND +]. In order to block it from parsing A-NOT-A sentences where the A elements are both single-character words, COMPS is given an additional constraint of *more-than-one* to its LENGTH feature. Finally, as with the basic form, both MOD and COMPS are indicated as [LIGHT +] to treat it as a single lexical item instead of a phrase.

2.2.5 AO-NOT-AO

The constraints for this type is shown in (29):

(29)
$$\begin{bmatrix} \textit{ao-not-ao-adv-lex} & \\ \text{MOD} & \left\langle \begin{bmatrix} \textit{verb} & \\ \text{LIGHT} & - \\ \text{WCHAR} & \boxed{1} \end{bmatrix} \right\rangle \\ \text{COMPS} & \left\langle \begin{bmatrix} \textit{verb} & \\ \text{LIGHT} & - \\ \text{WCHAR} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

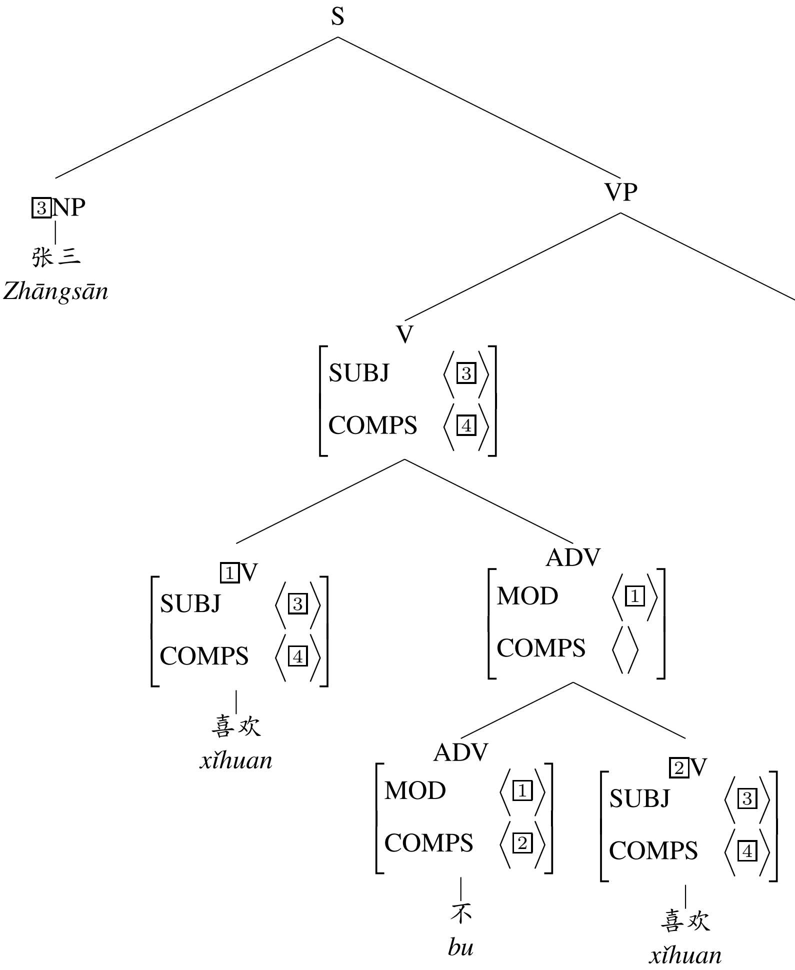
The AO-NOT-AO form's A elements are restricted to being verbs, and they are phrases instead of words. As with the basic form, the WCHAR value of the two

A elements' verb heads are identical. But unlike the basic form, the *AO*-NOT-*AO* structure is treated as a phrase, and is thus constrained to [LIGHT −].[3]

2.2.6 Sample Derivation

Using the sentence 张三 喜欢 不 喜欢 狗 ? 'Does Zhangsan like dogs?', we derive the tree in (30):

(30)



We see the NOT element (the ADV) selecting for MOD (A_1) and COMPS (A_2). It first combines with its COMPS via the *head-comp-phrase* rule, and then with the MOD via the *head-adj-scop-phrase* rule. As we indicate the *A*-NOT-*A* structure to be [LIGHT +], it combines to form only a V (instead of VP). The SUBJ and COMPS of both *A* elements are identical, and the *A*-NOT-*A* structure combines with the object *gǒu* via *head-comp-phrase*, before finally combining with *Zhāngsān* via *subj-head-phrase*.

[3] The current analysis does not constrain the objects (*O* in *AO*-NOT-*AO*) to be identical. The current analysis sometimes provides unwanted over-generation. These are left to future work.

(31)

$$\begin{bmatrix}
\text{INDEX} & \boxed{2}\begin{bmatrix}\text{SF} & \textit{ques}\\ \text{ASPECT} & \textit{non-aspect}\end{bmatrix}\\
\text{RELS} & \left\langle \begin{array}{l}
\begin{bmatrix}\textit{named\_rel} & \\ \text{LBL} & \boxed{4}\\ \text{CARG} & \text{”张三”}\\ \text{ARG0} & \boxed{6}\end{bmatrix}, \begin{bmatrix}\textit{proper\_q\_rel} & \\ \text{LBL} & \boxed{7}\\ \text{ARG0} & \boxed{6}\\ \text{RSTR} & \boxed{8}\\ \text{BODY} & \boxed{9}\end{bmatrix},\\
\begin{bmatrix}\textit{\_喜欢\_v\_1\_rel} & \\ \text{LBL} & \boxed{10}\\ \text{ARG0} & \boxed{2}\\ \text{ARG1} & \boxed{6}\\ \text{ARG2} & \boxed{11}\end{bmatrix}, \begin{bmatrix}\textit{\_不\_r\_rel} & \\ \text{LBL} & \boxed{1}\\ \text{ARG0} & \boxed{12}\\ \text{ARG1} & \boxed{13}\end{bmatrix}, \begin{bmatrix}\textit{\_喜欢\_v\_1\_rel} & \\ \text{LBL} & \boxed{14}\\ \text{ARG0} & \boxed{2}\\ \text{ARG1} & \boxed{6}\\ \text{ARG2} & \boxed{11}\end{bmatrix},\\
\begin{bmatrix}\textit{\_狗\_n\_1\_rel} & \\ \text{LBL} & \boxed{15}\\ \text{ARG0} & \boxed{11}\end{bmatrix}, \begin{bmatrix}\textit{exist\_q\_rel} & \\ \text{LBL} & \boxed{16}\\ \text{ARG0} & \boxed{11}\\ \text{RSTR} & \boxed{17}\\ \text{BODY} & \boxed{18}\end{bmatrix}
\end{array}\right\rangle\\
\text{HCONS} & \left\langle \begin{bmatrix}\textit{qeq} & \\ \text{HARG} & \boxed{8}\\ \text{LARG} & \boxed{4}\end{bmatrix}, \begin{bmatrix}\textit{qeq} & \\ \text{HARG} & \boxed{13}\\ \text{LARG} & \boxed{10}\end{bmatrix}, \begin{bmatrix}\textit{qeq} & \\ \text{HARG} & \boxed{17}\\ \text{LARG} & \boxed{15}\end{bmatrix}\right\rangle\\
\text{ICONS} & \left\langle \begin{bmatrix}\textit{focus} & \\ \text{IARG1} & \boxed{2}\\ \text{IARG2} & \boxed{2}\end{bmatrix}\right\rangle
\end{bmatrix}$$

The semantic relations are indicated in the MRS. A_1 and A_2 are given the same indexes: ARG0 for the verb itself, ARG1 for the subject *Zhāngsān*, and ARG2 for the object *gǒu*. This means that they share the same argument structure. The second element in the HCONS list is responsible for the scope of negative operator *bù*: HARG is co-indexed with the ARG1 of the scopal modifier (i.e. $\boxed{13}$), and LARG is co-indexed with the label of of A_1 (i.e. $\boxed{10}$). The element in the ICONS list is specified as *focus*, and the values of IARG1 and IARG2 are both co-indexed with the verb's INDEX. This means that the A-NOT-A structure is associated with focus within the clause. Finally, the semantic head $\boxed{2}$ has [SF *ques*], which indicates that the sentence is interrogative.

2.2.7 Limitations

This approach demonstrates the underlying syntactic rules of the A-NOT-A structure. However, in doing so, it offers a semantic analysis that contains two predicates — each of the A elements are treated as a separate predicate, with constraints to make them identical. This was found to be an unsatisfactory semantic analysis of the structure, which should instead have only a single predicate since it is not a true disjunctive. This analysis, however, does not permit the presence of only one predicate.

Secondly, the present approach does not block the modification of the *A*-NOT-*A* structure itself, and so cause the over-generation of sentences such as 张三 [很 [喜欢 狗 不 喜欢]] 狗? , where the *A*-NOT-*A* structure is modified by the degree adverb 很 'very'.

Thirdly, like the first approach, this approach does not have the ability to restrict the *O* (object) elements in the *AO*-NOT-*AO* structure to be identical due to limitations of the system, and would thus cause over-generation of structures like 张三 喜欢 **狗**$_{O1}$ 不 喜欢 **猫**$_{O2}$? .

3 Conclusion

In this paper, we provided two HPSG accounts for the *A*-NOT-*A* structure in Mandarin Chinese, approaching it syntactically as well as morphologically, and looked at their respective strengths and weaknesses. Overall, the morphological approach is preferred, as it provides more accurate semantics, and we are better able to restrict the modification of the entire structure, even if the formation of the *A*-NOT-*A* structure is not as transparently illustrated with this approach.

Both methods are unable to reliably account for the *AO*-NOT-*AO* structure as it remains non-trivial to constrain *O* to be identical, or to account for the arbitrariness in length that *O* can be. It is hoped that future work will be able to cover this particular pattern.

Acknowledgements

We would like to thank participants of the Linguistic Analysis Design (LAD) sessions (DELPH-IN '15 summit), attendees of the HPSG 2015 conference as well as all others who had given their useful input.

This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Abeillé, Anne & Daniele Godard. 2001. A Class of "Lite" Adverbs in French. In Joaquim Camps & Caroline R. Wiltshire (eds.), *Romance syntax, semantics and l2 acquisition: Selected papers from the 30th linguistic symposium on romance languages, gainesville, florida, february 2000*, 9–26. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan A. Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language & Computation* 3(4). 281–332.

Fan, Zhenzhen, Sanghoun Song & Francis Bond. 2015. Building Zhong [|], a Chinese HPSG meta-grammar. In *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, 97–110.

Liing, Woan-Jen. 2014. *How to Ask Questions in Mandarin Chinese*: City University of New York dissertation.

McCawley, James D. 1994. Remarks on the Syntax of Mandarin Yes-No Questions. *Journal of East Asian Linguistics* 3(2). 179–194.

Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, IL: The University of Chicago Press.

Song, Sanghoun. 2014. *A Grammar Library for Information Structure*: University of Washington dissertation.

Tseng, Wen-Hsin Karen. 2009. A Post-syntactic Approach to the A-not-A Questions. *UST Working Papers in Linguistics, Graduate Institute of Linguistics* 5(National Tsing Hua University).

Part II
Contributions to the Workshop

Degrees of affectedness and verbal prefixation in Abui (Papuan)

František Kratochvíl
Nanyang Technological University, Singapore

Benidiktus Delpada
Nanyang Technological University, Singapore

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 216–233

http://csli-publications.stanford.edu/HPSG/2015

Abstract

This paper deals with the encoding of affectedness in Abui, a Papuan language of Indonesia. Abui is a head-marking language of the rare type where the verbs are marked for their undergoer arguments (S_o, O) formally split into several subtypes. This marking has been previously analyzed as a type of semantic alignment sensitive among others to affectedness. Affectedness is understood here as a scalar property delimiting the predicate (following Tenny 1987 and Beavers 2011). The paper explores the structure of the affectedness scale for Abui, comparing the functions and meaning of three types of person prefix paradigms. We show that verbs with similar meaning, encoding the same type of change (in Beavers' terms) can differ in their entailments. We also show that there may be additional dimensions in which affectedness can be measured, such as affected agents, and that the interpretation of the degree on the affectedness scale interacts with instigator's (source of force) status on the referential hierarchy. While human agents in some cases allow lower degrees of affectedness, the inanimate forces select the maximal degree reading. We conclude, that despite a considerable amount of fluidity of marking (Fedden et al. 2013, 2014), the shifts in degree of affectedness can be predicted as lowering of the degree stipulated for the predicate.

1 Introduction

Abui is a Papuan language of the Alor-Pantar Archipelago of Eastern Indonesia (Alor branch of the Timor-Alor-Pantar family, Holton et al. 2012) spoken by over 17,000 people. Abui has a relatively simple phonemic inventory, with phonemic vowel length, lexical and grammatical tone. The tone system is presently not fully understood and the tones are not marked here. The language is head-marking, verb-final, and moderately agglutinative. Negation particles occur post-verbally and verb serialisation and clause chaining are extensive. The grammatical relations have been described as semantic alignment detected in both free pronouns and person prefixes (Kratochvíl 2007, 2011, 2014). Fedden et al. (2013, 2014) show that Abui verbs are highly fluid in argument selection and indexing, compared to related languages. The system is complex, and we do not presently fully understand the features predicting the distribution of person marking prefixes. The system interacts with the Abui aspectual system, expressed through a variety of morphosyntactic operations such as stem modification, suffixation and verb serialisation.

[†] We gratefully acknowledge the hospitality of the Abui community, as well as research funding from Nanyang Technological University (Singapore) through a Singapore Ministry of Education Tier 2 Grant MOE2013-T2-1-016. Some of the reported research has been done earlier with funding from La Trobe University (Australia), and Leiden University (the Netherlands). The data presented in this paper comes from the Abui corpus (roughly 200,000 words) compiled by the authors. Delpada is a native speaker of Abui, Kratochvíl has worked on Abui since 2003. Glosses follow the Leipzig Glossing Conventions with the following additions: AD addressee-perspective, AGT agentive pronoun, ASSOC associative, CONT continuative, EVID evidential, INC inchoative, LNK linker, MOD modal, PRIOR priorative, SEQ sequential, SIM simultaneous, SPC specific determiner.

Abui verbs agree in person and number with their undergoer arguments (O, S_o). Person prefixes are listed in Table 1. Number is distinguished in the first and second person only. Distributive forms have both distributive and reciprocal reading. The third person is split between the *d-* series prefixes (indexing the A argument), and the *h-* series (non-A argument). Five prefix paradigms distinguish five basic types of undergoer arguments. For more details about their use, see Kratochvíl (2011, 2014). The glosses are not to be taken to literally indicate the semantic role of the argument.

Table 1: Abui person prefixes

PERSON	I	II	III	IV	V
1SG	*na-*	*no-*	*ne-*	*noo-*	*nee-*
2SG	*a-*	*o-*	*e-*	*oo-*	*ee-*
3UND	*ha-*	*ho-*	*he-*	*hoo-*	*hee-*
3AGT	*da-*	*do-*	*de-*	*doo-*	*dee-*
DISTR	*ta-*	*to-*	*te-*	*too-*	*tee-*
1PL.EXCL	*ni-*	*nu-*	*ni-*	*nuu-*	*nii-*
1PL.INCL	*pi-*	*pu-/po-*	*ni-*	*puu-/poo-*	*pii-*
2PL	*ri-*	*ru-/ro-*	*ri-*	*ruu-/roo-*	*rii-*
GLOSS	PAT	REC	LOC	GOAL	BEN

Although a number of predicates are rigid in their argument selection, most verbs are quite fluid. To illustrate the fluidity, the paradigm of the verb *wik* ∼ *wit* 'carry in hands/arms' is given in (1-6). Each person prefix series indicates a different degree/degrees of affectedness (Kratochvíl 2014: 558-559). The prefixes *ha-* (3UND.PAT) and *he-* (3UND.LOC) index the carried theme in (1, 2, 4). The prefix *do-* (3AGT.REC) indexes the carrier, who is affected by his own action (3), *hee-* (3UND.BEN) the benefactor (4), *hoo-* (3UND.GOAL) someone who is given something to carry (5). Note also that the theme does not have to be indexed, although it is definite, when the sentence does not contain an agent argument (6).

(1) Bui kaai ha-wik
PN [dog]$_{\text{PAT}}$ 3UND.PAT-carry.in.arms.IPFV
'Bui is carrying her dog in her arms.'

(2) Bui bataa tuku mii de-wiil hee-r ba
PN wood piece take.PFV 3AGT.AL-child 3UND.BEN-reach SIM
ha-wik
3UND.PAT-carry.in.arms.IPFV
'Bui made a doll from a piece of wood and carries it around.'

(3) akuun nuku, dikang di de-naamang
morning one again 3AGT 3AGT.AL-cloth
do-witi, pun namei he-yaari
3AGT.REC-carry.in.arms.PFV field prepare.field 3UND.LOC-go.PFV-PFV
'one morning, he again took his clothing and went to work in the field.'

(4) a-taáng do mi he-wik,
2SG.INAL-hand PROX use 3UND.LOC-carry.in.arms.IPFV
hee-wik-e!
3UND.BEN-carry.in.arms.IPFV-PROG
'carry it in your hands, carry (it) for him!'

(5) na ara mii hoo-wik
1SG.AGT firewood take.PFV 3UND.GOAL-carry.in.arms.IPFV
'I give him firewood to carry.'

(6) sura foka do baai wik-e?
book big PROX also carry.IPFV-PROG
'should this big book be carried too?'

The basic meaning of the root is not a good predictor of its inflectional behavior, as shown in (7-12). The root *rumai* can be interpreted as a state (7) or as an inchoative (8-9). Further, some of the combinations may be used in an idiomatic way (10-12), where the basic meaning is extended based on a metaphor (here STRONG >RELY ON, PUT FAITH IN, DERIVE STRENGTH FROM):

(7) di rumai natet hare eel baai rumai
3AGT be.strong stand.up.PFV so 2SG.TOP also be.strong
'He is firm, so you too be strong!' [E14BD.A63]

(8) ni-maama wee lik ha-rumai
1PL.EXCL.AL-father ASSOC platform 3UND.PAT-strengthen
'my father and his friends are strengthening the platform.' [E14BD.A64]

(9) no-rumai
1SG.REC-be.strong
'I feel strong (and I took the decision to feel so).' [E14BD.A65]

(10) he-tanga nu a he-rumai naha!
3UND.AL-word SPC 2SG.AGT 3UND.LOC-strengthen not
Do not put your trust in his words!' [E14BD.A66]

(11) moku kaik loku di needo noo-rumai
kid orphan PL 3AGT 1SG.FOC 1SG.GOAL-rely
'The orphaned children rely on me, have support in me.' [E14BD.A67]

(12) na ama wala hee-rumai naha
1SG.AGT person just 3UND.BEN-have.strength not
'I don't expect any support from anyone.' [E14BD.A68]

2 Affectedness

Affectedness has been invoked by typologists to define prototypical undergoers and it is undestood as the property of simply undergoing change (literature dealing with alignment, case, transitivity). In formal semantic work affectedness is understood as a scalar property delimiting the predicate, starting from Tenny (1987).

The most influential among the typological approaches is Tsunoda (1981), who identified typical verb class boundaries through what he termed 'verb effectiveness hierarchy' (p. 395), shown in (13):

(13) effective action > perception > pursuit > knowledge > feeling > relation

If the transitive frame (construction) can be used for classes on the right of that hierarchy, it can be used also for classes on the left of them. Tsunoda's proposal has been modified by Christian Lehman, who highlighted the two-dimensional nature of affectedness and its internal scale: total ∼ partial ∼ minimal (1991:221). Importantly, the internal scale is not consistent across verb classes (effected object are created, and therefore show no grades of existence). The most recent elaboration of the hierarchy comes from the ValPal Project (Hartmann et al., 2013), which has revised the hierarchy (Malchukov & Comrie, 2015).

The semantic approaches have discussed affectedness in relation to pre-posed NPs, middles and subsume it under aspect (for example Tenny (1987)):

> Affectedness may be defined as the property of a verb, such that it describes a situation or happening that can be delimited by the direct argument of the verb. Affectedness verbs describe events, which are 'measured out' and delimited by their direct arguments. Affectedness defined in this way as an aspectual property more adequately characterizes the verbs that allow middles and noun phrase passives than the definition of affectedness based on the notion of 'undergoing change'. (Tenny 1987:75)

In Tenny's framework, there are five verb classes for which the notion affectedness is relevant. These are (i) *verbs of creation, consumption and path-motion*, (ii) *verbs of physical change*, (iii) *verbs of abstract change*, (iv) *achievement verbs*, and (v) *verbs of locomotion* (1987:105).

Beavers (2011) reorganised Tenny's framework in a two-dimensional space for the encoding of affectedness. One dimension represents the types of change, and the other the degree of change. With respect to the types of change, Beavers identifies the following six types of change, restricting the discussion to transitive verbs:

(a) *x* changes in some observable property (clean/paint/delouse/fix/break *x*)

(b) *x* transforms into something else (turn/carve/change/transform *x* into *y*)

(c) *x* moves and stays at some location (move/push/angle/roll *x* into *y*)

(d) *x* is physically impinged (hit/kick/punch/rub/slap/wipe/scrub/sweep *x*)

(e) *x* goes out of existence (delete/eat/consume/reduce/devour *x*)

(f) *x* comes into existence (build/design/construct/create *x*)

Beavers measures the degree of affectedness along a scale in his Affectedness Hierarchy, shown in (14) and proposes a number of semantic tests characterising each degree. Predicates listed in (a-c, e, f) combine with patients and entail a resulting state. Predicates in (d) are non-patient force-recipients which do not always entail a resulting state.

(14) Affectedness Hierarchy (Beavers 2011:359)

TEST	quantized	non-quantized	potential	unspecified
telic	+	-	-	-
change entailed	+	+	-	-
result XP	+	+	+/-	-
happened to ***x***	+	+	+	-
dynamic	+	+	+	+/-
result variation	low	low/high	high	n.a.

Importantly, Beavers excludes intransitive predicates from his discussion of affectedness, while in Tenny's framework includes preposed NPs (middles and DP-passives).

3 Affectedness in Abui

Affectedness is relevant for several of the person prefix paradigms listed in Table 1.[1] We will restrict our discussion to transitive verbs that combine with three paradigms. Section 3.1 will discuss verbs that combine exclusively with the PAT prefix paradigm. Section 3.2 examines the meaning of the PAT∼ LOC alternation. Section 3.3 discusses another type of affectedness, not covered in Beavers' framework, but very common in Abui, where the scale of affectedness is applied to the agent who is in some way also experiencing or affected by the action.

[1] The data discussed in this paper is drawn mostly from a purpose built database of Abui inflectional paradigms (v. 2015). The database contains attested combinations of over 300 verbal roots and person prefixes. The database contains the most frequent verbs from the corpus and also the 80 verbs covered by the VALPAL database. In this paper we selected verbs that are compatible with the person prefix that marks affected undergoers (PAT) and whether the verb allows the LOC-prefix alternation.

3.1 Abui PAT-verbs

A high degree of affectedness is marked by the first prefix paradigm (PAT), however, the relationship between the marking of the degree of affectedness and the prefix is not straightforward. We will start out discussion with the verbs of OBSERVABLE CHANGE, following Beavers' classification discussed above. Verbs denoting OBSERVABLE CHANGE that are compatible with the PAT prefix fall apart into two formally-defined subclasses. The PAT-subclass does not allow the alternation of the person indexing prefix (as shown in (15)), but the verbs of the PATLOC-subclass do alternate (16).

(15) OBSERVABLE CHANGE (PAT-type)

ha-basa	3UND.PAT-brush.off.IPFV	'brush him off, dust it'
h-iel	3UND.PAT-roast.IPFV	'roast it'
ha-weel	3UND.PAT-bathe	'wash him, bathe him'
ha-tamadia	3UND.PAT-repair.IPFV	'repair it'
ha-kuol	3UND.PAT-shave.IPFV	'shave it'

(16) OBSERVABLE CHANGE (PAT-LOC-type)

he-komangdi	'make it blunter'	∼	*ha-komangdi*	'make it blunt'
he-lilri	'warm it up'	∼	*ha-lilri*	'boil it'
he-siki	'split it'	∼	*ha-siki*	'separate it'
he-kol	'tie it'	∼	*ha-kol*	'tie it up'
he-kuya	'peel it'	∼	*ha-kuya*	'expose it'

The above two subclasses differ in their inflectional possibilities. PAT-subclass verbs belong to the 12% of the 300-verb sample which are lexicalized with the PAT prefix and incompatible with other prefixes. PAT∼LOC-subclass verbs belong to an additional 28% of the same sample, compatible with both PAT as well as other prefixes (LOC, REC etc.).

Both subclasses also differ in the specification of the degree of affectedness in the root. Verbs belonging to OBSERVABLE CHANGE (PAT)-class entail a change, which can be characterized with a result description in the subsequent clause (17), but this change is not necessarily maximal, as can be seen in (18). In Beavers' terms, this change can be characterized as *non-quantized*. This is true with human agents, but when the acting force originates in an inanimate participant, the situation is different. We will return to this problem in section 3.4.

(17)

Na	h-ier-i,	#haba	ara	diyei	naha.
1SG.AGT	3UND.PAT-roast.PFV-PFV	but	fire	burn	not

'I roasted it, #but the fire didn't burn it.' [E15BD.27]

(18)

Na	h-ier-i,	haba	dara	kowa.
1SG.AGT	3UND.PAT-roast.PFV-PFV	but	still	be.raw

'I roasted it, but it remains raw.' [E15BD.26]

Beavers (2011:359) lists the variation of the result XP as a diagnostic feature of the *non-quantized* degree, which seems to match the Abui data. Some variation of the results are shown for the verb *ha-wel* 'wash him' in (19-20).

(19) Na ha-wel-i, haba sanra naha.
1SG.AGT 3UND.PAT-wash-PFV but become.clean.IPFV not
'I washed him, but he is not clean.' [E15BD.25]

(20) Na ha-wel-i, haba he-isi de-i
1SG.AGT 3UND.PAT-wash-PFV but 3UND.AL-body 3AGT.LOC-have
dakuni.
be.dirty
'I washed him but he is still dirty.' [E15BD.22]

Similarly to the other verbs of the same class, some minimal degree of affectedness is entailed, as shown by the implausible entailments in (21-22).

(21) Na ha-wel-i, #haba nala da-lakda
1SG.AGT 3UND.PAT-wash-PFV but something 3AGT.PAT-happen.IPFV
naha.
not
'I washed him, #but nothing happened.' [E15BD.23]

(22) Na ha-wel-i, #haba yokda naha.
1SG.AGT 3UND.PAT-wash-PFV but become.wet.IPFV not
'I washed him, #but he didn't get wet.' [E15BD.24]

The second type of change identified by Beavers (2011:339) is TRANSFORM INTO SOMETHING ELSE. Our 300-verb sample does not contain any verbs marked with the PAT prefix belonging to this type. The third type - MOVE AND STAY AT SOME LOCATION - is common and some examples are listed in (23).

(23) MOVE AND STAY AT SOME LOCATION (PAT-type)

ha-fik	3UND.PAT-pull	'pull it, pull him'
ha-suonra	3UND.PAT-push.IPFV	'push it'
ha-kuoila	3UND.PAT-topple.IPFV	'topple it'
ha-kai	3UND.PAT-drop.IPFV	'drop it, trip him'
ha-ai	3UND.PAT-add.IPFV	'add it'
ha-reng	3UND.PAT-turn.to.IPFV	'turn to it'
ha-bi	3UND.PAT-lean.PFV	'lean against it'

Beavers considers this type of change to be compatible with the non-quantized degree of affectedness (2011:245). However, in Abui, no change is necessarily entailed with human agents and the forms may also describe failed attempts, if forced by the context, as shown in (24-26).

(24) Na ha-fik-i haba burook naha.
1SG.AGT 3UND.PAT-pull-PFV but move not
'I pulled it but it didn't move.' [E15BD.34]

(25) Na ha-fik-i haba sik naha.
1SG.AGT 3UND.PAT-pull-PFV but snap not
'I pulled it but it didn't snap.' [E15BD.35]

(26) Na ha-fik-i haba dara de-yal mia.
1SG.AGT 3UND.PAT-pull-PFV but still 3AGT.AL-place be.in
'I pulled it but it is in its place (it's too heavy).' [E15BD.36]

It seems that the PAT-subclass of MOVE AND STAY AT SOME LOCATION verbs in Abui alternates between taking patients with non-specific result (*non-quantized degree*) and non-patient force recipients (*potential degree*).

The fourth type of change BE PHYSICALLY IMPINGED is encoded by the following verbs in Abui:

(27) BE PHYSICALLY IMPINGED (PAT-type)

ha-balak	3UND.PAT-punch	'punch him'
ha-paakda	3UND.PAT-slap.IPFV	'slap him'
h-uol	3UND.PAT-hit.IPFV	'hit/strike him'
ha-taak	3UND.PAT-shoot.IPFV	'shoot him'
ha-laanga	3UND.PAT-grope.IPFV	'grope him'

These verbs are similar to the previous type in allowing the failed readings with human agents, shown in (28). This is consistent with Beavers' classification (2011:345) in which this class of verbs combines with non-patient force-recipients only compatible with certain types of result XPs.

(28) Di n-uol mai ne-l=ha-yei
3AGT 1SG.PAT-strike.IPFV and.then 1SG.LOC-GIVE=3UND.PAT-hit.IPFV
naha.
not
'He struck at me, but didn't hit me.' [E15BD.45]

The fifth type of change is GO OUT OF EXISTENCE (Beavers 2011:339). Examples of Abui PAT-marked belonging to this type are given in (29).

(29) GO OUT OF EXISTENCE (PAT-type)

ha-al	3UND.PAT-burn.IPFV	'burn it'
ha-fuul	3UND.PAT-swallow.IPFV	'swallow it'
ha-pok	3UND.PAT-cover.IPFV	'cover it'
ha-yol	3UND.PAT-bury.IPFV	'bury it'

The verbs in (29) are incompatible with the constructions of the type *x but nothing happened* and entail therefore a minimal change, as shown in (30-31). These verbs are classified as taking patient arguments and entailing a change, matching Beaver's (2011:345) *non-quantized* degree.

(30) Na ha-ar-i, haba dara on-a.
1SG.AGT 3UND.PAT-burn.PFV-PFV but still make.PFV-CONT
'I burned it, but there was still some left.' [E15BD.50]

(31) Na ha-ar-i, #haba ara diyei naha.
1SG.AGT 3UND.PAT-burn.PFV-PFV but fire burn not
'I burned it, #but the fire didn't burn it.' [E15BD.49]

The last type of change is COME INTO EXISTENCE. Our sample contained a single PAT-marked verb of this type:

(32) COME INTO EXISTENCE (PAT-type)

ha-yaal	3UND.PAT-give.birth.IPFV	'give birth to him'

The verb *yaal* 'give birth' combines with a patient, but is not necessarily telic, as shown in (33), describing a failed birth where some complications prevented the baby from being born.

(33) Di moku ha-yaar-i, haba moku sei
3.AGT child 3UND.PAT-give.birth.PFV-PFV but child come.down.IPFV
naha.
not
'She gave birth to the child, but the child was not delivered. ' [E15BD.80]

Besides the above types, Abui PAT-marked verbs also include psych-verbs, and intransitives such as 'hurt', or 'fall'. It should also be noted that the above change type classes include other verbs, which do not require the PAT prefix. The discussion of those is beyond the scope of this paper.

3.2 Abui PAT∼LOC alternation

Verbs belonging to OBSERVABLE CHANGE (PAT∼LOC) class shown in (16) and repeated below as (34) use prefix alternation to distinguish different degrees on the affectedness scale.

(34) OBSERVABLE CHANGE (PAT-LOC-type)

he-komangdi	'make it blunter'	∼	*ha-komangdi*	'make it blunt'
he-lilri	'warm it up'	∼	*ha-lilri*	'boil it'
he-siki	'split it'	∼	*ha-siki*	'separate it'
he-kol	'tie it'	∼	*ha-kol*	'tie it up'
he-kuya	'peel it'	∼	*ha-kuya*	'expose it'

The LOC prefix entails a minimal change, of a lower degree than indicated by the PAT prefix, which usually takes the maximum degree. In terms of Beavers' typology, we consider this alternation as an overt marking of the patient as either involved in a *telic* event, where the final point is know, or in an atelic event where a change progresses in the specifiied direction, but the final point is not specified.

In (35), the LOC prefix attached to the verb *kol* 'bind' implies that there is still some thatching grass left that could be bound, although the binding has stopped (the verb is perfective). The PAT prefix allows the same entailment only if we imagine another agent undoing the binding, such as children or animals scattering the grass after it has been bound, as in (36). Note also, that in the second sense, the verb *kol* has a distinct perfective stem *kor*, while in the first sense, the stem does not have a perfective counterpart (Kratochvíl, 2015).

(35) Na ameng he-kol-i haba dara
1SG.AGT coarse.grass 3UND.LOC-bind-PFV but still
kata-kata-di ba iti.
be.scattered-GET.PFV PROG
'I tied the thatching grass, but some is still scattered around.' [E15BD57]

(36) Na ameng ha-kor-i haba dara
1SG.AGT coarse.grass 3UND.PAT-bind.up.PFV-PFV but still
kata-kata-di ba iti.
be.scattered-GET.PFV PROG
'I tied up the thatching grass, but there is (again) some scattered around (by chickens, or children).' [E15BD58]

The verb *-komangdia* is derived from the state *komang* 'blunt' with the inchoative suffix *-di*. The verb therefore contains the description of the final result. When combined with the LOC prefix, this action does not indicate the maximum degree on the affectedness scale, but a minimal degree of change towards the state denoted by the root (37). When the same root combines with the PAT prefix, the result matches the description in the root (38).

(37) Di kawen he-komangdii, haba de-i
3AGT machete 3UND.LOC-make.blunter.PFV but 3AGT.LOC-have
bula.
be.sharp
'He made the knife blunter, but it's still sharp.' [E15BD51]

(38) Di kawen ha-komangdii, #haba de-i
3AGT machete 3UND.PAT-make.blunter.PFV but 3AGT.LOC-have
bula.
be.sharp
'He made the knife blunt, #but it's still sharp.' [E15BD52]

As mentioned above, the second type of change (TRANSFORM INTO SOMETHING ELSE) proposed by Beavers (2011:339) is not found in our sample. The third type (MOVE AND STAY AT SOME LOCATION) also contains verbs compatible with the PAT∼LOC alternation, as shown in (39).

(39) MOVE AND STAY AT SOME LOCATION (PAT-LOC-type)

he-taang	'pass it along'	∼	*ha-taang*	'give it away'
he-fil	'pull on it'	∼	*ha-fil*	'pull it'
he-bel	'pluck it'	∼	*ha-bel*	'pull it out'
he-baang	'put on shoulder'	∼	*ha-baang*	'put on (its lid)'
he-kil	'put it out'	∼	*ha-kil*	'turn it upside down'

The MOVE AND STAY AT SOME LOCATION verbs marked with the PAT prefix take patient arguments and entail change. On the other hand, verbs marked with the LOC prefix are ambiguous and allow for readings compatible with a non-patient, force-recipient for whom a change is not necessarily entailed (Beavers 2011:345). This is illustrated with the LOC-marked verb denoting a failed attempt with *fil* 'pull' (40), a reading which is not compatible with the PAT-marked form (41).

(40) Ata di bataa he-fil-i, haba burook naha.
PN 3AGT wood 3UND.LOC-pull.on-PFV but move not
'Ata pulled on the log, but it didn't move.' [E15BD59]

(41) Ata di bataa ha-fil-i, #haba burook naha.
PN 3AGT wood 3UND.PAT-pull.on-PFV but move not
'Ata pulled the log, #but it didn't move.' [E15BD60]

BE PHYSICALLY IMPINGED type is well represented in our sample, with examples listed in (42).

(42) BE PHYSICALLY IMPINGED (PAT-LOC-type)

he-dik	'stab at it'	∼	*ha-dik*	'pierce it'
he-rel	'plant it in'	∼	*ha-ril*	'ram it in'
he-taakda	'skewer it'	∼	*ha-taakda*	'stab to death'
he-keila	'block it'	∼	*ha-keila*	'plug it'
he-afui	'scoop it'	∼	*ha-afuui*	'scoop it up'
he-ahii	'select it, pick it'	∼	*ha-ahii*	'remove it'
he-fuuidi	'made it flatter'	∼	*ha-fuuidi*	'flatten it'

The function of the PAT∼LOC alternation is the same as with the OBSERVABLE CHANGE verbs. The PAT marked verb describes a telic event reaching the result described by the the predicate. The LOC marked verb describes an atelic event entailing a minimal change. In both cases, the verb combines with a patient argument, as illustrated in (43-44). Note that we use the gloss 'stab' in both cases, although, it would be equally accurate to gloss the PAT-marked form as 'pierce'.

(43) Na baleei fooi he-dik-i, haba dara tukoladi
1SG.AGT banana stem 3UND.LOC-stab-PFV but still have.hole.PFV
naha.
not
'I stabbed the banana stem, but didn't perforate it.' [E15BD63]

(44) Na baleei fooi ha-dik-i, #haba dara tukoladi
1SG.AGT banana stem 3UND.PAT-stab-PFV but still have.hole.PFV
naha.
not
'I stabbed through the banana stem, #but didn't perforate it.' [E15BD64]

The last type of change is encoded by the GO OUT OF EXISTENCE verbs, exemplified in (45). The verbs belonging to this class pattern in the same way as the verbs of OBSERVABLE CHANGE: PAT-marked forms are telic, LOC-marked do entail a change, but are atelic.

(45) GO OUT OF EXISTENCE (PAT-LOC-type)

he-lak	'demolish it'	∼	*ha-lak*	'destroy it'
he-akung	'shade it'	∼	*h-akung*	'extinguish it'

There are no examples of the PAT∼LOC alternation with COME INTO EXISTENCE verbs in our sample. The PAT∼LOC alternation is also found with some psych-verbs and with state-causatives pairs. The state verb is marked with LOC; the causative verb takes the PAT prefix. This type is quite common, with many examples in our database. Some examples are listed in (46).

(46) STATE∼CAUSATIVE alternation (PAT-LOC-type)

he-rumai	'it is strong'	∼	*ha-rumai*	'strengthen it'
he-poku	'it hatched'	∼	*ha-poku*	'crack it'
he-lika	'it is stuck'	∼	*ha-lika*	'stick it in'
he-mong	'it is dead'	∼	*ha-mong*	'extinguish it'
he-liikda	'it leans sideways'	∼	*ha-liikda*	'bend it'

We are listing these verbs, because they are not treated in Beavers (2011) account, but in our view show that the affectedness space is multidimensional. Perhaps the PAT∼LOC alternation could be in this case thought of as a detransitivising process, where the absence of an external force is marked in this way.

Another type of PAT∼LOC alternation involves an intransitive process verb (LOC) and a transitive causative verb (PAT). Some examples are listed in (47). We included the LOC-marked punctual verbs, such as 'explode' and 'break off'.

(47) PROCESS∼CAUSATIVE alternation (PAT-LOC-type)

he-lai	‘it diffuses’	∼	*ha-lai*	‘squeeze it out’
he-buida	‘it’s getting short’	∼	*ha-buida*	‘shorten it’
he-takda	‘it’s getting empty’	∼	*ha-takda*	‘empty it’
he-fokda	‘it’s getting big’	∼	*ha-fokda*	‘enlarge it’
he-peekdi	‘it came near’	∼	*ha-peekdi*	‘put it near’
he-melri	‘it got flavor’	∼	*ha-melri*	‘season it’
he-fuunri	‘it piled up’	∼	*ha-fuunri*	‘pile it up’
he-fuuisi	‘it exploded’	∼	*ha-fuuisi*	‘blow it up’
he-tukdi	‘it broke off’	∼	*ha-tukdi*	‘break it off’

3.3 Abui transitive REC-verbs

The REC alternation is used when the involvement of the agent in the situation is at the centre of attention rather than the resulting state of the undergoer. The agent, which is always human, is usually acting in an involuntary or uncontrolled fashion, driven by some internal need which cannot be controlled. However, paraphrases with ‘want’ or ‘must’ are not quite precise, showing that this alternation is not a type of modality. Multiple results XPs are possible suggesting that a minimal change is always entailed for the undergoer argument (49-50).

(48) Ata ama he-baleei do-takaafi, haba mingwaha wala
PN person 3UND.AL-banana 3AGT.REC-steal.PFV but some only
mii.
take.PFV
‘Ata stole (for himself) bananas of those people, but he took just a few.’ [E15BD69]

(49) Ata ama he-baleei do-takaafi taaqdi.
PN person 3UND.AL-banana 3AGT.REC-steal.PFV exhaust.PFV
‘Ata stole (for himself) all the bananas of those people.’ [E15BD76]

(50) Ata ama he-baleei do-takaafi, #haba nuku baai
PN person 3UND.AL-banana 3AGT.REC-steal.PFV but one also
mii naha.
take.PFV not
‘Ata stole (for himself) bananas of those people, #but he didn’t take any.’ [E15BD68]

Example (51) shows that the agent remains the acting force but that his control is reduced in a way detectable for the speaker. The agent can perform the event, and still be dissatisfied with the result (51).

(51) Ata ama he-baleei do-takaafi, haba
PN person 3UND.AL-banana 3AGT.REC-steal.PFV but
ho-ming kaanri naha.
3UND.REC-satisfied.PFV not

‘Ata stole for himself enough bananas of those people, but he is not satisfied.’ [E15BD.70]

The control over the event is cannot be transferred to another agent, as shown with the quasi-causative construction in (52). Note also, that the presence or absence of an agentive pronouns is encodes the presence or absence of control with the agent (Kratochvíl 2014:561-563).

(52) A panen-te di ko ama he-baleei
2SG.AGT make.PFV-PRIOR 3.AGT IRR person 3UND.AL-banana
do-takaafi
3AGT.REC-steal.PFV
‘Do something that he would steal bananas of those people.’ [E15BD.77]

This type of affectedness is not included in Beavers’ framework, but the REC alternation is very common in Abui. In our database, more than 75% of the verbs are compatible with the REC prefix. The REC paradigm is also used to mark experiencers of some psych-verbs in Abui, and so the REC-marked agent can be thought of as similar to an experiencer, or in Beavers’ terms as a non-force recipient (2011:358). As we said above, we do not consider this alternation a type of modality.

We conclude that the REC alternation is neither a simple detransitivising process, but rather a construction indicating a temporary absence of control, which is regained through performing the action. In future research, we will explore, whether the ‘change’ on the agent’s side can be expressed with a result XP and whether the scale can be named more precisely to answer the question whether this type of change could be considered an additional degree of affectedness (agent-oriented).

The REC-alternation shares some similarities with the Slavic dispositional reflexives which emphasise a different aspect of the agent, usually casting it not only as instigating but also as experiencing the action in a particular way, positive or negative (Fried 2007:743-744).

3.4 Agents, forces, and degrees of affectedness

The degree of affectedness can vary for PAT-marked verbs depending on whether the cause of the event is a human agent or an inanimate cause. As shown in section 3.1, various degrees of affectedness are compatible with human agents. However, as shown in (53-55), inanimate force is compatible with the maximum degree of change and does not allow for failed readings.

(53) Na ha-kaai haba ha-yei naha.
1SG.AGT 3UND.PAT-make.fall.PFV but 3UND.PAT-fall.IPFV not
‘I tripped him but he didn’t fall.’ [E15BD.37]

(54) Na ha-kaai haba da-kai naha.
1SG.AGT 3UND.PAT-make.fall.PFV but 3AGT.PAT-make.fall.IPFV not
'I tripped him but he didn't trip.' [E15BD.38]

(55) Wii foka ha-kaai, #haba da-kai
stone be.big 3UND.PAT-make.fall.PFV but 3AGT.PAT-make.fall.IPFV
naha.
not
'The large stone made it fall, #but it didn't fall.' [E15BD.39]

These examples show, that the degree of affectedness interacts in subtle ways with agency, which in turn can combine with affectedness, as shown in section 3.3. This effect also suggests that the basic predicate meaning should be modelled with the maximum degree of affectedness available (marked with the PAT paradigm) and that the lower degrees of affectedness may be derived from there. Such approach would fit well with the comparative and diachronic pattern within the Alor-Pantar family (Klamer 2014).

4 Discussion

Abui PAT-compatible verbs of the five types of change - (i) OBSERVABLE CHANGE, (ii) MOVE AND STAY AT SOME LOCATION, (iii) PHYSICAL IMPINGED, (iv) GO OUT OF EXISTENCE, and (v) COME INTO EXISTENCE fall apart into two subclasses. The exclusively PAT-marked subclasses contain of verbs that take patient arguments and typically entail a change, but show flexibility with human agents. With inanimate force responsible for the action, the degree of change is maximal and matches the *quantized* degree in Beavers' Affectedness Hierarchy. On the other hand, the PAT∼LOC-subclasses marks specifically, whether the degree of change is maximal (a change is entailed) or not (failed attempt-compatible).

The REC alternation presents a possibility of an additional dimension of affectedness, applied to the agent, in some way affected by the action, possibly, as a non-force recipient, although it is presently unclear whether the change can be described with a result XP and the scale clearly identified. The above examples showed that the alternation does not have any consequences for the amount of change affecting the undergoer and cannot be rephrased in terms of modality.

The mapping between the Abui prefix paradigms to the Affectedness Hierarchy (Beavers 2011:359) is not simple, although the discussed prefix paradigms are clearly involved in encoding of affectedness and its degree. It should also be noted that some psych-verbs and some verbs of communication are compatible with the PAT prefix and with the PAT∼LOC alternation. Finally, the PAT∼LOC alternation also admits intransitive states (LOC) and causatives (PAT).

We conclude that verbs with very similar meaning may still differ in ways in which they can be manipulated and enter various constructions, pointing to a fine distribution of labour between lexicon and grammar and between morphology and

syntax. In other words, Abui verbs show an interesting pattern of lexical stipulation of the verbal root, as discussed by Fedden et al. (2013, 2014). Although the interpretation of the predicate is co-determined by the argument agreement selection, verbs with the same marking lexicalised a different maximal degree of affectedness in terms of Beavers' (2011) hierarchy. For the PAT∼
textscloc alternation the shifting is always one degree lower for the LOC-marked form, which suggests that systematic encoding of the maximum degree of affectedness for each verb (class) in the lexicon may be sufficient to determine the meaning of this alternation.

References

Beavers, John. 2011. On affectedness. *Natural Language & Linguistic Theory* 29(2). 335–370.

Fedden, Sebastian, Dunstan Brown, Greville Corbett, Gary Holton, Marian Klamer, Laura C Robinson & Antoinette Schapper. 2013. Conditions on pronominal marking in the Alor-Pantar languages. *Linguistics* 51(1). 33–74.

Fedden, Sebastian, Dunstan Brown, František Kratochvíl, Laura C Robinson & Antoinette Schapper. 2014. Variation in pronominal indexing: lexical stipulation vs. referential properties in Alor-Pantar languages. *Studies in Language* 38(1). 44–79.

Fried, Mirjam. 2007. Constructing grammatical meaning: Isomorphism and polysemy in Czech reflexivization. *Studies in Language* 31(4). 721–764.

Hartmann, Iren, Martin Haspelmath & Bradley Taylor. 2013. Valency Patterns Leipzig. http://valpal.info/.

Holton, Gary, Marian Klamer, František Kratochvíl, Laura C Robinson & Antoinette Schapper. 2012. The historical relations of the Papuan languages of Alor and Pantar. *Oceanic Linguistics* 51(1). 86–122.

Klamer, Marian. 2014. The role of Affectedness in the emergence of Differential Object Marking in Alor-Pantar languages. Paper read at the 1st Workshop on Affectedness 2014: Manifestation of Affectedness in Natural Languages (17-20 June). Nanyang Technological University, Singapore.

Kratochvíl, František. 2007. *A grammar of Abui: a Papuan language of Alor*. Utrecht: LOT.

Kratochvíl, František. 2011. Transitivity in Abui. *Studies in Language* 35(3). 588–635.

Kratochvíl, František. 2014. Differential argument realization in Abui. *Linguistics* 52(2). 543–602.

Kratochvíl, František. 2015. Aspectual pairing in Abui. Manuscript. Nanyang Technological University, Singapore.

Lehmann, Christian. 1991. Predicate classes and participation. In H. Seiler W. Premper (ed.), *Partizipation: Das sprachliche erfassen von sachverhalten*, 183–239. Tübingen: Gunter Narr Verlag.

Malchukov, A. & B. Comrie. 2015. *Valency Classes in the World's Languages - Introducing the Framework, and Case Studies from Africa and Eurasia*, vol. 1 Comparative Handbooks of Linguistics. De Gruyter.

Tenny, Carol Lee. 1987. *Grammaticalizing aspect and affectedness*: Massachusetts Institute of Technology dissertation.

Tsunoda, Tasaku. 1981. Split case-marking patterns in verb-types and tense/aspect/mood. *Linguistics* 19(5-6). 389–438.

Scalarity and the Cantonese post-verbal particle *can1*

Joanna Ut-Seong Sio
Nanyang Technological University

Proceedings of the 22nd International Conference on
Head-Driven Phrase Structure Grammar

Nanyang Technological University (NTU), Singapore

Stefan Müller (Editor)

2015

CSLI Publications

pages 234–241

http://csli-publications.stanford.edu/HPSG/2015

Abstract

This paper provides an analysis of the Cantonese post-verbal particle *can1*. We argue that *can1* is a resultative particle encoding the meaning of ‘a small degree’. It is only compatible with (i) verbs that entail a specific resulted state of the theme argument and (ii) verbs that encode a potential change of the theme argument (Beavers, 2011, 2013). Assuming that change of state verbs involve a property scale (Hay et al., 1999), we propose that *can1* makes the property scale bounded by providing an end-point. This end-point, however, is not precise. It consists of a range of values on the lower end of the scale.

1 Introduction

Cantonese has a very rich inventory of post-verbal particles (Matthews and Yip, 2011). Some examples are given below:

Aspectual particles: *gan2* ‘progressive’, *zo2* ‘perfective’, etc.
Directional particles: *hei2* ‘up’, *dai1* ‘down’, *zau2* ‘away’, etc.
Resultative particles: *bao2* ‘full’, *dou2* ‘arrive’, *sei2* ‘dead’ etc.
Quantifying particles: *saai3* ‘completely’, *maai4* ‘also’, etc.
Adversative/habitual particle: *can1*

The last particle listed above, *can1*, has two different senses. It can mean (i) ‘being mildly and negatively affected’, as in (1) or (ii) ‘whenever’, as in (2). Matthews and Yip (2011) calls the former ‘adversative’ and the latter ‘habitual’.

(1) Ngo5 zong6-can1 zek3 maau1 aa3
1SG bump.into-CAN CL cat SFP
‘I bumped into the cat (and as a result the cat was mildly hurt).’

(2) Keoi5 coeng3-can1 go1 dou1 ham3 ga3
3SG sing-CAN song always cry SFP
‘S/He cries whenever s/he sings.’

This paper focuses on the adversative sense of the particle *can1*. We will discuss its grammatical properties and propose an analysis that captures its selectional restriction.

[†] The research reported here is supported by the project ’Grammar Matrix Reloaded: Syntax and Semantics of Affectedness’ (MOE 2013-T2-1-016) funded by the Ministry of Education (MOE) in Singapore. We would like to thank all the participants of the HPSG 2015 conference as well as others who had given their useful input.

1.1 The grammatical properties of *can1*

Can1 is a post-verbal particle. It is placed after the verb. Stacking of post-verbal particles is possible, subject to semantic compatibility. Though it is hard to find cases with more than 2 post-verbal particles in a row. An example of *can1* followed by the aspectual particle *zo2* is given below:

(3) Ngo5 zong6-can1-zo2 zek3 maau1 aa3
1SG bump.into-CAN-PERF CL cat SFP
'I bumped into the cat (and as a result the cat is mildly hurt).'

When *can1* appears in transitive sentences, the affected argument is the object. The affected argument has to be sentient. *Can1* is not compatible with an inanimate object.

(4) * Ngo5 zong6-can1 bun2 syu1 aa3
1SG bump.into-CAN CL book SFP
Intending reading:'I bumped into a book (and as a result the book was mildly hurt).'

Physical contact is not required for *can1* to be used:

(5) Lei5 haak3-can1 keoi5 laa3
2SG scare-CAN 3SG SFP
'You scared her/him (and as a result she/he was frightened mildly).'

Can1 is also compatible with intransitive sentences. As observed by Gu and Yip (2004), it is compatible with unaccusatives, but not unergatives:

(6) a. unaccusative
Keoi5 dit3-can1 aa3
3SG fall-CAN SFP
'S/He fell (and as a result s/he was mildly hurt).'

b. unergative
* Zek3 maau1 tiu3-can1 aa3
CL cat jump-CAN SFP
Intended reading: 'The cat jumped (and as a result it was mildly hurt).'

The negative effect on the participant has to be small. In example (1), repeated here as (7), if the result of the event is that the cat ends up dead, the use of *can1* would not be appropriate.

(7) Ngo5 zong6-can1 zek3 maau1 aa3
1SG bump.into-CAN CL cat SFP
'I bumped into the cat (and as a result the cat was mildly hurt).'

In addition to the reading of ‘a small degree’, *can1* is also adversive. It has to mean being negatively affected to a small degree but not positively affected to a small degree. In fact, when *can1* is used with a verb with a positive connotation, the sentence is either ungrammatical, (8) or it would be interpreted negatively, (9):

(8) * Lei5 zan3-can1 Siu2koeng4 aa3
2SG praise-CAN Siukoeng SFP
Intended reading: ‘You praised Siukeong (and as a result Siukoeng was mildly annoyed).’

(9) Lei5 caat3-can1 keoi5 haai4 aa3
2SG polish-CAN 3SG shoe SFP
‘You flattered her/him (and as a result s/he was mildly annoyed).’

Zan3 ‘praise’ is a positive thing. It cannot be combined with *can1*, as in (8). *Caat3 haai4* literally means ‘polish shoes’. It has the meaning of ‘’trying hard to flatter someone’. When used with *can1*, as in (9), it gives rise to the interpretation of over-doing the flattering and generating annoyance on the receiving end.

Gu and Yip (2004) observe that verb-*can1* complexes are not compatible with *hai2dou6* ‘right now’ or the progressive aspectual particle *gan2*:

(10) * Keoi5 hai2dou6 haak3-can1 go3 bi4bi1
3SG right.now scare-CAN CL baby
Intended reading: ‘S/He is now scaring the baby.’

(11) * Keoi5 haak3-can1-gan2 go3 bi4bi1
3SG scare-CAN-PROG CL baby
Intended reading: ‘S/He is now scaring the baby.’

Verb-*can1* complexes act like achievements, which are punctual events. Since punctual events have exactly two atomic parts, a beginning and an end, but have no middle (Dowty, 1979). *Can1* is expected to be incompatible with the progressive aspect, *gan2*, or adverbs that modify the middle of an event, *hai2dou6* ‘right now’.

2 Analysis

2.1 Verb selection

Beavers (2011, 2013) identifies 4 classes of verbs which encode different degrees of affectedness on the event participant x (in descending order):

(i) x undergoes a quantized change (e.g. *peel*, *kill*, *shatter* x).
(ii) x undergoes a non-quantized change (e.g. *cut*, *widen*, *lengthen* x).
(iii) x has potential for change (e.g. *hit*, *wipe*, *rub* x).
(iv) x is unspecified for change (e.g. *see*, *smell*, *ponder* x)

For verbs of type (i), the participant reaches a precise result state. The result is encoded as part of the semantics of the verb (e.g. being *killed* means the victim results in death). For verbs of type (ii), a result on the participant is entailed, but it is not uniquely specified (e.g. a piece of dough can be *flattened* into different degrees) . For verbs of type (iii), a change on the participant is possible, but there does not have to be one (e.g. being *hit* by a baby may not result in any observable change). For verbs of type (iv), there is no change (e.g. being *seen* would not cause any change).

The Cantonese *-can1* is only compatible with verbs of type (ii) and (iii), non-quantized change and potential for change, but not (i) and (iv), quantized change and unspecfied for change. The relevant data are given below:

(12) Quantized change:

* Siuming saat3-can1 Siukoeng aa3
  Siuming kill-CAN Siukoeng SFP

Intended reading: 'Siuming killed Siukeong (and as a result Siukoeng was mildly hurt).'

(13) Non-quantized change:

Siuming cap3-can1 Siukoeng aa3
Siuming stab-CAN Siukoeng SFP

'Siuming stabbed Siukoeng (and as a result Siukoeng was mildly hurt).'

(14) Potential for change:

Siuming daa2-can1 Siukoeng aa3
Siuming hit-CAN Siukoeng SFP

'Siuming hit Siukeong (and as a result Siukoeng was mildly hurt).'

(15) Unspecified for change:

* Siuming tai2-can1 Siukoeng aa3
  Siuming see-CAN Siukoeng SFP

Intended reading: 'Siuming saw Siukeong (and as a result Siukoeng was mildly hurt).'

In (13) and (14), the object was both mildly hurt. It is natural to assume that being stabbed is more severe than being hit in general. Thus, being mildly hurt from being stabbed could be more sever than being mildly hurt from being hit. The 'mildly' interpretation is calculated according to the range of possibles effect of the action, but not a general standard that applies across the board.

2.2 Scalarity

There are three type of incremental themes (Tenny, 1994):

(16) a. Creation/Consumption predicates
John ate the fish.

b. Motion predicates
John walked to the store.

c. Change of state predicates
John scrubbed the sink clean.

Each of the example above encodes a three-way relation between an event, a theme and a scale. The type of scales differs depending on the verb type (Hay, Kennedy and Levin 1999). For creation and consumption predicates, the scale is the spatial content of the theme argument (ascending for creation or descending for consumption). For motion predicates, the scale is the path of motion of the theme argument (a path from the original location of the theme to the final location of the theme). For change of state predicates, the scale is the gradable property (of the resulted state) of the theme argument.

Affectedness encodes a change of property of the theme argument. Different degrees of affectedness on the theme argument can be expressed using a property scale model (Beavers 2013):

kill: theme x undergoes a quantized change on a scale and reaches a specific point in the scale.

stab: theme x undergoes a non-quantized change on a scale and reaches some unspecified point in the scale.

hit: theme x might change but there might not be any actual change. (latent scale)

see: x is not specified for change as it is just an event participant. (no scale)

The post-verbal particle *can1* has no lexical meaning on its own. But when interpreted with a verb, it means ‘a small degree’ (the degree interpretation of ‘mildly’). We claim that *can1* is only compatible with verbs that involve a scale that is unbound, i.e. with no end-point. *Can1* provides an end-point for the scale, making it bounded. For quantized change, the scale is already bounded. The extra value that *can1* provides will lead to ungrammaticality. For verbs that are unspecified for change, there is no scale, and are thus not compatible with *can1* either. For verbs that encode non-quantized change on the theme argument, *can1* provides an end-point for that scale (out of the many possible end-points). For verbs encode potential change on the theme argument, the use of *can1* indicates that there is indeed a change in the theme argument (i.e. there is a change of state) and again *can1* provides an end-point for that property scale.

To be precise, *can1* does not provide just one value for the scale. *Can1* indicates that the theme argument is negatively affected to a small degree. ‘A small degree’ is compatible with many possible values, as long as they are close to the lower end of the scale. vanden Wyngaerd (2001) claims that resultative predicates are subject to a boundedness requirement: they are telic. Gu and Yip (2004) argues that such boundedness, however, can be non-precise (a range of values), as in the case of *can1*.

2.3 *Can1* and other resultative particles

Gu and Yip (2004) treat *can1* as a resultative particle. *Can1*, however, is different from the other resultative particles. Unlike the other resultative particles, it does not have a clear lexical meaning (unlike *sei2* 'dead' for example), and as a consequence we think, it does not provide a precise end-point.

The lack of precision has consequences on *can1*'s distribution. Its appearance is more restricted than the other regular resultative particles that encode a precise end-point. As discussed earlier on, Gu and Yip (2004) claim that *can1* is not compatible with unergatives because *can1* does not provide a precise enough end-point, (17). When the resultative particle provides a precise end-point, it is compatible with unergatives verbs, (18).

(17) * Zek3 maau1 tiu3-can1-zo2 aa3
CL cat jump-CAN-PERF SFP
Intended reading:'The cat jumped (and as a result it was mildly hurt.'

(18) Zek3 maau1 tiu3-wan4-zo2 aa3
CL cat jump-faint-PERF SFP
'The cat jumped to the extent that it fainted.'

Regular resultative particles are compatible with *dou3*, which means 'to the extent', (20). *Can1*, however, is not, (19). This could be due to the fact that *can1* does not give a precise end-point and thus it is unclear what the extent is.

(19) * Lei3 daa2 dou3 keoi5 can1 laa3
2SG hit to.the.extent 3SG CAN SFP
Intended reading:'You are hitting him to the extent that s/he is going hurt a little bit.'

(20) Lei3 daa2 dou3 keoi5 sei2 laa3
2SG hit to.the.extent 3SG dead SFP
'You are hitting him to the extent that s/he is going to die.'

3 Conclusion

In this paper, we have provided an overview of the grammatical properties of the Cantonese post-verbal particle *can1*. We follow Beavers (2011, 2013) in classifying verbs into four classes with respect to affectedness. *Can1* is compatible with verbs that encode a non-quantized change with an entailed result and verbs that encode a potential result. We propose that *can1* specifies a result state that is not precise. It provides a range of value denoting a small degree on a property scale.

Even though our analysis, adopting Beavers (2011, 2013), accounts for the selectional restriction of *can1*, the analysis does not account for its advertive reading. It is imaginable that a theme argument is positively affected to a small degree, but *can1* cannot encode that.

Beavers (2011, 2013) makes a prediction on affectedness in general which is contrary to the behaviour of *can1*. Beavers claims that the relevant degrees of affectedness fall into an implicational Affectedness Hierarchy based on monotonically weakening truth conditions: quantized >non-quantized >potential >unspecified. He claims that no grammatical phenomenon picks out a discontinuous range on the hierarchy, or picks out a continuous range that excludes quantized change. This is not true. In fact, *can1* does exactly that. *Can1* picks out the middle range, non-quantized and potential, excluding the edges, quantized and unspecifed.

References

Beavers, J. (2011). On affectedness. *Natural Language & Linguistic Theory*, 29(2):335–370.

Beavers, J. (2013). Aspectual classes and scales of change. *Linguistics*, 51(4):681–706.

Dowty, D. (1979). Word meaning and montague grammar: The semantics of verbs and times in generative semantics and in montague's ptq, d. *Reidel, Dordrecht*.

Gu, Y. and Yip, V. (2004). On the cantonese resultative predicate v-can. *Concentric: Studies in Linguistics*, 30:35–67.

Hay, J., Kennedy, C., and Levin, B. (1999). Scalar structure underlies telicity in" degree achievements". In *Semantics and linguistic theory*, pages 127–144.

Matthews, S. and Yip, V. (2011). *Cantonese: A comprehensive grammar*. Routledge.

Tenny, C. L. (1994). *Aspectual roles and the syntax-semantics interface*. Springer.

vanden Wyngaerd, G. J. (2001). Measuring events. *Language*, 77(1):61–90.