Unique lexical entries in a subconstructional grammar

Petter Haugereid
Bergen University College

Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar
Nanyang Technological University (NTU), Singapore
Stefan Müller (Editor)
2015
CSLI Publications
pages 110–121
http://csli-publications.stanford.edu/HPSG/2015

Abstract

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

1 Introduction

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

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

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

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

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

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

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

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

*I would like to thank four anonymous reviewers, the INESS group in Bergen and the audience at the HPSG 2015 conference in Singapore for very useful comments and suggestions.*
Table 1: Pairing of categories and the number of words assigned to both categories in NorGram.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>ADV</th>
<th>ADVdeg</th>
<th>ADVs</th>
<th>Cadv</th>
<th>P</th>
<th>PRT</th>
<th>Psel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psel</td>
<td>0</td>
<td>38</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>53</td>
<td>31</td>
<td>-</td>
</tr>
<tr>
<td>PRT</td>
<td>5</td>
<td>39</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>38</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>P</td>
<td>5</td>
<td>66</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cadv</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ADVs</td>
<td>6</td>
<td>15</td>
<td>31</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ADVdeg</td>
<td>43</td>
<td>15</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ADV</td>
<td>13</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

2 Multiple lexical items

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

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

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

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

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

- Adverbs and prepositions also can be complementizers.

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

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

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

### 3 Incremental parsing with left-branching structures

Instead of assuming that lexical entries are specific to the extent that multiple lexical entries are needed for the same form (where the basic meaning is the same), I suggest an approach where lexical items are allowed to be underspecified with
regard to what function they fill. This approach depends on three factors: (i) under-
specification, (ii) multiple inheritance, and (iii) category specific phrase structure
rules that access the words in question. While the the first two factors are common
practice in HPSG, the third factor is an innovation. It can be achieved by means of
incremental parsing with left-branching structures.

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

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

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

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

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

4 Analysis of prepositions as unique lexical entries

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

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

My approach to prepositions is inspired by the treatment of particles and selected prepositions in the English Resource Grammar (ERG) (Flickinger, 2000), where the lexical entry for on as a particle or selected preposition is shown in (4).
The ERG lexical entry for selected prepositions/particles has an empty RELS list, which means it is semantically empty. Still, it has specified a KEYREL with a PRED value (\_on\_p\_sel\_rel) that will be required by the verb that selects it. But this relation does not end up on the RELS list.

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

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

\(^1\)The predicate names also indicate the number of arguments as well as their function. This is discussed in Haugereid (2014).
It is the function på has in the clause that determines which predicate it will end up with. If it functions as a particle of logge (‘log’), på_prd will be unified with the pred value of logge (logge_v), and the resulting relation will be _logge-på_1_rel. If it functions as a selected preposition of fokusere (‘focus’), på_prd will be unified with fokusere_v, yielding the predicate _fokusere*på_14_rel. And if it functions as a modifier, på_prd will be unified with the type mod+, which gives the predicate _på_p_rel.

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

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

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

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{Rule for attaching verbs}
\end{figure}

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

## 5 Implementation

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

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

\[ _{\text{abdisere}}_{\text{1_rel}} := \text{abdisere}_v & \text{prt–} & \text{1np} & \text{arg2–} & \text{arg3–} & \text{arg4–}. \]

Here, the type \textit{abdisere}_v is the type that is specified on the verb.\footnote{Since the verb only has one frame associated with it, it could also have been specified with its only subtype, \textit{abdisere}_1_rel.} The lexical entry for \textit{abdisere} is given in (6). Note that, as with prepositions, the \texttt{RELS} list of the verb is empty. It is rather the subconstruction rule for adding verbs, \textit{vbl-struct},...
that enters the KEYREL value of the verb onto the RELS list. (See Figure 5.) In this way we are not committing ourselves to the existence of a specific verbal relation if a verb appears in a sentence. The verb may for example be a part of an idiom or function as a light verb in a serial verb construction.

(6)  
\[
\begin{array}{l}
\text{verb-word}
\hline
\text{ORTH} & \langle "abdisere" \rangle \\
\text{CAT} & \text{HEAD verb} \\
\text{CONT} & \text{RELS} \langle !! \rangle \\
\text{KEYREL} & \text{PRED abdisere_v}
\end{array}
\]

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

(7)  
\[
\begin{array}{l}
\text{verb-word}
\hline
\text{ORTH} & \langle "få" \rangle \\
\text{CAT} & \text{HEAD verb} \\
\text{CONT} & \text{RELS} \langle !! \rangle \\
\text{KEYREL} & \text{PRED få_v}
\end{array}
\]

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

_ få_12_rel := få_v & prt- & 1np & 2np & arg3- & arg4-.
_ få-bort_12_rel := få_v & bort_prt & 1np & 2np & arg3- & arg4-.

The subtype _ få_12_rel allows få to be realized as a regular transitive verb with an NP subject (1np) and an NP object (2np).

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

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

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

---

3As mentioned in Section 2, the NorGram lexicon, which is more developed, lists på with 38 frames.
takes 1 hour and 43 minutes to compile with ACE (http://sweaglesw.org/linguistics/ace/). However, once the grammar is compiled, the size of the hierarchy of subconstruction types does not seem to have a serious effect on the efficiency of the parser. The parsing time of a sentence parsed when a small lexicon with 2,000 lexical entries is loaded is 0.01534 seconds, and the parsing time for the same sentence when the full lexicon (75,000 lexical entries) is loaded is 0.01778 seconds. Whether the increase is due to the size of the lexicon or the size of the hierarchy of subconstruction types is unknown.

6 Future work

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

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

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

References


