Proceedings of the 26th International Conference on Head-Driven Phrase Structure Grammar

University of Bucharest

Stefan Müller, Petya Osenova (Editors)

2019

CSLI Publications

http://csli-publications.stanford.edu/HPSG/2019

Contents

Editor's note

The 26th International Conference on Head-Driven Phrase Structure Grammar (2019) was held at the University of Bucharest.

The conference featured 2 invited talks and 10 papers selected by the program committee (Anne Abeillé, Doug Arnold, Emily Bender, Felix Bildhauer, Olivier Bonami, Francis Bond, Gosse Bouma, Antonio Branco, Rui Chaves, Philippa Cook, Berthold Crysmann, Dan Flickinger, Antske Fokkens, Petter Haugereid, Fabiola Henri, Anke Holler, Gianina Iordăchioaia, Jong-Bok Kim, Jean-Pierre Koenig, David Lahm, Bob Levine, Nurit Melnik, Philip Miller, Stefan Müller, Tsuneko Nakazawa, Rainer Osswald, Petya Osenova (chair), Gerald Penn, Frank Richter, Louisa Sadler, Manfred Sailer, Pollet Samvellian, Jesse Tseng, Frank van Eynde, Stephen Wechsler, Shûichi Yatabe, Eun-Jung Yoo).

There was a workshop on *Romance languages* with five talks and one invited talk.

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

Thanks go to Gabriela Bîlbîie and Emil Ionescu, who were in charge of local arrangements.

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 of the conference except the ones by Anne Abeillé & Elodie Winckel, Gabriel Aguila-Multner & Berthold Crysmann, Antonio Machicao y Priemer & Paola Fritz-Huechante, Nurit Melnik & Bracha Nir, Stefan Müller, Jong-Bok Kim & Alain Kihm, and Frank Van Eynde, who will submit their papers to journals. The workshop contributions will be published in 2020 in issue 43:1 of Lingvisticæ Investigationes.

The complexities of the Welsh copula

Bob Borsley
University of Essex and Bangor University

Proceedings of the 26th International Conference on
Head-Driven Phrase Structure Grammar

University of Bucharest

Stefan Müller, Petya Osenova (Editors)

2019

CSLI Publications

pages 5–25

http://csli-publications.stanford.edu/HPSG/2019

Keywords: copula, syntax-morphology mismatches, Welsh

Borsley, Bob. 2019. The complexities of the Welsh copula. In Müller, Stefan, & Osenova, Petya (Eds.), *Proceedings of the 26th International Conference on Head-Driven Phrase Structure Grammar, University of Bucharest*, 5–25. Stanford, CA: CSLI Publications.

Abstract

The Welsh copula has a complex set of forms reflecting agreement, tense, polarity, the distinction between main and complement clauses, the presence of a gap as subject or complement, and the contrast between predicative and equative interpretations. An HPSG analysis of the full set of complexities is possible given a principle of blocking, whereby constraints with more specific antecedents take precedence over constraints with less specific antecedents, and a distinction between morphosyntactic features relevant to syntax and morphosyntactic features relevant to morphology.

1. Introduction

It is probably a feature of most languages that the copula is more complex in various ways than standard verbs. This is true in English, and it is very definitely true in Welsh. The Welsh copula has a complex set of forms reflecting agreement, tense, polarity, the distinction between main and complement clauses, the presence of a gap as subject or complement, and the contrast between predicative and equative interpretations. In this paper, I will set out the facts and develop an analysis within the Head-Driven Phrase Structure Grammar (HPSG) framework. I will draw here on the proposals of Borsley (2015) and especially Bonami, Borsley, and Tallerman (2016). In particular, I will utilize two mechanisms which are employed in the latter. Firstly, I will assume a principle of blocking, whereby if the antecedents of two constraints stand in a subsumption relation, only the more specific constraint may apply. Secondly, I will assume that there is a distinction between two sets of morphosyntactic features, one relevant to syntax and another relevant to morphology. For most words the two sets will be identical, but in some cases there will be a mismatch. These two mechanisms will be crucial for ensuring the correct form of the copula.

The paper is organized as follows. In section 2, I develop an analysis of the basic argument selection properties of the Welsh copula. Then, in section 3, I consider agreement and tense. I go on in section 4 to look at the relevance of polarity and the main-complement distinction. Then, in section 5, I consider the influence of first subject and then complement gaps. In section 6, I look at the distinction between predicational and identity uses. Finally, in section 7, I summarize the paper.

* I am grateful to Bob Morris Jones for help with the data, and to Olivier Bonami, David Willis, Ian Roberts, and Marieke Meelen for helpful discussion of some of the ideas presented here. I am also grateful to various anonymous reviewers and the audience at HPSG19 for their comments and discussion. I alone am responsible for what appears here.

2. Argument selection

Like its counterpart in many languages, the Welsh copula *bod* allows a number of different complements.[1] Perhaps the simplest case is a PP complement, as in (1).

(1) Mae Gwyn yn yr ardd.
be.PRES Gwyn in the garden
'Gwyn is in the garden.'

(This and subsequent examples show that Welsh is a VSO language with verb-subject order in all finite clauses.) It can also have what I will call a Perfect Phrase (PerfP), consisting of the perfect particle *wedi* and a non-finite VP, and what I will call a Progressive Phrase (ProgP), consisting of the progressive particle *yn* and a non-finite VP, as in the following:[2]

(2) Mae Gwyn wedi cysgu.
be.PRES Gwyn PERF sleep.INF
'Gwyn has slept.'
(3) Mae Gwyn yn cysgu.
be.PRES Gwyn PROG sleep.INF
'Gwyn is sleeping.'

Progressive *yn* derives historically from the preposition *yn*, but it triggers no mutation, whereas the preposition *yn* triggers so-called nasal mutation, giving e.g. *yn Neiniolen* for 'in Deiniolen' (a village in North Wales). Finally, it can have what I will call a Predicative Phrase (PredP), consisting of the predicative particle *yn* and an AP or NP, as in the following:

(4) Mae Gwyn yn glyfar.
be.PRES Gwyn PRED clever
'Gwyn is clever.'
(5) Mae Gwyn yn feddyg.
be.PRES Gwyn PRED doctor
'Gwyn is a doctor.'

Unlike progressive *yn*, predicative *yn* triggers soft mutation. The basic forms of *glyfar* and *feddyg* are *clyfar* and *meddyg*, respectively.

---

[1] For general discussion of Welsh syntax, see Borsley, Tallerman, and Willis (2007).
[2] Welsh has a number of other aspectual particles, most of which are homophonous with prepositions, e.g. *ar* 'on', *heb* 'without', and *am* 'about'. See Jones (2010: Chapter 9) for discussion.

As with *be*, coordinations of different phrase types suggest that there is a single verb here.

(6) Mae Gwyn yn ddiog ac yn cysgu.
be.PRES Gwyn PRED lazy and PROG sleep.INF
'Gwyn is lazy and sleeping.'
(7) Mae Gwyn yn sâl ac yn y gwely.
be.PRES Gwyn PRED ill and in the bed
'Gwyn is ill and in bed.'
(8) Mae Gwyn yn ieithydd ac yn astudio Cymraeg.
be.PRES Gwyn PRED linguist and PROG study.INF Welsh
'Gwyn is a linguist and studying Welsh.'

The facts can be handled like similar facts in English and elsewhere by assuming that the Welsh copula takes a [PRED +] complement and that all these phrase types are [PRED +].

Bod takes as its subject whatever its complement requires, including an expletive subject, as the following illustrate:[3]

(9) Mae (hi) 'n bwrw glaw.
be.PRES she PRED strike.INF rain
'It's raining.'
(10) Mae (hi) 'n amlwg bod Mair wedi dod yn ôl.
be.PRES she PRED obvious be Mair PERF come.INF back
'It is obvious that Megan has come back.'

Thus, it appears to be a raising verb.[4] This means an ARG-ST feature of the following form:

(11) $\left[\text{ARG-ST} < [1], \begin{bmatrix}\text{HEAD}[\text{PRED}+] \\ \text{SUBJ} , < [1] >\end{bmatrix} >\right]$

I am assuming here that the subject of a [PRED +] element appears in its SUBJ list. However, I will assume below, following Borsley (1989), that all the arguments of finite verbs, subjects as well as compements, appear in their COMPS lists. Among other things, this accounts for the fact that the subject of a finite verb is always post-verbal.

[3] As Joan Maling has emphasized to me, Welsh is rather unusual in using a feminine pronoun as an expletive.
[4] Cf. Pollard and Sag (1994, 147) and Bender (2001, 48) on *be*.

3. Agreement and Tense

It is not surprising that the Welsh copula has forms reflecting agreement and tense. However, in both areas, it has interesting properties.

Unlike the English copula, but like standard Welsh verbs, the copula only shows agreement with a pronominal subject. Here are examples with third person singular and plural pronouns.

(12) a. Mae o / hi yn y gegin.
be.PRES he she in the kitchen
'He/She is in the garden.'
b. Maen nhw yn y gegin.
be.PRES.3PL they in the kitchen
'They are in the garden.'

With a non-pronominal subject, singular or plural, the form in (12a) appears and not that in (12b).

(13) Mae 'r bachgen / bechgyn yn y gegin.
be.PRES the boy boys in the kitchen
'They boy is/The boys are in the garden.'
(14) *Maen y bechgyn yn y gegin.
be.PRES.3PL the boys in the kitchen

The form in (12a) is sometimes seen as a third person singular form, but I will argue that it is a form unspecified for agreement (hence the gloss).

Borsley (2009) argues that verb-subject agreement is one instance of agreement between a head and an immediately following pronoun. Prepositions show agreement the form of a suffix with a following pronominal object, non-finite verbs show agreement in the form of a preceding clitic with a following pronominal object, and nouns show agreement in the form of a preceding clitic with a following pronominal possessor. In all cases, we also have agreement with a pronominal first conjunct of a coordinate NP in the relevant position. Borsley (2009) proposes that all these heads have an AGR(EEMENT) feature whose value is the relevant index when followed by a pronoun and otherwise *none*.

To capture the distinctive agreement behavior of finite verbs, we can propose that they have five forms in each tense specified for agreement with first and second person singular and plural and third person plural pronouns, and a form in each tense which is not specified for agreement. Following Bonami, Borsley, and Tallerman (2016), I assume that the morphological features which are responsible for the form of verbs and other parts of speech are the value of a feature INFL. Given this, assumption, we can propose constraints like the following, where, following a variety of earlier work, LID

is a feature whose value is unique to each distinct lexeme, the words that realise it, and the phrases that they head.

(15) $\left[\begin{array}{l}\text{INFL}\left[\begin{array}{l}\text{LID } \textit{bod}\\ \text{VFORM } \textit{fin}\\ \text{TENSE } \textit{pres}\\ \text{AGR } [\textit{3rd}, \textit{plur}]\end{array}\right]\end{array}\right] \rightarrow$ [PHON *maen*]

We will have similar constraints for first and second person singular and plural forms. We will also have a constraint of the following form:

(16) $\left[\begin{array}{l}\text{INFL}\left[\begin{array}{l}\text{LID } \textit{bod}\\ \text{VFORM } \textit{fin}\\ \text{TENSE } \textit{pres}\end{array}\right]\end{array}\right] \rightarrow$ [PHON *mae*]

Notice that this does not specify a value for AGR. Given the principle of blocking, (16) will not apply where a constraint specifies a specific value for AGR. Hence, *mae* will not appear with third person plural pronouns or first and second person singular or plural pronouns. But it will appear with a third person singular pronoun and with a non-pronominal NP, singular or plural. This is what we have in (12a) and (13). We will see later that slightly more complex constraints are in fact necessary.

The Welsh copula is just like other verbs where agreement is concerned, but with tense it is different. While standard verbs have three tenses, past, future, and conditional, the copula has five tenses, these three and two more, present and imperfect. Table 1 illustrates the third person singular forms of a standard verb and the copula.

	Cerdded 'walk'	*Bod* 'be'
Future	*cerddith*	*bydd*
Past	*cerddodd*	*buodd*
Conditional	*cerddai*	*byddai*
Present	----------	*mae*
Imperfect	----------	*roedd*

Table 1: Third person forms of *cerdded* 'walk' and *bod* 'be'

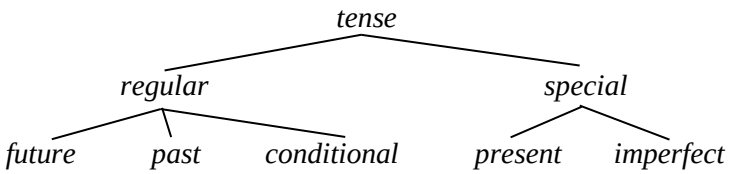
The present and imperfect of *bod* are used to express present and imperfect meanings with standard verbs, as the following illustrate:

(17) Mae Megan yn gadael.
be.PRES Megan PROG leave.INF
'Megan is leaving.'
(18) Roedd Megan yn gadael.
be.IMPF Megan PROG leave.INF
'Megan was leaving.'

One might propose that these are complex or periphrastic present and imperfect forms of the copula. However, all tenses of *bod* can take a ProgP complement. What we have here, then, is not periphrasis but an independent construction which allows the language to express the meanings that certain non-existent forms would have if they existed.[5]

It is not difficult to deal with this contrast between *bod* and standard verbs with respect to tense. Following Bonami, Borsley, and Tallerman (2016), I assume the following system of values for the feature TENSE:[6]

(19)



The following constraint will ensure that standard verbs only have past, future, and conditional forms:

(20) $\begin{bmatrix} \text{LID } \textit{standard-verb} \\ \text{VFORM } \textit{fin} \end{bmatrix} \rightarrow$ [TENSE *regular*]

I assume that *standard-verb* is a supertype of the LID values of all standard verbs. Thus, (20) will ensure that the finite forms of standard verbs are never present or imperfect. There will be no comparable constraint on finite forms of *bod*, and so all five tenses will be possible.

4. Polarity and the main–complement distinction

Some further complexities involve polarity and the distinction between main and complement clauses. The former just involve the third person present tense. The latter are more widespread.

[5] See Brown et al. (2012) for discussion of the nature of periphrasis.

[6] Bonami, Borsley, and Tallerman (2016) call this feature TMA (TENSE-MOOD-ASPECT). What it is called is of no real importance.

As earlier examples indicate, in affirmative declarative clauses, the basic present tense form of *bod* is *mae*. Different forms appear in negative declarative, and interrogative or conditional clauses.[7]

(21) Dydy Gwyn ddim yn yr ardd.
be.PRES Gwyn NEG in the garden
'Gwyn is not in the garden.'

(22) a. Ydy Gwyn yn yr ardd?
be.PRES Gwyn in the garden
'Is Gwyn in the garden?'

b. os ydy Gwyn yn yr ardd
if be.PRES Gwyn in the garden
'if Gwyn is in the garden'

These examples have definite subjects. Different forms appear with an indefinite subject, as the following show:

(23) Does neb yn yr ardd.
be.PRES nobody in the garden
'Nobody in the garden.'

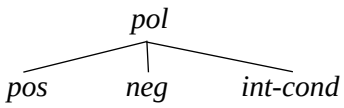
(24) a. Oes unrhyw un yn yr ardd?
be.PRES anybody in the garden
'Is anybody in the garden?'

b. os oes unrhyw un yn yr ardd
if be.PRES anybody in the garden
'if anybody is in the garden'

Clearly, there are some important complexities here.[8]

The facts suggest that we need a POL(ARITY) feature with three values: *pos*(*itive*), *neg*(*ative*), and *int*(*errogative*)-*cond*(*itional*). With *pol*(*arity*) as an unspecified value, this gives us the following values:

(25)



[7] A few ordinary verbs have distinct negative forms in some varieties (see Borsley and Jones 2005: 50-52), but most ordinary verbs take the same form in the three types of sentence that we are distinguishing here.

[8] *Dydy* and *does* are morphologically negative but not semantically negative. As discussed in Borsley and Jones (2005) and Borsley (2006), negative sentences must contain a prominent semantically negative constituent. This entails that *dydy* must co-occur with a negative post-subject adverb such as *ddim* and that *does* must co-occur with a negative subject such as *neb*.

Mae will be [POL *pos*], *dydy* and *does* [POL *neg*], and *ydy* and *oes* [POL *int-cond*]. This means the following constraint for *mae* instead of (16):

(26) $\begin{bmatrix} \text{INFL} \begin{bmatrix} \text{LID } \textit{bod} \\ \text{VFORM } \textit{fin} \\ \text{TENSE } \textit{pres} \\ \text{POL } \textit{pos} \end{bmatrix} \end{bmatrix} \rightarrow$ [PHON *mae*]

Assuming that the subject of a finite verb is the first member of its COMPS list, *dydy* and *ydy* will have NP[DEF +] as the first member of their COMPS list, and *does* and *oes* will have NP[DEF –]. For *dydy* and *does*, this means the following constraints:

(27) $\begin{bmatrix} \text{INFL} \begin{bmatrix} \text{LID } \textit{bod} \\ \text{VFORM } \textit{fin} \\ \text{TENSE } \textit{pres} \\ \text{POL } \textit{neg} \end{bmatrix} \\ \text{COMPS} < \text{NP[DEF} +], \ldots > \end{bmatrix} \rightarrow$ [PHON *dydy*]

(28) $\begin{bmatrix} \text{INFL} \begin{bmatrix} \text{LID } \textit{bod} \\ \text{VFORM } \textit{fin} \\ \text{TENSE } \textit{pres} \\ \text{POL } \textit{neg} \end{bmatrix} \\ \text{COMPS} < \text{NP[DEF} -], \ldots > \end{bmatrix} \rightarrow$ [PHON *does*]

Ydy and *oes* will be a result of similar constraints with [POL *int-cond*] instead of [POL *neg*].

There is more to be said here. There is evidence that the values *pos* and *neg* form a natural class. Both [POL *pos*] and [POL *neg*] forms appear in many contexts, especially declarative main clauses and many complement clauses. This suggests that they should be grouped together. But there is also evidence that *neg* and *int-cond* form a natural class. Both [POL *int-cond*] and [POL *neg*] forms forms appear in interrogatives and conditionals. The following illustrate the latter:

(29) a. Dydy ’r ddafad ddim yn yr ardd?
be.PRES the sheep NEG in the garden
‘Is the sheep not in the garden?’

b. os dydy 'r ddafad ddim yn yr ardd?
if be.PRES the sheep NEG in the garden
'if the sheep is not in the garden'

Moreover, *bod* has certain reduced forms which can appear where both [POL *neg*] and [POL *int-cond*] forms appear. Thus, (30a) has *dy* where *dydy* might appear, and (30b) and (30c) have it where *ydy* might appear:

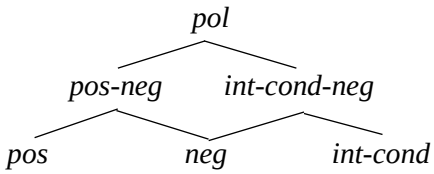
(30) a. Dy 'r ddafad ddim yn yr ardd.
be.PRES the sheep NEG in the garden
'The sheep is not in the garden.
b. Dy 'r ddafad yn yr ardd?
be.PRES the sheep in the garden
'Is the sheep in the garden.'
c. os dy 'r ddafad yn yr ardd?
if be.PRES the sheep in the garden
'if the sheep is in the garden'

Similarly, (31a) has *'s* where *does* might appear and (31b) has it where *oes* might appear.

(31) a. 'S neb yn yr ardd.
be.PRES nobody in the garden
'Nobody in the garden.'
b. 'S unrhyw un yn yr ardd.
be.PRES anybody in the garden
'Is anybody in the garden?'

We can treat both *pos* and *neg* and *neg* and *int-cond* as natural classes by proposing the following system of values:

(32)



With this system we can say that declarative main clauses and many complement clauses are [POL *pos-neg*] and that interrogatives and conditional clauses are [POL *int-cond-neg*]. We can also say that reduced forms like *dy* and *'s* are [POL *int-cond-neg*].

We turn now to the effects of the main-complement distinction. Certain pre-verbal particles are relevant here. In affirmative declarative main clauses,

the copula, like standard verbs, may be preceded by a particle, *mi* in North Wales or *fe* in South Wales. The following illustrates:

(33) Mi/Fe fydd Gwyn yn yr ardd.
AFF be.FUT Gwyn in the garden
'Gwyn will be in the garden.'

In negative complement clauses, verbs, including the copula, may be preceded by a particle *na* (*nad* before a vowel).

(34) Dywedodd Megan [na fydd Gwyn ddim yn yr ardd].
say.PAST Megan NEG be.FUT Gwyn NEG in the garden
'Megan said Gwyn will not be in the garden.'

Harlow (1983), Willis (1998: 70-71) and Borsley and Jones (2005: 57) argue that these particles form a constituent with the following verb. It is not clear whether they are separate words or prefixes, but much the same analytic issues arise on either assumption. In either case, the facts can be handled by labelling bare verbs as [MARKING *unmarked*] and particle + verb combinations as [MARKING *marked*]. *Mi*/*fe* will then combine with an *unmarked* form which is [POL *pos*, ROOT +] and *na*(*d*) will combine with an *unmarked* form which is [POL *neg*, ROOT –].[9]

For some speakers. *mi*/*fe* only occurs with past, future, and conditional forms of the copula, and not with the present and imperfect forms. For such speakers, we can say that the particles only combine with [TENSE *regular*] forms. Other speakers allow *mi*/*fe* with present and imperfect forms of *bod* but not with the third person present tense forms. For these speakers, we can assume that *mi*/*fe* combines with any [MARKING *unmarked*] form but that third person present tense forms are [MARKING *marked*].[10]

Also relevant here are some facts discussed in Bonami, Borsley and Tallerman (2016). As they note, present forms of *bod* and, for some speakers, imperfect forms too are ungrammatical in complement clauses:

---

[9] Bonami, Borsley, and Tallerman (2016) propose that there is a three-way distinction between main clauses, complement clauses, and unbounded dependency clauses and employ a three-valued STATUS feature rather than a two-valued ROOT. Whether this is necessary is not clear to me.

[10] Southern dialects have certain special negative present tense forms of the copula. Here is an example:

(i) So 'r ddafad yn yr ardd.
be.NEG.PRES the sheep in the garden
'The sheep is not in the garden.'

These forms are confined to main clauses and hence must be [POL *neg*, ROOT +].

(35) *Dywedodd Megan [mae Gwyn yn yr ardd].
say.PRES Megan be.PRES Gwyn in the garden
'Megan said Gwyn is in the garden.'
(36) %Dywedodd Megan [roedd Gwyn yn yr ardd].
say.PRES Megan be.IMPF Gwyn in the garden
'Megan said Gwyn was in the garden.'

Instead of present forms of *bod* and for some speakers imperfect forms as well, what looks like the non-finite form *bod* appears.

(37) Dywedodd Megan [bod Gwyn yn yr ardd].
say.PRES Megan be.INF Gwyn in the garden
'Megan said Gwyn is/was in the garden.'

Bod shows agreement in the form of a clitic with a following pronoun like an ordinary non-finite verb. Thus, we have the same agreement in (38) and (39).

(38) Dywedodd Megan [ei fod o yn yr ardd].
say.PRES Megan 3SGM be.INF he in the garden
'Megan said he is/was in the garden.'
(39) Dylai Megan ei weld o.
ought Megan 3SGM see.INF he
'Megan ought to see him.'

The only difference is that the clitic marks agreement with a subject in (38) and with an object in (39). Thus, *bod* seems to be morphologically non-finite. But there is evidence that it is syntactically finite. Only finite verbs precede their subject, as *bod* does here. Moreover, only finite verbs are negated by the negative adverb *ddim*, and *bod* has this property:

(40) Dywedodd Megan [bod Gwyn ddim yn yr ardd].
say.PRES Megan be.INF Gwyn NEG in the garden
'Megan said Gwyn is/was not in the garden.'

It seems, then, that *bod* in these clauses is a form of the copula which is syntactically finite but morphologically non-finite. Thus, we need an approach which distinguishes between morphological and syntactic finiteness.

Before we outline an analysis, we should note that there is one situation in which present and imperfect forms of *bod* may appear in complement clauses. This is in complement clauses affected by an unbounded dependency such as the following (Willis 2000, 2011, Borsley 2013):[11]

[11] Some speakers have *bod* in such sentences, but others prefer present and imperfect forms.

(41) Beth mae Aled yn credu [mae Elen yn
what be.PRES Aled PROG believe.INF be.PRES Elen PROG
ei ddarllen]?
3SGM read.INF
'What does Aled believe that Elen is reading?'

(42) Beth mae Aled yn credu [roedd Elen yn
what be.PRES Aled PROG believe.INF be.IMPF Elen PROG
ei ddarllen]?
3SGM read.INF
'What does Aled believe that Elen was reading?'

It seems, then, that present and imperfect forms of *bod* are only morphologically non-finite when they are not affected by an unbounded dependency. On standard HPSG assumptions, this means when they are [SLASH {}].

Bonami, Borsley and Tallerman (2016) show that it is easy to accommodate the facts given a distinction between morphosyntactic features relevant to syntax (the value of HEAD) and morphosyntactic features relevant to morphology (the value of INFL). Normally, HEAD and INFL will have the same value as a result of the following constraint:

(43) $[\ ] \rightarrow \begin{bmatrix} \text{HEAD}\ [1] \\ \text{INFL}\ [1] \end{bmatrix}$

In [ROOT –] clauses which are [SLASH {}], the positive present tense of *bod* will be [HEAD [VFORM *fin*]] but [INFL [VFORM *inf*]] as a result of the following constraint:

(44) $\begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{LID}\ \textit{bod} \\ \text{VFORM}\ \textit{fin} \\ \text{ROOT}\ - \\ \text{TENSE}\ \textit{pres} \\ \text{POL}\ \textit{pos} \end{bmatrix} \\ \text{SLASH}\ \{\} \end{bmatrix} \rightarrow$ [INFL [VFORM *inf*]]

For speakers who have *bod* instead of imperfect forms as well the constraint will refer to [TENSE *special*].

Notice that the constraint in (44) refers to [POL *pos*] forms. What about [POL *neg*] and [POL *int-cond*] forms? [POL *neg*] forms may be *bod* (as in (40)) but may also be the ordinary present tense forms. This suggests that they

require a constraint with a disjunctive consequent. [POL *int-cond*] are ordinary present tense forms. So nothing special is required here.

5. The effect of gaps

We can turn now to examples where one of the arguments of *bod* is an unbounded dependency gap. In some cases, we see the forms of *bod* that appear in ordinary affirmative or negative clauses, but in others, we have something different.

The simplest of these cases is where a gap appears in a present tense subject position. We have examples like the following:

(45) y dyn [*mae / sy(dd) yn yr ardd]
the man be.PRES in the garden
'the man who is in the garden'

(46) y dyn [*dydy / sy(dd) ddim yn yr ardd]
the man be.PRES NEG in the garden
'the man who is not in the garden'

Here, we have not the expected forms *mae* and *dydy* but a special form *sy*(*dd*). To accommodate such examples, the constraints that are responsible for *mae* and *dydy* must be constrained to require a canonical subject. In the case of *mae*, this means the following constraint:

(47)
$$\begin{bmatrix} \text{INFL} \begin{bmatrix} \text{LID } \textit{bod} \\ \text{VFORM } \textit{fin} \\ \text{TENSE } \textit{pres} \\ \text{POL } \textit{pos} \end{bmatrix} \\ \text{COMPS} < [\textit{canon}], \ldots > \end{bmatrix} \rightarrow \quad [\text{PHON } \textit{mae}]$$

Sydd can then be analyzed as the product of the following constraint, which requires the subject to be a gap:

(48)
$$\begin{bmatrix} \text{INFL} \begin{bmatrix} \text{LID } \textit{bod} \\ \text{VFORM } \textit{fin} \\ \text{TENSE } \textit{pres} \\ \text{POL } \textit{pos} - \textit{neg} \end{bmatrix} \\ \text{COMPS} < [\textit{gap}], \ldots > \end{bmatrix} \rightarrow \quad [\text{PHON } \textit{sydd}]$$

This assumes, following Borsley (2009, 2013), that gaps appear in VALENCE lists and not just in ARG-ST lists.

We turn now to complement gaps. The copula takes the expected form if the gap is a PP, PerfP, or ProgP. The following are emphatic counterparts of (1) and (2) with a PP gap and a PerfP gap in complement:

(49) Yn yr ardd mae Gwyn.
In the garden be.PRES Gwyn
'Gwyn is IN THE GARDEN.'
(50) Wedi cysgu mae Gwyn.
PERF sleep.INF be.PRES Gwyn
'Gwyn has SLEPT.'

In both, the copula is *mae*, as we would expect. I assume the following is an emphatic counterpart of (3) with a ProgP gap in complement position:

(51) Cysgu mae Gwyn.
sleep.INF be.PRES Gwyn
'Gwyn is SLEEPING.'

There is no progressive *yn* here. But *yn* appears when the ProgP has some sort of adverbial element in initial position, as the following illustrates:

(52) Wrthi yn golchi 'r car mae Mair.
at.3SGF PROG wash.INF the car be.PRES Mair
'Mair is in the process of washing the car.'

Borsley (2015) proposes that predicative *yn* is normally deleted or suppressed when it is in initial position, hence its absence from (51). In the present context, however, the important point about (51) (and (52)) is that the copula is *mae*, as expected. The situation is different if the gap is a PredP. The following are emphatic counterparts of (4) and (5):

(53) Clyfar *mae/ydy Gwyn.
clever be.PRES Gwyn
'Gwyn is CLEVER.'
(54) Meddyg *mae/ydy Gwyn.
doctor be.PRES Gwyn
'Gwyn is A DOCTOR.'

There is no predicative *yn* in these examples just as there is no progressive *yn* in (51). However, like progressive *yn*, it appears when the PredP has some sort of adverbial element in initial position:

(55) Bron yn barod *mae/ydy Mair.
almost PRED ready be.PRES Mair
'Mair is ALMOST READY.'

(56) Bron yn fradychwr *mae/ydy o.
almost PRED traitor be.PRES he
'He is ALMOST A TRAITOR.'

But in all these examples, the copula is not *mae*, which is expected in an affirmative declarative clause, but *ydy*, which is normally confined to interrogatives and conditionals.

These examples appear to be affirmative declarative clauses. In fact they must be affirmative clauses. They have no ordinary negative counterparts.[12] The only way to negate such sentences is by negating the initial constituent with *nid*/*dim*. Thus, (57a) is ungrammatical, and only (57b) is possible:[13]

(57) a. *Cysgu dydy Gwyn ddim.
sleep.INF be.PRES Gwyn NEG
'Gwyn is SLEEPING.'

b. Nid/dim cysgu mae Gwyn.
NEG sleep.INF be.PRES Gwyn
'Gwyn is not SLEEPING.'

This suggests that these clauses are [POL *pos*], and one would expect the verb that heads them to be the same. But the verb looks like a [POL *int-cond*] form. This seems to be a second case where HEAD and INFL have different values, in this case for the feature POL. We can attribute the facts to the following constraint:

(58) $\begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{LID } \textit{bod} \\ \text{VFORM } \textit{fin} \\ \text{TENSE } \textit{pres} \\ \text{POL } \textit{pos} \end{bmatrix} \\ \text{COMPS} < [], \begin{bmatrix} \textit{gap} \\ \text{PredP} \end{bmatrix} > \end{bmatrix} \rightarrow$ [INFL [POL *int-cond*]]

[12] It seems that complement gaps are generally bad with negated forms of *bod*.

[13] Notice that *yn* does not appear here although it would not be in initial position if it did. See Borsley (2015) for some discussion.

6. Identity interpretations

We turn finally to sentences in which the copula has an identity interpretation. As discussed in Zaring (1996) and Borsley (2015, section 3), it has some distinctive properties in this use. The following is a typical example:[14]

(59) Y meddyg ydy Gwyn.
the doctor be.PRES Gwyn
'Gwyn is the doctor.'

Here, the initial constituent is understood as a complement, and there is presumably an NP gap in the normal complement position. Again, the form is *ydy*, and *mae* is not possible.

(60) *Y meddyg mae Gwyn.
the doctor be.PRES Gwyn

Examples like (59) have no verb-initial counterparts. Hence, (61) is not possible with either *mae* or *ydy*.

(61) *Mae/ydy Gwyn y meddyg.
be.PRES Gwyn the doctor

This suggests that there is a separate identity copula with a distinctive syntax. However, all its forms are identical to forms of the predicational copula, and a satisfactory analysis needs to take account of this.

Before we outline an analysis, we should note a further fact about the identity copula. As we might expect, sentences with the identity copula have no ordinary negative counterparts, and can only be negated by negating the initial constituent with *nid*/*dim*.

(62) *Y meddyg ydy Gwyn ddim.
the doctor be.PRES Gwyn NEG
'Gwyn is not the doctor.'
(63) Nid/dim y meddyg ydy Gwyn.
NEG the doctor be.PRES Gwyn
'It's not the doctor that Gwyn is.'

[14] The very different syntax of identity sentences such as (59) and sentences with a predicative nominal such as (5) argues against the approach of Van Eynde (2015), in which the latter are analysed as examples of the former.

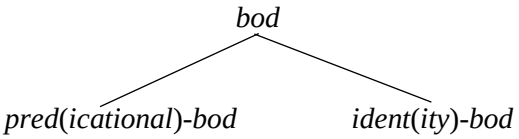
However, the identity copula can appear in both interrogatives and conditionals:[15]

(64) a. Y meddyg ydy Gwyn?
the doctor be.PRES Gwyn
'Is Gwyn the doctor?'
b. os y meddyg ydy Gwyn.
if the doctor be.PRES Gwyn
'if Gwyn is the doctor'

This suggests that the identity copula must be [POL *pos*] or [POL *int-cond*].

The facts that we are concerned with here can be handled by assuming that the two copulas are two forms of a single copula, i.e. by assuming an index *copula* with two subtypes, as follows:[16]

(65)



The syntactic and semantic properties of the two subtypes can be attributed to the following constraints:

(66) [LID *pred-bod*] $\rightarrow$

$$\begin{bmatrix} \text{ARG-ST} < [1], \begin{bmatrix} \text{HEAD}[\text{PRED}+] \\ \text{SUBJ}, < [1] > \\ \text{CONTENT}[2] \end{bmatrix} > \\ \text{CONTENT}[2] \end{bmatrix}$$

[15] Some speakers would have *mai*, which is generally viewed as complementizer, after *os* in a conditional clause, but assuming *os* combines with a [POL *int-cond*] clause, it seems reasonable to assume that *ydy* is [POL *int-cond*] in (66b).

[16] A rather similar approach is taken to the Arabic copula in Alotaibi and Borsley (forthcoming).

(67) [LID *ident-bod*] →

$$\begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{VFORM } \textit{fin} \\ \text{POL } \textit{pos} \vee \textit{int-cond} \end{bmatrix} \\ \text{ARG-ST} < [\text{INDEX}[1]], \begin{bmatrix} \textit{gap} \\ \text{INDEX}[2] \end{bmatrix} > \end{bmatrix} \\ \text{CONT} \begin{bmatrix} \textit{identity-rel} \\ \text{ARG}[1] \\ \text{ARG}[2] \end{bmatrix} \end{bmatrix}$$

The constraint in (66) ensures that the predicational copula takes a [PRED +] complement, has a subject which is the subject of its complement, and has the same interpretation as its complement. The constraint in (67) ensures that the identity-copula is finite and not negative, has a complement which is a gap, and has an identity interpretation.

But what about the forms of the two versions of the copula? In earlier discussion I have attributed the forms of the copula to constraints referring to [LID *bod*]. I will assume that all forms of the copula are the product of such constraints. With no further assumptions this would entail that parallel slots in the paradigms of two versions of the copula are filled by the same form. This is overwhelmingly what we find. The following imperfect tense examples illustrate the typical situation:

(68) Oedd Gwyn yn yr ardd.
be.IMPF Gwyn in the garden
'Gwyn was in the garden.'

(69) Yr athro oedd Gwyn.
the teacher be.IMPF Gwyn
'Gwyn was the teacher.'

But an issue obviously arises in the present tense, where identity *bod* has *ydy* and not *mae*. I propose that this is a third case where HEAD and INFL have different values, again in the value of POL. This can be attributed to the following constraint:

(70) $$\begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{LID } \textit{identity-bod} \\ \text{TENSE } \textit{pres} \\ \text{AGR}[1] \\ \text{POL } \textit{pos} \end{bmatrix} \end{bmatrix} \rightarrow \begin{bmatrix} \text{INFL} \begin{bmatrix} \text{LID } \textit{identity-bod} \\ \text{TENSE } \textit{pres} \\ \text{AGR}[1] \\ \text{POL } \textit{int-cond} \end{bmatrix} \end{bmatrix}$$

As a result of this constraint the present tense of the identity-copula will have *ydy* not only when it is [HEAD [POL *int-cond*]], as in (64a, b), but also when it is [HEAD [POL *pos*]], as in (59). Elsewhere, the identity-copula will have the same value for INFL as HEAD, and its forms will be identical to the corresponding forms of the predicational copula.

There is one further point to note about the identity-copula. This is that it does not take the form *bod* in complement clauses. We have example like the following:

(71) Dywedodd Megan [mai/taw y meddyg ydy Gwyn].
say.PAST Megan COMP the doctor be.PRES Gwyn
'Megan said that Gwyn is the doctor.'

This suggests that the constraint in (44) should be revised to refer not to [LID *bod*] but to [LID-*pred*-*bod*].

8. Concluding remarks

In the preceding pages I have developed an HPSG analysis for all the main complexities of the Welsh copula *bod*. I have assumed a variety of features, some very familiar, others less so, and I have proposed a variety of constraints to ensure that just the right forms appear. Following Bonami, Borsley, and Tallerman (2016), I have assumed a principle of blocking, whereby if the antecedents of two constraints stand in a subsumption relation, only the more specific constraint may apply. I have also made crucial use of a distinction between morphosyntactic features relevant to syntax, which are the value of HEAD, and morphosyntactic features relevant to morphology, which are the value of INFL. Normally these features have the same value, but I have proposed that there are three situations were forms of *bod* have different values for these features, one where *bod* appears rather than expected finite forms of the copula, and two where what looks like an interrogative-conditional form of *bod* appears rather than the expected positive declarative form. In all these situations, the principle of blocking ensures that certain unexpected forms appear and not the expected forms. The principle of blocking also allows a simple account of the way that what looks like the third person singular form of the verb appears with a non-pronominal subject, singular or plural.

REFERENCES

Alotaibi, A. and R. D. Borsley (forthcoming), The copula in Modern Standard Arabic, in A. Abeillé and O. Bonami (eds.), *Constraint-based Syntax and Semantics: Papers in Honor of Danièle Godard*, Stanford: CSLI Publications.

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

Bonami, O., R. D. Borsley, and M. O. Tallerman (2016), On pseudo-non-finite clauses in Welsh, *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*, 104–124.

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

Borsley, R. D. (2006), A linear approach to negative prominence, in S. Müller (ed.), *Proceedings of the HPSG06 Conference*, Stanford: CSLI Publications, 60–80.

Borsley, R. D. (2009), On the superficiality of Welsh agreement, *Natural Language and Linguistic Theory* 27, 225-265.

Borsley, R. D. (2013), On the nature of Welsh unbounded dependencies, *Lingua* 133, 1–29.

Borsley, R. D. (2015), Apparent filler–gap mismatches in Welsh, *Linguistics* 53-995-1029.

Borsley, R. D. and B. M. Jones (2005), *Welsh Negation and Grammatical Theory*, Cardiff: University of Wales Press.

Borsley, R. D., M. Tallerman and D. Willis (2007), *The Syntax of Welsh*, Cambridge: Cambridge University Press.

Brown, D., M. Chumakina, G. G., Corbett, G. Popova and A. Spencer, (2012), Defining 'periphrasis': key notions, *Morphology*, 22, 233–275.

Harlow, S. (1983), Celtic relatives, *York Papers in Linguistics* 10, 77–121.

Jones, B. M. (2010), *Tense and Aspect in Informal Welsh*. Berlin & New York: Mouton de Gruyter.

Pollard, C. and I. A. Sag (1994), *Head-driven Phrase Structure Grammar*, Chicago: University of Chicago Press.

Van Eynde, F. (2015), *Predicative constructions: from the Fregean to a Montagovian treatment*, Stanford: CSLI Publications.

Willis, D. W. E. (1998), *Syntactic Change in Welsh: A Study of the Loss of Verb-Second*, Oxford: Clarendon Press.

Willis, D. W. E. (2000). On the distribution of resumptive pronouns and wh-trace in Welsh, *Journal of Linguistics* 36, 531–573.

Willis, D. W. E. (2011). The limits of resumption in Welsh wh-dependencies. In: Rouveret, A. (Ed.), *Resumptive Pronouns at the Interfaces*, Amsterdam: Benjamins, pp. 189–222.

Zaring, L. (1996), Two 'be' or not two 'be': Identity, predication and the Welsh copula, *Linguistics and Philosophy* 19, 103–142.

Analogy-based Morphology: The Kasem number system

Matías Guzmán Naranjo
Université Paris Diderot

Proceedings of the 26th International Conference on
Head-Driven Phrase Structure Grammar

University of Bucharest

Stefan Müller, Petya Osenova (Editors)

2019

CSLI Publications

pages 26–41

http://csli-publications.stanford.edu/HPSG/2019

Guzmán Naranjo, Matías. 2019. Analogy-based Morphology: The Kasem number system. In Müller, Stefan, & Osenova, Petya (Eds.), *Proceedings of the 26th International Conference on Head-Driven Phrase Structure Grammar, University of Bucharest*, 26–41. Stanford, CA: CSLI Publications.

Abstract

This paper presents a formalization of proportional analogy using typed feature structures, which retains all key elements of analogical models of morphology. With the Kasem number system as an example, I show that using this model it is possible to express partial analogies which are unified into complete analogies. This paper is accompanied by a complete TRALE implementation.

Proportional analogy (PA) approaches to morphology are grounded on the idea that inflection systems are made up of relations between fully inflected items of a paradigm (Blevins, 2006, 2007, 2008, 2016; Neuvel, 2001; Singh et al., 2003; Singh & Ford, 2003) instead of individual morphemes, positions classes, morphological processes, rule blocks, etc. Proportional analogies are usually written as A:B::C:D, meaning that A is to B as C is to D. For example, a number of Kasem nouns exhibits the following relation between singular and plural: *agsɪ*:*agsa* ('candy'), which, modulo ATR harmony, can be generalize to: (sg)*Xɪ*:*Xa*(pl). Using this analogy, we can deduce the singular form *alapʊlɪ* ('airplane') from its plural *alapʊla*. This analogy has the property that it is a non-directional relation, i.e. there is no stem from which the singular and the plural are formed, nor does the singular serve as the base for the plural and vice versa.

Analogical models of morphology are attractive for several reasons. First of all, they make very few assumptions and are conceptually very simple. In PA models, there is no need for stems, bases, morphemes, or other sublexical elements besides those needed in the phonology. Second, PA can capture relations between any two cells in a paradigm, something which realizational approaches sometimes struggle with. Despite those advantages, there have been no serious attempts at formalizing proportional analogy. Additionally, the lack of formalization has the consequence that we do not know what the limits of PA are. It is unclear whether or not morphological systems which cannot be captured analogically exist. Neither do we know what the formal properties of PA are in morphology.

This paper presents a formalization of a purely analogical model of morphology in HPSG. The system uses reentrancies and `append` to express analogies between the cells of a paradigm. Combined with the use of underspecification and multiple inheritance, this model is able to express partial analogies for various morphological processes. As a case study, I present a partial analysis of the Kasem number system. This paper is accompanied by a full implementation in TRALE (Meurers et al., 2002; Penn, 2004; Müller, 2007).[1]

1 Kasem number classes

I will focus on the Kasem (Howard, 1969, 1970; Niggli & Niggli, 2007) number system as an illustrative example of complex multiple inheritance in inflectional

[†]I thank the anonymous reviewers and conference participants for their helpful comments.

[1]The code can be found at https://gitlab.com/abm-collection/kazem.

morphology (Guzmán Naranjo, 2019). Kasem nouns inflect for singular and plural; the challenge consists in the large number of inflection classes. Number inflection in Kasem can be analyzed as being composed of two non-suffixal (*stem*) processes, one or two suffixal singular markers, and one or two suffixal plural markers.

Like other West African languages, Kasem has ATR harmony with five +ATR vowels (ə, e, i, o, u), and five -ATR vowels (a, ɛ, ɪ, ɔ, ʊ). Contrasts are shown in (1). Besides a small number of exceptions, all vowels in a word must have the same ATR value as shown in (1). However, ATR harmony does not need to hold across members of a compound, as can be seen in (2). To abstract away from ATR harmony, I will use capital letters to represent Kasem vowels (A, E, I, O, U).

(1)

	singular	plural	gloss	
a.	colo	cwəəlu	'kilogram'	+
b.	cɔlɔ	cwaalʊ	'girl that likes going out with men'	-
c.	peeli	peelə	'shovel, spade'	+
d.	pɛɛlɪ	pɛɛla	'bean cake'	-
f.	vəlu	vələ	'traveller'	+
e.	valʊ	vala	'farmer'	-
g.	yiri	yirə	'type, kind'	+
h.	yɪrɪ	yɪra	'name'	-

(2)

	singular	gloss
a.	tɔn-yeenu	'scholar, scientist'
b.	tapwal-bu	'kidney'
c.	kaloŋ-zɔŋɔ	'Martial Eagle'
d.	bugə-sɔŋɔ	'tree species'

The singular is marked by a vowel and sometimes also by a consonant in the final syllable. There are at least 10 different singular vowel markers shown in (3)[2].[3] There is no obvious systematicity between singular and plural vowel marker combinations.

[2]Since tone is identical for singular and plural forms, tone marking is omitted in the present paper.

[3]I base the analysis on the dictionary by Niggli & Niggli (2007). Some speakers report forms different from those in the dictionary (Zaleska, 2017).

(3)

	singular	plural	sg marker	gloss
a.	banyɪɪr*ʊ*	banyɪɪrʊ	∅	‘guinea-corn’
b.	vw*e*	vwə	E	‘shelter’
c.	nabar*a*	nabarɪ	A	‘river’
d.	t*ɛɛ*	taa	EE	‘sling’
e.	nu-nakw*ɪ*	nu-nakwa	I	‘grandmother’
f.	surb*ɪa*	surbɛ	IA	‘kind of plant’
g.	pʊpɔn*ɔ*	pʊpwaanʊ	O	‘manure’
h.	diin*u*	diinə	U	‘rodent’
i.	kay*aa*	kayɛ	AA	‘round straw basket’
j.	b*ii*	biə	II	‘marble, ball’

Singular consonant markers are shown in (4). There are two types of consonant markers: onset consonants in the final syllable and coda consonants in the final syllable. Nouns can only use one of the those two strategies.

(4)

	singular	plural	sg marker	gloss
a.	ŋwam-pʊ*g*ʊ	ŋwam-pʊrrʊ	-g-	‘scale of wound’
b.	gwa*k*a	gwagsɪ	-k-	‘luggage rack’
c.	nato*ŋ*o	nantwəənu	-ŋ-	‘roof vent’
d.	co*r*o	ceeni	-r-	‘hen, fowl, chicken’
e.	kʊkɔ*n*ɔ	kʊkwarʊ	-n-	‘kind of fish’
f.	lu-sɪʊ*n*	lu-sɪɪrʊ	-n	‘metal sponge’
g.	mɪ*m*	mɪna	-m	‘millet’
h.	do*ŋ*	donnə	-ŋ	‘mate, fellow’

As singular forms, plural forms are marked by a vowel and sometimes by a consonant in their final syllable. The examples in (5) and (6) show vowel and consonant markers for the plural, respectively. Although there are some striking similarities between singular and plural markers, there is more variety in the singular than in the plural.

(5)

	singular	plural	pl marker	gloss
a.	manduru	mandurr*u*	∅	‘spoon’
b.	manlaa	manl*ɛ*	E	‘chamaleon’
c.	tɪgagɪrʊ	tɪgagɪr*a*	A	aardvark
e.	tɛɛ	t*aa*	AA	‘sling’
d.	gwala	gwal*ɪ*	I	‘slave rider’
e.	bu	b*iə*	IA	‘fruit, grain’
f.	kogo	kor*u*	U	‘kind of shrub’

(6)

	singular	plural	pl marker	gloss
a.	sʊgʊ	sʊ*m*	-m	'knife, razor'
b.	vɔsaŋa	vɔsɛ*n*	-n	'type of shrub'
c.	nʊŋʊ	nʊ*nn*ʊ	-n-	'marrow'
d.	balogo	balwə*r*u	-r-	'lizard'
e.	karga	kar*s*ɪ	-s-	'mite, bug'

Finally, there are two non-affixal processes which mark the plural: lengthening of the vowel of the penultimate syllable, gemination of the onset of the final syllable, and diphthongization of the vowel of the penultimate syllable. As shown in (7),[4] these two processes can occur either separately (a-e) or together (f-k).

(7)

	singular	plural	gloss
a.	lampo	lamp*oo*ru	'tax'
b.	lemu	lem*uu*ru	'orange'
c.	kalenziu	kalenz*ii*ru	'basket for fishing'
d.	tokunu	toku*nn*u	'seeds of baobab fruit'
e.	suru	su*rr*u	'shrub species'
f.	pɔlɔ	p*wall*ʊ	'saddle, seat'
g.	tasɔrɔ	tas*waa*rʊ	flint lighter
h.	soro	s*wəə*ru	mucilaginous herb used in soup
i.	yolo	y*wəll*u	'bag, sack'
j.	ni-viu	ni-v*wee*ru	'mouth breath'
k.	niu	n*wee*ru	'mirror, glass'

Besides the segmental markers discussed so far, the singular is related to the plural by one of six possible alternations shown in (8).[5] The alternations Xσ-X (a-c) and X-Xσ (d-f) are the mirror image. In σ-0, the singular has one syllable more than the plural, whereas in X-Xσ, the plural has one syllable more than the singular. There is a correspondence between the syllables denoted by X, although this correspondence is mediated by non-suffixal markers such as lengthening and diphthongization. In the alternation Xσ-Xσ (g-h), the singular and the plural have the same number of syllables, but there is no strict correspondence between the final syllable. The following three alternations form subtypes of this alternation. The alternation X-X (i-j) applies when the singular and the plural are identical (modulo lengthening and diphthongization). In XV-XV (k-l), only the vowel of the final syllable varies, while in XOY-XOY, only the onset of the final syllable varies (again modulo lengthening and diphthongization).

[4]There are two additional vowel mutations which I will not address in this paper.

[5]There are some additional fixed singular-plural alternations which do not interact with any individual affixal marker or non-affixal process. I do not discuss those in this paper.

(8)

	singular	plural	pattern	gloss
a.	zʊŋ*a*	zwɪ	Xσ-X	‘calabash’
b.	si*gə*	si	Xσ-X	‘Hartebeest’
c.	kapa-sɪŋ*a*	kapa-sɪn	Xσ-X	‘Cobra’
d.	kalanjoo	kalanjoo*ru*	X-Xσ	‘clam’
e.	kɔn	kɔɔ*na*	X-Xσ	‘Antelope’
f.	tangwam	tangwa*na*	X-Xσ	‘earth shrine’
g.	kaman-poŋ*o*	kaman-pwən*nu*	Xσ-Xσ	‘white maize’
h.	cɔ*g*ɔ	cɔ*r*ʊ	Xσ-Xσ	‘pond’
i.	kantwana	kantwana	X-X	‘sp. of fruit’
j.	suru	surru	X-X	‘sp. of shrub’
k.	lampo-joŋnu	lampo-joŋnə	XV-XV	‘tax-collector’
l.	kog-zono	kog-zwəənu	XV-XV	‘sp. of shrub’
m.	cɪŋʊ	cɪnnʊ	XOY-XOY	‘tapeworm’
n.	tasugu	tasuru	XOY-XOY	‘covering lid’

Affixal markers, non-affixal markers, and alternations being simple on their own, the system shows considerable complexity in that it has around 150 classes which arise from the combinations of individual markers and alternations. Most of the singular markers can appear together with most of the plural markers, and in several different singular–plural relations. Although many combinations are not attested, it is not evident whether these gaps are accidental or caused by hard grammatical constraints. I do not attempt to explain these gaps in this paper.

The previous discussion of Kasem is not complete, and there are additional non-affixal and affixal markers in the system. However, the classes described in this paper account for around 80% to 85% of Kasem nouns listed in Niggli & Niggli (2007).

2 Analogy-based Morphology: Kasem

The basic assumption of AbM (Analogy-based Morphology) is that lexemes list all their inflected forms.[6] This comes directly from the idea in PA models that lexemes are the set of inflected forms in a paradigm (Blevins, 2016).[7] In the case of Kasem, nouns list their singular and plural forms as in Figure 1.[8] Unlike the representations used by Bird & Klein (1994) and Monachesi (2005) which avoid the use of explicit syllable trees, both SINGULAR and PLURAL are lists of syllables.

[6] Or at least all forms which take part in analogical relations.

[7] I use attribute-value pairs to represent each paradigm cell. While there are possible alternatives which might be compatible with the general HPSG architecture, this approach is the most straightforward for making it computationally implementable in TRALE.

[8] I will sometimes omit the PARADIGM feature to save space in the AVMs.

The representations of phonemes, vowels, and syllables are given in Figures 2–6. Although more complex representations are possible, the distinctions made here are sufficient to capture the Kasem number system. The CORE feature in Figure 2 is a shorthand notation for the complete specification of place and manner of articulation of a segment (Bird & Klein, 1994), which does not play a direct role in the morphological analogies. These structures are organized as in the partial hierarchy in Figure 7.

$$\begin{bmatrix} \textit{lexeme} & \\ \textsc{paradigm} & \begin{bmatrix} \textsc{singular} & \textit{list(syllable)} \\ \textsc{plural} & \textit{list(syllable)} \end{bmatrix} \end{bmatrix}$$

Figure 1: Lexeme

$$\begin{bmatrix} \textit{phoneme} & \\ \textsc{core} & \textit{phon} \end{bmatrix}$$

Figure 2: Phoneme

$$\begin{bmatrix} \textit{vowel} & \\ \textsc{long} & \textit{bin} \\ \textsc{atr} & \textit{bin} \end{bmatrix}$$

Figure 3: Vowel

$$\begin{bmatrix} \textit{syllable} & \\ \textsc{nucleus} & \textit{vowel} \end{bmatrix}$$

Figure 4: Syllable

$$\begin{bmatrix} \textit{cv-syll} & \\ \textsc{onset} & \textit{consonant} \end{bmatrix}$$

Figure 5: CV Syllable

$$\begin{bmatrix} \textit{vc-syll} & \\ \textsc{coda} & \textit{consonant} \end{bmatrix}$$

Figure 6: VC Syllable

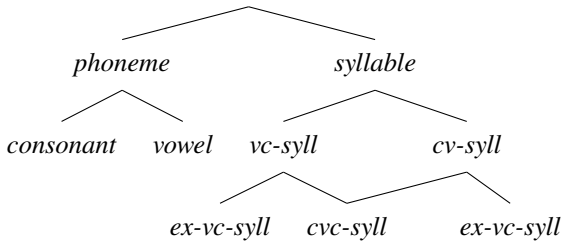


Figure 7: syllable-phoneme hierarchy

Given those simple assumptions, we can express complete analogical relations as constraints on the SINGULAR and PLURAL features. For instance, the complete analogy for nouns such as *agsɩ–agsa* ('candy'), which have non-alternating stems and *-I/-A* markers, is shown in Figure 8.

However, from the perspective of traditional PA models, a particularly challenging aspect of Kasem is the existence of number markers that behave independently of each other. To give an example, we need to be able to express the fact

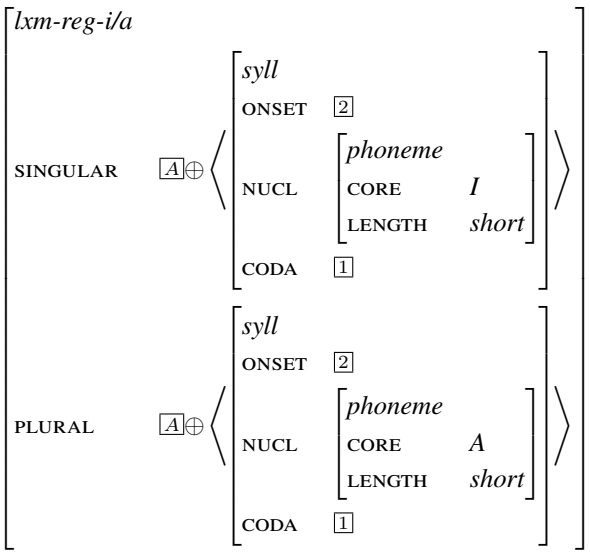


Figure 8: Non-alternating-A–I

that *-O* and *-I* are singular markers independently of the plural marker they appear in opposition to. This generalization runs opposite to PA models, which usually claim that morphological systems rely exclusively on oppositions. However, we would miss an important generalization without being able to express these partial patterns. Similarly, we need to be able to express non-affixal markers (lengthening and diphthongization) as independent processes, which can occur together, and with different suffix combinations. To model these facts, we need to decompose complete analogical relations into partial analogies.

We start by defining non-affixal relations. Figure 9 describes the analogy which ensures no lengthening. The reentrancies in the feature LONG ensure that there are no discrepancies between the length of the singular and the plural vowels of the penultimate syllable,[9] while the constraints of the coda ensures that there is no gemination of the consonant.

The opposite, vowel and consonant lengthening, is achieved by the constraints in Figures 10 and 11, respectively. In Figure 10, we impose the constraint that the nucleus of the penultimate syllable of the plural must be long. The constraint in Figure 11 ensures that the coda of the penultimate and onset of the final syllables of the plural are identical to the onset of the final syllable of the singular, and that the penultimate syllable of the singular is CV.

[9] I treat cases where both the singular and the plural have a long penultimate syllable as cases of no lengthening.

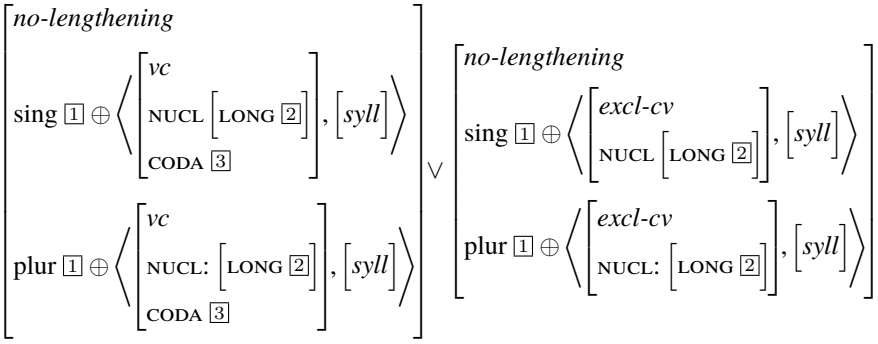


Figure 9: No lengthening

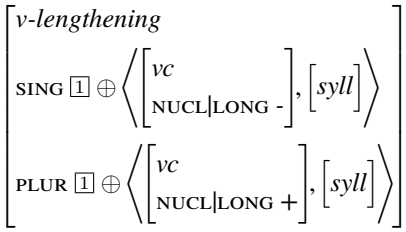


Figure 10: V-lengthening

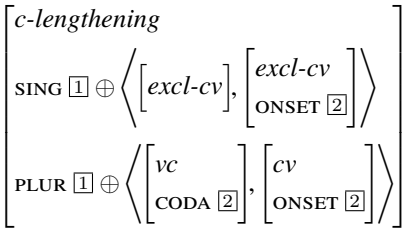


Figure 11: C-lengthening

Figures 12 and 13 ensure no diphthongization and diphthongazation to occur, respectively. No diphthongization is achieved by enforcing that the CORE of the nucleus of the penultimate syllables of the singular and the plural are identical. Diphthongization is expressed by directly specifying the CORE value of the singular as /O/ and the CORE value of the plural as /WE/.[10]

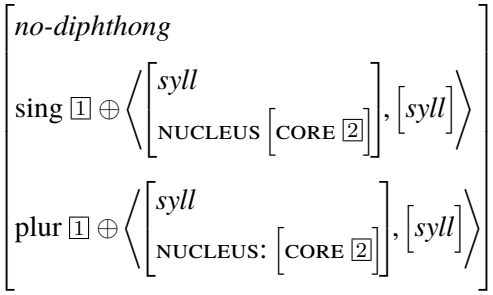


Figure 12: No diphthongization

The partial hierarchy in 14 captures the possible combinations of non-affixal processes.

Vowel and consonant suffixal markers can be captured in a straightforward way

[10] Similar constraints must be introduced for other cases of diphthongization.

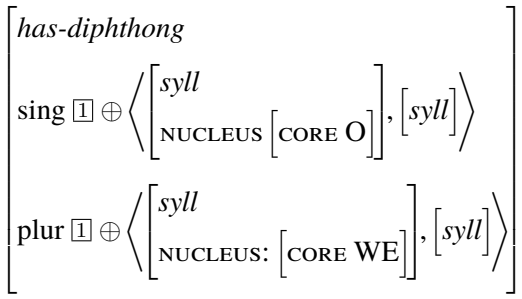


Figure 13: Diphthongization

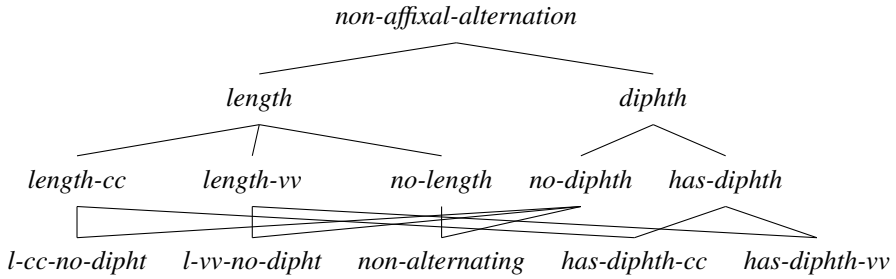


Figure 14: Non-affixal hierarchy

as well. Figures 15 and 16 give examples of vowel markers, and Figures 17 and 18 show different consonant markers.

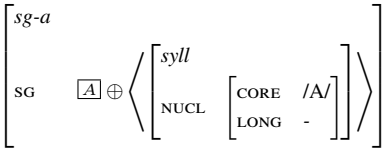


Figure 15: Suffixal marker -A

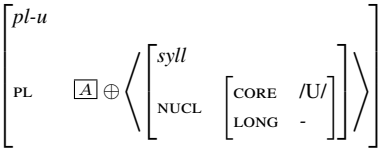


Figure 16: Suffixal marker -U

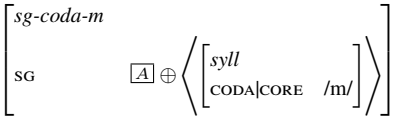


Figure 17: Suffixal marker -m

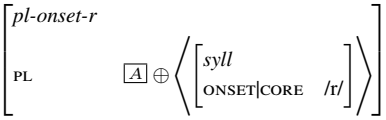


Figure 18: Suffixal marker -r-

Finally, the 6 analogical relations are what links the singular to the plural. Figures 19 to 24 present those patterns. Relation Xσ–Xσ states that both the singular and the plural have the same number of syllables and the onsets of their penultimate syllables are identical. This relation also states that the ATR value of the singular

and the plural must be identical on a syllable-by-syllable basis.[11] As mutation only occurs in the final two syllables, we state that all preceding syllables are identical for the singular and the plural.

$$\left[\begin{array}{l} \textit{X}\sigma\textit{–X}\sigma\textit{-relation} \\ \\ \textsc{paradigm} \end{array}\right.\left[\begin{array}{ll} \textsc{sg} & \boxed{A} \oplus \left\langle \left[\begin{array}{ll} \textit{syll} & \\ \textsc{onset} & \boxed{1} \\ \textsc{nucl}|\textsc{atr} & \boxed{2} \end{array}\right], \left[\begin{array}{ll} \textit{syll} & \\ \textsc{nucl}|\textsc{atr} & \boxed{3} \end{array}\right] \right\rangle \\ \\ \textsc{pl} & \boxed{A} \oplus \left\langle \left[\begin{array}{ll} \textit{syll} & \\ \textsc{onset} & \boxed{1} \\ \textsc{nucl}|\textsc{atr} & \boxed{2} \end{array}\right], \left[\begin{array}{ll} \textit{syll} & \\ \textsc{nucl}|\textsc{atr} & \boxed{3} \end{array}\right] \right\rangle \end{array}\right]$$

Figure 19: Relation Xσ–Xσ

Relation Xσ–X states that the singular has all the syllables of the plural plus one additional syllable. Because this relation does not allow for vowel lengthening in the plural the CORE of the penultimate syllables in both cells are identical. However, this relation does allow for additional consonant markers in the plural. Relation X–Xσ is almost the mirror image of relation Xσ–X, allowing for diphthongization and lengthening in the plural.

$$\left[\begin{array}{l} \textit{X}\sigma\textit{–X-relation} \\ \\ \textsc{paradigm} \end{array}\right.\left[\begin{array}{ll} \textsc{sg} & \boxed{A} \oplus \left\langle \left[\begin{array}{ll} \textit{syll} & \\ \textsc{onset} & \boxed{2} \\ \textsc{nucl} & \boxed{4}\left[\begin{array}{ll} \textsc{core} & \boxed{1} \\ \textsc{atr} & \boxed{3} \end{array}\right] \end{array}\right], \left[\begin{array}{ll} \textit{syll} & \\ \textsc{nucl}|\textsc{atr} & \boxed{3} \end{array}\right] \right\rangle \\ \\ \textsc{pl} & \boxed{A} \oplus \left\langle \left[\begin{array}{ll} \textit{syll} & \\ \textsc{onset} & \boxed{2} \\ \textsc{nucl} & \boxed{4} \end{array}\right] \right\rangle \end{array}\right]$$

Figure 20: Relation Xσ–X

Relations XV–XV and XOY–XOZ are subtypes of relation Xσ–Xσ; however, they impose additional constraints. Relation XV–XV states that the onset of the final syllable of both cells must be identical, while relation XOY–XOZ requires that the nucleus of the final syllable of both cells be identical. Finally, relation X–X simply states that, modulo lengthening and diphthongization, the singular and

[11] Since compounds can break ATR harmony, we cannot state that the final and penultimate syllables have the same ATR value.

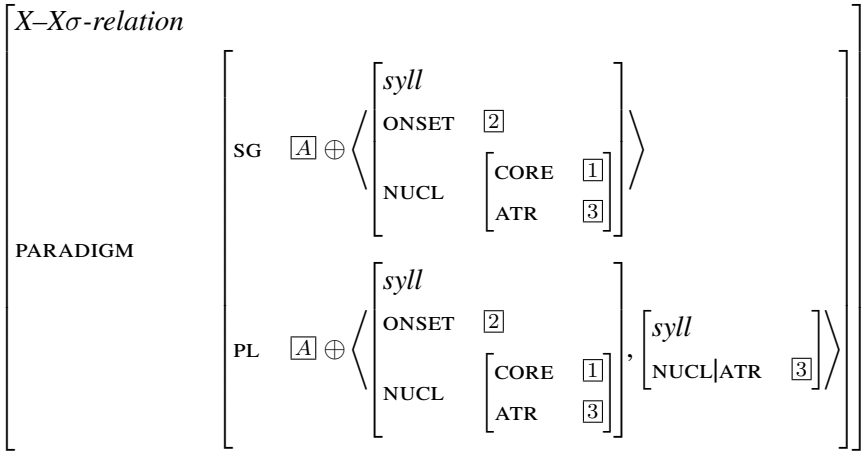


Figure 21: Relation X–Xσ

plural cells are identical.

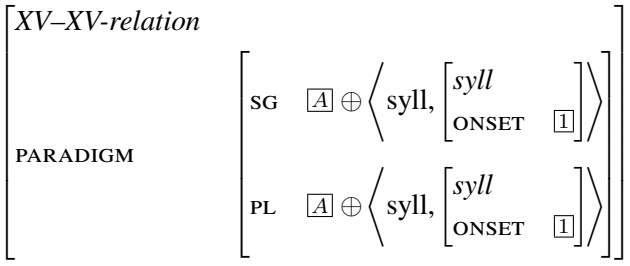


Figure 22: Relation XV–XV

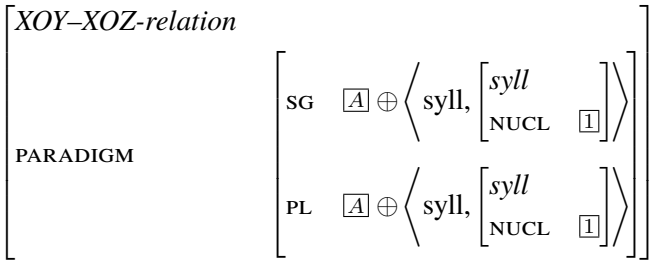


Figure 23: Relation XOY–XOZ

These constraints work together to build full inflectional classes. For example, the singular-plural pair *laancɪga–laancɪ* ('Flapped Lark') instantiates a *-g-* marker, a singular *-A*, no non-affixal mutations and the Xσ–X alternation. The complete structure of *laancɪga–laancɪ* is shown in Figure 25.

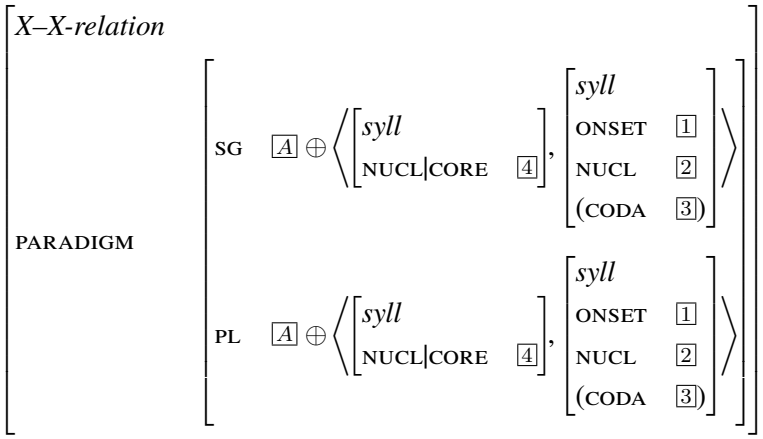


Figure 24: Relation X–X

3 Concluding remarks

The system proposed in this paper correctly captures the key aspects of PA approaches and at the same time allows for more abstract generalizations. The main advantage of this formalization over traditional PA models is that we can build complete analogies out of partial analogies, which allows us to express stem alternations without stems, and individual markers without morphemes. The advantage over realizational models like Information based Morphology (Bonami & Crysmann, 2015; Crysmann & Bonami, 2017) is that, since this system is simpler (it makes fewer assumptions), computational implementation and automatic induction (Beniamine and Guzmán Naranjo forth.) are easier to achieve. Additionally, unlike realizational models, PA models are completely non-directional. In AbM knowing the singular of a noun and its inflection class suffices to deduce its plural form, and vice versa.

This formalization is similar to the string unification approach taken by (Calder, 1989, 1991); however, there are three important differences. First, this approach does not assume that analogies are between strings, strictly speaking, but rather between phonological objects which can have as much structure as needed for the language in question (e.g. syllables, moras, etc.). The second main difference is that this model puts emphasis on being able to express partial analogies and partial descriptions to form complete analogies. Finally, while the system proposed by Calder made use of morphemes and was directional, the present implementation is neither. In the way that AbM is set up, there are no morphemes and no directional relations (at least they are not required).

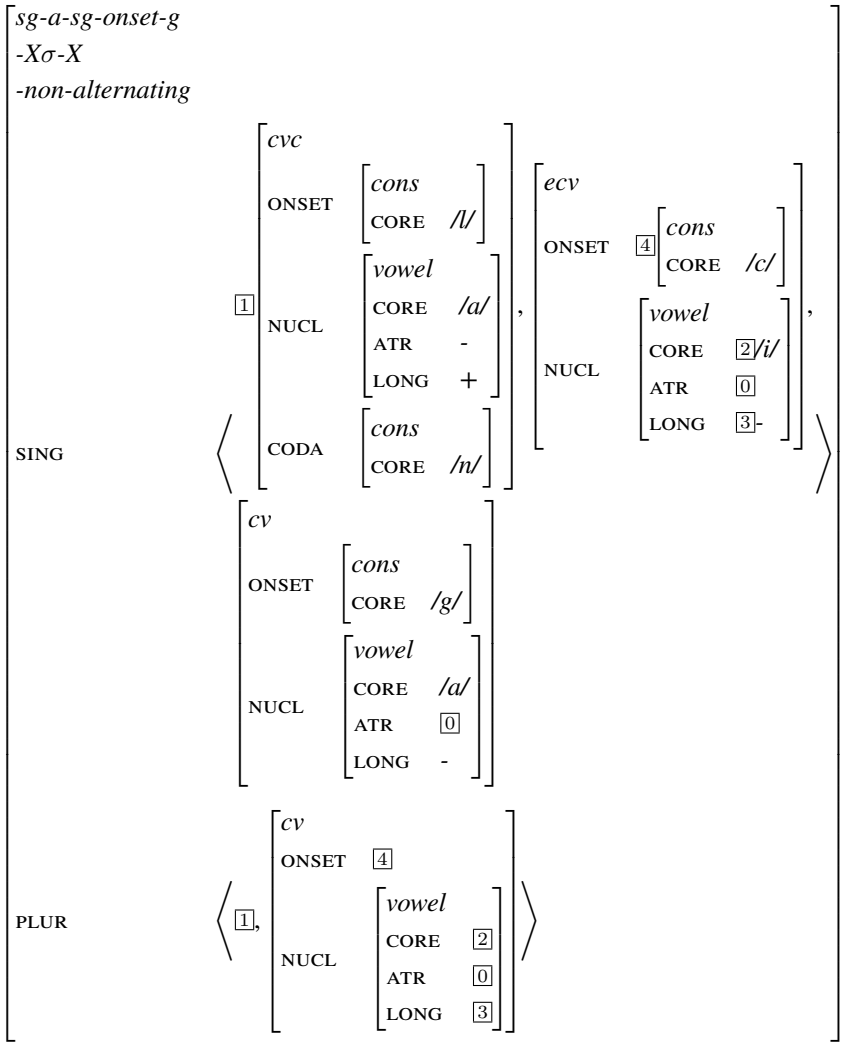


Figure 25: Full analogy for *laancıga*

References

Bird, Steven & Ewan Klein. 1994. Phonological analysis in typed feature systems. *Computational linguistics* 20(3). 455–491.

Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42(3). 531–573.

Blevins, James P. 2007. Conjugation classes in Estonian. *Linguistica Uralica* 43(4). 250–267.

Blevins, James P. 2008. Declension classes in Estonian. *Linguistica Uralica* 44(4). 241–267.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford, New York: Oxford University Press.

Bonami, Olivier & Berthold Crysmann. 2015. Morphology in Constraint-based lexicalist approaches to grammar. In Andrew Hippisley & Gregory T. Stump (eds.), *Cambridge Handbook of Morphology*, Cambridge: Cambridge University Press.

Calder, Jonathan. 1989. Paradigmatic morphology. In *Proceedings of the fourth conference on European chapter of the Association for Computational Linguistics*, 58–65. Association for Computational Linguistics.

Calder, Jonathan. 1991. *An Interpretation of Paradigmatic Morphology*. Edinburgh: University of Edinburgh dissertation.

Crysmann, Berthold & Olivier Bonami. 2017. Atomistic and holistic exponence in information-based morphology. In Stefan Müller (ed.), *Proceedings of the 24th International Conference on oceedings of the 24th International Conference on Head-Driven Phrase Structure Grammar*, 140–160. Stanford, CA: CSLI Publications.

Guzmán Naranjo, Matías. 2019. *Analogical classification in formal grammar* Empirically Oriented Theoretical Morphology and Syntax. Language Science Press.

Howard, Irwin. 1969. Kasem nominals revisited. *Working Papers in Linguistics - University of Hawaii* 10.

Howard, Irwin. 1970. Kasem nominals and the ordering of phonological rules. *Working Papers in Linguistics. University of Hawaii* 2(3). 112.

Meurers, Walt Detmar, Gerald Penn & Frank Richter. 2002. A Web-based Instructional Platform for Constraint-Based Grammar Formalisms and Parsing. In Dragomir Radev & Chris Brew (eds.), *Effective Tools and Methodologies for*

Teaching NLP and CL, 18–25. New Brunswick, NJ: The Association for Computational Linguistics. http://ling.osu.edu/~dm/papers/acl02.html. Proceedings of the Workshop held at the 40th Annual Meeting of the Association for Computational Linguistics. 7.–12. July 2002. Philadelphia, PA.

Monachesi, Paola. 2005. *The verbal complex in Romance: A case study in grammatical interfaces*, vol. 9. Oxford, New York: Oxford University Press.

Müller, Stefan. 2007. The Grammix CD Rom. a software collection for developing typed feature structure grammars. In Tracy Holloway King & Emily M. Bender (eds.), *Grammar engineering across frameworks 2007* Studies in Computational Linguistics ONLINE, 259–266. Stanford, CA: CSLI Publications. http://hpsg.fu-berlin.de/~stefan/Pub/grammix.html.

Neuvel, Sylvain. 2001. Pattern analogy vs. word-internal syntactic structure in West - Greenlandic: Towards a functional definition of morphology. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2000* Yearbook of Morphology, 253–278. Amsterdam: Springer.

Niggli, Idda & Urs Niggli. 2007. *Dictionaire bilingue kasɩm - français français - kassem*. Burkina Faso: Société Internationale de Linguistique. kassem-bf.webonary.org/.

Penn, Gerald. 2004. Balancing clarity and efficiency in typed feature logic through delaying. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, main volume*, 239–246. Barcelona, Spain: Association for Computational Linguistics.

Singh, Rajendra & Alan Ford. 2003. In Praise of Śākaṭāyana: Some Remarks on Whole Word Morphology. In Rajendra Singh, Stanley Starosta & Sylvain Neuvel (eds.), *Explorations in seamless morphology*, 66–76. New Delhi: SAGE.

Singh, Rajendra, Stanley Starosta & Sylvain Neuvel. 2003. *Explorations in seamless morphology*. New Delhi: SAGE.

Zaleska, Joanna. 2017. *Coalescence in without coalescence*: Universität Leipzig dissertation.

An incremental approach to gapping in Japanese

Petter Haugereid
Western Norway University of Applied Sciences

Proceedings of the 26th International Conference on
Head-Driven Phrase Structure Grammar

University of Bucharest

Stefan Müller, Petya Osenova (Editors)

2019

CSLI Publications

pages 42–57

http://csli-publications.stanford.edu/HPSG/2019

Keywords: HPSG, Gapping, Japanese, Incremental Parsing

Haugereid, Petter. 2019. An incremental approach to gapping in Japanese. In Müller, Stefan, & Osenova, Petya (Eds.), *Proceedings of the 26th International Conference on Head-Driven Phrase Structure Grammar, University of Bucharest*, 42–57. Stanford, CA: CSLI Publications.

Abstract

Gapping in Japanese, which is an SOV language, differs from gapping in SVO languages in that the conjuncts with the elided verbs appear in non-final position. In this paper I present an incremental approach to gapping in Japanese, where it is assumed that an argument structure type is constructed in the non-final clause(s) in the gapping construction. This type is unified with the construction type created by the final clause resulting in identical construction types for all conjuncts in the construction.

1 Introduction

Gapping is a phenomenon that poses a challenge to lexicalist approaches given the fact that the main verb of one or more of the conjuncts in these constructions is elided. Example (1) (from Sag *et al.* (1985)) shows the prototypical gapping construction with a transitive sentence in the first conjunct, and two arguments, but no verb, in the second conjunct.

(1) Kim likes Sandy, and Lee Leslie.

The constituents of the conjunct with the elided verb, *Lee* and *Leslie*, are referred to as the *remnants*, and the constituents that have their roles in the conjunct with the verb, *Kim* and *Sandy*, are referred to as *correlates*.

The examples in (2)–(6) demonstrate gapping in Japanese (from Kato (2006, p. 1–14)). (2) is a conjunction of two transitive sentences where the verb of the first conjunct is elided. (3) shows that the elided verb cannot be in the second conjunct, as in English. (4) shows that gapping also may occur with intransitive verbs. (5) shows that there may be four dependents in each conjunct, and (6) shows that there may be more than one conjunct with a gap.

(2) John-ga hon-o sosite Mary-ga hana-o katta.
John-NOM book-ACC and Mary-NOM flower-ACC bought
'John bought books, and Mary flowers.'

(3) * John-ga hon-o katta sosite Mary-ga hana-o.
John-NOM book-ACC bought and Mary-NOM flower-ACC

(4) John-ga kayobi-ni sosite Mary-ga doyoubi-ni hasiru.
John-NOM Tuesday-ON and Mary-NOM Saturday-ON run
'John runs on Tuesdays, and Mary on Saturdays.'

(5) John-ga kinou Fred-ni hon-o sosite Mary-ga kyou
John-NOM yesterday Fred-DAT book-ACC and Mary-NOM today
Susan-ni hana-o katta.
Susan-DAT flower-ACC bought

†I would like to thank three anonymous reviewers and the audience at the HPSG 2019 conference in Bucharest, Romania, for very useful comments and suggestions.

'John bought books for Fred yesterday, and Mary flowers for Susan today.'

(6) John-ga hon-o sosite Mary-ga hana-o sosite Fred-ga
John-NOM book-ACC and Mary-NOM flower-ACC and Fred-NOM
pen-o sosite Sue-ga kitte-o katta.
pen-ACC and Sue-NOM stamp-ACC bought
'John bought books, Mary flowers, Fred pens, and Sue stamps.'

According to Ross (1970), gapping operates forward in SVO languages like English. This is referred to as forward gapping (see (7)). And in SOV languages like Japanese, the verb appears in the last conjunct in gapping constructions (Ross, 1970). This is referred to as backward gapping (see (8)).

(7) a. SVO + SVO + SVO + ... + SVO ⇒
b. SVO + SO + SO + ... + SO

(8) a. SOV + SOV + SOV + ... + SOV ⇒
b. SO + SO + SO + ... + SOV

Gapping in Japanese is sometimes equaled to Right Node Raising (Kato, 2006, p. 55). Yatabe and Tanigawa (2018) claim that Japanese does not have gapping, only Right Node Raising. They base their argument on the fact that the apparent ellipsis only is at the right node of the conjunct, illustrated in (9), where it appears that the whole right node *nani o kau to yakusoku shita no* is gapped. According to Yatabe and Tanigawa (2018), the reading of (10), where the verb *kau* and the complementizer *to* are not gapped, should be the same as the reading of (9) if Japanese had gapping, but this reading is not available, and they present this as evidence that Japanese does not have gapping.

(9) [Masao wa] ashita, (soshite) [Hanako wa] asatte
[Masao TOP] tomorrow (and) [Hanako TOP] day after tomorrow
[nani o] kau to yakusoku shita no?
[what ACC] buy-PRES COMP promise do-PAST NML
'What has Masao promised to buy tomorrow, and what has Hanako promised to buy the day after tomorrow?'

(10) ?* [Masao wa] ashita kau to, (soshite) [Hanako wa]
[Masao TOP] tomorrow buy-PRES COMP (and) [Hanako TOP]
asatte [nani o] kau to yakusoku shita
day after tomorrow [what ACC] buy-PRES COMP promise do-PAST
no?
NML
'Same as (9)'

In this paper, I will not discuss whether or not gapping exists in Japanese. The aim will be to present an analysis of the examples in (2)–(6), where a verb is shared by two (or more) conjuncts. However, in the following I will refer to this phenomenon as gapping.

Gapping is a widely discussed phenomenon in the linguistic literature, and it is one of the hardest phenomena to handle in a grammar implementation. The analyses of gapping rarely find their way into grammar implementations. In this paper, the focus will be on implementability of accounts of gapping, and hence the perspective will be different from other, more theoretical, approaches. The hope is that it can complement the other approaches and show a way forward to how analyses of gapping can be implemented. Guided by limitations imposed by concerns about implementability and parser efficiency,[1] the account I will present is limited in scope, and only accounts for a fraction of the data on gapping found in the literature.[2]

2 Gapping in HPSG

In lexicalist theories, the syntactic structure is built up around heads which carry detailed information about the structure that will be built around them. This makes gapping constructions hard to account for, given that the verb, which is the head of the sentence, is missing.

Most HPSG approaches to gapping makes use of the linearization approach (Kathol, 1995; Beavers and Sag, 2004; Chaves, 2005; Crysmann, 2008; Kim and Cho, 2012). In this approach, the feature DOM(ain) (Reape, 1994) represents the linear order of phonological items, and this order is allowed to be different from the order in the constituent tree. This separation of linear order and constituent tree is powerful, and although relational constraints may be added to the grammar in order to impose restrictions on the order of the phonological items, it may put a heavy burden on the parser if it is not properly constrained.

Abeillé *et al.* (2014) present an alternative, construction-based HPSG approach to gapping. It is based on Mouret (2006), and does not make use of linearization. Instead it assumes that the constituents in the conjuncts with the elided verb, the *remnants*, form a non-headed constituent where the *synsem*s of the remnants are entered onto a CLUSTER list in HEAD. This constituent undergoes a unary rule *head-fragment-ph*. This *head-fragment* rule checks the HEAD values of the remnants (via the CLUSTER list) against the HEAD values of the correlates, which the rule accesses via the context SAL(ient)-(sub)UTT(erance) feature (SAL-UTT) (see (12)).

[1] The analysis presented is possible to implement with the LKB system (Copestake, 2002).

[2] More complex examples of gapping, for example including chains of control verbs as in (11) (from Sag *et al.* (1985)) and examples like (9), will be topic for future research.

(11) Pat wanted to try to go to Berne, and Chris to Rome.

(12) Syntactic constraints on *head-fragment-ph* (Abeillé *et al.*, 2014)

$$\textit{head-fragment-ph} \Rightarrow \left[\begin{array}{l} \text{CONTEXT}\,|\,\text{SAL-UTT}\left\langle \begin{bmatrix}\text{HEAD}\ \boxed{H_1}\\ \text{MAJOR}\ +\end{bmatrix}, \ldots, \begin{bmatrix}\text{HEAD}\ \boxed{H_n}\\ \text{MAJOR}\ +\end{bmatrix}\right\rangle \\ \text{CATEGORY}\,|\,\text{HEAD}\,|\,\text{CLUSTER}\left\langle \begin{bmatrix}\text{HEAD}\ \boxed{H_1}\\ \text{MAJOR}\ +\end{bmatrix}, \ldots, \begin{bmatrix}\text{HEAD}\ \boxed{H_n}\\ \text{MAJOR}\ +\end{bmatrix}\right\rangle \end{array}\right]$$

In example (1), repeated here as (13), the correlates are *Kim* and *Sandy*. Consequently, their *synsem*s can be accessed via the SAL-UTT feature. The *head-fragment* rule checks that their head values match with those of the remnants, *Lee* and *Leslie*. In this way, the subcategorization frames of the conjuncts with elided verbs are guaranteed to correspond to the subcategorization frames of the conjunct with the verb.

(13) Kim likes Sandy, and Lee Leslie.

The tree in Figure 1 is an illustration of how the features SAL-UTT and CLUSTER account for the matching of the argument frames of the initial conjunct and a conjunct with an elided verb.[3]

In addition to the syntactic constraint shown in (12), there is a separate constraint on the *head-fragment* rule that assigns the semantic predicate that was assigned to the correlates, to the remnants. In (13), this means that the semantics of the second conjunct is *like'(lee',leslie')*.

There are some challenges to the approach to gapping in Abeillé *et al.* (2014), and in particular the syntactic constraints on *head-fragment-ph* shown in (12). While it is possible to match the HEAD values of the synsems on the CLUSTER list with those on the SAL-UTT list, one needs to know the length of the SAL-UTT and CLUSTER lists, unless one introduces some extra functionality for list matching. If there are two correlates and two remnants, as in (13), one needs a *head-fragment-ph* type that has SAL-UTT and CLUSTER lists of length two, and which matches the HEAD values of the two first items and the HEAD values of the two second items. If there are three correlates and three remnants, one needs another *head-fragment-ph* type with lists of length three, and so on. In addition, if one allows the matching elements on the lists to come in different order, the number of matching rules required becomes large.

A more serious problem with the *head-fragment* rule is how to make the items on the SAL-UTT and CLUSTER lists accessible to the rule at the same time. The access to the correlates on the SAL-UTT list in the *head-fragment* rule presupposes that the conjunct with the correlates has been parsed when the *head-fragment* rule is applied, and that the the correlates has been put on a SAL-UTT list. This list will have to be made accessible to the coordination rule, which pushes it down into the *head-fragment* rule, via the *head-comps* rule, as shown in Figure 1. The fact

[3]The analysis presented in Abeillé *et al.* (2014) is more elaborate than the illustration shown here. It includes a functionality that allows them to match constituents with differing HEAD values, like *adv* and *prep*, and *noun* and *adj*.

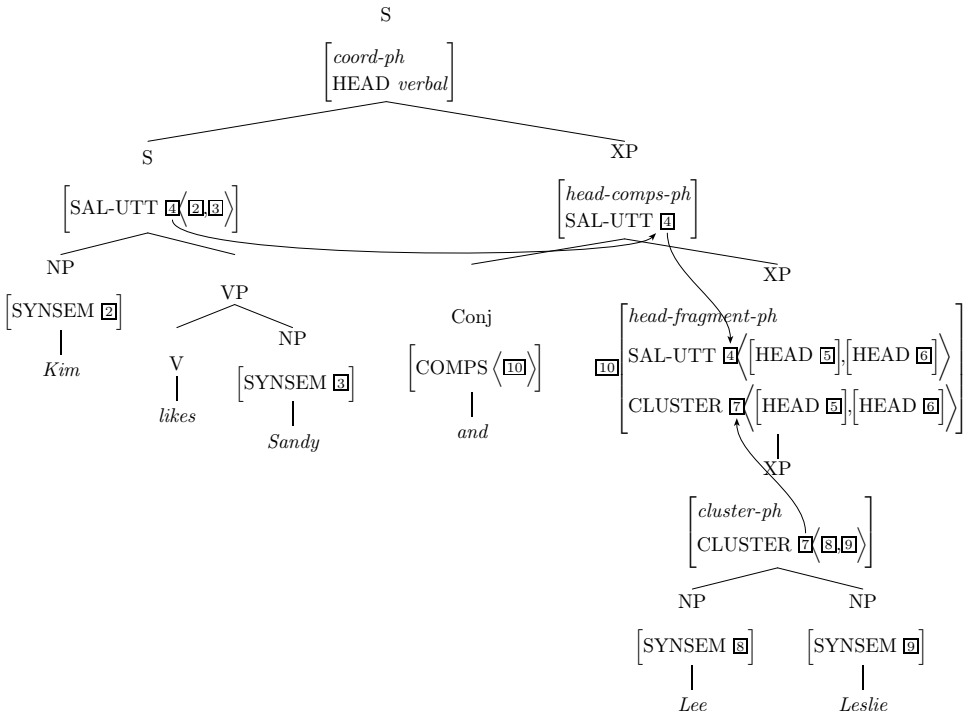


Figure 1: Simplified tree for (13) demonstrating the SAL-UTT and CLUSTER lists

that the CLUSTER list comes from below, and the SAL-UTT list comes from above, means that only one of the lists will be populated when the parser attempts to apply the *head-fragment* rule, irrespective of whether the parsing strategy is bottom-up or top-down. One of the lists will be empty until the whole coordination is parsed. This will lead to a large number of contexts where the *head-fragment* rules would be applicable, before both the lists are populated and the matching of the lists can be attempted, and it will lead to a massive burden of the parser.

Both the syntactic and semantic constraints assumed on the *head-fragment* rule in Abeillé *et al.* (2014) assume access to information in the conjunct with the verb. This makes sense in languages with forward gapping, like English, French and Romanian, but in a language like Japanese, where the remnants come before the correlates, one would have to wait for the final verb before the constraints required by the *head-fragment* rule would be made available. If one assumes a parser that works right-to-left, this can be accounted for, but it would be hard to defend from a psycholinguistic point of view.

In this paper, the incremental left-to-right approach to gapping in Haugereid (2017) will be adapted, and it will be shown how the left-to-right approach used to account for forward gapping also can be used for backward gapping, even though the verb only appears in the final conjunct. This is made possible given that the grammar is designed in such a way that a clause in principle can be parsed without

a verb. The argument structure is assumed to originate from the syntactic rules, and the verb is treated as a kind of obligatory modifier. If there is no verb, the parse will result in an underspecified construction type which only reflects the argument structure of the clause, but not the predicate of the main verb.

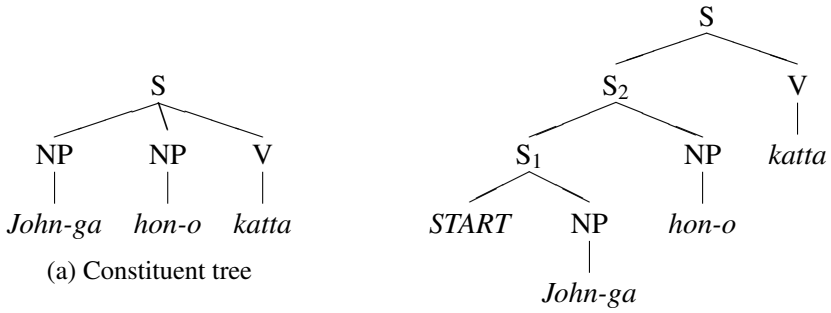
3 Analysis of gapping in Japanese

In Haugereid (2017), gapping in Norwegian, which is an SVO language, is accounted for by assuming that the predicate type of the first conjunct in a gapping construction is unified with predicates introduced by unary rules representing the elided verbs in the non-initial conjuncts. The predicate type reflects the argument structure of the clause, so the conjuncts with gapped verbs will have to realize the same type of arguments (for example a subject and an object) as the initial clause.

3.1 Incremental parsing and constituent structure

The constituent tree of the transitive sentence in (14) is assumed to be the flat structure in Figure 2a. The constituent structure is derived from the AVM of the parse tree, shown in Figure 2b. The step from parse tree to constituent tree involves the use of a feature STACK (Haugereid and Morey, 2012). In the following, the trees that will be presented, are parse trees, but they all have corresponding constituent trees.

(14) John-ga hon-o katta.
John-NOM book-ACC bought
John bought books.



(a) Constituent tree

(b) Parse tree

Figure 2: Illustration of constituent tree and parse tree of a transitive sentence in Japanese

The parse starts in the bottom left corner with a *START* symbol. This symbol has the features shown in (15). In the illustration, there are three valence features

that all have negative values *arg1–*, *arg2–*, and *arg3–*. In addition, the value of the feature VBL is *synsem*, which means that it requires a verb.

(15)
$$\begin{bmatrix} \textit{START} & \\ \text{VAL} & \begin{bmatrix} \text{CMP1}\,|\,\text{LINK} & \textit{arg1–} \\ \text{CMP2}\,|\,\text{LINK} & \textit{arg2–} \\ \text{CMP3}\,|\,\text{LINK} & \textit{arg3–} \end{bmatrix} \\ \text{VBL} & \textit{synsem} \end{bmatrix}$$

The *START* symbol is combined with the subject *John-ga* and the direct object *hon-o*. There are separate valence rules for each of these functions, and they switch a negative link type to a positive. The rule for the direct object is shown in (16). It changes the negative link value *arg2–* in the first daughter to a positive link value in the mother *arg2+*.

(16)
$$\begin{bmatrix} \textit{cmp2-struc} & \\ \text{VAL} & \begin{bmatrix} \text{CMP1} & \boxed{1} \\ \text{CMP2} & \boxed{2}\begin{bmatrix}\text{LINK}\ \textit{arg2+}\end{bmatrix} \\ \text{CMP3} & \boxed{3} \end{bmatrix} \\ \text{VBL} & \boxed{4}\textit{synsem} \\ \text{ARGS} & \left\langle \begin{bmatrix} \text{VAL} & \begin{bmatrix} \text{CMP1} & \boxed{1} \\ \text{CMP2}\,|\,\text{LINK}\ \textit{arg2–} \\ \text{CMP3} & \boxed{3} \end{bmatrix} \\ \text{VBL} & \boxed{4} \end{bmatrix}, \boxed{2} \right\rangle \end{bmatrix}$$

At the top of the tree in Figure 2b, the verb is realized. This is done by the verb rule shown in (17). The rule takes as its first daughter a structure that requires a verb, and as its second daughter a verb, and it produces a structure that has saturated the verb requirement (VBL *anti-synsem*). In addition, the rule unifies all the link types with the PRED type of the verb.

(17)
$$\begin{bmatrix} \textit{verb-struc} & \\ \text{VAL} & \boxed{1}\begin{bmatrix} \text{CMP1}\,|\,\text{LINK} & \boxed{2} \\ \text{CMP2}\,|\,\text{LINK} & \boxed{2} \\ \text{CMP3}\,|\,\text{LINK} & \boxed{2} \end{bmatrix} \\ \text{VBL} & \textit{anti-synsem} \\ \text{ARGS} & \left\langle \begin{bmatrix} \text{VAL} & \boxed{1} \\ \text{VBL} & \boxed{3} \end{bmatrix}, \boxed{3}\begin{bmatrix} \text{HEAD} & \textit{verb} \\ \text{LKEYS}\,|\,\text{KEYREL}\,|\,\text{PRED} & \boxed{2} \end{bmatrix} \right\rangle \end{bmatrix}$$

The lexical entry for the verb *ka* ('buy') is shown in (18). It only has an ORTH value, a HEAD value, and a PRED value. There are no VAL features or ARG-ST list.

(18)
$$\begin{bmatrix} \textit{verb-lxm} & \\ \text{ORTH} & \langle \text{ka} \rangle \\ \text{HEAD} & \textit{verb} \\ \text{LKEYS}\,|\,\text{KEYREL}\,|\,\text{PRED} & \textit{buy\_prd} \end{bmatrix}$$

Instead of the regular valence requirements associated with verb lexical items, the verb is given a PRED value *buy_prd*, and it is the position of this type in a type hierarchy of subconstruction types, that determines which argument frames that are possible for the verb. A simplified type hierarchy involving the type *buy_prd* is shown in Figure 3.

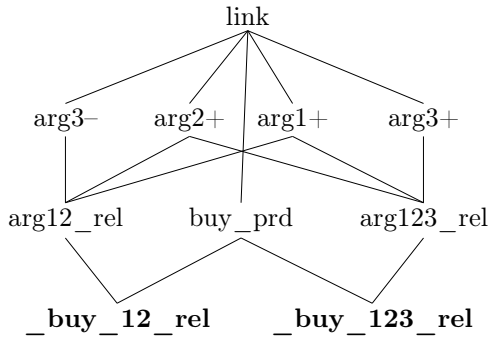


Figure 3: Type hierarchy of subconstruction types, argument frame types, and construction types

The hierarchy shows that the *buy_prd* type is compatible with two argument frames, a transitive frame *arg_12_rel*, and a ditransitive frame *arg_123_rel*. When the predicate is unified with one of these two frames, we get the construction types *_buy_12_rel* and *_buy_123_rel*, respectively. In this way, it is the type hierarchy of subconstruction types that determines which frames that are possible for a verb to enter.

The tree in Figure 4 shows how the linking types are changed from negative in the *START* node to positive in the top of the tree, and how the link types are unified with the PRED value of the verb. Since the types *arg1+*, *arg2+*, *arg3–*, and *buy_prd* are compatible (given the type hierarchy in Figure 3) the sentence is ultimately given a parse.

3.2 Analysis of gapping

SOV clause structure and backward gapping as demonstrated in (2)-(6) pose a challenge to the incremental left-to-right approach in Haugereid (2017). However, the

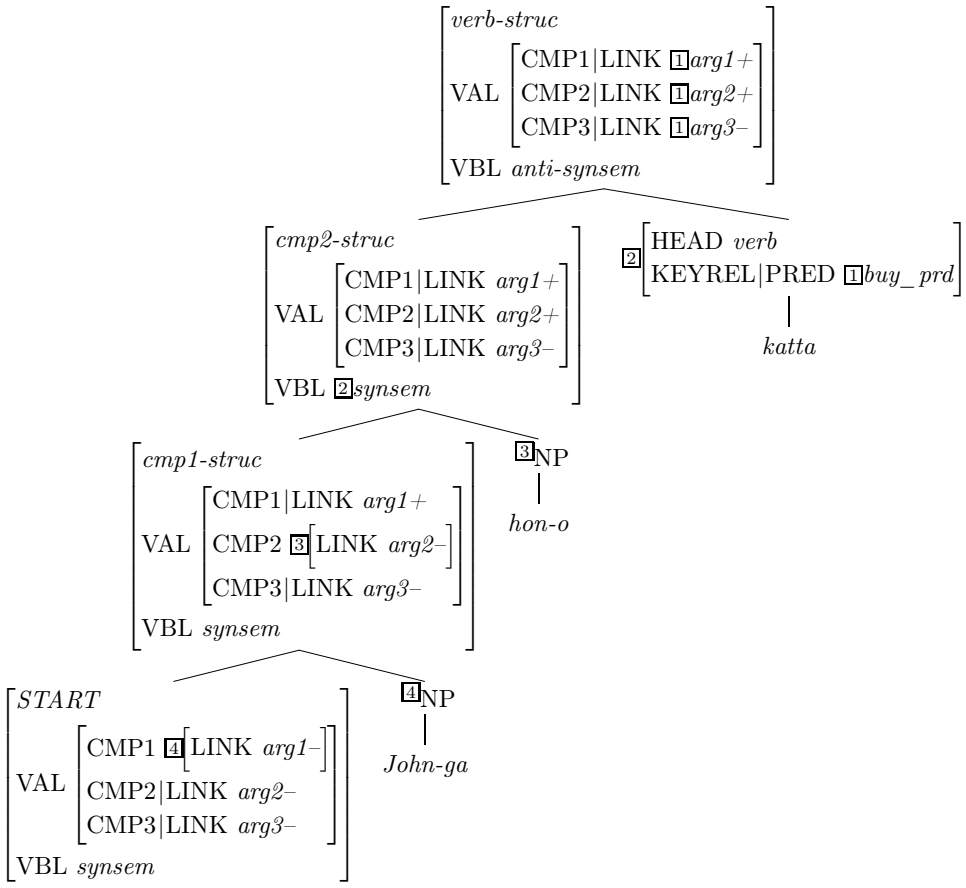


Figure 4: Parse tree for Japanese transitive sentence

constructional approach allows for the construction of an argument frame type that is underspecified with regard to the predicate of the main verb of the clause. This is shown in Figure 3, where the type *arg12_rel* is the result of the unification of three subconstruction types: *arg1+*, which is contributed by the rule that realizes the subject, *arg2+*, which is contributed by the rule that realizes the direct object, and *arg3–*, which shows that no indirect object has been realized. (If an indirect object is realized, the *arg3–* type will be replaced by *arg3+*, resulting in the argument frame type *arg123_rel*.) The tree in Figure 5 illustrates how the subconstruction types accumulate as the conjuncts in (2) are parsed.[4]

The parse starts in the bottom left corner with the structure *START* that has only negative subconstruction types (see (15)), represented in the tree in Figure 5 as an empty set. The rule that attaches the subject *John-ga* adds the subconstruction type *arg1+*, and the rule that attaches the object *hon-o* adds the type *arg2+*. When

[4]In order to make the representation compact, I have used sets to illustrate the accumulation of the subconstruction types in Figure 5. In reality, each subconstruction type is the value of a separate feature. The underlining of subconstruction types in the tree represents the unification of these types.

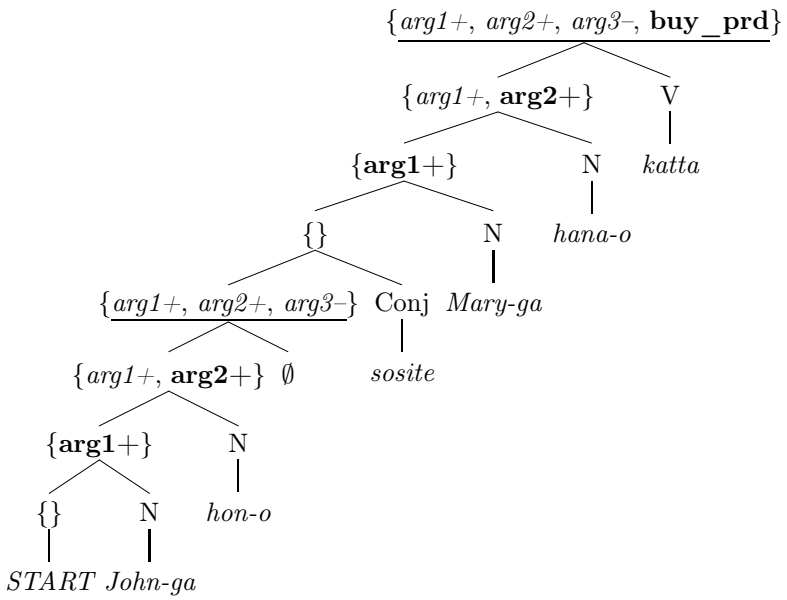


Figure 5: Accumulation of subconstruction types

no more arguments are attached, a unary gapping rule (see (19)) unifies the LINK values, here *arg1+*, *arg2+* and *arg3–*, with the PRED value of the KEYREL. (In the tree, this is marked by underlining the subconstruction types, and the elided verb is marked with the symbol ∅.) The KEYREL value is unified with the GAPREL value. The rule switches the VBL value from *synsem* in the daughter to *anti-synsem* in the mother. It also gets the HEAD feature GAPPING +.

(19)
$$\begin{bmatrix} \textit{verb-gapping-struc} & \\ \text{HEAD} & \left[\text{GAPPING} + \right] \\ \text{VAL} & \boxed{1}\begin{bmatrix} \text{CMP1}\,|\,\text{LINK}\ \boxed{2} \\ \text{CMP2}\,|\,\text{LINK}\ \boxed{2} \\ \text{CMP3}\,|\,\text{LINK}\ \boxed{2} \end{bmatrix} \\ \text{VBL} & \textit{anti-synsem} \\ \text{LKEYS} & \begin{bmatrix} \text{KEYREL} & \boxed{3}\left[\text{PRED}\ \boxed{2}\right] \\ \text{GAPREL} & \boxed{3} \end{bmatrix} \\ \text{ARGS} & \left\langle \begin{bmatrix} \text{VAL} & \boxed{1} \\ \text{VBL} & \textit{synsem} \end{bmatrix} \right\rangle \end{bmatrix}$$

At this point, the three subconstruction types are unified, resulting in the argument frame type *arg12_rel* (see the type hierarchy in Figure 3). The conjunct *sosite*

initiates a new conjunct (see (20)), and it carries into the new clause the argument frame type from the gapping rule (see (21)).

(20)
$$\begin{bmatrix} \textit{conj-word} \\ \text{ORTH} \quad \langle \textit{sosite} \rangle \\ \text{HEAD} \quad \begin{bmatrix} \textit{conj} \\ \text{GAPPING} \quad + \end{bmatrix} \end{bmatrix}$$

(21)
$$\begin{bmatrix} \textit{coord-struc} \\ \text{VAL} \quad \begin{bmatrix} \text{CMP1}\,|\,\text{LINK}\ \textit{arg1}– \\ \text{CMP2}\,|\,\text{LINK}\ \textit{arg2}– \\ \text{CMP3}\,|\,\text{LINK}\ \textit{arg3}– \end{bmatrix} \\ \text{VBL} \quad \textit{synsem} \\ \text{LKEYS} \quad \left[\text{KEYREL} \quad \boxed{1}\right] \\ \text{ARGS} \quad \left\langle \begin{bmatrix} \text{HEAD} \quad \left[\text{GAPPING}\ \boxed{2}\right] \\ \text{VBL} \quad \textit{anti-synsem} \\ \text{GAPREL} \quad \boxed{1} \end{bmatrix}, \begin{bmatrix} \textit{coord-word} \\ \text{HEAD}\,|\,\text{GAPPING} \quad \boxed{2} \end{bmatrix} \right\rangle \end{bmatrix}$$

The second clause is parsed in the same manner, and at the top of the tree, the rule that attaches the verb, unifies the predicate *buy_prd* with the subconstruction types of the second conjunct (*arg1*+, *arg2*+, *arg3*–), resulting in the predicate type *buy_12_rel*. (The unified subconstruction types are underlined at the top of the tree in Figure 5). The rule also unifies this predicate type with the argument frame type carried over from the first conjunct (*arg12_rel*). In this way, the identity of the two construction types is ensured, and the two clauses get the same predicate.

The MRS (Copestake *et al.*, 2005) for example (2), repeated below as (22) is given in Figure 6. The first *_buy_12_rel* predicate is the result of unifying the construction type of the first conjunct *arg_12_rel* with the construction type of the last conjunct *_buy_12_rel*.

(22) John-ga hon-o sosite Mary-ga hana-o katta.
John-NOM book-ACC and Mary-NOM flower-ACC bought
'John bought books, and Mary flowers.'

The incremental subconstructional approach assumed in this paper is similar to the approach in Abeillé *et al.* (2014) in that the argument frame of the conjunct with the verb is unified with the argument frame of the conjuncts with the elided verbs. However, it differs from Abeillé *et al.* (2014), as well as other lexicalist approaches, in several respects. Firstly, this account is an incremental account. It differentiates between a parse tree (which is left branching, as shown in Figure

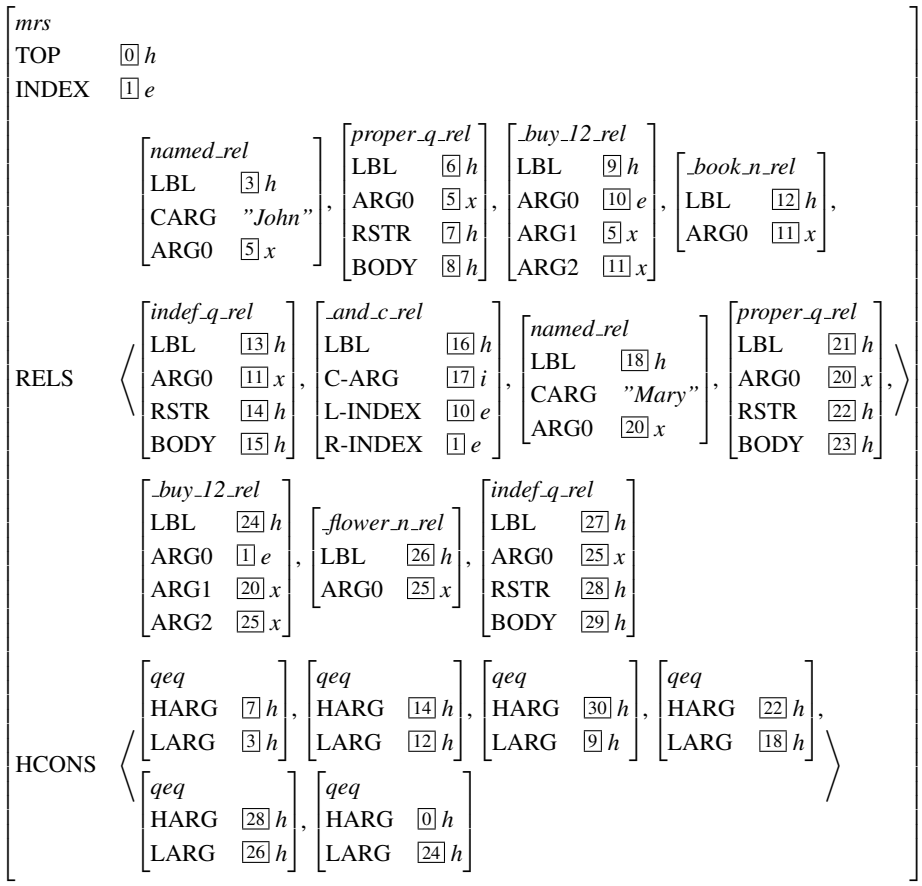


Figure 6: Semantic representation – Gapping

5) and a constituent tree, which is relatively standard (see Haugereid and Morey (2012)). Secondly, this approach assumes a hierarchy of subconstruction types (Haugereid, 2009), as illustrated in Figure 3. It is this hierarchy of subconstruction types that accounts for the argument frames in the grammar, not the constraints on the lexical items, and this makes it possible to parse a sentence without a verb, as illustrated in the first conjunct in Figure 5. In standard HPSG, including Abeillé *et al.* (2014), verbs are specified with an ARG-ST, and there is no generalization over ARG-ST lists that corresponds to the hierarchy of subconstruction types assumed in my approach. A third difference between the approach in this paper and Abeillé *et al.* (2014) is the semantics. In my approach, the construction type that results from the unification of the subconstruction types, becomes the predicate of the verb (and the elided verb). The semantics is in this way integrated with the syntax. In Abeillé *et al.* (2014) however, there are separate constraints accounting for the syntax and semantics of gapping.

4 Future work

The suggested method accounts for the data in (2)–(6). There will be some overgeneration with regard to adjuncts, since they are not reflected in the argument structure of the verb. One solution to that would be to let not only information about arguments, but also adjuncts be carried over to the next conjunct. This is a topic for further investigation. Another foreseeable problem with the approach is the fact that the verb does not appear in the first conjunct. This will increase the search space of the parser, although it will be constrained by the hierarchy of subconstruction types. The search space could be further restricted if the method were to be combined with some kind of statistical ”guesser” for each word that is added.

References

Abeillé, A., Bîlbîie, G., and Mouret, F. (2014). A romance perspective on gapping constructions. In H. C. Boas and F. Gonzálvez-García, editors, *Romance Perspectives on Construction Grammar*, pages 227–265. John Benjamins Publishing Company.

Beavers, J. and Sag, I. A. (2004). Coordinate ellipsis and apparent non-constituent coordination. In S. Müller, editor, *Proceedings of the HPSG-2004 Conference, Center for Computational Linguistics, Katholieke Universiteit Leuven*, pages 48–69. CSLI Publications, Stanford.

Chaves, R. P. (2005). A linearization-based approach to gapping. In G. Jäger, P. Monachesi, G. Penn, and S. Wintner, editors, *FG-MOL 2005: The 10th conference on Formal Grammar and The 9th Meeting on Mathematics of Language*, page 14. CSLI.

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

Copestake, A., Flickinger, D., Pollard, C. J., and Sag, I. A. (2005). Minimal recursion semantics: an introduction. *Research on Language and Computation*, **3**(4), 281–332.

Crysmann, B. (2008). An asymmetric theory of peripheral sharing in HPSG: Conjunction reduction and coordination of unlikes. In G. Penn, editor, *Proceedings of FGVienna : The 8th Conference on Formal Grammar, Aug 16—17 2003, Vienna*, pages 45–64, Stanford. CSLI publications.

Haugereid, P. (2009). *Phrasal subconstructions: A constructionalist grammar design, exemplified with Norwegian and English*. Ph.D. thesis, Norwegian University of Science and Technology.

Haugereid, P. (2017). An incremental approach to gapping and conjunction reduction. In S. Müller, editor, *Proceedings of the 24th International Conference on Head-Driven Phrase Structure Grammar, University of Kentucky, Lexington*, pages 179–198, Stanford, CA. CSLI Publications.

Haugereid, P. and Morey, M. (2012). A left-branching grammar design for incremental parsing. In S. Müller, editor, *Proceedings of the 19th International Conference on Head-Driven Phrase Structure Grammar, Chungnam National University Daejeon*, pages 181–194.

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

Kato, K. (2006). *Japanese Gapping in Minimalist Syntax*. Ph.D. thesis, University of Washington.

Kim, Y.-J. and Cho, S.-Y. (2012). Tense and honorific interpretations in korean gapping construction: A constraint- and construction-based approach. In S. Müller, editor, *Proceedings of the 19th International Conference on Head-Driven Phrase Structure Grammar, Chungnam National University Daejeon*, pages 388–408.

Mouret, F. (2006). A phrase structure approach to argument cluster coordination. In S. Müller, editor, *The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, pages 247–267, Stanford. CSLI Publications.

Reape, M. (1994). Domain union and word order variation in German. In J. Nerbonne, K. Netter, and C. J. Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, number 46 in CSLI Lecture Notes, pages 151–197. CSLI Publications, Stanford University.

Ross, J. R. (1970). Gapping and the order of constituents. In M. Bierwisch and K. Heidolph, editors, *Progress in Linguistics*, pages 249–259. The Hague: Mouton.

Sag, I. A., Gazdar, G., Wasow, T., and Weisler, S. (1985). Coordination and how to distinguish categories. *Natural Language & Linguistic Theory*, **3**(2), 117–171.

Yatabe, S. and Tanigawa, K. (2018). The fine structure of clausal right-node raising constructions in japanese. In S. Müller and F. Richter, editors, *Proceedings of the 25th International Conference on Head-Driven Phrase Structure Grammar, University of Tokyo*, pages 176–196, Stanford, CA. CSLI Publications.

What grammars are, or ought to be

Geoffrey K. Pullum
University of Edinburgh

Proceedings of the 26th International Conference on
Head-Driven Phrase Structure Grammar

University of Bucharest

Stefan Müller, Petya Osenova (Editors)

2019

CSLI Publications

pages 58–78

http://csli-publications.stanford.edu/HPSG/2019

Keywords: grammars, HPSG, constraints, model-theoretic syntax

Pullum, Geoffrey K. 2019. What grammars are, or ought to be. In Müller, Stefan, & Osenova, Petya (Eds.), *Proceedings of the 26th International Conference on Head-Driven Phrase Structure Grammar, University of Bucharest*, 58–78. Stanford, CA: CSLI Publications.

Abstract

Progress toward distinguishing clearly between generative and model-theoretic syntactic frameworks has not been smooth or swift, and the obfuscatory term 'constraint-based' has not helped. This paper reviews some elementary subregular formal language theory relevant to comparing description languages for model-theoretic grammars, generalizes the results to trees, and points out that HPSG linguists have maintained an unacknowledged and perhaps unintended allegiance to the idea of strictly local description: unbounded dependencies, in particular, are still being conceptualized in terms of plugging together local tree parts annotated with the SLASH feature. Adopting a description language with quantifiers holds out the prospect of eliminating the need for the SLASH feature. We need to ask whether that would be a good idea. Binding domain phenomena might tell us. More work of both descriptive and mathematical sorts is needed before the answer is clear.

1 Introduction

What sort of system should we employ to give a formal description of a human language? Two sharply distinct views compete for linguists' attention. One emerged in the mid 1950s, when Chomsky persuaded most younger linguists that grammars should be formalized as nondeterministic constructive set generators composed of an initial symbol (traditionally S) and a set of expansion-oriented rewriting rules that build derivations ultimately yielding terminal strings, the whole system being interpreted as a constructive definition of the set of all and only those strings that it could in principle construct. Chomsky (1959) is generally taken to be the foundational work on this type of system.

The other emerged later and made much more hesitant progress. Its mathematical foundations go back to a proposition which was proved independently by several mathematicians (Medvedev 1956 [1964], Büchi 1960, Elgot 1961, Trakhtenbrot 1962), but is most often called Büchi's Theorem (Büchi, 1960). It says a set of strings is finite-state if and only if it is the set of all finite string-like models of a closed formula of weak monadic second-order logic. This offers a new way of characterizing a set of strings: instead of asking what sort of a rewriting system can generate all and only the members of a certain set,

[†]I am grateful to the organizers of the 2019 HPSG conference for inviting me to present a paper, and especially to Gabriela Bîlbîie for her brilliant local organization. I thank the attendees for their questions and discussion. I owe a major debt to James Rogers, whose work is the source of the observations in sections 4 and 5 of this paper. Conversations with Mark Steedman before the conference were extremely useful to me, and after the conference I benefited from careful successive critiques of several drafts by Bob Levine, Bob Borsley, and especially Stefan Müller. They all helped me to correct serious errors I had made. The remaining faults and blunders are solely mine.

ask what sort of logic can express a statement that is true of all and only the sorts of things that are members of that set. Within theoretical computer science it has led to significant results such as that existential second-order logic characterizes stringsets recognizable in nondeterministic polynomial time (Fagin, 1974), and has spawned new subdisciplines such as descriptive complexity theory (Immerman, 1999).

Introducing the second kind of thinking into linguistics created model-theoretic syntax (MTS), but progress toward accepting it has been anything but straight and smooth. McCawley (1968) is widely thought to have presaged it, but did not (see §3). Lakoff (1971) groped toward it but botched the job (Soames, 1974). Kac (1978) clearly adumbrates it but has been overlooked. Johnson & Postal (1980) makes the most serious attempt at it, but contradicts itself with its misguided ideas (in Chapter 14) about formalizing transderivational constraints.

HPSG perhaps comes closer than any other framework to developing in purely MTS mode, but even there the progress has been hesitant. HPSG grew out of GPSG (Gazdar et al., 1985), which was developed as a kind of hybrid theory, a generative grammar with filters. It was only in 1987 that Gerald Gazdar realized that GPSG should have been conceptualized model-theoretically. In unpublished lectures at the 1987 LSA Linguistic Institute he showed how this could be done. By then Pollard and Sag had developed the first version of HPSG (Pollard & Sag, 1987), very much within mainstream generative thinking. By 1994 Pollard and Sag were laying more stress on principles which stated facts about about well-formed structures, but still employed 'schemata' written in the form 'A $\rightarrow$ B C' to outline the gross properties of syntactic constructions, and those are visibly like context-free (CF) rules. A formula like 'Clause $\rightarrow$ NP VP' (and it makes no difference if NP and VP are replaced by complex AVMs) cannot be construed as a statement that is truth-evaluable within the structure of a sentence.

By the second half of the 1990s, both Pollard and Sag had commenced using the term 'constraint-based grammar' to characterize their approach, and some have equated this with MTS. I do not favor this term, and try to explain why in what follows. I then discuss some relevant results concerning subregular families of stringsets, which I then generalizes to trees. I conclude by attempting to bring all this to bear on HPSG.

2 The 'constraint-based' label

The term 'constraint-based grammar' (henceforth CBG) figures prominently in works like Pollard (1996) and the textbook by Ivan Sag et al. (Sag & Wasow 1999; 2nd edition Sag et al. 2003). Müller (2019) takes it to be a syn-

onym for MTS, but it seems to me to have a mainly sociological import: the crucial requirement for membership in the CBG community is not positing transformations (hence not following Chomsky). The CBG membership roll according to Sag & Wasow (1999) includes GPSG, HPSG, LFG, functional unification grammar, dependency grammar, categorial grammar, construction grammar, and the framework Sag was working on in his last years, sign-based construction grammar (SBCG). Sag et al. (2003) adds brief sections on three other syntactic theories which are claimed not to fit into the typology: relational grammar (RG), tree-adjoining grammar (TAG), and optimality theory (OT). But this claim of failure to fit suggests incoherence in the typology.

RG uses no transformations and should surely be classed as CBG — its more highly mathematicized descendant arc pair grammar (APG) is correctly recognized by Sag et al. as 'the first formalized constraint-based theory of grammar to be developed' (Sag et al. 2003: 539).

TAG, by contrast, is straightforward composition-oriented (bottom-up) generative grammar, analogous to categorial grammar but founded on trees rather than strings, hence should surely be excluded (especially if the adjunction operation is taken to be analogous to a generalized transformation, as seems to be suggested by Chomsky 1993: 21).

And if OT is not based in constraints, no framework is: OT posits a universal set of constraints, different grammars being distinguished solely by different orders of application priority, so why does it not fit the classification?

Further puzzlement arises when Culicover & Jackendoff (2005) classify their work as CBG. They class categorial grammar and tree-adjoining grammar with 'mainstream generative grammar' (which seems correct to me), but count their 'simpler syntax' (along with LFG, HPSG, and Construction Grammar) as CBG, despite indications that it employs generative components for each of phonology, syntax, and semantics. They claim that in CBG theories:

> Each constraint determines or licenses a small piece of linguistic structure or a relation between two small pieces. A linguistic structure is acceptable overall if it conforms to all applicable constraints. There is no logical ordering among constraints, so one can use constraints to license or construct linguistic structures starting at any point in the sentence: top-down, bottom-up, left-to-right, or any combination thereof. Thus a constraint-based grammar readily lends itself to interpretations in terms of performance...

Note the locutions 'determines or licenses' and 'license or construct'. Which is it? Constructing sentences? Licensing them as having been constructed correctly? Or stating conditions on the structure they are permitted to have?

In a footnote they suggest that CBG 'was suggested as early as [McCawley (1968)], who referred to "node admissibility conditions"; other terms for this

formulation are “declarative”, “representational”, and “model-theoretic”.’ This embodies a confusion that is worth discussing in detail.

3 Node admissibility conditions

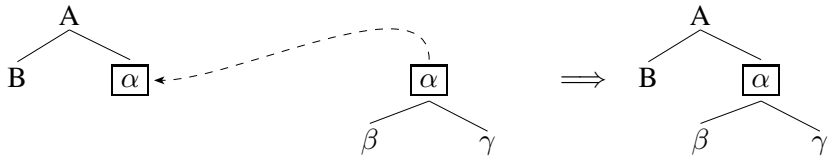
The novel interpretation of CF rules to which Culicover and Jackendoff allude was suggested to James McCawley by Richard Stanley in 1965. The idea was to reinterpret phrase structure rules as ‘node admissibility conditions’ (henceforth NACs). This meant treating a CF rule ‘A $\rightarrow$ B C’ not as meaning ‘if the current string has an A you may rewrite it with the A replaced by B C’, but rather as meaning ‘a subtree consisting of an A-labeled node with a first child labeled B and a second child labeled C is permissible’. Context-sensitive rules can also be thus reinterpreted: ‘A $\rightarrow$ B C / D $\_\_$ E’ would standardly be read as ‘if the current string has an A with a D preceding and an E following, you may rewrite it with the A replaced by B C’; the new interpretation would be: ‘a node labeled A with a first child labeled B and a second child labeled C is permissible if in the tree a D-node immediately precedes the replaced A and an E-node immediately follows it’.

McCawley observed that for context-sensitive rules there was a difference in expressive power between the two interpretations: he exhibited a tiny grammar which under one interpretation generated a single tree and under the other generated nothing. Clarifying this, a later mathematical result of Peters & Ritchie (1969) showing that the NAC interpretation yielded only context-free stringsets (CFLs). However, none of this has much to do with MTS. To start with, ‘node admissibility condition’ (henceforth NAC) was always a misnomer. NACs are not conditions on the admissibility of nodes or trees or anything else. An NAC saying ‘A $\rightarrow$ B C’ doesn’t place any condition on nodes, not even on nodes labeled A: it requires neither that a node labeled A should have the child sequence B C (there could be another NAC saying A $\rightarrow$ D E F) and it doesn’t require that a child sequence B C must have a parent node A (there could be another NAC saying ‘D $\rightarrow$ B C’).

The Stanley/McCawley interpretation makes trees directly answerable to the content of the grammar without the need for Chomsky’s procedure for constructing trees from the information in derivations, so the connection between rules and structures becomes far more transparent. A grammar becomes in effect a finite library of pictures of local regions of a tree, and a tree is well formed iff every region of appropriate size matches one of the pictures (see Rogers 1999, where CF grammars (CFGs) are introduced in this way). It was part of what motivated Gerald Gazdar to reconsider the descriptive value of CF rules.

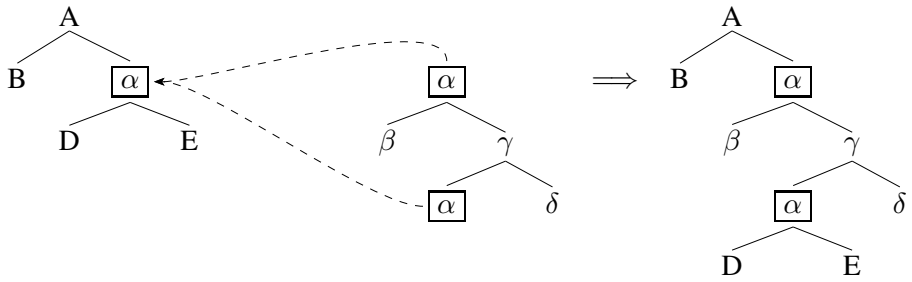
But grammars under the Stanley/McCawley NAC interpretation are purely

generative grammars, though of the composition-oriented type first exhibited in categorial grammar (Ajdukiewicz 1935). A grammar consists of some building blocks and a mode of composition for putting them together. The implicit composition operation for NAC grammars is what I will call *frontier nonterminal substitution*: wherever a frontier node in a developing tree is labeled with a nonterminal symbol α you can plug in some local tree in the grammar that has root node labeled α:



The set of trees generated is the set of all and only those trees with a root label S (or whatever root node label may be specified) and a frontier entirely labeled by terminal symbols (lexical items). The strings generated are all and only the frontiers of those trees.

Tree-adjoining grammar (TAG) also defines composition-oriented generative grammars, differing in that they feature a more complex kind of composition operation, based on *auxiliary trees*, which have a frontier nonterminal node label matching the root node label. These provide for operations of what I will call *internal nonterminal substitution*: a designated internal node labeled α is replaced by an auxiliary tree that has α both as root label and a frontier label, like this:



Neither NAC grammars nor TAGs are anything like MTS. Notice that MTS constraints express *necessary conditions on expression structure*, and well-formedness is determined by *satisfaction of all the constraints*. But no node can ever match more than one NAC, so there could never be a tree that satisfied two or more NACs.

McCawley himself makes an error on this point, saying that 'the admissibility of a tree is defined in terms of the admissibility of all of its nodes, i.e., in the form of a condition which has the form of a logical conjunction' (p. 248). It is in fact a *dis*junction. An NAC is a one-place predicate of nodes; for exam-

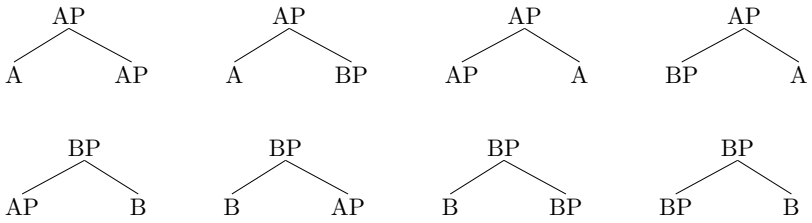
ple, the NAC corresponding to the rule 'A $\rightarrow$ B C', expressed as a property of a node x, says 'x is labeled A and has a left child labeled B and a right child labeled C'. For generality, it will be convenient to add a trivial one-node NAC of depth zero for each terminal symbol: the NAC '*eat*' will correspond to the statement 'x is a node labeled *eat*'. One might want to stipulate that that a node with no child (a node on the frontier, also known as a leaf) must be labeled with a member of the terminal vocabulary (though that rules out some theories of 'empty categories'), and perhaps also that a node that has no parent (the root) is labeled with some designated symbol such as S or Clause; but these are matters of detail. The main thing that has to be true in a tree to make it well-formed according to a set $\varphi_1, \ldots, \varphi_k$ of k NACs is the multiple disjunction shown in (1).

(1) $(\forall x)[\bigvee_{1 \leq i \leq k} \varphi_i(x)]$ 'Every node satisfies the disjunction of all the NACs.'

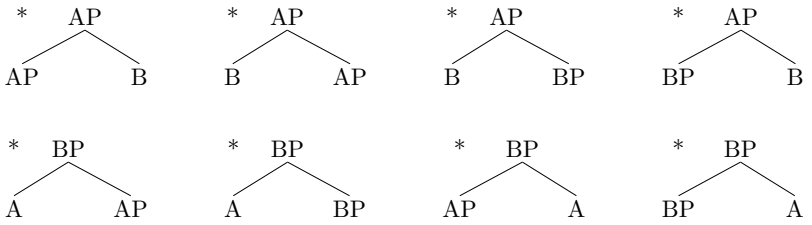
So NACs are in effect atomic propositions, each stating a sufficient condition for some specific local tree (of depth 0 or 1) to be a legitimate subpart of a well-formed tree, and written in a way that is isomorphic to such a subpart. The grammar is really just a finite list of local trees, generating all and only the trees that are entirely composed of the local trees on the list and could be constructed by combining them using frontier nonterminal substitution.

Stanley/McCawley CFG thus construed is a composition-oriented generative formalism employing a set of small building blocks and a set of operations for putting them together to make larger units. In that respect, it is just like categorial grammar, TAG, and Chomsky's minimalism. However, it will be relevant in the next section that there is a particularly easy way to turn a set of NACs into a model-theoretic description that is in many respects equivalent. It depends on the fact that there are only finitely many trees employing a given finite vocabulary V of node labels and given finite bounds on depth (d) and breadth (b). Hence for any finite subset of them interpreted as NACs we can take the complement of that set (relative to all trees of depth $\leq d$ and width $\leq w$ labeled from V), and interpret each as a subtree prohibition. Each local tree T_i in the complement set will be understood as the statement φ_i meaning 'the configuration T_i does not occur as a subtree'. Then a tree $\mathcal{T}$ is well-formed if and only if each φ_i is true in $\mathcal{T}$.

To illustrate, consider this set of NACs constituting a tree-generating grammar defining (in effect) a one-bar-level X-bar theory on binary trees with only two lexical categories, A and B:



Now consider the complement set (which happens in this case to be exactly the same size):



Making sure a tree does NOT contain any of the configurations in the second set yields exactly the same results as making sure it IS entirely composed of the local trees in the first set.

Now in the next section I want to begin to make clear the sense in which this yields an MTS description, though one stated in an extremely primitive description language. Various results were achieved after 1968 seem in retrospect highly relevant to clarifying this point. I will first review the results, and then, returning to my theme, relate them to HPSG.

4 Expressive power of description languages for strings

It was noted in Rogers (1997) that stating a CFG as a set of local trees is strongly analogous to a bigram description of a set of strings. A bigram description over an alphabet Σ is a finite list of 2-symbol sequences over Σ, and a string is grammatical according to it if every length-2 substring of the string (every *factor*, as the formal language theorists put it) is on the list. And as I have pointed out, using the complement of the set of bigrams instead permits the description to be construed in MTS terms.

But bigram descriptions define only a very small and primitive class of stringsets, the SL_2 stringsets. Using depth-1 local trees as building blocks for trees is analogous to using bigrams as building blocks for strings. When implemented in detail it employs a finite vocabulary Σ plus an additional symbol $\rtimes \notin \Sigma$ to mark of the beginning of a string and $\ltimes \notin \Sigma$ to mark the end. A string w is generated iff it begins with some symbol σ such that $\rtimes\sigma$ is one of the bigrams and it ends with some symbol σ such that $\ltimes\sigma$ is one of the bigrams,

and for every substring $\sigma_1\sigma_2$ the string $\sigma_1\sigma_2$ is one of the bigrams. Model-theoretically, it amounts to using a description language of atomic propositions interpreted as substring bans: a proposition ab means 'the substring ab does not occur'.

Letting $k = 3$ then yields the trigram stringsets, a proper superset; letting $k = 4$ yields the quadrigram stringsets, larger still; and so on upward: n-gram stringset for any for any positive integer n can be defined by letting $k = n$. So there is an infinite hierarchy of strictly local (SL) stringsets.

If we add in the results of taking unions, intersections, and complements of SL stringsets we get a strictly larger class known as the Locally Testable stringsets (LT). And again, there is a class LT_2 where the basis is SL_2 stringsets, a class LT_3 where the basis is SL_3 stringsets, and so on.

LT stringsets can be described model-theoretically by allowing not just atomic propositions like $\sigma_1\sigma_2$ (meaning 'the substring $\sigma_1\sigma_2$ does not occur') but also propositional calculus formulas which are conjunctions, disjunctions, or negations of such propositions. Now you can say things like 'either ab does not occur or bc and cd do not both occur', and so on. The class of stringsets describable is now larger, a proper superset of the SL stringsets known as the LT stringsets. For a simple example of a stringset that is LT but not SL, consider a^*ba^*. It contains all and only those strings over $\{a, b\}$ that contain just a single b, and it has no SL_k description for any k. So the apparently very simple notion 'contains a b' is not expressible in a language as primitive as the language of atomic propositions about n-gram presence or absence.

Allowing quantifiers adds considerably to expressive power. If we permit first-order quantification and assume a binary relation symbol $<_1$ intuitively meaning 'immediately precedes' (= 'left-adjacent to' = 'predecessor of') we can describe a larger class of stringsets known as the (locally) Threshold Testable (TT) sets (Thomas 1982: 372). This permits us to verify that a certain substring occurs at least a certain number of times in each string, up to a fixed threshold.

We can step up the expressive power yet more by choosing the binary relation '$<_1^*$', the reflexive transitive closure of $<_1$, interpretable as 'precedes (not necessarily immediately)'.[1] We get a larger family of stringsets, also obtainable by closing LT under union, intersection, complement, and concatenation, and thus known as the 'Locally Testable with Order' class in McNaughton & Papert (1971). But it is most commonly known under the name 'Star-Free' (SF), because the languages can be characterized by expressions very much like regular expressions except that they use the complement operator instead of asteration (Kleene star): every finite stringset is SF; every union of SF stringsets is SF; every concatenation of SF stringsets is SF; the complement of any SF stringset

[1] The $<_1$ relation is first-order definable from $<_1^*$, but not conversely.

is SF; and nothing else is.

An alternative characterization is as the class of *non-counting* stringsets, within which beyond a certain finite limit k there is no further possibility of counting whether there were k consecutive occurrences of some substring or more than k, so that if uv^kw is in a set $uv^{k+1}w$ is in as well.

An important result due to McNaughton & Papert (1971) shows that a language is SF iff it is the set of all finite stringlike structures satisfying a closed formula of first-order logic with '$<_1^*$'. This permits describing the set of all strings over $\{a, b, c\}$ satifying $\forall x[c(x) \Rightarrow \exists y[b(y) \wedge y <_1^* x]]$, in which any occurrence of c has to co-occur with a b somewhere earlier in the string.

As a final step up in expressive power, if we replace first-order logic by weak monadic second-order logic (wMSO), we have a theorem obtained independently by several researchers in the late 1950s (Medvedev 1956 [1964], Büchi 1960, Elgot 1961): using wMSO on string-like models, the describable stringsets are an even larger class, namely the regular (finite-state) stringsets.

Thus we have this tableau of progressively larger and larger families of stringsets:

(2)
a. Strictly Local $\boxed{\text{SL}}$ (finite n-gram lists); SL_2, SL_3, SL_4, ... SL_k
b. Locally Testable $\boxed{\text{LT}}$ (closure of SL under boolean connectives); LT_2, LT_3, LT_4, ... LT_k
c. Threshold Testable $\boxed{\text{TT}}$ (first-order logic with $<_1$); TT_2, TT_3, TT_4, ... TT_k
d. Star-Free $\boxed{\text{SF}}$ (star-free expressions; counter-free automata; FO with $<^*$)
e. Finite-State $\boxed{\text{FS}}$ (regular expressions or grammars; finite automata; wMSO)

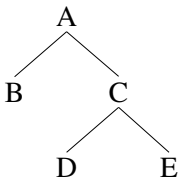
And they form a hierarchy (actually a hierarchy of hierarchies, because there are SL_3 stringsets that are not SL_2, LT_5 stringsets that are not LT_4, and so on).

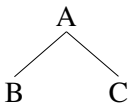
(3) $\boxed{\text{SL}} \subsetneq \boxed{\text{LT}} \subsetneq \boxed{\text{TT}} \subsetneq \boxed{\text{SF}} \subsetneq \boxed{\text{FS}}$

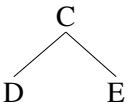
5 Expressive power of description languages for trees

The relevance of this material on the formal language theory of subregular stringsets (see Rogers & Pullum 2011 for more) will become clearer once we generalize to trees. The analog of n-grams are trees of depth $n + 1$ (where an isolated node has depth 0). Under the reformalization discussed above, a CFG can be given as simply a finite set of local trees, i.e. trees of depth 1. These correspond to bigrams: in the linear precedence dimension a bigram has a length

of two; a local tree has a corresponding measurement in the dominance dimension, from the root level to the child level. Just as you can decompose strings into the bigrams they contain, you can decompose a tree into the local trees it contains. There is an overlap of one symbol in each case. The string $abba$ is made up of the bigrams ab, bb, and ba. In an analogous way, he tree in (4a) is made up of the local trees (4b) and (4c).

(4) a. 

b. 

c. 

Adding a superscript τ to remind us that we are now talking about tree-sets rather than stringsets, the family of all tree-sets we can define using local trees as building blocks can be called the SL_2^τ tree-sets. But as with the stringset families SL_2, SL_3, and so on, we can use trees of greater and greater depth as building blocks to get an infinite hierarchy of larger and larger families SL_2^τ (the 2-local tree-sets), SL_3^τ (the 3-local tree-sets), and so on for all positive $n \geq 2$.

But there will be sets of trees we cannot describe with local trees of any finite maximum depth. Consider, for example, the set of binary trees in which most nodes are labeled α but there is *at least one* node somewhere that is labeled β, with α above and below it. This cannot be SL_k^τ for any k: no matter how large k is (i.e., no matter how deep the building-block trees are), a tree with no β will be indistinguishable from one containing a β more than k nodes above or below the bit you're currently checking.

I am simply using a tree analog of a simple theorem about SL string languages here. In an SL string language, a property that we could call Prefix-Blind Suffixing always holds, and the converse is also true. In a strictly k-local stringset (SL_k for some $k \geq 2$), given a string x of length $k-1$, if wxy and vxz are both in the set, then wxz has to be in the set. The occurrence of the suffix z cannot depend on anything before the x, because the x is too long.

The tree analog is that in a strictly k-local tree-set, once a tree contains a subtree T of depth $k-1$, the well-formedness of a tree formed by adding some further subtree below it (closer to the frontier) cannot depend on what occurs above it (closer to the root).

We can develop a more powerful description language by analogy with the LT string languages. Instead of just providing a list of depth-k trees as our grammar, we can close the defined sets under boolean operations by allowing grammars to say things like 'either T_1 does not occur or T_2 and T_3 do not both occur', and so on. This permits us to define for each k the tree-sets that are k-locally testable, which we can call LT_k^τ.

The family of LT_k^τ tree-sets is rich enough to contain the set of all and only those binary-branching trees in which there is at least one node labeled β, but every β has $alpha$ nodes both immediately above it and immediately below it. This is the intersection of three obviously SL_2^τ tree-sets: (i) the trees which have α as the root label and all the frontier labels, (ii) the trees which do not contain any cases of a β node with a β-labeled child, and (iii) the trees in which there is an α-labeled node with a β-labeled child. That proves it is LT_2^τ, because the LT_k^τ tree-sets can be described by propositional calculus formulas in which the primitive propositions say things like 'k-depth subtree T does not occur'. The set just mentioned is describable by the conjunction of three SL_2^τ descriptions couched in terms of primitive propositions.

We can make a yet more powerful description language by allowing ourselves first-order quantification over nodes in a language containing a binary relation symbol $<_2$ to denote 'immediately dominates'. I will call this description language $\text{FO}_{<_2}$. This gives us a still larger family, TT_k^τ, the tree analog of the k-locally threshold testable stringsets. With this as our description language, we finally have enough expressive power to describe the set of all trees in which all nodes are labeled α except for a unique β-labeled node which has α-labeled nodes both above it (its parent) and below it (all of its children).

We will use $x <_2 y$) for 'x immediately dominates y'. Let '$\text{ROOT}(x)$' (meaning intuitively that x is the root of the tree) be defined by '$\neg(\exists y)[y <_2 x]$' (which says that nothing immediately dominates x). Define '$\text{LEAF}(x)$' (meaning that x is on the frontier of the tree) by '$\neg(\exists y)[x <_2 y]$' (which says that x does not immediately dominate anything). And define '$\text{BINARY}(x)$' (meaning that x is a node with exactly two children) by this formula:

(5) $(\exists y, z)[x <_2 y \wedge x <_2 z \wedge y \neq z \wedge (\forall u)[x <_2 u \Rightarrow (u = y \vee u = z)]]$

'The node x has exactly two distinct children.'

Then the set of all purely binary-branching trees is the set in which BINARY-ONLY is true:

(6) $\text{BINARYONLY} \equiv_{\text{def}} (\forall x)[\text{ROOT}(x) \vee \text{LEAF}(x) \vee \text{BINARY}(x)]$

'Every node is either the root, or a leaf, or binary-branching.'

Let LONELYBETA be the proposition that all nodes are labeled α except for a unique node labeled β like this:

(7) $\text{LONELYBETA} \equiv_{\text{def}} (\exists x)[\beta(x) \wedge (\forall y)[(\beta(y) \Rightarrow (y = x)) \wedge (\neg\beta(y) \Rightarrow \alpha(y))]]$

'There is an x that is labeled β, and x is the only node labeled β (i.e., any y labeled β is identical with x), and any other node (i.e., any y not labeled β) is labeled α.'

Now the set we want is the set satisfying the conjunction of BINARYONLY and LONELYBETA.

A further increase in expressive power can be obtained (though to save space I won't illustrate) if we move to a language in which the relation $<_2$ is replaced by its reflexive and transitive closure $<_2^*$, so that we can say 'dominates' as well as 'immediately dominates'.

And we have still not reached maximum expressive power for languages describing tree-sets. It would still not be possible to describe a set of trees in which, for some fixed k, the depth in terms of k-depth subtrees is always an even number. To achieve that, we could move from first-order logic to wMSO, which is capable of describing such sets. It was proved in the 1960s (Thatcher 1967, Thatcher & Wright 1968) that a set of trees is recognizable by a finite-state tree automaton if and only if its string yield is a CFL, and by a fundamental result of Doner (1970), wMSO on finite trees yields exactly the expressive power of finite-state tree automata.

McCawley did not know about Doner's result, and may not have known Thatcher and Wright's work, but he did recognize that the string yield of a tree-set defined by NACs (i.e., defined using SL_2^τ) is always a CFL. This insight influenced Gerald Gazdar in devising what came to be known as generalized phrase structure grammar in 1978–1979. I think it influenced the creators of its direct heir HPSG as well. But it is natural to ask what sorts of stringset you can define by using more powerful description languages. And considering the string yields of the larger and larger tree-sets describable with the analogs of the more and more powerful description languages just briefly reviewed reveals something rather amazing — though the proofs are straightforward, some covered in textbooks like Libkin (2004) and others just basically trusted mathematical folklore:

(8)

TYPE OF TREE-SET DEFINITION	STRINGSET YIELD
strictly 2-local (local trees $\equiv$ NACs)	context-free
strictly 3-local (depth-2 trees)	context-free
strictly k-local (depth-$k-1$ trees, $k > 3$)	context-free
context-sensitive 2-local NACs	context-free
locally k-testable, $k \geq 2$	context-free
first-order logic with $<_2$	context-free
first-order logic with $<_2^*$	context-free
weak monadic second-order logic (wMSO)	context-free

We seem to have reached a plateau: no matter what description language you choose, from strictly 2-local all the way up to wMSO, you just get the CFLs over and over again.

Notice that the third line in this table tells us that no matter what the size of your tree-like building blocks, if you close the set of building blocks under the

plugging-in that I earlier called frontier nonterminal substitution, you get a set of trees that has a CFL as its string yield. Thus the data-oriented parsing proposed by Remko Scha and others, and developed in Bod (1998), where in effect the grammar is simply a (statistically annotated) treebank — a set containing all of some set of trees plus all of their subtrees — can only yield CFLs.

I should make it clear that this does *not* mean that MTS is doomed to remain within the context-free realm. James Rogers (2003) realized that if you settle on wMSO as your description language, you can define a hierarchy of classes of structures of increasing complexity that has a hierarchy of classes of strings going along with it, the classes of structures being differentiated by their number of dimensions. A sentence considered as an unanalyzable unit has zero dimensions. A string has one dimension, hence only one way in which two nodes can be adjacent, the one called linear precedence, denoted by $\leq_1$. A tree has two. One is $\leq_1$. The other, denoted by $\leq_2$, allows a node to be adjacent to an entire 1-dimensional string (its children). Tree-like objects with three dimensions can be defined by adding a third relation, $\leq_3$, in terms of which a single node can be adjacent to an entire 2-dimensional tree. And so on upward. Rogers proved that the following holds:

(9) Stringset classes definable by wMSO on models fo various dimensions

NUMBER OF DIMENSIONS	TYPE OF MODELS	RESULTING STRINGSET YIELD CLASS	PROOF
0	points	finite stringsets	(obvious)
1	strings	regular stringsets	(Büchi 1960)
2	trees	context-free stringsets	(Doner 1970)
3	3-d trees	tree-adjoining stringsets	(Rogers 2003)
4	4-d trees	(no name for the class)	(Rogers 2004)
...	...	...	...

The hierarchy continues without bound — though there is currently no terminology for the stringset yields of wMSO-characterizable sets of singly-rooted tree-like models of 4 dimensions, 5 dimensions, etc. Furthermore, it has been proved by Jens Michaelis that the infinite union of all the stringset classes in the Rogers hierarchy is the one characterized by minimalist grammars as formalized by Stabler (see Stabler, Jr. 1997, Michaelis 2001). In other words, minimalist grammars in Stabler's sense are the stringsets that are the string yields of model classes wMSO-characterizable sets of singly-rooted tree-like graph models of arbitrary finite dimensionality (and thus, surely, far more expressive than will be needed for describing human languages).

6 Expressive power of HPSG

Bringing this back to the issue of HPSG is made more difficult by the curious state of the current literature. The standard works introducing the basics of HPSG contain no discussion of phrasal reduplication (as has been claimed to exist in some African languages) or the sort of cross-serial dependencies found in certain subordinate clause constructions of Dutch and Züritüütsch (Zurich Swiss German). Züritüütsch is particularly important because the varying case marking governed by different verbs yields an argument that its stringset cannot be a CFL (Shieber, 1985). Yet Swiss German does not figure in Pollard & Sag (1994), or in any of the basic pedagogical works such as Sag & Wasow (1999), Sag et al. (2003), or Levine (2017).

A number of more technical works — more than I have space to review or even list here — do cover ways of giving HPSG accounts of non-CF phenomena of the sort Swiss German exemplifies. Reape (1992) is perhaps the most influential, but see Müller (2019), Chapter 9, for pointers to the rest of the literature. A variety of different techniques are involved: re-entrancy (structure sharing) is one; relational constraints of arbitrary power are sometimes alluded to; and what is particularly important is argument merger, allowing the list-valued valence feature COMPS to gather up arguments of a subordinate constituent and then break up the list to permit checking off its members in some desired sequence. This looks as if it has the power to use the COMPS as a queue rather than a stack, which immediately provides for greater than CFG power.

I do not think there is any unitary answer to the question of what generative power results from the different uses of these mechanisms that various linguists make. Some versions of HPSG may be limited to the weak generative capacity of combinatory categorial grammar (Steedman, 2000); some probably have Turing-machine power. I confess to not having enough understanding of the voluminous literature to adjudicate on such matters; it seems to me that there is much scope for mathematical linguists to do some focused work on the weak generative capacity of HPSG in various forms, and the ways in which the various mechanisms contribute.

Such issues are important. Consider, for example, what was discovered after Richter (2000) developed a language named RSRL explicitly for stating constraints of the sort presupposed by Pollard & Sag (1994). RSRL turned out to be so expressive that even its finite model checking problem is undecidable (Kepser, 2004). In other words, full HPSG structures can be so complex, and queries expressed in RSRL can be so rich, that the task of determining what finite structures are compliant with a given RSRL constraint can lose its way as if it were being evaluated in an infinite structure, with the result that no algorithm can guarantee to determine in finite time whether a given arbitrary structure

is grammatical according to a given arbitrary RSRL-expressed grammar. This result alone tells us, of course, that RSRL on HPSG structures is vastly more complex than wMSO on trees (which guarantees decidability not just for finite model-checking but also for satisfiability).

The only point I want to contribute here has to do with unbounded dependencies. Given that description languages of significantly more expressive power than strictly local ones are available for HPSG structures and have been explored, it is a curious fact that Pollard & Sag (1994) develop their analysis of unbounded dependencies using the SLASH feature inherited from GPSG. The SLASH mechanism, we can now see (though this was not clear to the developers of GPSG in 1980), is a way of sticking to SL_2^τ descriptions, or equivalently, modifying the nonterminal vocabulary so that simple unmodified CFGs can describe unbounded dependencies.

This seems odd to me. It is as if Pollard and Sag were following Gazdar (1981) in assuming that their description had to be couched in the most primitive description language possible, namely SL_2^τ. Gazdar's breakthrough observation about unbounded dependencies was that it only takes adding a finite number of slashed categories to the inventory (upper-bounded by the square of the number of full phrasal categories, since they are the ones that can be 'extracted') to cope with unbounded dependencies and island constraints using strictly 2-local tree description (i.e., Stanley/McCawley NACs). Pollard & Sag (1994) follows Gazdar point for point on the general theory, developing different analyses where the syntactic phenomena call for it but always assuming the basic 2-local-equivalent technology that Gazdar developed.

Pollard & Sag (1994) employ the full redundancy of Gazdar's system. Take the very simple case of 'topicalization' (unbounded complement preposing) as treated in their Chapter 4. The tree in their (18) on page 165 has (when the Nonlocal Feature Principle on p. 400 is consulted to flesh it out) 'SYNSEM|NONLOCAL|INHER|SLASH $\{\,\}$' on the root of the entire sentence; 'SYNSEM|NONLOCAL|TO-BIND|SLASH $\{\boxed{1}\}$' on the root of the topicalization construction; the same thing on the trace; and 'SYNSEM|NONLOCAL|INHER|SLASH $\{\boxed{1}\}$' on each of the eight head nodes in between them.

In addition, trace nodes have to have full details of the INHER and TO-BIND values of SLASH (and QUE and REL); there is a complex specification of the internal feature structure of a trace; there is a 'Nonlocal Feature Principle', given in only the most casually informal terms on p. 164 (I quote the different version on p. 400) saying that 'For each nonlocal feature, the value of SYNSEM|NONLOCAL|INHERITED|SLASH on the mother is the set difference of the union of the values on all the daughters and the value of SYNSEM|NONLOCAL|TO-BIND|SLASH on the head daughter'; and (fn. 5, p. 164) all other ID schemata introducing phrasal heads have to be modified to include '[TO-BIND|SLASH $\{\,\}$]' on

the head daughter. All of this highly redundant feature annotation is employed simply to guarantee that there has to be a trace in the clause that accompanies a preposed complement.

If instead we do not (implicitly) restrict ourselves to SL_2^τ rules, but allow the power of (say) first-order logic in our description language for trees, we can easily guarantee the presence of a 'trace' in some subconstituent accompanying a dislocated element, *without* using GPSG-style paths of slashed categories.

For simplicity, let me assume with Huddleston et al. (2002) (henceforth *CGEL*) that preposed ('topicalized') complements are distinguished by bearing the grammatical relation 'Prenucleus' in the main clause, and every Prenucleus phrase is accompanied by a following phrasal head bearing the relation Nucleus (which is really just a special case of the head relation). We can give a simple direct statement of the fact that the head clause accompanying a Prenucleus NP must contain an NP trace. I represent grammatical relations top-down to match dominance, so '$x <_2^* y$)' means 'x dominates y' and 'Prenucleus(x, y)' means 'x has a child y bearing the Prenucleus relation to it'. Here is the statement we need:

(10) $(\forall x, y)[(\text{Prenucleus}(x, y) \wedge \text{NP}(y)) \Rightarrow$
$(\exists z)[\text{Nucleus}(x, z) \wedge (\exists t)[(z >_2^* t) \wedge \text{Trace}(t) \wedge \text{NP}(t)]]]$

'If x has an NP child y in Prenucleus function, then x also has a child z in Nucleus (= head) function and z contains an NP trace.'

This shows that we do not need SLASH to guarantee that a clause accompanying a Prenucleus ('extracted') constituent must contain a trace, even in an entirely CF-restricted descriptive system like using first-order logic on labeled trees.

Various constructions with non-subject gaps (such as the so-called '*tough*-movement' construction) can be described in a similar way, though they do not call for a Prenucleus constituent; instead they involve a complement specifically required to contain an NP gap. That is, the complement is required to be rooted at a node z such that $(\exists t)[(z >_2^* t) \wedge \text{Trace}(t) \wedge \text{NP}(t)]$ (see the discussion of 'hollow VP' complements in *CGEL*, 1245–1251, for an informal survey of the several constructions at issue).

The foregoing remarks should not be taken as an argument that we *should* describe unbounded dependencies without the now familiar GPSG-style chains of nodes bearing SLASH values. I am only pointing out that it could easily be done, using a description language that is not very rich, and has a decidable satisfiability problem.

It will take more work before we can decide whether SLASH as used in Pollard & Sag (1994) is a valuable idea in the HPSG context or or an unnecessary hold-over from GPSG. The most interesting phenomena to study in this con-

text might be the binding domain phenomena discussed by Zaenen (1983) — syntactic phenomena in various languages that are encountered only between a left-extracted constituent and the gap in subordinate structure associated with it. Zaenen posits a feature [bnd] present on every node along the spine between the two constituents, just where Gazdar's work had posited a category with a SLASH value. Could these phenomena be insightfully described in a way that involves no SLASH or BND features? We do not know, because the unacknowledged bias toward strictly local treatments of phrase structure has meant that linguists have not been asking that question during the last four decades. It might be interesting to reopen the questions raised by the data that Zaenen considered.

References

Ajdukiewicz, Kazimierz. 1935. Die syntaktische Konnexität. *Studia Philosophica* 1. 1–27. English translation published in Storrs McCall (ed.), *Polish Logic 1920–1939*, 207–231, Oxford University Press.

Bod, Rens. 1998. *Beyond grammar: An experience-based theory of language*. Stanford, CA: CSLI Publications.

Büchi, J. Richard. 1960. Weak second-order arithmetic and finite automata. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 6. 66–92.

Chomsky, Noam. 1959. On certain formal properties of grammars. *Information and Control* 2(2). 137–167. Reprinted in *Readings in Mathematical Psychology*, Volume II, ed. by R. Duncan Luce, Robert R. Bush, and Eugene Galanter, 125–155, New York: John Wiley & Sons, 1965 (citation to the original on p. 125 of this reprinting is incorrect).

Chomsky, Noam. 1993. A minimalist program for linguistic theory. In Kenneth Hale & Samuel Jay Keyser (eds.), *The view from Building 20*, 1–52. Cambridge, Massachusetts: MIT Press.

Culicover, Peter W. & Ray S. Jackendoff. 2005. *Simpler syntax*. Oxford: Oxford University Press.

Doner, John. 1970. Tree acceptors and some of their applications. *Journal of Computer and System Sciences* 4. 406–451.

Elgot, Calvin C. 1961. Decision problems of finite automata and related arithmetics. *Transactions of the American Mathematical Society* 98. 21–51.

Fagin, Ronald. 1974. Generalized first-order spectra and polynomial-time recognizable sets. In *Complexity of computation: SIAM-AMS proceedings*, vol. 7, 43–73. American Mathematical Society.

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

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum & Ivan A. Sag. 1985. *Generalized phrase structure grammar*. Oxford: Basil Blackwell.

Huddleston, Rodney, Geoffrey K. Pullum et al. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Immerman, Neil. 1999. *Descriptive complexity*. New York: Springer.

Johnson, David E. & Paul M. Postal. 1980. *Arc pair grammar*. Princeton, NJ: Princeton University Press.

Kac, Michael B. 1978. *Corepresentation of grammatical structure*. London: Croom Helm.

Kepser, Stephan. 2004. On the complexity of RSRL. *Electronic Notes in Theoretical Computer Science (ENTCS)* 53. 146–162. In Proceedings of the Joint Meeting of the 6th Conference on Formal Grammar and the 7th Conference on Mathematics of Language; online at http://dx.doi.org/10.1016/S1571-0661(05)82580-0.

Lakoff, George. 1971. On generative semantics. In Danny D. Steinberg & Leon A. Jakobovitz (eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*, 232–296. Cambridge: Cambridge University Press.

Levine, Robert D. 2017. *Syntactic analysis: An HPSG-based approach*. Cambridge: Cambridge University Press.

Libkin, Leonid. 2004. *Elements of finite model theory* Texts in Theoretical Computer Science. Springer.

McCawley, James D. 1968. Concerning the base component of a transformational grammar. *Foundations of Language* 4. 243–269. Reprinted in James D. McCawley, *Grammar and Meaning*, 35–58 (New York: Academic Press; Tokyo: Taishukan, 1973).

McNaughton, Robert & Seymour Papert. 1971. *Counter-free automata*. Cambridge, MA: MIT Press.

Medvedev, Yu. T. 1956 [1964]. On the class of events representable in a finite automaton. In Edward F. Moore (ed.), *Sequential machines: Selected papers*, vol. II, 215–227. Reading, MA: Addison-Wesley. Originally published in Russian in *Avtomaty* (1956), 385–401.

Michaelis, Jens. 2001. Transforming linear context-free rewriting systems into minimalist grammars. In Philippe de Groote, Glyn Morrill & Christian Retoré (eds.), *Logical Aspects of Computational Linguistics: 4th international conference* (Lecture Notes in Artificial Intelligence 2099), 228–244. Berlin and New York: Springer.

Müller, Stefan. 2019. *Grammatical theory: From transformational grammar to constraint-based approaches*. Berlin: Language Science Press 3rd edn.

Peters, P. Stanley & Robert W. Ritchie. 1969. Context-sensitive immediate constituent analysis — context-free languages revisited. In *Proceedings of the ACM conference on the theory of computing*, 1–8. Republished in *Mathematical Systems Theory* 6 (1973), 324-333.

Pollard, Carl J. 1996. The nature of constraint-based grammar. Presented at Pacific Asia Conference on Language, Information, and Computation, Kyung Hee University, Seoul, Korea, December 20. Plain text draft available online at http://lingo.stanford.edu/sag/L221a/pollard-96.txt.

Pollard, Carl J. & Ivan A. Sag. 1987. *Information-based syntax and semantics, volume 1: Fundamentals*. Stanford, CA: CSLI Publications.

Pollard, Carl J. & Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Stanford, CA: CSLI Publications.

Reape, Mike. 1992. *A formal theory of word order: A case study in West Germanic*. Edinburgh, UK: University of Edinburgh dissertation.

Richter, Frank. 2000. *A mathematical formalism for linguistic theories with an application in Head-driven Phrase Structure Grammar*. Tübingen, Germany: Universität Tübingen dissertation.

Rogers, James. 1997. Strict LT_2 : Regular :: Local : Recognizable. In Christian Retoré (ed.), *Logical Aspects of Computational Linguistics: First international conference, LACL '96 (selected papers)* (Lecture Notes in Artificial Intelligence 1328), 366–385. Berlin and New York: Springer.

Rogers, James. 1999. The descriptive complexity of generalized local sets. In Hans-Peter Kolb & Uwe Mönnich (eds.), *The mathematics of syntactic structure: Trees and their logics* (Studies in Generative Grammar 44), 21–40. Berlin: Mouton de Gruyter.

Rogers, James. 2003. wMSO theories as grammar formalisms. *Theoretical Computer Science* 293. 291–320.

Rogers, James. 2004. Wrapping of trees. In Donia Scott (ed.), *Proceedings of the 42nd annual meeting of the association for computational linguistics*, Morristown, NJ: Association for Computational Linguistics. Article no. 558, doi 10.3115/1218955.1219026; online at http://portal.acm.org/citation.cfm?id=1219026#.

Rogers, James & Geoffrey K. Pullum. 2011. Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information* 20. 329–342.

Sag, Ivan A. & Thomas Wasow. 1999. *Syntactic theory: A formal introduction*. Stanford, CA: CSLI Publications 1st edn.

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

Shieber, Stuart. 1985. Evidence against the context-freeness of human language. *Linguistics and Philosophy* 8. 333–343.

Soames, Scott. 1974. Rule orderings, obligatory transformations, and derivational constraints. *Theoretical Linguistics* 1. 116–138.

Stabler, Jr., Edward P. 1997. Derivational minimalism. In Christian Retoré (ed.), *Logical Aspects of Computational Linguistics, LACL '96* (Lecture Notes in Artificial Intelligence 1328), 68–95. Berlin: Springer Verlag.

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

Thatcher, James W. 1967. Characterizing derivation trees of context-free grammars through a generalization of finite automata theory. *Journal of Computer and System Sciences* 1. 317–322.

Thatcher, James W. & J. B. Wright. 1968. Generalized finite automata theory with an application to a decision problem of second-order logic. *Mathematical Systems Theory* 2(1). 57–81.

Thomas, Wolfgang. 1982. Classifying regular events in symbolic logic. *Journal of Computer and Systems Sciences* 25. 360–376.

Trakhtenbrot, Boris A. 1962. Finite automata and monadic second order logic. *Sibirskii Matematicheskii Zhurnal* 3. 101–131. In Russian; English translation in AMS Translations 59:23–55, 1966.

Zaenen, Annie. 1983. On syntactic binding. *Linguistic Inquiry* 14. 469–504.

Representing scales: Degree result clauses and emphatic negative polarity items in Romanian

Monica-Mihaela Rizea
Solomon Marcus Center for Computational Linguistics, Bucharest

Manfred Sailer
Goethe-University Frankfurt

Proceedings of the 26th International Conference on
Head-Driven Phrase Structure Grammar

University of Bucharest

Stefan Müller, Petya Osenova (Editors)

2019

CSLI Publications

pages 79–99

http://csli-publications.stanford.edu/HPSG/2019

Keywords: negative polarity item, result clause semantics, LRS, Romanian

Rizea, Monica-Mihaela, & Sailer, Manfred. 2019. Representing scales: Degree result clauses and emphatic negative polarity items in Romanian. In Müller, Stefan, & Osenova, Petya (Eds.), *Proceedings of the 26th International Conference on Head-Driven Phrase Structure Grammar, University of Bucharest*, 79–99. Stanford, CA: CSLI Publications.

Abstract

The paper proposes a representational re-encoding of the scalar, pragmatic accounts of NPI licensing within the framework of *Lexical Resource Semantics* (LRS). The analysis focuses on a less researched distribution pattern: emphatic NPIs occurring in result clause constructions that receive an intensification reading. We will provide a scalar extension of a standard semantic account of result clauses to capture the high degree interpretations. Our investigation will also offer new insights on NPI licensing in embedded clauses. We will primarily consider Romanian data.

1 Introduction

While scalar analyzes play an important role in recent research in formal semantics and pragmatics, there has been no attempt to integrate them into a representational framework. In this paper, we will propose an implementation of the scalar theories within *Lexical Resource Semantics* (LRS, Richter & Sailer 2004) – a constraint-based underspecified semantic combinatorics for HPSG. In particular, we will discuss two phenomena for which a scalar approach is very natural: high degree readings of finite *result clause constructions* (RCX) and *emphatic negative polarity items* (E-NPI). We base our analysis on the patterns identified in Romanian.

The paper will proceed as follows: Section 2 describes the distributional properties of the E-NPIs occurring in the Romanian finite RCXs that receive a high degree interpretation. Section 3 defines some important characteristics of the LRS framework. We then propose an LRS-rendering of a scalar approach on NPI licensing, starting from the theory of Krifka (1995) (Section 4). In Section 5, we focus on the analysis of result clauses and on the interaction between emphatic NPIs and degree RCXs, while pointing out some important differences between Romanian and English; we also adapt the standard semantic analysis of degree result clauses from Meier (2003) and provide an LRS description. In Section 6, we develop an analysis of the fixed, idiomatic degree result clauses, which contribute a plain intensification reading, as *mixed expressives* with non-at-issue literal meaning (Gutzmann, 2011). Section 7 concludes the paper.

2 Data

In this paper, we focus on finite *result clause constructions* (RCXs), which express a primary predication in the main clause and a secondary predication in the result clause (RCl) – see example (1). We restrict ourselves to RCls modifying adjectives, where the RCl can be used to make a high degree statement for the

[†]Monica-Mihaela Rizea was supported by a DAAD research grant to Frankfurt a.M., January–March 2019. We thank the reviewers and the audience for their comments. All errors are ours.

matrix predicate. In (1), the RCX *(atât de) deasă de nu se vede om cu om*, just as its English correspondent *so thick (that) you can't see your hand in front of your face*, can receive the high degree interpretation of 'extremely thick'.

(1) Dimineața e o ceață [**RCX: atât de deasă**, [RCl: **de nu se vede om cu om**]].
Morning.DEF is a fog so thick.ADJ that not REFL see.3SG person with person
Intended: 'In the morning, the fog is [**RCX: so thick [RCl: you can't see your hand in front of your face**]].'

In (1), and in (2) below, the content of the RCl corresponds to an extreme outcome of the primary predicate (i.e., it makes an emphatic statement), which triggers the intensification reading of the modified predicate. A similar observation is made in Hoeksema & Napoli (2019), for Dutch and English. Note that in Romanian, unlike in English, high degree RCXs do not require a degree marker, *atât de/așa de* 'so'. We will discuss this in detail in Section 5.

(2) Ion e [**RCX: așa de prost [RCl: de nu știe cum îl cheamă (cu buletinul în mână)**].
lit.: Ion is so stupid that he does not know his own name (with the ID in hand).
Intended: 'Ion is [**RCX: so stupid [RCl: he can't see a hole in a ladder**]].'

We have analyzed a special type of degree RCXs, where the secondary predication in the RCl is an *emphatic negative polarity item* (E-NPI). E-NPIs, which are a prominently-studied case of emphatic statements (see Krifka 1995, Eckardt 2005, Chierchia 2006, and others), represent expressions that are excluded from positive environments. As shown below, a positive statement would make the expressions highly infelicitous:

(3) Dimineața e o ceață [atât de deasă, de **#(nu)** se vede om cu om].
Intended: 'In the morning, the fog is [so thick you can **#(not)** see your hand in front of your face].'.

(4) Ion e [așa de prost de **#(nu)** știe cum îl cheamă (cu buletinul în mână)].
Intended: 'Ion is [so stupid he can **#(not)** see a hole in a ladder].'.

Many E-NPIs are also *minimizers*, which typically denote minimal elements on a contextually-salient scale. In the examples above, *se vede om cu om* (lit. *see the person in one's immediate range of sight*), corresponding to the English *see one's hand in front of one's face*, emphatically expresses what for the speaker counts as a minimal range of visibility; *știe cum îl cheamă* (lit. *he knows his own*

name) stands for a minimal manifestation of one's knowledge, while the English correspondent, *see a hole in a ladder*, suggests a minimal manifestation of one's sensitivity to details. The observation is that negating some minimum pragmatic threshold on a contextual scale can lead to strong emphatic utterances (Krifka 1995, Eckardt 2005); this further proves that, when embedded in RCls, negated minimizers can be very naturally employed for triggering a high degree reading of the matrix predicates. For example, in *ceață atât de deasă de nu* ***se vede om cu om*** 'fog so thick that you can't **see your hand in front of your face**', the minimizer could be interpreted as emphatically indicating an extremely low degree of visibility, and, when negated, it triggers an inference related to the (extreme) intensity of the fog.

In what concerns the complementizers, in Romanian, RCls can be introduced with *încât* (which is the default case), *că*, or *de* (see GBLR, Pană Dindelegan 2010, 583). When it occurs in RCls, *de* seems to be restricted to emphatic sentences. In (5), a strongly favorable consequence of the quality of being elegant (i.e., being *admired*) is contrasted with a neutral consequence, where Ion is no more than *noticed*:

(5) Ion se îmbracă atât de elegant [**încât/de** lumea îl *admiră*]/
Ion REFL dresses so elegantly that people him *admire*/
[**încât/#de** lumea îl *observă*].
that people him *notice*
'Ion dresses so elegantly that people *admire* him/that people (no more than) *notice* him.'

Conventionalized finite RCls, many originating from RCls hosting regular word combinations associated with an extreme outcome, seem to represent a productive pattern for expressions that have been lexicalized into high-degree modifiers in Romanian – cases when the RCl expresses a high degree of intensity of the primary predicate, and the result interpretation is entirely replaced by an intensification reading (see (6)). Moreover, the most conventionalized expressions that have evolved into high-degree modifiers normally collocate with *de* and reject interchangeability with *încât*, the regular connector for the non-conventionalized RCls; this further proves that *de* is strongly associated with an intensification interpretation:

(6) a. *(frumoasă)* [*de/# încât nu se poate*]
(lit.: (so beautiful) that it cannot be) 'very beautiful'
b. *(frumoasă)* [*de/# încât mori*]
(lit.: (so beautiful) that one dies) 'very beautiful'.

Up to this point, we have made the following observations: RCXs can have a high degree interpretation (**OBS1**); *de*-RCXs require an emphatic statement inside the RCl (**OBS2**); there are lexicalized RCls that *only* have an intensification reading (**OBS3**).

In the rest of the chapter, we will present four tests (**T1–T4**) that we have designed in order to classify E-NPIs embedded in high-degree RCls. For Romanian, we have identified three main types of E-NPIs; each type will be illustrated with one example:

(7) a. E-NPI1: *a (nu) vedea la un pas* 'not see within a step'
(lit.: not to see a step ahead)
(referential, result reading: 'there is no visibility at all')

b. E-NPI2: *a (nu) se vedea om cu om* 'not REFL see person with person'
(lit.: not to see the person in one's immediate range of sight)
(referential, result reading: 'there is no visibility at all')

c. E-NPI3: *a (nu) [te/vă] vedea* 'not CL.ACC.2SG/PL I.see'
(lit.: not to see you)

Our tests will show that E-NPI1s and E-NPI2s convey a result state of the primary predicate since they have a referential reading – in our examples, related to the *lack of visibility*, i.e., *there is no visibility at all*; E-NPI3s contribute a purely intensifier reading in relation to the matrix predicate, and do not assert a result meaning.

> **T1:** Can we change the RCX into a coordination without changing the meaning of the expression?

(8) E-NPI1 & E-NPI2

a. E o aglomerație pe străzi în timpul grevei [**de** nu *se vede la un pas*]/ [**de** nu *se vede om cu om*].
'There is a huge crowd in the streets during the strike.' (lit.:There is a crowd in the streets during the strike **that** one cannot see a step ahead/ **that** one cannot see the person in their immediate range of sight.)

b. = E o aglomerație pe străzi în timpul grevei [**și** nu *se vede la un pas*]/ [**și** nu *se vede om cu om*]. (lit.: There is a crowd in the streets during the strike **and** one cannot see a step ahead/ **and** one cannot see the person in their immediate range of sight.)

(9) E-NPI3

a. Emoțiile astea mi-au făcut foame [**de** nu *te văd*]. (CoRoLa)
'These emotions made me extremely hungry.'
(lit.: These emotions made me hungry **that** I cannot see you.)

b. ≠ Emoțiile astea mi-au făcut foame [**și** nu *te văd*].
(lit.: These emotions made me hungry **and** I cannot see you.)

In **T1**, we have started from an RCX and changed it into a coordination, where a result relation can still be inferred. Both E-NPI1 and E-NPI2 pass the

test – (8b); however, if the sentence hosting E-NPI3 is considered in isolation, the expression suffers a change in meaning since only the *literal* reading is available in coordination – see (9b) – i.e., **T1** distinguishes between the third type and the first two types of E-NPIs. E-NPI1 and E-NPI2 are felicitous according to **T1**, since their meaning, based on a scalar inference (i.e., *there is no visibility at all*), remains unchanged when used outside an RCX. In other words, E-NPIs such as *a se vedea la un pas* and *a se vedea om cu om* clearly have distinct *literal* meanings – one expressing visibility within the distance of a step, the other visibility to the nearest person in someone's immediate range of sight. Used as E-NPIs, however, both assert a minimal degree of visibility. By contrast, an E-NPI3 undergoes a change in meaning when used in a coordination structure – see the infelicity of (9b). Thus, the meaning that the expression would have in isolation does not contribute to the high degree reading of the entire RCX.

In **T1**, the RCX is changed into a coordination, and a result relation can be inferred in all the examples. In **T2**, we will look at cases in which no such relation can be inferred. Since E-NPI3 is already excluded by **T1**, we will only apply **T2** to E-NPI1 and E-NPI2:

> **T2**: Can the expression be used felicitously if the context does not permit the inference of a result relation?

(10) E-NPI1 & E-NPI2
Mergeam pe stradă [**și** nu *se vedea la un pas*]/
[#**și** nu *se vedea om cu om*].
(lit.: I was walking down the street **and** one could not see a step ahead/ **and** one could not see the person in their immediate range of sight.)

As shown in (10), E-NPI1 passes **T2**, whereas E-NPI2 cannot be used felicitously in the absence of a salient result relation in discourse. This shows that an E-NPI2 is collocationally restricted to a result relation.

The following test looks at the distribution of the possible complementizers of the RCls that occur in high degree result constructions:

> **T3:** Is variation possible with respect to the RCl complementizer without a change in the meaning of the expression in the RCl?

(11) E-NPI1 & E-NPI2
E așa de întuneric afară [**de/încât** nu *se vede la un pas*]/ [**de/încât** nu *se vede om cu om*].
(lit.: It's so dark outside that one cannot see a step ahead/ that one could not see the person in their immediate range of sight.)
'It is very dark outside.'

(12) E-NPI3
Emoțiile astea mi-au făcut foame [**de**/#**încât** nu *te văd*].
(lit.: These emotions made me hungry **that** I cannot see you.)
'These emotions made me extremely hungry.'

In (11), E-NPI1 and E-NPI2 allow for both *de* and *încât*, while the meaning of the RCl remains unchanged (i.e, *there is no visibility at all*); by contrast, E-NPI3 requires the presence of *de*, see (12). The use of *încât* in (12) triggers a change in meaning: the expression in the RCl can only be interpreted *literally*, which leads to infelicity.

T4 is intended to clarify what is the meaning contributed by RCl hosting the E-NPI to the overall RCX:

T4: Does the RCX entail the proposition in the result clause?

(13) E-NPI1 & E-NPI2
Ninge **a.** [de nu *se vede la un pas*]/**b.** [de nu *se vede om cu om*].
(lit.: It is snowing **a.** [that one cannot see a step ahead]/
b.[that one can't see the person in one's immediate range of sight].)
'It is snowing very hard.'
Entails: **a.** Nu *se vede la un pas*./**b.** Nu *se vede om cu om*.
(result reading: both **a.** and **b.** trigger the scalar inference *there is no visibility at all*)

(14) E-NPI3
Emoțiile astea mi-au făcut o foame [de nu *te văd*].
(lit.: These emotions made me hungry [that I cannot see you].)
'These emotions made me extremely hungry.'
Does not entail: Nu *te văd*. (no result reading)

Both expressions in (13) have a high-degree reading, and they entail the proposition in the RCl. In both cases, there is also a result reading since the expressions trigger a scalar inference: If it is snowing so hard that *one cannot see a step ahead*/that one *cannot see the person in their range of sight*, then *there might be no visibility whatsoever*. By contrast, the RCX with the interpretation of 'extremely hungry' in (14) does not entail the meaning of the sentence in the RCl. This shows that the sole meaning contribution of the expression to the RCX is *intensification* i.e., the RCl asserts high degree rather than its result reading.

The results of our tests are summarized in Table 1. They allow us to identify three types of E-NPIs that can occur in RCXs with high degree readings:

(15) a. E-NPI1s are only occasionally used in result clauses and act as intensifiers; there is also a result interpretation.

		T1	T2	T3	T4
E-NPI1:	*(*de) nu se vede la un pas	✓	✓	✓	✓
E-NPI2:	*de nu se vede om cu om*	✓	✗	✓	✓
E-NPI3:	*de nu [te/vă] văd*	✗	n/a	✗	✗

Table 1: Results of the tests

b. E-NPI2s require a result relation, being bound to the RCXs; they encode a high degree reading, while also keeping the notion of result.

c. E-NPI3s express nothing but intensification, being lexicalized into high-degree modifiers.

Having presented the core data, in the following chapter we will describe the general framework used in the analysis.

3 Framework: Lexical Resource Semantics (LRS)

Lexical Resource Semantics (LRS, Richter & Sailer 2004) is a constraint-based underspecified semantic combinatorics for HPSG – similar in some respects to *Minimal Recursion Semantics* (Copestake et al., 2005) or *Constraint Language on Lambda Structures* (Egg et al., 2001). The major difference is that LRS uses expressions of some standard semantic representation language for the semantic representation of a linguistic expressions – in the present paper, a version of higher order predicate logic. LRS has been successfully applied to a number of challenging phenomena at the syntax-semantics interface, including scope ambiguity (Richter & Sailer, 2004), negative concord (Iordăchioaia & Richter, 2015), gapping (Park et al., 2018), projective meaning (Hasegawa & Koenig, 2011; Sailer & Am-David, 2016), and others. We will use a version of the compact LRS notation introduced in Penn & Richter (2005), which can be transformed into the more explicit AVM-notation used in Richter & Sailer (2004) without loss of information.[1]

In LRS, linguistic signs contribute constraints on the semantic representation of the structure containing them. There are *contribution* constraints, which determine the constants, variables, and operators, and *embedding* constraints, which determine subexpression relationships within the larger semantic representation. LRS is *lexical* in the sense that only lexical items (signs licensed by lexical entries and lexical rules) may make contribution constraints. We use a semantic metalanguage to express LRS-constraints which enriches our representation language with metavariables ($\alpha, \beta, \ldots$).

[1] A complete list of LRS-related publications and other material can be found at https://www.lexical-resource-semantics.de, accessed 14.10.2019.

We briefly illustrate the system with the example in (16a). We show the constraints contributed by the words in (16b). The word *call* requires that the expression **call**(x) occur in the semantic representation of the clause. The negated auxiliary constrains the semantic representation to contain a negation, but it does not commit to what is in the scope of the negation. This is expressed with the metavariable α. The quantified NP *everyone* requires a universal quantifier, the variable it binds, and information on the restrictor. The scope is largely underspecified, indicated with the metavariable β. There is an embedding constraint, $\beta[x]$ requiring that whatever expression β is, it must contain an occurrence of x.

(16) a. [S: Everyone [VP: didn't call]].

b. Lexical constraints:

call:	**call**(x)	*didn't*:	$\neg\alpha$
everyone:	$\forall x(\textbf{person}(x) \rightarrow \beta[x])$		

When the words combine, these constraints are collected and embedding constraints will be added, depending on the kind of syntactic combination. For the VP, it is required that a semantic contribution of the verb *call* be in the scope of negation, see (17a).[2] When the VP combines with the quantified NP, the expression **call**(x) needs to be in the scope of the universal quantifier, (17b).

(17) Phrasal constraints on scoping:

a. $\alpha[\textbf{call}(x)]$ (*call* is in the scope of negation)

b. $\beta[\textbf{call}(x)]$ (*call* is in the scope of the subject quantifier).

This leads to the overall constraint in (18). The metavariable γ is constrained to contain a universal quantifier and a negation, both of which will have **call**(x) in their scope. The relative scoping of the universal and the negation is, however, not constrained.

(18) $\gamma[\forall x(\textbf{person}(x) \rightarrow \beta[\textbf{call}(x)], \neg\alpha[\textbf{call}(x)]]$

At the level of the overall utterance, there is a closure constraint. This means that the semantic representation of the sentence is any expression that satisfies all the constraints and does not contain any additional elements. We can see which readings there are by looking for *pluggings* (Bos, 1996), i.e., mappings from metavariables to subexpressions of the meta-expression in (18). For our example, there are exactly two possible pluggings, which accounts for the two readings of the sentence. We show these pluggings in (19).

[2]The expression **call**(x) is called the *internal content* of the verb, which will be inherited by the auxiliary. Phrase-level embedding constraints typically make reference to the internal content of one of the daughters. See Richter & Sailer (2004) and Penn & Richter (2005) for details.

(19) a. Plugging: $\alpha = \mathbf{call}(x); \beta = \neg\alpha; \gamma = \forall x(\mathbf{person}(x) \rightarrow \beta)$:
Reading: $\forall x(\mathbf{person}(x) \rightarrow \neg\mathbf{call}(x))$

b. Plugging: $\alpha = \forall x(\mathbf{person}(x) \rightarrow \beta); \beta = \mathbf{call}(x); \gamma = \neg\alpha$:
Reading: $\neg\forall x(\mathbf{person}(x) \rightarrow \mathbf{call}(x))$

We will say a few words on our treatment of presuppositions and conventional implicatures. We largely follow Sailer & Am-David (2016), changing some attributes. All combinatorial and projective semantics information is collected in the value of an attribute LRS. The semantic constraints of a sign are given as meta-expressions on a PARTS-list. A sign's at issue content corresponds to the value of an AT-ISSUE attribute. There are two additional list-valued attributes, PRESUP(POSITIONS) and CI. The PARTS list contains (at least) the meta-expression in the AT-ISSUE value and everything on the PRESUP and CI lists. The final semantic representation of an utterance, i.e. the value of the EX(TERNAL)-CONT(ENT) attribute, contains all meaning components, integrating all presuppositions and CIs. Projective content that appears as part of the EX-CONT value is removed from the PRESSUP and CI lists (Sailer & Am-David, 2016, 653).

Our feature geometry is illustrated in (20), which is an adaptation the analysis of the definite article from Sailer & Am-David (2016). The EXC-CONT is underspecified. The PARTS list contains all meta-expressions of the remaining semantic features. The AT-ISSUE value is just a variable. The existence requirement of definites is encoded as a presupposition in the PRESUP list, and uniqueness is assumed to be a CI and, consequently, included in the CI value.

(20) Semantic constraints of the definite article:

$$\left[\begin{array}{ll} & \left[\begin{array}{ll} \textsc{ex-cont} & \delta \\ \textsc{parts} & \langle x \rangle \oplus \boxed{1} \oplus \boxed{2} \\ \textsc{at-issue} & x \\ \textsc{presup} & \boxed{1}\langle \exists x(\alpha[x] \wedge \beta[x]) \rangle \\ \textsc{ci} & \boxed{2}\langle \gamma \wedge (\exists x \alpha) \rightarrow (\exists! x(\alpha[x])) \rangle \end{array}\right] \\ \textsc{lrs} & \end{array}\right]$$

The distinction between presuppositions and CIs is useful as these meaning components have distinct projective properties (see Karttunen & Peters 1979; Bach 1999; Potts 2005; Tonhauser et al. 2013, among others). Presuppositions can be integrated into the at issue content in the scope of operators, CIs need to project until the level of a speech act operator.[3]

4 Analysis 1: NPIs

Having established our framework, we can now propose an LRS-rendering of a scalar theory of emphatic NPIs in the spirit of Krifka (1995). Example (21), which we use for illustration, contains the minimizer NPI *a thing*. We include the at issue content of the sentence.

[3] As we do not use speech act operators here, CIs will be integrated into the highest EX-CONT.

(21) Alex didn't see a thing. $\neg\exists x(\textbf{minimal-thing}(x) \wedge \textbf{see}(\textbf{alex}, x))$

The NPI *a thing* refers to a minimal thing one could perceive visually, for which we use the constant **min(imal)-thing**. Krifka (1995) builds his analysis on a background-focus structure. The focus is determined by the descriptive content of the NPI, here the predicate **min-thing**. Minimizer NPIs trigger larger, scalar alternatives, i.e. alternatives that contain the meaning of the NPI. For our example the set of alternatives is $\{P|\textbf{min-thing} \subseteq P\}$. These alternatives are context dependent. Being on an African safari and trying to spot some animals, for example, the alternatives would include an antelope, a lion, a herd of elephants, etc. – but not trees, photographic equipment or others.

According to Krifka, a minimizer NPI has to be used in an emphatic statement. He expresses this by requiring that what is asserted in a sentence with an NPI must entail what would have been asserted had any of the alternatives been used instead. Example (21) is well formed because it entails all alternatives, i.e., not seeing an antilope, a lion, etc. Without a negation (or another scale-reversing operator), the entailment would not hold, i.e., seeing a minimal thing does not entail seeing an antelope, etc. Krifka (1995) expresses this requirement with a speech-act operator, **ScalarAssert**, that takes a background-focus-alternatives structure as its argument. An NPI triggers a set of alternatives and must be used in an utterance that makes a scalar assertion.

This theory has been widely adapted. Eckardt (2005) refines the semantics of the NPIs, and Chierchia (2004, 2006) shows how this theory can be integrated into Mainstream Generative Grammar. To name just two examples.

While very attractive, the original approach faces some serious problems. First, as NPI licensing is connected to the speech act operator **ScalarAssert**, it is unclear how NPI licensing works in embedded clauses. Our data on NPIs in RCl are a case in point. Second, not all NPIs are emphatic, such as *ever* or unstressed uses of *any*. Third, Eckardt & Csipak (2013) show that the proposal cannot capture the varieties of types of NPIs found in languages.

Previous HPSG-approaches to NPIs, such as Richter & Soehn (2006) or Sailer (2007), address some of these problems, but do not capture the intuitive connection between the minimal semantics of many NPIs and their NPI-hood.

In this paper, we will present a representational rendering of basic ideas from Krifka (1995). The main component of our theory is an operator **ScAs**. It is defined in such a way that it has the same effect as Krifka's **ScalarAssert** when used with highest scope in an unembedded utterance. It is, however, an ordinary operator and can, therefore, be used in embedded contexts as well. This operator is defined in (22).

(22) For each formula β with subexpression ϕ_τ, and each expression $\Sigma_{\tau t}$,
ScAs(β, ϕ, Σ) is an *emphatic expression*, where
$[\![\textbf{ScAs}(\beta, \phi, \Sigma)]\!] = [\![\beta \wedge \forall P \in \Sigma(\beta \rightarrow \beta'))]\!]$,
where β' is just like β but with P replacing ϕ.

In this definition, the expression ϕ has the function of Krifka's focus. The formula β corresponds to Krifka's background applied to the focus. Σ is the set of alternatives to ϕ. **ScAs**(β, ϕ, Σ) is a complex expression whose truth conditions are defined holistically instead of compositionally. Such an emphatic expression is true iff β is true and for each alternative P in Σ, β implies the result of replacing every occurrence of ϕ in β with P.

We use this operator in our analysis of a Romanian E-NPI. The semantic specification of the E-NPI are given in (23), followed by an example sentence in (24) for which we provide the relevant semantic attributes as well. The at issue content of the sentence, $\boxed{1}$, contains only its basic truth conditions that Maria lacks visibility. The NPI triggers a set of alternatives as a presupposition, $\boxed{2}$. The PRESUP value specifies that the alternatives are such that each of them must entail the minimal range of visibility. The NPI also contributes a **ScAs** operator. The first argument of this operator is the at issue content. The second argument is the focus element, i.e. the basic semantic predicate contributed by the NPI. Here, it is a minimal range of visibility, **min(imal)-range**. The third argument is the presupposed set of alternatives. As the variable A occurs freely inside the **ScAs** expression, the presupposition needs to take scope over it.

(23) LRS value of an E-NPI1: *vede la un pas*

$$\left[\text{LRS}\left[\begin{array}{ll}\text{PARTS} & \langle \boxed{1}, \boxed{2}, \textbf{ScAs}(\alpha, \textbf{min-range}, A)\rangle \\ \text{AT-ISSUE} & \boxed{1}\, \alpha[\exists x(\textbf{min-range}(x) \wedge \textbf{see}(x, y))] \\ \text{PRESUP} & \langle \boxed{2}\, \exists A(\forall P \in A(P \subseteq \textbf{min-range}) \wedge \beta)\rangle\end{array}\right]\right]$$

(24) Maria nu vede la un pas.
Maria not sees within a step 'Maria doesn't have any visibility.'

$$\left[\text{LRS}\left[\begin{array}{ll}\text{EX-C} & \boxed{2}\, \exists A(\forall P \in A(P \subseteq \textbf{min-range}) \wedge \textbf{ScAs}(\boxed{1}, \textbf{min-range}, A))) \\ \text{AI} & \boxed{1}\, \neg\exists x(\textbf{min-range}(x) \wedge \textbf{see}(\textbf{maria}, x)) \\ \text{PRESUP} & \langle\rangle\end{array}\right]\right]$$

Our analysis captures the scalar effect of the E-NPI correctly: it presupposes a set of alternatives and is true if the asserted content entails any alternative if used instead of the NPI.

It is important that the **ScAs** expression is not part of the at issue content. This means that it is backgrounded in the sense of Potts (2005). Potts argues that if backgrounded material is not true, the sentence cannot be interpreted properly. The **ScAs** expression is similar to CIs in that its truth value is independent of that of the at issue content. However, it is not a CI, as CIs are outside the scope of presuppositions (Potts, 2005), whereas the **ScAs** expression needs to be in the scope of the presupposition, as explained above.

Let us assume we use our NPI without a licensing operator, i.e., we remove *nu* 'not' in (24). In this case, the sentence would not be ungrammatical and the at issue content could even be true. However, the **ScAs** expression would not be true and we get a similar effect as for untrue CIs. Consequently, just as in Krifka's and other pragmatic theories, we do not need to specify the negation

in the lexical specification of an E-NPI, as it will follow from the requirements of the **ScAs** operator. Not being a CI, however, the **ScAs** expression might turn out as part of the at issue content of a higher clause in a structure.

In this section, we showed how an NPI-licensing theory based on scalar inference can be expressed within a representational framework. Our LRS encoding has at least the two advantages: First, the NPI can be lexically specified to contribute the predicate **min-range** and the **ScAs** operator at the same time, see (23). This last aspect has remained unaddressed in the purely semantic-pragmatic literature and solved by some syntactic feature mechanism in Chierchia (2004). Second, while **ScAs** is an ordinary operator, it is backgrounded but neither presupposed nor a CI.

5 Analysis 2: Result clauses

We will adopt the analysis of result clauses from Meier (2003) and, again, provide an HPSG/LRS rendering. We will, then, point out some differences between RCXs in English and Romanian and discuss the lexical entries for the Romanian RCl-complementizers *încât* and *de*.

Meier (2003) uses a *degree parameter*, d, for gradable adjectives. The degree – or *extent* – is an interval denoting the *extent* of a property. The semantic representation of a simple sentence with a gradable adjective is given in (25). The sentence is true iff the maximal extent of darkness of the room is higher than or equal to some contextually given standard.

(25) The room was dark. $\mathbf{Max}(\{d|\mathbf{dark}(d,\mathbf{the\text{-}room})\}) \geq \mathbf{standard}$

Meier analyzes RCXs as a comparison of extents. She also observes that there is a modal component. Sentence (26) is true iff the maximal extent of darkness of the room is at least as high as the minimal extent of the room's darkness that is necessary for Alex not to seen anything.

(26) The room was so dark that Alex didn't see anything.
$\mathbf{Max}(\{d|\mathbf{dark}(d,\mathbf{the\text{-}room})\})$
$\geq \mathbf{Min}(\{d|\mathbf{dark}(d,\mathbf{the\text{-}room}) \rightarrow \Box\neg\exists x(\mathbf{see}(\mathbf{alex},x))\})$

There are two occurrences of the formula $\mathbf{dark}(d,\mathbf{the\text{-}room})$ in (26). For convenience, we define the more compact notation in (27) and use it for sentence (26) in (28).

(27) For each extent variable d and each formulæ α and β,
$[\![\mathbf{ResOp}\, d\, (\alpha:\beta)]\!] = [\![\mathbf{Max}(\{d|\alpha\}) \geq \mathbf{Min}(\{d|\alpha \rightarrow \Box\beta\})]\!]$

(28) $\mathbf{ResOp}\, d\, (\mathbf{dark}(d,\mathbf{the\text{-}room}) : \neg\exists x(\mathbf{see}(\mathbf{alex},x)))$ (= (26))

In English, the degree particle *so* is obligatory, so we can assume that it contributes the result clause meaning. The RCl starts with the ordinary, optional complementizer *that*, see (29).

(29) The room was *(so) dark [(that) Alex couldn't see anything].
ResOp d (**dark**(d, **the-room**) : $\neg\exists x$(**see**(**alex**, x)))

This contrasts with Romanian, see (30). There, the degree particle is optional. However, we find a meaningful difference between the possible complementizers *de* and *încât*. This leads us to the assumption that, in Romanian, both the degree particle and the RCl-complementizer contribute a result meaning.

(30) Camera este (atât de) întunecată [*(încât) Alex nu vede nimic].
room.the is so dark that Alex not sees nothing
'The room is so dark that Alex doesn't see anything.'
ResOp d (**dark**(d, **the-room**) : $\neg\exists x$(**see**(**alex**, x)))

We can now provide the lexical specification for the result complementizers *de* and *încât* in (31), and for the degree particle *atât de* in (32).

The complementizer in (31) contributes the operator **ResOp**. It takes a clausal complement, the RCl and requires that its complement's semantics, β^*, be integrated into the second part of **ResOp**. The RCl will be integrated into a larger sentence as a modifier, selecting its head with the SELECT feature. The semantics of the modified element, α^*, occurs in the first argument of **ResOp**.

(31) Lexical entry of the result complementizers:

$$\begin{bmatrix} \text{PHON} & \langle \textit{de/încât} \rangle \\ \text{SYNS} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{RCl-complementizer} \\ \text{SELECT } A\begin{bmatrix} \text{INDEX } d \\ \text{MAIN } \alpha^* \end{bmatrix} \end{bmatrix} \\ \text{VAL} & \left[\text{COMPS } \left\langle \text{S}\left[\text{MAIN } \beta^*\right]\right\rangle\right] \\ \text{CONT} & \begin{bmatrix} \text{INDEX } d \\ \text{MAIN } \textbf{ResOp} \end{bmatrix} \end{bmatrix} \\ \text{LRS} & \left[\text{AT-ISSUE } \textbf{ResOp}\, d\, (\alpha[\alpha^*] : \beta[\beta^*])\right] \end{bmatrix}$$

The lexical entry of the degree particle is given in (32).

(32) Lexical entry of the degree particle *atât de*

$$\begin{bmatrix} \text{PHON} & \langle \textit{atât de} \rangle \\ \text{SYNS} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{degree-particle} \\ \text{SELECT } \boxed{1} A\begin{bmatrix} \text{INDEX } d \\ \text{MAIN } \alpha^* \end{bmatrix} \end{bmatrix} \\ \text{VAL} & \left[\text{COMPS } \left\langle \left(\text{CP} \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{RCl-compl} \\ \text{SELECT } \boxed{1} \end{bmatrix} \\ \text{CONT} & \boxed{2} \\ \text{EXTRA} & + \end{bmatrix} \right) \right\rangle\right] \\ \text{CONT} & \boxed{2}\begin{bmatrix} \text{INDEX } d \\ \text{MAIN } \textbf{ResOp} \end{bmatrix} \end{bmatrix} \\ \text{LRS} & \left[\text{AT-ISSUE } \textbf{ResOp}\, d\, (\alpha[\alpha^*] : \beta)\right] \end{bmatrix}$$

The degree particle similar to the result complementizer, but selects an optional RCl. If present, the RCl must be extraposed and has the same semantics as the particle, 2. If there is no RCl, the comparison needed for the result clause operator, β, is inferred from context.

We can now turn to the properties of RCXs from Section 2. There, we saw that an RCl with an emphatic content can be interpreted as an intensification of the matrix predicate. We provide a modelling of this observation which makes use of different features of projective meaning introduced in Section 3. The result complementizers come with an additional CI, which states that if the RCl is emphatic, the main clause predicate is also interpreted as emphatic, i.e., as being intensified. We can use the **ScAs** operator to formalize this CI, see (33). The formula should appear on the CI list in the lexical entry in (31).

(33) a. At issue: $\mathbf{ResOp}\, d\, (\alpha : \beta)$

b. CI content of the result construction:
$\exists A(\mathbf{ScAs}(\beta, \gamma, A)) \rightarrow \exists A' \mathbf{ResOp}\, d\, (\alpha : \mathbf{ScAs}(\alpha, d, A'))$

This CI is a formal encoding of a generalization in Hoeksema & Napoli (2019) according to which, if the matrix predicate has an extreme result, it holds to an extreme degree.

While both *încât* and *de* can be found with intensifying RCls, result-*de* is restricted to them (**OBS2**). We capture this by adding a presuppostion that the content of the RCl expresses something emphatic. In (34), the CI from (33) is added to the CI-list together with the above-mentioned presupposition.

(34) Lexical entry of the RCl-complementizer *de*:

$$\left[\begin{array}{ll}
\textsc{phon} & \langle de \rangle \\
\textsc{syns} & \left[\begin{array}{ll}
\textsc{lid} & \textit{result-de} \\
\textsc{head} & \left[\begin{array}{ll} \textit{RCl-complementizer} & \\ \textsc{select}\ A & \left[\begin{array}{ll} \textsc{index} & d \\ \textsc{main} & \alpha^* \end{array}\right] \end{array}\right] \\
\textsc{val} & \left[\textsc{comps}\ \left\langle \textsc{s}\left[\textsc{main}\ \beta^*\right]\right\rangle\right] \\
\textsc{cont} & \left[\begin{array}{ll} \textsc{index} & d \\ \textsc{main} & \mathbf{ResOp} \end{array}\right]
\end{array}\right] \\
\textsc{lrs} & \left[\begin{array}{ll}
\textsc{at-issue} & \mathbf{ResOp}\, d\, (\alpha[\alpha^*] : \beta[\beta^*]) \\
\textsc{presup} & \left\langle \exists A(\mathbf{ScAs}(\beta'[\beta^*], \gamma, A)) \right\rangle \\
\textsc{ci} & \left\langle \exists A(\mathbf{ScAs}(\beta', \gamma, A)) \rightarrow \exists A' \mathbf{ResOp}\, d\, (\alpha : \mathbf{ScAs}(\alpha, d, A')) \right\rangle
\end{array}\right]
\end{array}\right]$$

We can, now, combine the analyzes of E-NPIs and RCXs. For free E-NPIs, i.e. E-NPI1s, we use the encoding from Section 4 inside an RCl. In (35) we use the NPI *vede la un pas*.

(35) E un întuneric afară de Maria nu vede la un pas.
there.is a darkness outside that Maria not sees within a step
'It is so dark outside that Maria can't see anything.'

The LRS-value of the RCX in (35) is given in (36). The semantic representation of the RCl was already given in (24). The existential presupposition of the set of alternatives can, however, project out of the RCl and take widest scope. The resulting at issue content of the sentence is given in the AT-ISSUE-value. The PRESUP-value contains the definition of the set of alternatives. As we use the complementizer *de*, it also contains the information that the content of the RCl is interpreted emphatically, i.e. a **ScAs** expression. This condition is trivially fulfilled since the E-NPI contributes this operator. This explains why E-NPIs are well fit for use in *de*-RCXs. Finally, the CI-list contains the CI from (33). Given the presupposition that the content of the RCl is emphatic, this makes a high degree, i.e., intensification reading available.

(36) LRS-value of the RCX in (35):

$$\begin{bmatrix} \text{AI} & \textbf{ResOp}\, d\, (\textbf{dark}(d, \textbf{outside}) : \boxed{1}\textbf{ScAs}(\neg\exists x(\textbf{m-range}(x) \wedge \textbf{see}(y, x)), \textbf{m-range}, A)) \\ \text{PR} & \langle \boxed{1},\ \exists A(\forall P(P \in A \rightarrow P \subseteq \textbf{m-range}) \wedge \ldots) \rangle \\ \text{CI} & \langle \ldots(\boxed{1} \rightarrow \exists A' \textbf{ResOp}\, d\, (\textbf{dark}(d, \textbf{outside}) : \textbf{ScAs}(\textbf{dark}(d, \textbf{outside}), d, A'))) \rangle \end{bmatrix}$$

Our analysis of E-NPI1s captures their behavior with respect to our four tests: As the NPI can occur outside result clauses, we get the free exchangeability with coordination (**T1**). It also follows that the NPI can be used even if there is no salient result relation (**T2**). Since the NPI contributes **ScAs** there can be free variation with respect to the complementizer (**T3**). We think that *de* is nonetheless preferred with E-NPI1s. The result clause makes a real descriptive contribution to the meaning of the overall construction (**T4**).

We can briefly turn to E-NPI2s. They are very much like the first type of E-NPIs, but they are bound to a result semantics. We can express this by using a collocational module as proposed for HPSG in Soehn (2009) and the reference therein. Soehn (2009) assumes a feature COLL. The value of COLL contains an attribute LIC(ENSER), whose value is a list of objects that describe under which circumstances the lexical sign is licensed.

We sketch this restriction in (37), which represents the relevant parts of the lexical description of the expression. In this AVM, the expression is restricted to occur in the scope of a result clause operator, **ResOp**.

(37) Specification of an E-NPI2: se vede om cu om

$$\begin{bmatrix} & \begin{bmatrix} \text{PARTS} & \langle \boxed{1}, \boxed{2}, \textbf{ScAs}(\alpha, \textbf{min-range}, A) \rangle \\ \text{AT-ISSUE} & \boxed{1}\, \alpha[\exists x(\textbf{min-range}(x) \wedge \textbf{see}(x, y))] \\ \text{PRESUP} & \langle \boxed{2}\, \exists A(\forall P \in A(P \subseteq \textbf{min-range}) \wedge \beta[\boxed{1}]) \rangle \end{bmatrix} \\ \text{COLL} & \left[\text{LIC} \left\langle \left[\text{EX-CONT} \ [\kappa[\textbf{ResOp}\, d\, (\alpha : \beta[\textbf{min-range}(x)])]]\right] \right\rangle\right] \end{bmatrix}$$

The only difference between the two first types of E-NPIs lies in the collocational restriction, we, thus, predict the attested behavior of E-NPI2s. (**T1**) Alternation with coordination is possible as long as the result relation is salient in discourse. This means that the required **ResOp** operator can be contributed by the words in the sentence (as in overt RCXs), or it can be accommodated. (**T2**)

Consequently, if there is no – explicit or implicit – result relation, the E-NPI2 cannot be used. (**T3**) As the E-NPI2 contributes a **ScAs** operator, it is compatible with both *încât* and *de*. Finally, (**T4**), the referential reading of the idiom is present – in our case, the lack of visibility.

6 Analysis 3: Plain high degree readings

After this general discussion of NPIs and result clauses, we can turn to our third type of E-NPIs. Our analysis of this type will be analogous to that of *mixed expressives* such as slurs in Gutzmann (2011) and Gutzmann & McCready (2016), i.e., we will make use, again, of the difference between at issue content and CIs. Gutzmann & McCready's analysis is sketched in (38). The word *kraut* has as its at issue content the information that someone is German. However, the word triggers a CI that the speaker has a negative attitude towards Germans.

(38) Dan is a Kraut.
at issue: Dan is German.
CI: I have a negative attitude towards Germans.

We can adapt this theory to data on fixed RCls: such RCls, like *de mori* 'that one dies' – see (40) below – contribute an intensification as their at issue content, i.e., they basically mean the same as the particle *foarte* 'very'. At the same time, they trigger a CI that is based on the expression's literal meaning.

Let us look at the at issue semantics first. In (39), we add the intensification particle *foarte* 'very' to the Romanian version of example (25). We provide the EX-CONT value of the sentence, underlining its at issue content.

(39) Camera este foarte întunecată.
room.the is very dark 'The room is very dark.'
$\exists A(A = \{d' | \Diamond \textbf{dark}(d', \textbf{the-room})\}$
$\wedge \underline{\textbf{ResOp}\, d\, (\textbf{dark}(d, \textbf{the-room}) : \textbf{ScAs}(\textbf{dark}(d, \textbf{the-room}), d, A))})$

The particle *foarte* 'very' triggers a presupposed set of contextually relevant alternatives around some standard. The degree particle, then, adds a semantics that expresses exactly what was inferred for the other two types of E-NPIs above (see (33)), i.e., that the extent d to which the room is dark is at least as high as the minimal extent of darkness that is higher than all relevant alternatives.

We can apply this to fixed idiomatic phrases. We use the expression with a generic reading *de mori* 'that one dies' (lit.: that you.die) in (40):

(40) E [RCX: frumoasă [RCl: de mori]].
She.is beautiful that you.die 'She is very beautiful.'

In addition to an intensification at issue content, there is a CI component, parallel to mixed expressives such as in (38). In our case, however, the CI states

that whenever some predicate's extent results in someone dying, this extent must be very high. We sketch the lexical entry of idiomatic *mori* in (41).

(41) Lexical entry of *mori* 'you.die' as used in *de mori*:

$$\left[\begin{array}{ll}
\textsc{phon} & \langle \textit{mori} \rangle \\
\textsc{syns} & \left[\textsc{cont}\left[\textsc{main}\ \textbf{die}\right]\right] \\
\textsc{lrs} & \left[\begin{array}{ll}
\textsc{ai} & \boxed{1}\ \textbf{ScAs}(\alpha[\alpha^*], d, A) \\
\textsc{pres} & \left\langle \exists A(A = \{d' | \Diamond[\lambda d.\alpha](d')\} \wedge \gamma[\boxed{1}]) \right\rangle \\
\textsc{ci} & \left\langle \delta \wedge \forall P\, \exists A(\alpha \approx P(x) \rightarrow (\textbf{ResOp}\, d\, (P(x) : \textbf{die}(x)) \rightarrow \textbf{ScAs}(P(x), d, A)))) \right\rangle
\end{array}\right] \\
\textsc{coll} & \left[\textsc{lic}\left\langle\left[\begin{array}{ll}
\textsc{lid} & \textit{result-de} \\
\textsc{head} & \left[\textsc{sel}|\textsc{cont}\left[\begin{array}{ll}\textsc{index} & d \\ \textsc{main} & \alpha^*\end{array}\right]\right]
\end{array}\right]\right\rangle\right]
\end{array}\right]$$

The AT-ISSUE only consists of an emphatic expression (a **ScAs** expression). The word is collocationally restricted to occur in an RCX, i.e., it must be dominated by a phrase that is headed by the *de*-complementizer and modifies some element with a basic meaning α^*, which is exactly the content that is used in the **ScAs** expression. The set of contextually relevant alternatives is presupposed.

The CI value says: for any predicate P such that $P(x)$ is similar to the matrix proposition α, if $P(x)$ results in dying, then $P(x)$ is an emphatic statement. This shows that the CI allows us to integrate the literal meaning of the RCl without committing to the factivity of the result clause, i.e., in (40), the speaker does not factually die from another person's beauty.

We can apply this analysis to E-NPI3s, i.e., to E-NPIs with a purely intensifier meaning, see (42). Our analysis is just like for *de mori* above. The NPI-licensing requirement is satisfied in the representation of the referential reading of the RCl, i.e., the lack of visibility. This reading, however, is not asserted but occurs inside the CI-value, encoding a speaker's knowledge that this RCX can be used for high degree statements for the matrix predicate.

(42) Mi-e foame de nu te văd.
(lit.: I am hungry that I cannot see you.) 'I am extremely hungry.'

(43) Sketch of the lexical entry of *văd*:

$$\left[\begin{array}{ll}
\textsc{phon} & \langle \textit{văd} \rangle \\
\textsc{syns} & \left[\textsc{cont}\left[\textsc{main}\ \textbf{see}\right]\right] \\
\textsc{lrs} & \left[\begin{array}{ll}
\textsc{ai} & \boxed{1}\ \textbf{ScAs}(\alpha[\alpha^*], d, A) \\
\textsc{pres} & \left\langle \exists A(A = \{d' | \Diamond[\lambda d.\alpha](d')\} \wedge \gamma[\boxed{1}]) \right\rangle \\
\textsc{ci} & \left\langle \begin{array}{l} \delta \wedge \forall P\, \exists A(\alpha \approx P(x) \rightarrow \\ (\textbf{ResOp}\, d\, (P(d,x) : \textbf{ScAs}(\beta[\textbf{see}, \textbf{min-range}, A']) \rightarrow \textbf{ScAs}(P(d,x), d, A))))) \end{array} \right\rangle
\end{array}\right] \\
\textsc{coll} & \left[\textsc{lic}\left\langle\left[\begin{array}{ll}
\textsc{lid} & \textit{result-de} \\
\textsc{head} & \left[\textsc{sel}|\textsc{cont}\left[\begin{array}{ll}\textsc{index} & d \\ \textsc{main} & \alpha^*\end{array}\right]\right]
\end{array}\right], \ldots\right\rangle\right]
\end{array}\right]$$

We can check that this analysis captures the expression's behavior with respect to our tests. The collocational requirement of the E-NPI blocks it from occurring outside a *de*-marked RCX (**T1**). Consequently, (**T2**) is not applicable. The use of *încât* is excluded by the COLL value as well (**T3**). Finally, (**T4**) says that the referential reading of the NPI is not asserted. This is clearly the case as the referential reading is integrated into a CI.

The analysis of E-NPI3s combines our treatment of NPIs and RCXs with an analysis of mixed expressives. We use the NPI-licensing mechanism from Section 4 on the referential reading of the RCl. However, the referential reading does not contribute to the at issue content, which is just a plain intensification.

7 Conclusion

This paper looked at the distribution of NPIs in Romanian RCXs. We identified three main types of NPIs. We introduced some aspects of LRS and provided a representational re-encoding of a scalar theory of NPI licensing. We also adapted the semantic analysis of RCXs from Meier (2003) and added the refinements necessary for Romanian. For E-NPI1s, it was enough to provide a scalar NPI analysis. From this, it followed immediately that these NPIs can be used in high-degree RCXs, as they contribute the **ScAs** operator, which is required for high degree readings. The only difference for E-NPI2s is that they need to specify a collocational requirement to ensure that they can only be used in the scope of an **ResOp** operator.

For E-NPI3s, this collocation requiremement is not about a semantic operator, but about a particular lexical item, the complementizer *de*. In addition, these expressions are mixed expressives in the sense that they make a non-trivial meaning contribution both to the at issue content and to the CI content.

References

Bach, Kent. 1999. The myth of conventional implicature. *Linguistics and Philosophy* 22(4). 327–366.

Bos, Johan. 1996. Predicate logic unplugged. In Paul Dekker & Martin Stokhof (eds.), *Proceedings of the 10th Amsterdam Colloquium*, 133–143. Amsterdam: ILLC/Department of Philosophy, University of Amsterdam.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In Adriana Belletti (ed.), *Structure and beyond: The cartography of syntactic structures*, vol. 3, 39–103. Oxford: Oxford University Press.

Chierchia, Gennaro. 2006. Broaden your views: Implicatures of domain widening and the "logicality" of language. *Linguistic Inquiry* 37(4). 535–590.

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

Eckardt, Regine. 2005. Too poor to mention: Subminimal events and negative polarity items. In Claudia Maienborn & Angelika Wöllstein (eds.), *Event arguments: Foundations and applications*, 301–330. Tübingen: Niemeyer.

Eckardt, Regine & Eva Csipak. 2013. Minimizers: Towards pragmatic licensing. In Eva Csipak, Mingya Liu, Regine Eckardt & Manfred Sailer (eds.), *Beyond "any" and "ever". New explorations in negative polarity sensitivity*, 267–298. Berlin: De Gruyter.

Egg, Markus, Alexander Koller & Joachim Niehren. 2001. The constraint language for lambda structures. *Journal of Logic, Language and Information* 10(4). 457–485.

Gutzmann, Daniel. 2011. Expressive modifiers & mixed expressives. *Empirical Issues in Syntax and Semantics* 8. 123–141. http://www.cssp.cnrs.fr/eiss8/gutzmann-eiss8.pdf.

Gutzmann, Daniel & Eric McCready. 2016. Quantification with pejoratives. In Rita Finkbeiner, Jörg Meibauer & Heike Wiese (eds.), *Pejoration* (Linguistics Today 2016), 75–102. Amsterdam and Philadelphia: Benjamins.

Hasegawa, Akio & Jean-Pierre Koenig. 2011. Focus particles, secondary meanings, and Lexical Resource Semantics: The case of Japanese *shika*. In Stefan Müller (ed.), *Proceedings of the 18th International Conference on HPSG, University of Washington*, 81–101. Stanford, CA: CSLI Publications. http://cslipublications.stanford.edu/HPSG/2011/hasegawa-koenig.pdf.

Hoeksema, Jack & Donna Jo Napoli. 2019. Degree resultatives as second-order constructions. *Journal of Germanic Linguistics* 31(3). 225–297.

Iordăchioaia, Gianina & Frank Richter. 2015. Negative concord with polyadic quantifiers. *Natural Language and Linguistic Theory* 33. 607–658. doi:10.1007/s11049-014-9261-9.

Karttunen, Lauri & Stanley Peters. 1979. Conventional implicature. In C. Oh & D. Dinneen (eds.), *Presupposition*, vol. 11 Syntax and Semantics, 1–56. New York: Academic Press.

Krifka, Manfred. 1995. The semantics and pragmatics of weak and strong polarity items. *Linguistic Analysis* 25(3–4). 209–257.

Meier, Cécile. 2003. The meaning of *too*, *enough*, and *so ... that*. *Natural Language Semantics* 11(1). 69–107.

Park, Sang-Hee, Jean-Pierre Koenig & Rui P. Chaves. 2018. A semantic underspecification-based analysis of scope ambiguitites in gapping. Paper presented at SuB 2018, Barcelona.

Penn, Gerald & Frank Richter. 2005. The other syntax: Approaching natural language semantics through logical form composition. In Henning Christiansen, Peter Rossen Skadhauge & Jørgen Villadsen (eds.), *Constraint solving and language processing* (Lecture Notes in Computer Science 3438), 48–73. Springer.

Potts, Christopher. 2005. *The logic of conventional implicatures* Oxford Studies in Theoretical Linguistics. Oxford: Oxford University Press.

Richter, Frank & Manfred Sailer. 2004. Basic concepts of Lexical Resource Semantics. In Arne Beckmann & Norbert Preining (eds.), *ESSLLI 2004: Course material I* (Collegium Logicum 5), 87–143. Vienna: Kurt Gödel Society.

Richter, Frank & Jan-Philipp Soehn. 2006. *Braucht niemanden zu scheren*: A survey of NPI licensing in German. In Stefan Müller (ed.), *Proceedings of the 13th International Conference on HPSG, Varna*, 421–440. http://cslipublications.stanford.edu/HPSG/2006/richter-soehn.pdf.

Sailer, Manfred. 2007. NPI licensing, intervention and discourse representation structures in HPSG. In Stefan Müller (ed.), *Proceedings of the 14th International Conference on HPSG, Stanford, 2007*, 214–234. Stanford, CA: CSLI Publications. http://cslipublications.stanford.edu/HPSG/14/sailer.pdf.

Sailer, Manfred & Assif Am-David. 2016. Definite meaning and definite marking. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King & Stefan Müller (eds.), *Proceedings of the Joint 2016 Conference on HPSG and LFG, Polish Academy of Sciences, Warsaw, Poland*, 641–661. Stanford, CA: CSLI Publications. http://cslipublications.stanford.edu/HPSG/2016/headlex2016-sailer-am-david.pdf.

Soehn, Jan-Philipp. 2009. Lexical licensing in formal grammar. Universität Tübingen. http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-42035.

Tonhauser, Judith, David Beaver, Craige Roberts & Mandy Simons. 2013. Toward a taxonomy of projective content. *Language* 89(1). 66–109.

Syntactic haplology and the Dutch proform "er"

Gert Webelhuth
University of Frankfurt

Olivier Bonami
Universite de Paris, LLF, CNRS

Proceedings of the 26th International Conference on
Head-Driven Phrase Structure Grammar

University of Bucharest

Stefan Müller, Petya Osenova (Editors)

2019

CSLI Publications

pages 100–119

http://csli-publications.stanford.edu/HPSG/2019

Keywords: HPSG, Dutch, Haplology, ARG-ST, COMPS

Webelhuth, Gert, & Bonami, Olivier. 2019. Syntactic haplology and the Dutch proform "er". In Müller, Stefan, & Osenova, Petya (Eds.), *Proceedings of the 26th International Conference on Head-Driven Phrase Structure Grammar, University of Bucharest*, 100–119. Stanford, CA: CSLI Publications.

Abstract

Dutch has four pronouns 'er' which show an intriguing pattern of syntactic haplology when a finite verb has more than one 'er' dependent. We present a theory that captures this pattern by relying on two central aspects of HPSG: (i) the distinction between ARG-ST and COMPS and (ii) the distinction between canonical and non-canonical synsem objects. No deletion rules of the kind used in transformational analyses of 'er' are necessary.

1 Introduction

Dutch has four expressions spelled 'er' with different syntactic and semantic functions and syntactic distributions that display unusual and intriguing interdependencies. We give an overview of the major data and show that it can be captured through the interaction of a small number of constraints on argument realization[1]

2 The Data

The sentences in (1) each contain a single example of each type of *er*.[2] The first example features existential *er*, which cooccurs with indefinite subjects and is the only *er* that can fill the first position of a Dutch main clause. In (1b), pronominal *er* expresses the obligatory complement of the preposition *op*. The example shows that pronominal *er* does not need to be adjacent to its selector. (1c) contains *er* in its function as a locational expression, comparable to the English referential locational adverb *there*. Finally, (1d) illustrates quantitative *er*: it serves as the complement of the numeral *drie* in this example and performs a function similar to the partitive elements *en* in French or *ne* in Italian.

(1) a. *Er*$_X$ loopt een man op straat.
there walks a man in the.street

b. Jan wacht *er*$_P$ al tijden **op**.
Jan waits there for ages for

c. Jan staat *er*$_L$ al.
Jan stands there already

d. Jan heeft *er*$_Q$ [$_{\text{NP}}$ **drie** [*e*]]
Jan has there three

[1] We are greatly indebted to Hans Broekhuis for patiently providing his expertise about the subject matter of this article and for making us see the data and the theoretical issues involved more clearly. Without his help, this article and the talk it is based on would probably not exist! We would also like to thank Gosse Bouma, Fenna Bergsma, Ruby Sleeman, Manfred Sailer, Frank Richter, Frank Van Eynde, and three anonymous reviewers for their help and suggestions. Any errors in this paper are our responsibility alone.

[2] We use the following system to label the four *er*s: *er*$_X$ = existential, *er*$_P$ = pronominal, *er*$_Q$ = quantitative, *er*$_L$ = locational.

There are many previous analyses of *er* in the literature. Space limitations make it impossible to do anything other than listing the most important ones here: Bech (1952), van Riemsdijk (1978), Bennis (1986), Odijk (1993), and Broekhuis (2013).

As non-native speakers of Dutch we are faced with the problem that on *er* "conflicting judgments can be found in the literature" (Broekhuis (2013, p. 338)). We chose to handle this problem by citing the data and judgments of just a single author. With the exception of (12), which was supplied to us in a personal communication by Hans Broekhuis, all examples are drawn from Broekhuis (2013), which is an extremely comprehensive and detailed treatment of *er*.

We decided to develop a new analysis of *er* in HPSG, as we felt that we can improve on existing HPSG analyses. Bouma (2001), Van Eynde & Augustinus (2014), and Van Eynde (2019) all do not cover quantitative *er*, which behaves differently from the three remaining *er*s, as will be demonstrated below. Campbell-Kibler (2001) was meant to account for different judgments than those considered here. Moreover, since the author draws examples from different works in the literature, it is unclear that this data reflects a consistent set of judgments.

2.1 Linear structure of finite clauses in Dutch

We assume that Dutch sentences can be analyzed as consisting of a number of topological (= linear) fields, as follows:[3]

Subordinate clause:		C	Middle field	Verb(s)
Main clause:	Prefield	V_{finite}	Middle field	(Verb(s))

The prefield is limited to a single constituent whereas the middle field can contain zero, one, or several constituents. The next two sections will describe the distribution of *er* in these two fields.

2.2 Clauses without existential *er* in the prefield

As the linear field schema above illustrates, both main and subordinate clauses contain a middle field. The present section deals with *er* in subordinate clauses and in those main clauses whose prefield is *not* filled by er_X, i.e. with main clauses like (1b)-(1d). The middle field in these kinds of sentences satisfies the simple generalization that it can contain at most one overt *er*.[4]

(2) illustrates that an overt existential *er* cannot cooccur with any of the three other other *er*s:

[3] On topological fields, see Drach (1937), Reis (1980), Höhle (1983), Höhle (1986). Note that our analysis in this article is restricted to the occurrences of *er* in finite sentences.

[4] Neeleman & van de Koot (2006) accept certain sentences with two *er*s in the middle field as long as the *er*s are not adjacent. Hans Broekhuis and the native speakers we were able to consult consider these examples ungrammatical (personal communication). The theory developed below is only meant to cover Broekhuis' judgments.

(2) a. * dat er_X er_L gedanst wordt.
that there there danced is
Intended reading: 'People are dancing there.'

b. * dat er_X er_P over gesproken wordt.
that there there about spoken is
'Intended reading: 'People are talking about it."

c. * dat er_X er_Q [NP twee *e*]] gestolen zijn.
that there there two stolen have.been
'Intended reading: 'Two [e.g., computers] have been stolen."

All three sentences become grammatical if one of the two *er*s is dropped.

(3a)-(3b) show that overt pronominal *er* cannot cooccur with an overt locational or quantitative *er* and (4) provides evidence that the remaining potential combination of overt *er*s is impossible as well:

(3) a. * dat Jan er_P er_L over praatte.
that Jan there there about talked
'that Jan talked about it there.'

b. * dat Jan er_P er_Q drie in stopte.
that Jan there there three into put
'that Jan put three [e.g., cigars] in it.'

(4) * dat Jan er_Q er_L [NP twee [*e*]] gezien heeft.
that Jan there there two seen has
'that Jan saw two [e.g., rats] there.'

Again, these sentences become grammatical, if only a single overt *er* appears.

2.3 Sentences with an overt and an implicit *er*

The data presented in the previous subsection jointly illustrate the generalization that in the ideolect studied here the middle field of sentences without an expletive *er* in the prefield can contain only a single overt *er*. Interestingly, however, when one overt *er* appears, one or more additional *er*s can be understood. The sentences in (5) demonstrate this. The existential subordinate clause (5a) contains an overt er_X, an indefinite subject NP, and an object PP.

(5) a. dat er_X gisteren [NP drie potloden] [PP op tafel] lagen.
that there yesterday three pencils on the.table lay
'that there were three pencils lying on the table yesterday.'

b. dat er_{XP} gisteren [NP drie potloden] [PP op] lagen.
that there yesterday three pencils on lay
'that there were three pencils lying on it yesterday.'

c. dat er_{XQ} gisteren [NP drie] [PP op tafel] lagen.
that there yesterday three on the.table lay
'that there were three lying on the table yesterday.'

d. dat er_{XL} veel mensen wonen.
that there many people live
'that many people live there.'

In (5b), the object of the preposition *op* gets a deictic interpretation 'there', even though the object is unexpressed. If the sentence did not contain an expletive *er*, then the object of the preposition would need to be expressed as pronominal *er*. In (5c) the quantitative *er* of the partitive NP *drie* remains implicit. (5d), finally, illustrates the case where the adverbial complement of the verb *wonen* with the sense of 'reside' can remain unexpressed in the presence of an overt expletive *er* in the middle field.

Intriguingly, but in light of the examples just provided perhaps no longer surprising, it is also possible for a single pronominal *er* to represent the objects of two separate prepositions in a sentence. This is shown in (6). The first sentence contains two PPs with non-pronominal NPs. The second and third examples show that pronomional *er* can serve as the object of each preposition:[5]

(6) a. Jan heeft de sleutel [**met een tang**] [**uit het slot**] gehaald
Jan has the key with a pair.of.tongs out.of the lock taken
'Jan took the key out of the lock with pliers.'

b. Jan heeft er_P de sleutel [**mee**] [uit het slot] gehaald.

c. Jan heeft er_P de sleutel [met een tang] [**uit**] gehaald.

d. Jan heeft er_{PP} de sleutel [**mee**] [**uit**]] gehaald.

e. * Jan heeft er_P er_P de sleutel [**mee**] [**uit**] gehaald.

(6d)-(6e) demonstrate what happens when both prepositions are stranded at the same time: the objects of the prepositions must be represented by a single *er*, as two *er*s in the Dutch middle field are forbidden.

The same pattern occurs with quantitative *er*. The second conjunct of the following example contains two partitive NPs but only a single quantitative *er*:[6]

(7) Iedere student heeft een onvoldoende gekregen …
every student has an unsatisfactory mark gotten
'Every student got an unsatisfactory mark …'

a. … en [NP **drie** *e*] hebben er_Q zelfs [NP **twee** *e*].
and three have there even two
'… and three even got two.'

[5] *er* occurs in a position for clitics in these examples, thus stranding the prepositions. Also note that when the preposition *met* is stranded, it takes on the allomorphic form *mee*.

[6] Observe that the partitive subject precedes the quantitative clitic *er* in this example.

The final examples of this section show that a single overt *er* can represent four different functions in a single sentence. The initial example contains expletive *er*, as the sentence is existential:

(8) a. dat er$_X$ [**twee studenten**] [**drie boeken**] [**uit de boekkast**] gehaald
that there two students three books out.of the bookcase fetched
hebben.
have

b. dat er$_{XQQ}$ [$_{\text{NP}}$ **twee** e] [$_{\text{NP}}$ **drie** e] **uit de boekkast** gehaald hebben.

c. dat er$_{XQQP}$ [$_{\text{NP}}$ **twee** e] [$_{\text{NP}}$ **drie** e] **uit** gehaald hebben.

In (8b), the single *er* in addition represents the quantitative *er*s of the two partitive noun phrases *twee* and *drie*. Finally, in (8c), the object of the preposition *uit* is interpreted as pronominal *er*, leading in sum to the single overt *er* carrying four different functions within that sentence.

In sum, the examples in this section support the following two descriptive generalizations about sentences without existential *er* in the prefield:

1. Only one overt *er* can occur in the middle field.
2. When one overt *er* is present in the middle field, additional *er*s may be understood.

2.4 Clauses with existential *er* in the prefield

We now turn to sentences like (1a), repeated for convenience below, which contain an existential *er* in the prefield:

(9) *Er*$_X$ loopt een man op straat.
there walks a man in the.street

These structures need to be discussed separately because unlike the clauses without *er*$_X$ in the prefield, they permit more than a single overt *er* in a single clause under some circumstances. All of these clauses are verb-second main clauses and existential *er* is the only *er* permitted to fill the prefield. Moreover, *er*$_X$ in the prefield can co-occur with all other *er*s in the middle field, however the latter differ in whether they are allowed to be overt or not.

The behavior of locational and pronominal *er* is simple: both have to remain unexpressed when expletive *er* fills the prefield, as the examples below demonstrate:

(10) a. *Er*$_X$ wordt (**er*$_L$) morgen gedanst.
there is there tomorrow danced

b. *Er*$_X$ wordt (**er*$_P$) morgen *over* gesproken.
there is there tomorrow about spoken

The examples become grammatical if the *er* in the middle field does not appear.

Quantitative *er* shows the opposite behavior: it cannot remain implicit but must be spelled out separately from the initial existential *er* in the middle field:

(11) *Er*$_X$ zijn *er*$_Q$ gisteren [$_{NP}$ **twee** [*e*]] gestolen.
there have.been there yesterday two stolen

In sentences with two quantitative NPs in the middle field, only one quantitative *er* can be spelled out, however (Hans Broekhuis, p.c.):

(12) [Er hebben veel studenten een onvoldoende gekregen] en
there(E) have many students an unsatisfactory_mark gotten and
er hebben **er**$_{QQ}$ [**een paar** *e*] zelfs [**twee** *e*] gekregen.
there(E) have there(QQ) a couple even two gotten

We sum up the generalizations for sentences with existential *er* in the prefield:

1. Only existential *er* can occur in the prefield, the other ones cannot.
2. When existential *er* occupies the prefield, then
   - an additional single overt quantitative *er* can appear in the middle field
   - implicit locational and pronominal *er*s are possible.

3 The Analysis

As we saw above, both main and subordinate clauses in Dutch show the phenomenon that one or more *er*s can remain implicit when at least one *er* is expressed overtly. In order to capture this in a grammatical theory, a mechanism is needed that makes it possible for an overt *er* to influence whether additional *er*s can or must be expressed. Moreover, this mechanism must be sensitive to the location of the overt *er* in phrase and/or linear structure.

The guiding ideas of our analysis are the following. We assume that existential and locational *er* are arguments of finite verbs, perhaps directly or through argument extension. Moreover, finite verbs attract to their ARG-ST the quantitative and pronominal *er*-complements of their NP and PP arguments that have not been realized within these phrases. Thus, all the *er*s that in principle can be realized at the sentence level appear in one place, namely the ARG-ST of the finite verb heading the sentence. The haplological effect then arises through the interaction of a number of constraints on the ARG-ST and COMPS lists of finite verbs.

In the remainder of the article, we will make these guiding ideas more precise and apply the resulting theory to representative examples.

3.1 Assumptions about Dutch phrase structure

We assume that the Dutch phrase structure system creates a number of linear fields and that in every sentence where they are realized, the fields occur in the left-to-right order that corresponds to their top-to-bottom ordering in Table 1:

Field	Description
pre-fld	the initial position in main clauses
lb	the left sentence bracket, filled by either a finite verb or a complementizer
mid-fld	the middle field contains the elements inbetween the two sentence brackets
rb	the right sentence bracket is made up of one or more verbs
fin-fld	the final field follows the right sentence bracket

Table 1: Description of Dutch topological fields

Except for the pre-field, which is restricted to main clauses, every field can occur in both main and subordinate clauses. We postulate an attribute FLD appropriate for objects of type *synsem*. It is crucial to our account that phrase structure configurations as well as the lexicon may constrain the FLD value of signs.

The trees in Figure 1 sketch the phrase and linear structure of verb-second and subordinate clauses we assume. The units that are connected to their mothers by

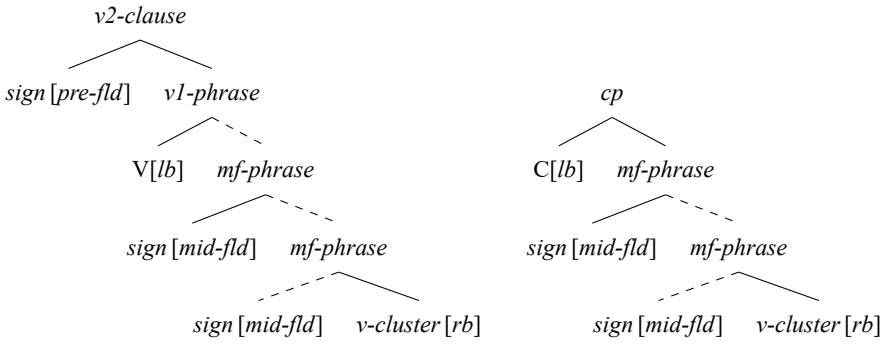


Figure 1: Basic phrase structure of verb-second and subordinate clauses

dashed lines are optional. The phenomenon we are dealing with in this paper reflects the appearance of expletive *er* in the pre-field of main clauses and of one or more *er*s in the middle field of both main and subordinate clauses.

3.2 Assumptions about *er*

In order to capture that the four *er*s on the one hand share properties and yet have different meanings and distributions, we postulate a general lexical identifier *er-lid* that all four *er*s share and a specific subtype for each different *er*: *er-X*, *er-Q*, *er-P*, and *er-L*, as shown in Figure 2.

Using these LID values, we can impose field constraints on the four *er*s lexically. The partial lexical entries in figure 3 permit existential *er* to occur in the pre-field and the middle field (see Broekhuis (2013, p. 337, 338) for this constraint) whereas the three remaining *er*s are restricted to the value *mid-fld* for the FLD attribute.

Second, quantitative *er* must be prevented from being realized within its partitive NP, as it always occurs outside of that NP when it is realized overtly. The

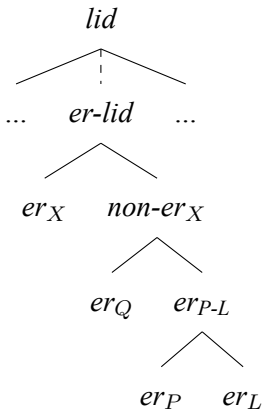


Figure 2: Partial hierarchy of lexeme identifier types

(13)

$$\begin{bmatrix} word \\ \text{SYNSEM} \begin{bmatrix} \text{HEAD} & er_X \\ \text{FLD} & pre\text{-}fld \vee mid\text{-}fld \end{bmatrix} \end{bmatrix} \qquad \begin{bmatrix} word \\ \text{SYNSEM} \begin{bmatrix} \text{HEAD} & non\text{-}er_X \\ \text{FLD} & mid\text{-}fld \end{bmatrix} \end{bmatrix}$$

Existential *er* **Non-existential *er*s**

Figure 3: Lexical entries for *er*

following constraint has the desired consequence by ruling out noun phrases with er_Q as a non-head daughter:

(14) $$\begin{bmatrix} hd\text{-}comp\text{-}ph \\ \text{HEAD} \ \text{noun} \end{bmatrix} \longrightarrow \begin{bmatrix} \text{NON-HD-DTR} \begin{bmatrix} \text{LID} \ \neg ER_Q \end{bmatrix} \end{bmatrix}$$

Like other units, *er*s can be canonical and noncanonical synsems. It will be important for our analysis that like clitics in languages such as French (Miller & Sag (1997)), *er*s in Dutch have the option of the *synsem* value *pro-synsem*, as shown in Figure 4. This causes them to remain unrealized in phrase structure.

Next, we state argument realization constraints for finite verbs and nouns. Finite verbs map all and only their canonical arguments to their COMPS list. Note that it follows from this constraint that *pro-synsem er* arguments of finite verbs cannot appear on the verbs' COMPS list:

(15) $$\begin{bmatrix} \text{HEAD} \ \text{V}[fin] \end{bmatrix} \longrightarrow \begin{bmatrix} \text{SUBJ} & \langle\rangle \\ \text{COMPS} & \boxed{1} \ list(canon\text{-}ss) \\ \text{ARG-ST} & \boxed{1} \bigcirc list(noncanon\text{-}ss) \end{bmatrix}$$

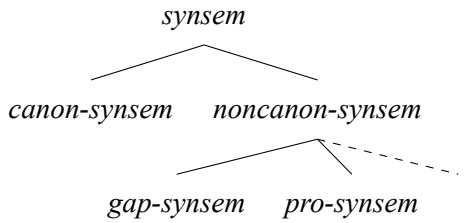


Figure 4: Partial hierarchy of *synsem* types

Nouns and prepositions differ from (finite) verbs in one crucial respect. Recall from the introduction to this section that we are assuming that finite verbs attract to their ARG-ST the quantitative and pronominal *er*-complements of their NP and PP arguments that have not been realized within these phrases. Whether or not such a raised *er* is expressed at the sentence level is a function of constraints on the ARG-S and the COMPS lists of the finite verb heading the sentence. As that determination is made only after the argument raising of quantitative and pronominal *er*-complements, such raising must be possible, no matter whether the *er*s' *synsem* type is canonical or non-canonical. Therefore, nouns and prepositions map all their *er*-arguments to their COMPS list, independent of the *er*'s canonicality. Below we present the constraint on nouns. The *er* (which is optional, since not every use of a noun is partitive) carries the tag $\boxed{2}$:

(16) $$\left[\text{HEAD } \text{N}\right] \longrightarrow \left[\begin{array}{ll} \text{SUBJ} & \langle\rangle \\ \text{COMPS} & \boxed{1}\ \textit{list(canon-ss)} \bigcirc \boxed{2} \\ \text{ARG-ST} & \boxed{1}\textit{list}\left(\neg\left[\text{LID } \textit{er}\right]\right) \bigcirc \boxed{2}\left\langle(\textit{er}_Q)\right\rangle \bigcirc \textit{list(noncanon-ss)} \end{array}\right]$$

The constraint on prepositions is analogous.

3.3 Constraints on the Argument Structures and COMPS Lists of Finite Verbs

With these preliminaries out of the way, we are now in a position to state the constraints that will interact to create the haplology effects illustrated in section 2. All constraints regulate the occurrence or co-occurrence of *er*s on the argument structure or COMPS lists of finite verbs.

3.3.1 *er*-Expression Constraints

The first constaint simply states that at least one *er*-argument of a finite verb must appear on the verb's COMPS list and be overtly expressed:

(17) *er*-EXPRESSION CONSTRAINT

$$\begin{bmatrix} \text{HEAD} & \text{V}[\textit{fin}] \\ \text{ARG-ST} & \left\langle \begin{bmatrix} \text{LID} & \textit{er} \end{bmatrix} \right\rangle \bigcirc \textit{list} \end{bmatrix} \longrightarrow \begin{bmatrix} \text{COMPS} & \left\langle \begin{bmatrix} \text{LID} & \textit{er} \end{bmatrix} \right\rangle \bigcirc \textit{list} \end{bmatrix}$$

Given that the system will permit *er* arguments to remain implicit, the constraint above is epistemologically plausible, as it requires at least one of the *er* arguments of a verb to be expressed. The expression of this *er* can thus serve as a signal to the possibility of implict *er*s.

The next two constraints contribute to the opposing behavior of quantitative *er* on the one hand and pronominal and locational *er* on the other in main clauses like (10)-(12), whose prefield is filled by existential *er*. (18) requires that verbs with a quantitative *er* argument must realize an *er* complement in the middle field:

(18) MIDDLE FIELD *er*-EXPRESSION CONSTRAINT

$$\begin{bmatrix} \text{HEAD} & \text{V}[\textit{fin}] \\ \text{ARG-ST} & \left\langle \begin{bmatrix} \text{LID} & \textit{er}_{Q} \end{bmatrix} \right\rangle \bigcirc \textit{list} \end{bmatrix} \longrightarrow \begin{bmatrix} \text{COMPS} & \left\langle \begin{bmatrix} \text{LID} & \textit{er} \\ \text{FLD} & \textit{mid-fld} \end{bmatrix} \right\rangle \bigcirc \textit{list} \end{bmatrix}$$

The next constraint applies to verbs which have a canonical pronominal or locational *er* argument. The COMPS list of these verbs is well formed only if it does not contain an expletive *er* with field value *pre-fld*.

(19) P-L *er*-EXPRESSION CONSTRAINT

$$\begin{bmatrix} \text{HEAD} & \text{V}[\textit{fin}] \\ \text{ARG-ST} & \left\langle \begin{bmatrix} \textit{canon-synsem} \\ \text{LID} & \textit{er}_{P\text{-}L} \end{bmatrix} \right\rangle \bigcirc \textit{list} \end{bmatrix} \longrightarrow \begin{bmatrix} \text{COMPS} & \textit{list}\left(\neg \begin{bmatrix} \text{LID} & \textit{er} \\ \text{FLD} & \textit{pre-fld} \end{bmatrix} \right) \end{bmatrix}$$

Finally, we state the constraint that creates the syntactic haplology effect of *er*. (20) says that a finite verb selects at most one *er*-complement in the middle field:

(20) MIDFIELD SINGLE-*er* CONSTRAINT

$$\begin{bmatrix} \text{HEAD} & \text{V}[\textit{fin}] \\ \text{COMPS} & \left\langle \begin{bmatrix} \text{LID} & \textit{er} \\ \text{FLD} & \textit{mid-fld} \end{bmatrix} \right\rangle \bigcirc \textit{list} \end{bmatrix} \longrightarrow \begin{bmatrix} \text{COMPS} & \left\langle \begin{bmatrix} \text{LID} & \textit{er} \\ \text{FLD} & \textit{mid-fld} \end{bmatrix} \right\rangle \bigcirc \textit{list}\left(\neg \begin{bmatrix} \text{LID} & \textit{er} \\ \text{FLD} & \textit{mid-fld} \end{bmatrix} \right) \end{bmatrix}$$

4 Illustration of the major cases

We now illustrate the interplay of the lexical, phrasal, and linear constraints that we have formulated in the previous section. We discuss five cases in detail.

4.1 Case 1: two overt *er*s in the middle field are ruled out

(21) * dat er_X er_L gedanst wordt.
that there there danced is
Intended reading: 'People are dancing there.'

According to our assumptions, the presence of two overt *er*s in the middle field of this sentence would require the verb *wordt* to have two *er* complements (in addition to its verbal complement), as shown in Figure 5. This structure is obviously not licensed by our approach, as *wordt* violates the MIDFIELD SINGLE-*er* CONSTRAINT (20), which permits verbs to have at most one *er* complement with field value *mid-fld*.

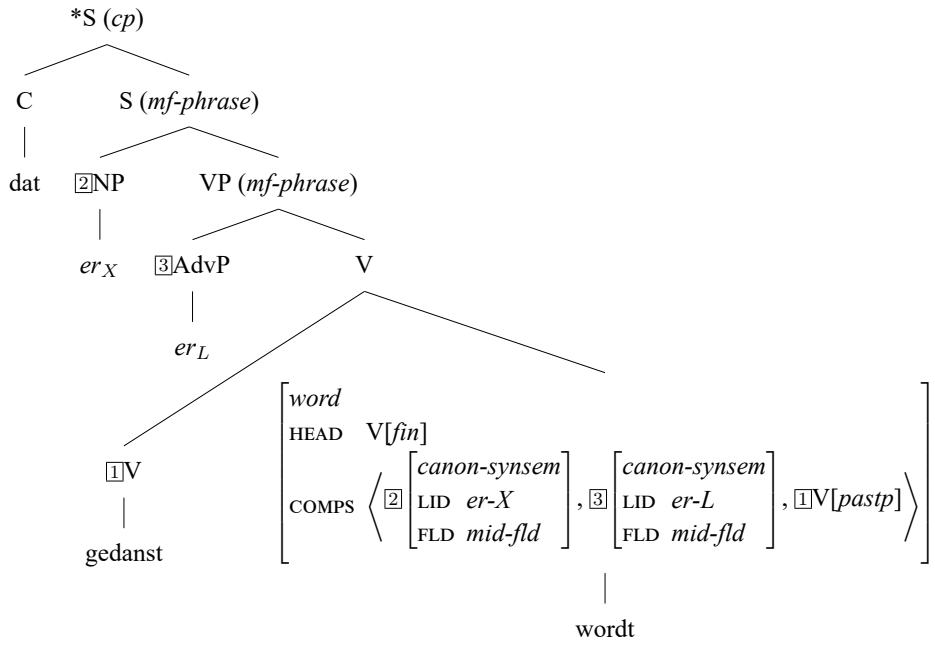


Figure 5: Analysis for example (21)

4.2 Case 2: one overt and one implicit *er* in the middle field

The next case differs from the previous one in that it contains a single overt *er* in the middle field and a second understood *er*, as the verb *wonen* selects a locational complement. The example is grammatical.

(22) dat er_{XL} veel mensen **wonen**.
that there many people live
'that many people live there.'

To license the structure above, in addition to its two overt complements *er* and *veel mensen*, the verb *wonen* must have an implicit er_L argument which is not mapped to the verb's COMPS list, as shown in Figure 6. Unlike the finite verb in Case

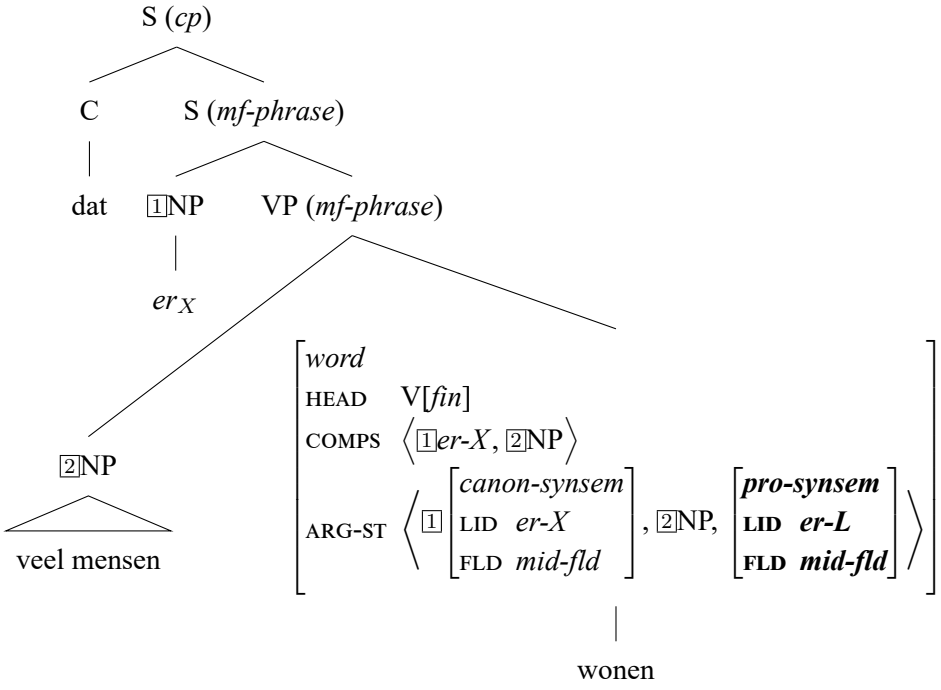


Figure 6: Partial analysis for example (22)

1, the word *wonen* in the tree immediately above satisfies all of our constraints:

1. *er*-EXPRESSION CONSTRAINT: *wonen* selects an *er* complement.
2. MIDDLE FIELD *er*-EXPRESSION CONSTRAINT: is vacuously satisfied, as *wonen* doesn't have an er_Q argument.
3. P-L *er*-EXPRESSION CONSTRAINT: is satisfied, as there is no *er* in the prefield.
4. MIDFIELD SINGLE-*er* CONSTRAINT: *wonen* has no more than a single mid-field *er* complement.

4.3 Case 3: an expletive *er* in the prefield and a pronominal *er* in the middle field

The next two cases deal with main clauses whose prefield is filled by expletive *er*. Recall that quantitative *er* parts ways with locational and pronominal *er* in this

sentence type. When present, the latter two have to remain implicit. This is why (23) with an overt pronominal *er* in the middle field is ungrammatical:

(23) * Er_X wordt ***er***$_P$ morgen *over* gesproken.
there is there tomorrow about spoken

For the string above to be licensed, it would need to have the structure shown in Figure 7.

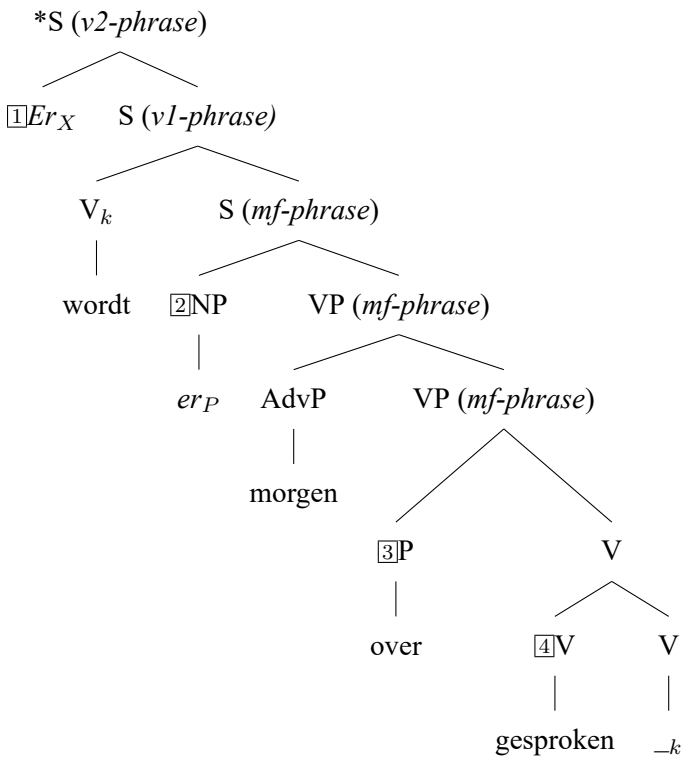


Figure 7: Partial analysis of example (23)

In this structure, the gap of the verb *wordt* would need to have four complements: (i) the expletive *er* occuring in the prefield, (ii) the overt pronominal *er* in the middle field which the verb has inherited from (iii) the preposition *over*, and (iv) the passive participle *gesproken*:

(24)
$$\begin{bmatrix} \textit{word} \\ \textsc{head} \quad \text{V}[\textit{fin}] \\ \textsc{comps} \quad \left\langle \boxed{1} \begin{bmatrix} \textbf{\textit{canon-synsem}} \\ \textbf{LID} \quad \textbf{\textit{er-X}} \\ \textbf{FLD} \quad \textbf{\textit{pre-fld}} \end{bmatrix}, \boxed{2} \begin{bmatrix} \textbf{\textit{canon-synsem}} \\ \textbf{LID} \quad \textbf{\textit{er-P}} \\ \textbf{FLD} \quad \textbf{\textit{mid-fld}} \end{bmatrix}, \boxed{3} \begin{bmatrix} \textit{canon-synsem} \\ \textsc{head} \quad \text{P} \\ \textsc{comps} \quad \langle \boxed{2} \rangle \end{bmatrix}, \boxed{4}\text{V}[\textit{pass}] \right\rangle \end{bmatrix}$$

According to our approach, the COMPS list above is illicit:

The P-L *er*-EXPRESSION CONSTRAINT is violated, as a pronominal *er* is incompatible with an *er* in the prefield. (10a) with an overt locational *er* in predicted to be ungrammatical for the same reason.

Since (23) violates only one of our constraints, it is correctly predicted that it becomes grammatical when the offending pronominal *er* remains implicit:

(25) Er_{XP} wordt morgen *over* gesproken.
there is tomorrow about spoken

The constraint profile of this structure is as follows:

1. *er*-EXPRESSION CONSTRAINT: the verb selects an *er* complement.
2. MIDDLE FIELD *er*-EXPRESSION CONSTRAINT: is vacuously satisfied, as the verb doesn't have an er_Q argument.
3. P-L *er*-EXPRESSION CONSTRAINT: is vacuously satisfied, as there is no canonical P-L *er* argument in the middle field.
4. MIDFIELD SINGLE-*er*: CONSTRAINT the verb has no more than a single midfield *er* complement.

4.4 Case 4: an expletive *er* in the prefield and a quantitative *er* in the middle field

Quantitative *er* differs from locational and pronominal *er* in that it must appear overtly in the middle field of sentences introduced by existential *er*. Without the second *er* in the middle field, (26) is ungrammatical.

(26) Er_X zijn *(er_Q) gisteren [$_{NP}$ **twee** [*e*]] gestolen.
there have.been there yesterday two stolen

Under our assumptions, the structure of this sentence is as shown in Figure 8. Expletive *er* appears in the pre-field and quantitative *er* in the middle field. The main verb *gestolen* has inherited the quantitative *er* from the noun *twee* and the auxiliary *zijn* has interited all the arguments of *gestolen*. Altogether, this requires the auxiliary (and its gap) to have two *er* and one N-complement (= *twee*), plus the passive participle of the main verb, as indicated in example (27).

(27)
$$\begin{bmatrix} \textit{word} \\ \text{HEAD} \quad \text{V}[\textit{fin}] \\ \text{COMPS} \quad \left\langle \boxed{1} \begin{bmatrix} \textbf{\textit{canon-synsem}} \\ \textbf{LID} \quad \textbf{\textit{er-X}} \\ \textbf{FLD} \quad \textbf{\textit{pre-fld}} \end{bmatrix}, \boxed{2} \begin{bmatrix} \textbf{\textit{canon-synsem}} \\ \textbf{LID} \quad \textbf{\textit{er-Q}} \\ \textbf{FLD} \quad \textbf{\textit{mid-fld}} \end{bmatrix}, \boxed{3} \begin{bmatrix} \textit{canon-synsem} \\ \text{HEAD} \quad \text{N} \\ \text{COMPS} \quad \langle \boxed{2} \rangle \end{bmatrix}, \boxed{4}\text{V}[\textit{pass}] \right\rangle \end{bmatrix}$$

With quantitative *er* expressed in the middle field, the sentence obeys all constraints:

1. *er*-EXPRESSION CONSTRAINT: the verb selects an *er* complement.

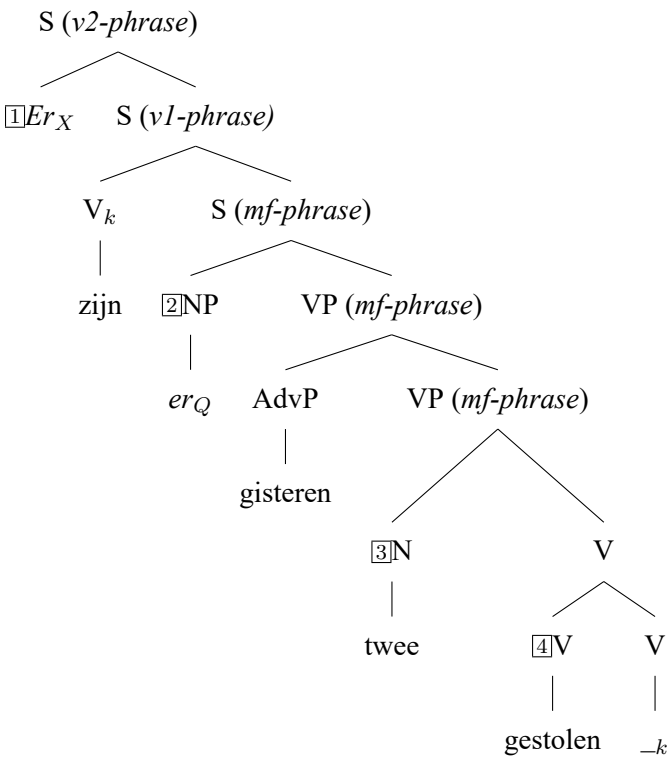


Figure 8: Partial analysis of example (26)

2. MIDDLE FIELD *er*-EXPRESSION CONSTRAINT: is satisfied, as there is an *er* complement in the middle field.
3. P-L *er*-EXPRESSION CONSTRAINT: is vacuously satisfied, as there is no P-L *er* argument in the middle field.
4. MIDFIELD SINGLE-*er*: CONSTRAINT the verb has no more than a single mid-field *er* complement.

Without the *er* in the middle field, the sentence becomes ungrammatical, as theMIDDLE FIELD ER-EXPRESSION CONSTRAINT is now violated because the constraint requires a verb with an *er*$_Q$ argument to have an overt mid-field *er* complement.

4.5 Case 5: an *er* with four functions

This brings us to the final case. We will demonstrate that our constraints predict the following sentence to be grammatical, in which a single overt *er* expresses four functions at once. As the sentence is existential, the existential function must be present, the two partitive NPs *twee* and *drie* each require a quantitative function, and the stranded preposition *uit* requires the pronominal function.

(28) dat er$_{XQQP}$ [NP **twee** e] [NP **drie** e] **uit** gehaald hebben.
that there two students three books out.of fetched have

(28) has the structure shown in Figure 9. The head of the finite sentence *hebben*

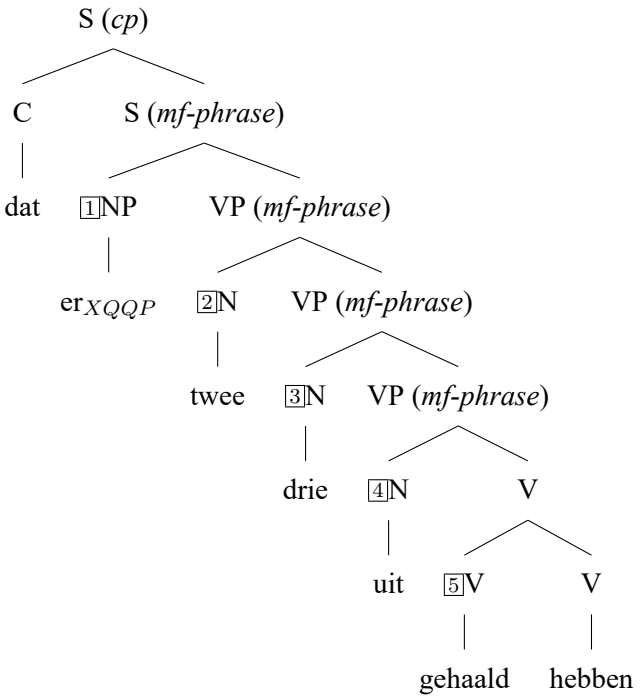


Figure 9: Partial analysis of example (28)

has the following COMPS list in the structure above. From left to right, the verb's complements are as follows: (i) the existential *er*, (ii)-(iii) the partitive nouns *twee* and *drie*, (iv) the preposition *uit*, and (v) the main verb *gehaald*.

(29)
$$\left[\begin{array}{l} \textit{word} \\ \textsc{head}\ \ \text{V}[\textit{fin}] \\ \textsc{comps}\ \left\langle \begin{array}{l} \boxed{1}\left[\begin{array}{l}\textit{canon-synsem}\\ \textsc{lid}\ \ \textit{er-X}\\ \textsc{fld}\ \ \textit{mid-fld}\end{array}\right], \boxed{2}\left[\begin{array}{l}\textit{canon-synsem}\\ \textsc{head}\ \ \text{N}\\ \textsc{comps}\ \ \langle \textit{ER}_Q \rangle\end{array}\right], \boxed{3}\left[\begin{array}{l}\textit{canon-synsem}\\ \textsc{head}\ \ \text{N}\\ \textsc{comps}\ \ \langle \textit{ER}_Q \rangle\end{array}\right], \\ \boxed{4}\left[\begin{array}{l}\textit{canon-synsem}\\ \textsc{head}\ \ \text{P}\\ \textsc{comps}\ \ \langle \textit{ER}_P \rangle\end{array}\right], \boxed{5}\text{V}[\textit{pastp}] \end{array} \right\rangle \end{array}\right]$$

Note that the two stranded partitive nouns and the preposition each have an *er*-complement on their COMPS lists which is inherited by *gehaald* and ultimately by the head *hebben* of the whole structure. These *er*s are not visible in the COMPS list of *hebben*, since then the verb would have more than a single *er* on its COMPS list in

violation of the MIDFIELD SINGLE-*er* CONSTRAINT. But they are present on the verb's ARG-ST, where they immediately precede their source.

(30)
$$\begin{bmatrix} \textit{word} \\ \text{HEAD} \quad \text{V}[\textit{fin}] \\ \text{ARG-ST} \quad \left\langle \begin{array}{l} \boxed{1}\begin{bmatrix}\textit{canon-synsem} \\ \text{LID} \quad \textit{er-X} \\ \text{FLD} \quad \textit{mid-fld}\end{bmatrix}, \boxed{6}\begin{bmatrix}\textit{pro-synsem} \\ \text{LID} \quad \textit{er-Q} \\ \text{FLD} \quad \textit{mid-fld}\end{bmatrix}, \boxed{2}\begin{bmatrix}\textit{canon-synsem} \\ \text{HEAD} \quad \text{N} \\ \text{COMPS} \quad \left\langle \boxed{6}\textit{ER}_Q \right\rangle\end{bmatrix}, \boxed{7}\begin{bmatrix}\textit{pro-synsem} \\ \text{LID} \quad \textit{er-Q} \\ \text{FLD} \quad \textit{mid-fld}\end{bmatrix}, \\ \boxed{3}\begin{bmatrix}\textit{canon-synsem} \\ \text{HEAD} \quad \text{N} \\ \text{COMPS} \quad \left\langle \boxed{7}\textit{ER}_Q \right\rangle\end{bmatrix}, \boxed{8}\begin{bmatrix}\textit{pro-synsem} \\ \text{LID} \quad \textit{er-P} \\ \text{FLD} \quad \textit{mid-fld}\end{bmatrix}, \boxed{4}\begin{bmatrix}\textit{canon-synsem} \\ \text{HEAD} \quad \text{P} \\ \text{COMPS} \quad \left\langle \boxed{8}\textit{ER}_P \right\rangle\end{bmatrix}, \boxed{5}\text{V}[\textit{pastp}] \end{array} \right\rangle \end{bmatrix}$$

Despite the relative complexity of this argument structure and its relation to the verb's COMPS list, *hebben* satisfies all of the constraints we formulated in section 3.

1. *er*-EXPRESSION CONSTRAINT: the verb selects an *er* complement.
2. MIDDLE FIELD *er*-EXPRESSION CONSTRAINT: is satisfied, as there is an *er* complement in the middle field.
3. P-L *er*-EXPRESSION CONSTRAINT: is satisfied, as there is no *er* in the pre-field.
4. MIDFIELD SINGLE-*er* CONSTRAINT: the verb has no more than a single mid-field *er* complement.

(28) is thus correctly predicted to be grammatical with one overt *er* that carries four different functions.

5 Conclusion

Dutch has four pronouns *er* which show an intriguing pattern of syntactic haplology when a finite verb has more than one *er* argument. We presented a theory that captures this pattern by relying on two central aspects of HPSG:

1. the distinction between ARG-ST and COMPS
2. the distinction between canonical and non-canonical *synsem*.

No deletion rules of the kind used in transformational analyses of *er* are necessary. We are not aware of any other formal theory that captures all the data presented in this paper. It remains to be seen whether other cases of syntactic haplology are susceptable to the kind of analysis used here.

References

Bech, Gunnar. 1952. Über das niederländische Adverbialpronomen *er*. *Traveaux du Cercle Linguistique de Copenhague* 8. 5–32.

Bennis, Hans. 1986. *Gaps and Dummies*. Dordrecht: Foris.

Bouma, Gosse. 2001. Argument Realization and Dutch R-Pronouns: Solving Bech's Problem without Movement or Deletion. In Ronnie Cann, Claire Grover & Philip Miller (eds.), *Grammatical Interfaces in HPSG*, Stanford: CSLI Publications.

Broekhuis, Hans. 2013. *Syntax of Dutch: Adpositions and Adpositional Phrases*. Amsterdam: Amsterdam University Press.

Campbell-Kibler, Kathryn. 2001. Bech's Problem, Again: Using Linearization on Dutch R-Pronouns. In Frank Van Eynde, Lars Hellan & Dorothee Beermann (eds.), *The Proceedings of the 8th International Conference on Head-Driven Phrase Structure Grammar*, 87—102. Stanford: CSLI Publications.

Drach, Erich. 1937. *Grundgedanken der deutschen Satzlehre*. Frankfurt: Diesterweg.

Höhle, Tilman N. 1983. Topologische Felder. Ms.

Höhle, Tilman N. 1986. Der Begriff "'Mittelfeld"', Anmerkungen über die Theorie der topologischen Felder. In Walter Weiss, Herbert Ernst Wiegand & Marga Reis (eds.), *Akten des VII. Kongresses der Internationalen Vereinigung für germanische Sprach- und Literaturwissenschaft. Göttingen 1985. Band 3. Textlinguistik contra Stilistik? – Wortschatz und Wörterbuch – Grammatische oder pragmatische Organisation von Rede?* (Kontroversen, alte und neue 4), 329–340. Tübingen: Max Niemeyer Verlag.

Miller, Philip & Ivan A. Sag. 1997. French Clitic Movement without Clitics or Movement. *Natural Language and Linguistic Theory* 15. 573–639.

Neeleman, A. & H. van de Koot. 2006. Syntactic Haplology. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, vol. 4, 685—710. London: Blackwell.

Odijk, Jan. 1993. *Compositionality and Syntactic Generalizations*: Tilburg University dissertation.

Reis, Marga. 1980. On Justifying Topological Frames: 'Positional Field' and the Order of Nonverbal Constituents in German. *Documentation et Recherche en Linguistique Allemande Contemporaine* 22/23. 59–85.

van Riemsdijk, Henk C. 1978. *A Case Study in Syntactic Markedness: The Binding Nature of Prepositional Phrases*. Dordrecht: Foris.

Van Eynde, Frank. 2019. Clustering and Stranding in Dutch. *Linguistics* 57(5). 1025–1071.

Van Eynde, Frank & Liesbeth Augustinus. 2014. Complement Raising, Extraction and Adposition Stranding in Dutch. In Stefan Müller (ed.), *Proceedings of the 21st International Conference on Head-Driven Phrase Structure Grammar, University at Buffalo*, 156–175. Stanford, CA: CSLI Publications. http://cslipublications.stanford.edu/HPSG/2014/vaneynde-augustinus.pdf.