VP idioms in Norwegian: A subconstructional approach

Petter Haugereid
Bergen University College

Proceedings of the 21st International Conference on Head-Driven Phrase Structure Grammar

University at Buffalo
Stefan Müller (Editor)
2014
CSLI Publications
pages 83–102

Abstract

This paper presents a brief overview of idiomatic expressions in the Norwegian LFG grammar NorGram and shows how the rich lexical information of the LFG grammar can be reused in an HPSG-like grammar with a radically different approach to alternating argument frames. Rather than accounting for idioms by means of special idiom lexical entries, which is the standard approach in LFG and HPSG, a constructional approach is taken where the verbs of the idioms are left underspecified with regard to whether they are idioms or not. A hierarchy of subconstruction types is assumed, which for each piece of evidence provided by the words and rules of the sentence, narrows down the possible frames of the verb to just one.

1 Introduction

The Norwegian LFG grammar NorGram (Dyvik, 2000; Butt et al., 2002) has 56 VP idioms in the lexicon, distributed over 20 templates. Abstracting away from whether the selected object of the idiom is definite or indefinite, and what kind of argument the selected preposition has (NP, subordinate clause or infinitival clause), we are left with four main kinds of idioms.1

The first two kinds of idioms are semantically intransitive, hence they only take one argument, namely the subject. In the first kind of intransitive idioms the main verb selects an object, as shown in (1), and the second kind the main verb selects a PP, as shown in (2).

(1) **Han** gikk **konkurs**.
    he went bankrupt
    *He went bankrupt.*

(2) **De** løftet i **flokk**.
    They lifted in flock
    *They worked together.*

The last two kinds of idioms are semantically transitive, hence they take two arguments. They differ in that in one kind the main verb selects an object and the preposition of a PP, see (3), while in the other the main verb selects a PP and takes an object as an argument, see (4).

(3) **Han** la ikke **skjul** på **sin glede**.
    he laid not hiding on his joy
    *He did not hide his joy.*

---

1 I would like to thank two anonymous reviewers, the INESS group in Bergen, the audience at the HPSG 2014 conference in Buffalo, and the participants at the 2014 PARSEME meeting in Frankfurt, for very useful comments and suggestions.

1 Three idioms (**ta på kreftene** (‘tax one’s strength’), **sende ord** (‘send a message’), and **komme på kant med** (‘fall out with’)), do not fall into any of the four categories, and they are left out of the present discussion.
A verb that is part of a VP idiom is assigned an idiom frame in the lexicon in addition to the other frames that it appears with. For example the verb bringe (‘bring’) is listed with the following frames:

\[(V\text{-SUBJ-POBJrefl-OBJ bringe med})\]
\[(V\text{-SUBJ-PRT-OBJ bringe inn})\]
\[(V\text{-SUBJ-OBJ-OBJ bringe})\]
\[(V\text{-SUBJ-OBJ-OBLBEN bringe})\]
\[(V\text{-SUBJ-OBJ bringe})\]
\[(VPIDIOM-PSEL-OBJ-OBJ bringe på bane)\]

A lexical entry is allowed to have more than one argument frame by using disjunctions of frames. Disjunctions are expanded into full lexical entries during parsing. This means that a lexical entry with 6 disjunctive argument frames is computationally equivalent to six lexical entries.

In this paper I will present a new way of representing information about argument frames, including the different kinds of VP idioms presented in this section. The account shifts the burden from the lexicon to a carefully designed hierarchy of subconstruction types. The transfer is achieved by means of phrasal subconstructions (see Haugereid & Morey (2012); Haugereid (2012)), which are construction parts that, when put together in a way that conforms with a constraint on the verb, form full constructions. The analysis is implemented in an HPSG-like grammar of Norwegian within the LKB system (Copestake, 2001).

2 Treatment of idioms in Sag et al. (2003)

In (Sag et al., 2003, 347–355), idioms are assumed to have special lexical entries for the words that constitute them. The idiom keep tabs on is analyzed by means of a lexical entry for keep (see (6)) with three items on the SUBCAT list; (i) the NP subject, (ii) an idiomatic noun tabs, and (iii) a constituent marked by the preposition on.
As (6) shows, the relation of the idiom *keep tabs on* (observe) has two arguments, *OBSERVER* and *OBSERVED*, and they are linked to the subject of *keep* and the constituent marked by the preposition *on*. Both the idiomatic noun *tabs* and the selected preposition *on* are semantically empty.

Given the degree of detail required in the lexicon, one is forced to assume separate lexical entries for idiomatic verbs. From a semantic point of view, this is motivated, considering how the meaning of idioms deviates from the compositional meaning. However, there is no morphological evidence indicating that idiomatic verbs should have separate lexical entries. They share the stem with their compositional versions and have the same inflections.

In section 3 I will present an account that allows us to have a single lexical entry for verbs that alternates between argument frames, including idiomatic frames.

### 3 Analysis

Instead of a lexical approach to subcategorization, a fully constructional approach is taken. In an analysis of a sentence, a *START* sign is assumed at the beginning of the sentence. Each word of the sentence is attached to this sign in an incremental, left-branching fashion (see Haugereid & Morey (2012)). A simplified structure of a sentence with three words is given in Figure 1.

The relation of the sentence is not contributed by the main verb, but rather by the *START* sign. Instead of contributing a relation, the verb is assumed to have a feature *FORM*, and the value of this feature is unified with the *PRED* value of the relation.²

The *VFORM* value of the verb is by itself not enough to determine the predicate of the event expressed. In order for it to be fully specified, the predicate needs to be unified with other pieces of information stemming from the attachment of

---

²The assumption that the relation of the sentence is introduced by the *START* sign rather than the main verb is motivated by the fact that some languages have empty copula constructions, where there is no verb to contribute the relation.
potential arguments. This is illustrated by means of the LINK features of word1 and word3 in Figure 1. Together, the LINK values here contribute the information that the predicate is a two-place predicate.

The motivation behind the demoted role of the verb is the fact that it is possible for verbs to alternate between different argument frames. Additionally, the approach lends itself nicely to the treatment of multiword expressions.

3.1 Lexical representation

In addition to the idiom frame shown in (4), the verb bringe also has a transitive and a ditransitive frame, as shown in (7).

(7) a. Han bragte maten.
   he brought food.the
   *He brought the food.*

   b. Han bragte henne maten.
   he brought her food.the
   *He brought her the food.*

It also has frames that involve particles, prepositions and reflexives, as shown in (8).

(8) a. Han bragte med seg maten.
   he brought with himself food.the
   *He brought the food.*

   b. Filmen bragte inn masse penger.
   movie.the brought in lots-of money
   *The movie brought in lots of money.*

   c. Han bragte maten til henne.
   he brought food.the to her
   *He brought the food to her.*
Even though we have six argument frames for the verb bringe, I assume only one lexical entry, shown in (9). The lexical entry has information about the STEM of the lexeme, the HEAD value and the HEAD value of its (potential) arguments; C(ONSTRUCTION)-ARG1, C-ARG2, C-ARG3, and C-ARG4. These four argument features correspond to external subject, (deep) direct object, (deep) indirect object, and oblique object, respectively. Note that there is no linking of the C-ARGS to the semantics. Rather, the linking is done in what I refer to as phrasal subconstructions.

The lexical entry also has a feature FORM, and it is the value of this feature that determines which constructions the verb is compatible with.

\[(9) \begin{array}{c}
\text{stem-verb} \\
\text{STEM } "bringe" \\
\text{HEAD verb} \\
\text{C-ARG1 [HEAD noun]} \\
\text{C-ARG2 [HEAD noun]} \\
\text{C-ARG3 [HEAD noun]} \\
\text{C-ARG4 [HEAD compl-noun]} \\
\text{FORM } [PRED bringe_v]
\end{array}\]

### 3.2 Phrasal subconstructions

One example of a phrasal subconstruction is the rule that links (external) subjects, arg1-struct, illustrated in (2). In this rule, the value of C-ARG1|LINK is switched from arg1– in the mother to arg1+ in the first daughter. At the same time, the argument (the second daughter of the rule) is linked to the ARG1 of the KEYREL. The grammar also has subconstructions that in the same fashion link (deep) direct objects arg2-struct, (deep) indirect objects arg3-struct, and oblique objects arg4-struct.

The grammar has a rule vbl-struct which adds the verb. (See Figure 3). The verb is selected via the VBL feature of the first daughter, and the VBL value of the verb is transferred to the mother. In this way, the added verb is able to constrain the following verb, if there is one. The rule also unifies its KEYREL|PRED value with the FORM value of the verb. The verb does not contribute the full predicate, just a predicate type which, when unified with types contributed by the other subconstructions, yields the predicate of the clause.

The tree in Figure 4 shows how a transitive sentence is analysed. At the top node, the subconstruction constraints are negative. Three subconstructions apply, the vbl-struct, which adds the verb bragte (‘brought’), the arg1-struct, which adds the subject han (‘he’), and the arg2-struct, which adds the direct object maten (‘the food’). Each subconstruction contributes a type; vbl-struct adds the FORM value of
the verb, bringe,v, arg1-struct switches arg1– in the mother to arg1+ in the first
daughter, and arg2-struc switches arg2– to arg2+. As for the subconstructions
do not apply, their respective values stay negative. In this way, the START
node reflects which subconstructions have applied, and which have not applied.

The result of unifying the subconstruction types arg1+, arg2+, arg3–, arg4–,
prt–, and bringe,v in the START sign in Figure 4 is the predicate bringe_12_rel.
This is shown in the type hierarchy in Figure 5, which will be discussed in Section
3.3.

3.3 Valence alternations

The valence alternations of the verb bringe (see (4), (7) and (8)) are accounted for
by means of a hierarchy of predicate types. The type hierarchy in Figure 5 shows
all the subconstruction types employed in order to account for the alternations of
FIGURE 4: Analysis of the transitive sentence Bragte han maten? (‘Did he bring the food?’)

bringe.

The function of the subtypes of link in the hierarchy is to show whether a subconstruction has applied or not. For example, arg1− means that the arg1 subconstruction has not applied, while arg1+ means that it has applied. The type vrb+ has as immediate subtypes the FORM value of all verbs in the lexicon. (In Figure 5, only the FORM value of the verb bringe (‘bring’) is shown.) The subtypes of verb FORM values decide what frames a verb can appear in. As the hierarchy indicates,
bring can appear in 6 frames, since it has 6 subtypes (ignoring the intermediate types). The type \(prt+\) has as immediate subtypes the FORM value of all the particles in the lexicon. (In Figure 5, only the FORM value of the particle inn ('in') is shown.) The type \(prp\) has as immediate subtypes the FORM value of all prepositions in the lexicon. (In Figure 5, only the FORM value of the prepositions til ('to'), med ('with'), and på ('on') are shown.) The type \(mwe\) has as immediate subtypes the FORM value of the reflexive (refl) and the FORM value of all the idiomatic nouns. (In Figure 5, only the FORM value of the reflexive and the idiomatic noun bane ('track') are shown.)

The subconstruction types are possible values of the features shown in Figure 6, and in order for a sentence to parse, these values need to unify. The features have different kinds of types as values before they are unified.

```plaintext
[ C-ARG1|LINK □
  C-ARG2|LINK □
  C-ARG3|LINK □
  C-ARG4|LINK □
  PART □
  KEYREL|PRED □ ]
```

Figure 6: Unification of subconstruction types.

The value of \(C-ARG1|LINK\) is binary; either \(arg1–\), which means that no (external) subject has been realized, or \(arg1+\), which means that it has been realized. In the case of bringe, all the frames require the \(arg1+\) type, which means that they are all agentive.
The feature C-ARG2|LINK can have three different kinds of values. It can be the type arg2–, which means that no (deep) direct object has been realized. It can have the value arg2+, which means that a (deep) direct object is realized, and that it has a semantic role (see (7a), repeated below as (10a)). It is then not part of an MWE. Finally, it can have a subtype of mwe_n as value. In this case, the direct object is either a reflexive, as in (10b), or it constitutes a part of an idiom, as in (1), repeated below as (10c).

(10) a. Han bragte maten.  
   he brought food.the  
   *He brought the food.*

b. Han barberer seg.  
   he shaves himself  
   *He shaves.*

c. Han gikk konkurs.  
   he went bankrupt  
   *He went bankrupt.*

If the value is a subtype of mwe_n, the direct object is not assumed to have a semantic role, as regular direct objects. Instead, it is added by the arg2-mwe-struct rule, which, rather than linking the object to the ARG2 role of the KEYREL, unifies the FORM value of the object with the PRED value of the KEYREL. This is shown in Figure 7.

Similar to the feature ARG2|LINK, the feature ARG3|LINK can have a negative value arg3–, which means that no (deep) indirect object has been realized, and a positive value arg3+, which means that a (deep) indirect object has been realized (with its own semantic role) (see (7b), repeated below as (11a)). It can also have an indirect object that is a part of an MWE, exemplified with a reflexive in (11b). This object is not assumed to have a semantic role and is added by the rule arg3-mwe-struct, which is similar to the arg2-mwe-struct rule.

(11) a. Han bragte henne maten.  
   he brought her food.the  
   *He brought her the food.*

b. Han nærmer seg en løsning.  
   he nears himself a solution  
   *He is closing in on a solution.*

The feature ARG4|LINK can have four types of values. It can have a negative value arg4–, which means that no oblique argument is realized. It can have a

3Currently, no distinction is made between frames with NPs, CPs, or IPs as direct objects. It is possible to account for this distinction by letting arg2+ have subtypes such as arg2_np, arg2_cp and arg2_ip, however this has not yet been implemented. Instead, the ARG2|HEAD value is constrained in the lexicon.
positive value \textit{arg4+}, which means that an oblique argument is realized, and that it has a semantic role (see (8c), repeated here as (12a)). In case the oblique argument does not have a semantic role, but constitutes a part of an MWE, the value is a subtype of \textit{mwe.n}, for example \textit{refl.n} in the case of reflexives (see (8a), repeated below as (12b)) or the FORM value of a oblique object that constitutes a part of an idiom (see (4), repeated below as (12c)). In the case of the idiom in (12c), the FORM value of the oblique object is \textit{bane.n}. The FORM value of the prepositions that mark the oblique objects are the fourth type of value that the ARG4|LINK feature can have. They are unified with the \textit{arg4+} type if the oblique object has a semantic role, or the relevant subtype of \textit{mwe.n} if the oblique object is a part of an MWE. In (12a)–(12c), the FORM value of the prepositions marking the oblique object are \textit{til.prp}, \textit{med.prp}, and \textit{på.prp}.

(12) a. Han bragte maten til henne.
   he brought food.the to her
   \textit{He brought the food to her.}

b. Han bragte med seg maten.
   he brought with himself food.the
   \textit{He brought the food.}

c. Han bragte temaet på bane.
   he brought topic.the on track
   \textit{He brought up the topic.}

The subconstruction rule that adds the preposition that marks the oblique object is the \textit{prepmark-struct} rule. (See Figure 8.) The rule unifies the FORM value of the preposition with the C-ARG4|LINK value, and switches the ARG4|MARKED
value from ‘–’ in the (first) daughter to ‘+’ in the mother. Once $\text{ARG4|MARKED}$ is switched to positive, the oblique argument can be attached.

Figure 8: The $\text{prepmark-struct}$ rule for prepositions marking oblique objects

If the oblique object is a part of an MWE (either a reflexive or an idiomatic noun), it is added by the $\text{arg4-mwe-struct}$ rule shown in Figure 9.

Figure 9: The $\text{arg4-mwe-struct}$ rule for oblique objects that are a part of an MWE.

The feature $\text{PART}$ has a negative value ($\text{part–}$) if the frame does not involve a
particle, and it has a subtype of part+ if the frame involves a particle, as in (8c), repeated below as (13). The subtype will then be the FORM value of the selected particle (here innprt).

(13) Filmen bragte inn masse penger.
    movie.the brought in a-lot-of money
    
    The movie brought in a lot of money.

The feature KEYREL has as value the FORM value of the main verb, which is a subtype of vrb+. 4

Figure 10 shows the subconstruction types that are unified in order to arrive at the frame type bringe_12_rel (arg1+, arg2+, arg3–, arg4–, prt–).

Figure 10: Unification of subconstruction types that result in the type bringe_12_rel.

Similarly, Figure 11 shows the subconstruction types that are unified in order to arrive at the frame type bringe_123_rel (arg1+, arg2+, arg3+, arg4–, prt–, and bringe_v).

Figure 11: Unification of subconstruction types resulting in the type bringe_123_rel.

The unifications resulting in the other frame types of bringe are given in Figures 12–14.

4In a language with empty copula constructions, one can also introduce a type vrb– for clauses without verbs. Norwegian, however, does not have this construction.
Figure 12: Unification of subconstruction types resulting in the type *bringe--inn_12_rel*.

Figure 13: Unification of subconstruction types resulting in the type *bringe*--*til_124_rel*.

Figure 14: Unification of subconstruction types resulting in the type *bringe*--*med-seg_12_rel*.

Figure 15: Unification of subconstruction types resulting in the type *bringe*--*på-bane_12_rel*.
Table 1: Subconstructions involved in the different VP idiom types.

<table>
<thead>
<tr>
<th></th>
<th>vbl-struct</th>
<th>arg1-struct</th>
<th>arg2-struct</th>
<th>arg2-mwe-struct</th>
<th>prepmark-struct</th>
<th>arg4-struct</th>
<th>arg4-mwe-struct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrans. with</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>idiomatic noun</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrans. with</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>idiomatic PP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans. with</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>idiomatic noun</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans. with</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>idiomatic PP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.4 Analysis of VP idioms

The analysis of VP idioms includes the subconstruction rule for prepositions marking oblique objects \textit{prepmark-struct} (see Figure 8) and two subconstructions rules for MWE nouns; \textit{arg2-mwe-struct} and \textit{arg4-mwe-struct} (see Figures 7 and 9).

An analysis of a sentence with a VP idiom (\textit{Bragte han temaet på bane} ‘Did he bring up the topic’) is illustrated in Figure 16. Five subconstruction apply. The first subconstruction \textit{vbl-struct} adds the verb \textit{bragte} and unifies the \textit{FORM} value of the verb with the \textit{KEYREL/PRED} value. The second subconstruction \textit{arg1-struct} adds the subject \textit{han}, and links its index to \textit{KEYREL/ARG1}. The third subconstruction \textit{arg2-struct} adds the direct object \textit{temaet}, and links its index to \textit{KEYREL/ARG2}. The fourth subconstruction \textit{prepmark-struct} adds the preposition marking the oblique object \textit{på} and unifies the \textit{FORM} value of the preposition with the \textit{KEYREL/PRED} value (and the \textit{C-ARG4/LINK} value of the first daughter). The fifth subconstruction adds the idiomatic noun \textit{bane} and unifies its \textit{FORM} value with the \textit{KEYREL/PRED} value (and the \textit{C-ARG4/LINK} value of the first daughter).

In the top node \textit{arg4-mwe-struct}, all \textit{LINK} values are constrained to be negative, and at the bottom of the tree, in the \textit{START} node, marks from all the subconstructions that have applied can be found, and they are unified. When the subconstruction types in the \textit{START} sign are unified, we get the type \textit{bringe*på-bane_12.rel}.

The four kinds of idiomatic expression types introduced in Section 1 are accounted for by the combinations of subconstructions shown in Table 1

4 Implementation

The most common templates in the NorGram LFG grammar are given in Table 2. The table shows how the information encoded in these frames can be broken down
into subconstruction types. For example, the most common template $V-$SUBJ-OBJ is associated with the subconstruction types $arg1+$, $arg2+$, $arg3-$, $arg4-$, and $prt-$, which are the types that come from a standard transitive sentence.

In addition to the types shown in Table 2, the FORM value of the verb, and the FORM values of prepositions and particles (if applicable), which are part of the LFG frames, are added to the subconstruction types. Given a table that maps templates to subconstruction types as shown in Table 2, the LFG frames in (5) can be translated into the following types:

5The inquit template (Vinq-SUBJ-COMP) for sentences like “Jeg kommer”, sa han. (“I’m coming”, he said.) is not included, as inquit frames currently are not handled by the grammar.

6The preposition of the template ‘V-SUBJ-OBJ-OBLBEN’ is specified in the template to be til,
The hierarchy of relation types and subconstruction types above is the same as the hierarchy in Figure 5. This shows how a type hierarchy of subconstruction types can be generated, given a conversion table. The program that generates the a type hierarchy from an LFG lexicon and a conversion table can be conceived of as a compiler.

The NorGram lexicon has 15,776 verb frames. I have tested the procedure on a slightly smaller version of the lexicon, the open source NKL lexicon with 13,069 verb frames, and loaded it into the LKB system. Loading the grammar now obviously takes more time, but the efficiency of the parser does not seem to be affected by the large number of subconstruction types (almost 20,000 in all).

The MRSs resulting from parsing the four idiomatic examples in (1), (2), (3), and (4) are given in Figure 17–20.

and is not specified in the frame.
Figure 17: MRS of the sentence *Han gikk konkurs.* (*He went bankrupt*)

Figure 18: MRS of the sentence *De løftet i flokk.* (*They worked together*)

5 Discussion and future work

The analysis presented in this paper is not restricted to idioms, but includes several kinds of MWEs, like particle verbs, verbs with selected prepositions, reflexive verbs, and combinations of these. It can also be expanded to nouns and adjectives with selected complements.

I have dealt only with idiomatic nouns that are indefinite, although idiomatic expressions also may consist of definite idiomatic nouns, like *øynene* (*eyes.the*) in *ta øynene fra* (*look away from*) or even idiomatic nouns modified by an adjective, like *et godt øye* (*a good eye*) in *ha et godt øye til* (*have a preference for*). Examples like these suggest that the predicates in the hierarchy of link types not only need to reflect the base form of idiomatic nouns, but also other features like definiteness and adjuncts.

The flexibility of the approach comes from the fact that it is a subconstructional approach. While lexicalist approaches need to be very specific about the argument structure of a verb, and need to use disjunctions of frames in lexical entries (LFG)
or multiple lexical entries/lexical rules (HPSG) in order to account for valence alternations, the subconstructional approach allows for precise underspecification using the hierarchy of subconstruction types. Only one lexical entry per verb is
needed. And while constructional approaches are forced to assume relatively flat syntactic structures in order to have access to the arguments of a construction, and hence risk ending up with an unmanageable amount of phrase structure rules, the subconstructional approach allows for binary structures and the number of phrase structure rules is kept relatively small (about 80). The combination of lexical underspecification and binary structures is achieved by means of the type hierarchy of subconstruction types which includes types for all verbs, prepositions, particles and idiomatic nouns in the lexicon and types for the frames they occur in. The hierarchy is designed in such a way that a verb is only allowed to combine with selected combinations of constituents. The hierarchy is huge, but finite. And it is interesting in that it reflects what kinds of subconstructions are needed in order to express all grammaticalized concepts in a grammar.

References


