Linguistically enriched corpora for establishing variation in support verb constructions

Begoña Villada Moirón
Alfa-Informatica
University of Groningen
P.O.Box 716
9700 AS Groningen
M.B.Villada.Moiron@rug.nl

Abstract

Many NLP tasks that require syntactic analysis necessitate an accurate description of the lexical components, morpho-syntactic constraints and the semantic idiosyncracies of fixed expressions. (Moon, 1998) and (Riehemann, 2001) show that many fixed expressions and idioms allow limited variation and modification inside their complementation.

This paper discusses to what extent a corpus-based method can help us establish the variation and adjectival modification potential of Dutch support verb constructions. We also discuss what problems the data poses when applying an automated data-driven method to solve the problem.

1 Introduction

We aim at finding methods that facilitate the description of the linguistic behavior of multiword expressions. Empirical evidence and generalizations about the linguistic properties of multiword expressions are required to further a theory of fixed expressions (or multiword expressions) as well as to expand the coverage of NLP lexical resources and grammars.

This paper describes an attempt to develop automated methods for induction of lexical information from a linguistically enriched corpus. In particular, the paper discusses to what extent can an automated corpus-based approach be useful to establish the variation potential of support verb constructions. The experimental work applies to Dutch expressions, however the issue is widely relevant in the development of lexical resources for other languages.

1.1 Partially lexicalized expressions

Corpus-based studies showed that certain fixed expressions and idioms allow limited variation and adjectival modification (Moon, 1998; Riehemann, 2001).[1] Riehemann (2001) investigated various types of multiword expressions in English and observed that around 25% of idiom occurrences in a corpus allow some variation. By way of example, among the occurrences of the idiom *keep tabs on* '(fig.) watch', variation affects verb tense inflection, adjective modifiers (*close, better, regular, daily*), noun number morpheme (*tab(s)*) and the location of the *on* complement phrase that may be separate from the object NP. The above example is by no means an isolated case.

Variation has an effect not only on the representation of the syntactic structure but also on the semantic interpretation of the multiword expression (Sag et al., 2001; Baldwin et al., to appear). The presence of variation in multiword expressions brings up two scenarios: (a) the loss of the peculiar meaning or (b) the modification of the original meaning. Returning to the example above, modifiers of *tabs* affect the interpretation of the event predicate as a whole. Thus, *keep*

[1] From now onwards, we use 'variation' to refer to morphological productivity or alternation of specifiers or prenominal modifiers.

close tabs on s.o. means 'watch s.o. closely'. A different effect has been reported of some VERB NP idioms in which the adjectival modification affects only the complement NP (Nicolas, 1995). For a correct interpretation, such idiomatic expressions require internal semantic structure.

These observations suggest that: (i) not all fixed expressions and idioms are frozen word combinations given that, parts of the expression participate in syntactic operations; (ii) some lexemes (in 'fixed' expressions) are subject to morphological processes; and (iii), some fixed expressions still preserve underlying semantic structure. A description that captures the previous facts needs to allow variable slots so that the mentioned variants of the expression are licensed by the grammar. In sum, variation is a property that should not be neglected while deciding the lexical representation of multiword expressions in computational resources.

1.2 Support verb constructions

Support verb constructions are made up out of a light verb (aka. support verb) and a complement (e.g. *take into account*). The predicational complement may be realized as a noun, an adjective or a prepositional phrase. The light verb and its complement form a complex predicate, in which the complement itself supplies most of the semantic load (Butt, 1995). The verb performs a 'support' function, i.e. it serves to 'further structure or modulate the event described by the main predicator' (Butt, 1995). Most researchers agree that the light verb adds aspect, tense and 'aktionsart' information to the predicate. Since the support verb's meaning differs from the meaning of the (main) verb lexeme, the meaning of the support verb construction is not fully compositional. Due to the similarities with other idiosyncratic expressions, support verb constructions (LVCs) belong to the group of lexicalized multiword expressions (Sag et al., 2001).

We limit this study to support verb constructions for two practical reasons. First, there seems to be a group of core light verbs that exist cross-linguistically. Thus, we can concentrate on a small set of verbal lexemes. Second, these light verbs are based on main verbs still in active use in the language (Butt, 1995). Concerning Dutch, nine verbs that can function as main but also as light verbs are *brengen* 'bring', *doen* 'do', *gaan* 'go', *geven* 'give', *hebben* 'have', *komen* 'come', *krijgen* 'get', *maken* 'make', *nemen* 'take' and *stellen* 'state' (Hollebrandse, 1993). Establishing the lexical properties of light verb predicates is necessary so that parsers do not misanalyze main verb and light verb uses.

Before we describe a corpus-based method to extract evidence of variation from a syntactically annotated corpus, we enumerate some research assumptions and highlight the types of variation and modification object of this study. Section 3 presents the automated method and the evaluation of its merits. Section 4 describes a proposal of the required lexical annotation drawn from a working implementation. Our conclusions and further improvements are summarised in section 6.

2 Base form, variation and modification

In addition to a subject, some prepositional support verb constructions select an additional complement. This may be realized by an accusative, dative or reflexive NP. Prior to applying the corpus-based method described in section 3, we partly ignore the lexical content within the PP complement; this is also why we want to establish the variation potential within LVCs. For the above two reasons, we assume that the *minimum required lexemes* (i.e. common to all prepositional LVCs) include the argument PP and the support verb and represent each expression as a triple of the form [PREPOSITION NOUN VERB] (P N V). (Thus, determiners and modifiers are left out).

Some further assumptions must be introduced, namely, what we understand as a *base form* and as a *variant* of a support verb construction. The base form includes the mentioned triple and may include other lexicalized arguments. In expressions that allow no morphosyntactic variation or modification within the required arguments, tense inflection is usually possible. The base form shows the infinitive verb form. The base form of the expression *voet bij stuk houden* 'stick to one's guns (fig)' includes the noun *voet*, the PP *bij stuk* and the verb *houden*; tense inflection is possible (1-b).

(1) a. VOET BIJ STUK HOUDEN

b. De verzekeraars *hielden* echter *voet bij*
the insurers kept really foot by
stuk.
piece
'The insurance companies really sticked to their guns (fig.)'

Any instance of an LVC whose NP within the PP argument differs from the NOUN lexeme is considered a variant. The expression *uit zijn dak gaan* 'go crazy' has as base form (2-a) with the noun *dak* allowing various possessive determiners (2-b).

(2) a. UIT DAK GAAN
b. Het publiek *ging uit zijn dak*.
the audience went out his roof
'The audience went crazy.'

We study variation observed within the expression. We focus on two levels:

lexeme level productive inflectional and derivational morphology.

phrase level variability in specifiers and modifiers.

The evidence we seek to extract is the following: (a) use of diminutive in nominal lexemes; (b) singular and plural alternation in nouns. Evidence of derivational morphology, for example, instances of compounding (another noun or an acronym prefixed to the head noun) or a genitive noun modifier; (c) alternation in specifiers. Among the specifiers: zero determiner, definite, indefinite, reciprocals, possessives, demonstratives and quantifiers; (d) NPs that are realized by reflexives. Reflexives may instantiate either open argument slots or an NP within complement PPs; and (e), among modification, we explore prenominal adjectives, past participles, gerunds and other intervening material.

In addition, some expressions allow relative clauses and PP post-nominal modifiers. Relative clauses are observed less often than PP postnominal modifiers. So far, we ignore these two types of modification because we extract the evidence from an automatically annotated corpus and with automated means. It is well-known that disambiguating a syntactic attachment site, e.g. a PP–attachment site, is one of the hardest problems for present-day parsing technology. Needless to say, the parser (`Alpino`) also encounters difficulties with this problem. In this work, we did not investigate syntactic flexibility at the sentence level, that is, processes such as passive, topicalization, control, clefting, coordination, etc.

3 A corpus-based method to infer variation

With access to automatically parsed data, subcategorization frames and a standard search query language such as `dt_search`, we can extract all instances of an LVC that satisfy rather specific morphosyntactic features and head-complement dependencies; these requirements – expressed as `dt_search` queries – are applied to XML-encoded syntactic dependency trees. For a more detailed description of the corpus-based method refer to (Villada Moirón, 2005).

3.1 Corpus annotation

A list of P N V triples was automatically acquired from a syntactically annotated corpus using collocation statistics and linguistic diagnostics (Villada Moirón, 2004). A P N V triple represents an abstraction of a support verb construction (LVC).

For each automatically extracted triple, all sentences containing the three component lexemes found in the Twente Nieuws Corpus (TWNC) (Ordelman, 2002) were collected in a subcorpus. For example, for the expression *uit zijn dak gaan* 'go crazy', all sentences that include the preposition *uit* 'out', the noun *dak* 'roof' and the verb *gaan* 'go' or one of its inflectional variants are collected in a subcorpus.

The `Alpino` parser (van der Beek et al., 2002) was used to annotate the subcorpora. This is a wide-coverage parser for Dutch. Based on a lexicalist constraint-based grammar framework (Head-Driven Phrase Structure Grammar) (Pollard and Sag, 1994), the Alpino grammar licenses a wide variety of syntactic constructions.

All parsed data is stored as XML-dependency trees. To illustrate the annotation, the result of parsing example (2-b) is the dependency structure tree shown in figure 1.

Among the information contained in the parsed trees, we use: (i) categorical information (phrasal

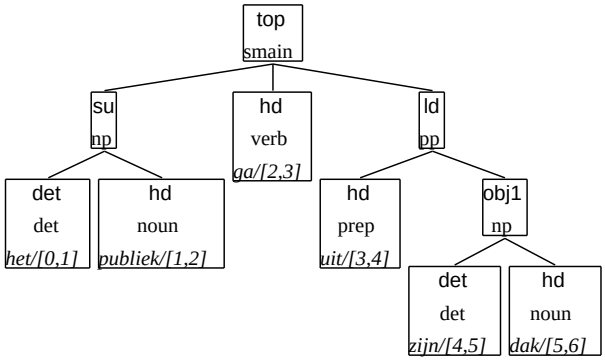

Figure 1: This syntactic dependency tree corresponds to the parsed sentence in (2-b).

(np, pp) and lexical (det, noun)), (ii) syntactic information (grammatical function or dependency relation (subject su, direct object obj1, locative or directive complement ld, head hd, determiner det)) and (iii) lexical information (lexemes and word forms). Dependency nodes are crucial in stating daughter–ancestor relations between constituents and sub-constituents in an LVC.

3.2 Extraction

dt_search (Bouma and Kloosterman, 2002), a treebank query tool based on XPATH,[2] is used to extract evidence from the annotated subcorpora. A dt_search query applied on the corresponding parsed subcorpus searches for all LVC instances. Two types of queries are needed: *narrow search* and *wide search* queries. Narrow search queries seek instances of a head-dependent relation between a VERB and a PP sibbling, given necessary lexical restrictions as input. Wide searches state that the PP is embedded (somewhere) under a clausal node whose head is VERB. Wide searches are needed because the parser may wrongly attach the sought PP to a previous noun. (Thus, in the annotated data the PP and VERB do not satisfy a head-dependent relation). Finally, the vaguest search states that a given PP needs to occur within the same sentence as the verb. This type of search is used in case the other two types fail to retrieve any evidence. The query in figure 2 seeks NP-internal adjectival modification.

[2]Nevertheless, other XML-based query tools are also freely available, e.g. XSLT or the TIGERSearch kit.

```
dt_search
 '//node[@cat="np" and
   ./node[@cat="ap"] and
   ./node[@rel="hd" and
          @root="gedachte"] and
   ../node[@rel="obj1"] and
   ../node[@rel="hd" and @word="op"
   and
   (../../../node[@rel="hd" and
                  @root="breng"] or
   ../../node[@rel="hd" and
              @root="breng"] or
   ../node[@rel="hd" and
           @root="breng"]) ] ]'
   breng.opgedachten/*.xml
```

Figure 2: Query to extract adjectives in the expression *iemand op gedachten brengen*.

Among the constraints expressed in the search queries there are: parent-child relations between nodes, phrase category (@cat), dependency relation (@rel), word base form (@root) or surface form (@word). Queries need to capture deeply embedded LVCs. A verbal complement embedded under several modal or auxiliary verbs is rather common. To allow uncertainty about the location of the PP argument node with respect to its head verb, disjunctive constraints are introduced in the queries (figure 2).

3.3 Retrieved corpus evidence

A search query retrieves each LVC realization that satisfies the query requirements, as well as the LVC frequency in the subcorpora.

Figure 3 gives an excerpt from the observed adjectival modification in *iemand op gedachten brengen* ‘give s.o. the idea’. *Op andere gedachten brengen* ‘change s.o.’s idea’ is the most frequent realization with 634 out of a total of 682 occurrences. This suggests that the adjective *andere* is almost frozen in the expression.

The method extracts evidence of morphological productivity, variation of specifiers and adjectival modification, i.e. positive and negative evidence. A description of the positive evidence

```
1 aangename gedachten
1 amoureuze gedachten
634 andere gedachten
1 andere politieke gedachten
1 andere, redelijke gedachten
1 beeldende gedachten
1 bepaalde gedachten
2 betere gedachten
1 duivelse gedachten
1 heel andere gedachten over...
1 hitsige gedachten
1 hogere gedachten
1 kritische gedachten
1 meer poëtische gedachten
```

Figure 3: Observed adjectival modification in the LVC *iemand op gedachten brengen*.

follows. We investigated 107 Dutch LVCs: 94 expressions that require a PP argument among which some show an NP_{acc} open slot; lexical restrictions affect the verb and the PP argument; in addition, 13 other expressions are made up of a (partially) lexicalized NP and a PP argument.

LVCs fall in one of three groups: (a) totally fixed, (b) semi-fixed and (c) flexible. **Fixed** LVCs show no variation and no modification in the lexicalized NP (if present) and PP constituent(s). 42% of the LVCs studied are fixed. **Semi-fixed** LVCs show partially lexicalized constituent(s) (20.5% of the studied LVCs). Rarely, a singular noun appears in plural. Variation affects the lexeme's morphology and/or the specifiers slot. Expressions whose lexicalized argument requires a reflexive are included into this group. **Flexible** LVCs allow adjectival modification (37.5% of the studied LVCs). The data is rather varied. There are LVCs that show: (i) non-productive morphology and no specifier variation but they show a limited number of adjectives and, (ii) specifier variation (some show compounding) and limited adjectival variation. Border-line cases exhibit no morphological productivity and either definite/possessive determiner alternation or no specifier variation; modification involves a unique adjective (e.g. *in (verzekerde) bewaring stellen* 'put into custody').

Negative evidence (noise) typically includes sentences where the VERB and the PP occur within the same clause but not in the LVC context (in its literal use). Often, the PP is an adjunct or a complement of another verb. The reason for this noise can be attributed to the uncertainty in the search queries or errors in the annotated data.

3.4 Discussion

We argue that the corpus-based method is efficient in extracting the linguistic contexts where variation and internal modification are found inside LVCs. Examining the evidence retrieved by the corpus-based method, a researcher quickly forms an impression about which expressions are totally fixed and which expressions allow some variation and/or modification. One also has direct access to the realizations of the variable slots, the LVC frequency and relevant examples in the corpus. Next, we discuss some limitations posed by the corpus annotation, extraction procedure and the nature of the idiosyncratic data.

Finding specific constructions in corpora of free word order languages such as Dutch is not trivial. Corpus annotation enriched with grammatical functions and/or dependency relations facilitates the search task.[3] Thus, we are able to explore LVC occurrences in any syntactic structure (main or subordinate sentence, questions, etc.) without stating linear precedence constraints. Furthermore, in most sentences, the annotation correctly identifies the clause containing the LVC thus, granting access to all sibblings of the head verb.

In general, knowledge of the grammar and the lexicon used by the parser is helpful. In particular, knowing whether some LVCs or idiosyncratic phrases are already annotated in the lexicon as *lexicalized phrases* helps. In the event that an LVC were described in the lexicon, the parser either analyzes the expression as an LVC or as a regular verb phrase. This uncertainty needs to be taken into account in the extraction queries.

The corpus-based method requires information about the subcategorization requirements of the LVCs. This information was manually entered for each expression. Once we have a list of PREPOSITION NOUN VERB triples, methods described in the literature on automatic acquisition of subcategorization information might be successful in finding out the remaining LVC syntactic requirements. This is an open issue for future re-

[3] Preliminary experiments were done on chunked data. A corpus-based method applied on phrasal chunks was impractical. A lot of noise needed to be manually discarded.

search, but a starting point would be the approach by (Briscoe and Carroll, 1997).

The success of the search queries is dependent on parsing accuracy. Sometimes extracted evidence shows the specific PP we seek but misanalyzed as a dependent of another verb. Parsing accuracy introduces another shortcoming: evidence of relative clauses and PP post-nominal modifiers cannot be automatically retrieved. Because of structural ambiguity, attachment decisions are still a hard parsing problem. This led us to ignore these two types of modification in our research.

Some limitations due to the nature of the support verb constructions emerged. Specifier changes or insertion of modification may destroy the LVC reading. The queries could extract evidence that looks like a variant of the LVC base form; in practice, the LVC interpretation does not apply. For example, in most of the instances of the expression *de hand boven het hoofd houden* 'to protect s.o.' (lit. the hand above the head hold), *hoofd* is preceded by the definite determiner; there are also a few instances with a reciprocal *elkaars* 'each other's' and some instances with possessive determiners. The query results suggest that all three specifiers are possible; however, the instances with possessive determiners are literal uses. Occasionally, a PREPOSITION NOUN VERB triple clusters homonymous expressions. A search that specifies the triple base form IN HAND HOUDEN could match any of the following: *iets in één hand houden* 'to be the boss', *het heft in handen houden* 'remain in control', *de touwtjes in handen houden*, *iets in handen houden* 'have control over sth' or *iets in de handen houden* 'to hold sth in one's hands (lit.)'. Access to the subcategorization requirements of the LVC use (that differs from those of the regular phrase) (e.g. *iemand van de straat houden* 'keep s.o. off the street' vs. *van de straat houden* 'to love the street') would solve some cases.

The corpus-based method cannot be fully automated; that is, extraction of variation and modification evidence cannot be done fully automatically. Instead, the evidence retrieved needs to be manually inspected. This brings up a last limitation of the method. At least one instance of each variation and modification type requires manual inspection. The researcher needs to establish whether the LVC interpretation is present or only a literal reading applies. Yet, all the tools we used facilitated this process and they provide plenty of relevant linguistic empirical evidence.

A last limitation affecting most corpus-based research is that having found no evidence of variation and modification does not mean that it is not possible in LVCs. Some LVCs are rare in the corpus; LVCs that exhibit variation and/or modification are even more infrequent. A larger corpus is desirable.

4 Lexicon representation in Alpino

The Alpino lexicon entries specify (if applicable) subcategorization frames enriched with dependency relations and some lexical restrictions. Support verb constructions and idiomatic expressions are treated similarly; neither of these expressions constitute a lexical entry on their own (cf. (Breidt et al., 1996)). We concentrate on the LVC annotation in the remainder.

Support verb constructions are lexicalized combinations of a support verb. Main verbs exhibit the same form (lemma) as their related support verb. We distinguish between a main verb and a support verb by specifying the distributional context of the support verb. This context is captured as an extended subcategorization frame. [4] An extended subcategorization frame consists of two parts: (a) list of syntactic dependents and (b) syntactic operations that the LVC (dis)allows. Among syntactic dependents, we include those lexemes and/or phrases necessary to derive the predicational content of the LVC. The syntactic dependents may be realized by three types of phrases: (i) fully lexicalized, (ii) partially lexicalized and (iii) variable argument slots. Next, the description of the phrase types is supported with expressions encountered earlier in the paper. [5]

Fully lexicalized phrases exist as individual lexical entries. No variation, modification nor extraction out of these phrases is possible. A fully

[4] This working implementation assumes that the verb selects the dependents of the LVC, thus, departing from other proposals (Abeill´e, 1995) where the complement noun selects the support verb. Although the semantics layer is left out, this approach echoes lexicalist HPSG proposals such as (Krenn and Erbach, 1994; Sailer, 2000).

[5] Each example displays the light verb followed by its syntactic dependents given within ⟨⟩. Subject is omitted.

lexicalized phrase is a string of lexemes – each in their surface form – and is represented within '[]':

```
houden   ⟨ dat,[de,hand],[boven,het,hoofd] ⟩
houden   ⟨ refl, [van,de,domme] ⟩
```

Partially lexicalized phrases declare the type of argument they introduce e.g. accusative, semi-fixed prepositional phrase, predicative argument. These phrases also specify lexical restrictions on the head lexeme and, allow alternation of specifiers and morphological productivity in nouns. Partially lexicalized PPs list the head preposition and its object NP head.

```
houden    ⟨ acc(rekening), pc(met) ⟩
brengen   ⟨ acc, pp(op,gedachten) ⟩
```

Finally, open argument slots state what sort of argument is required (e.g. acc(usative), refl(exive), dat(ive)). No lexical restrictions are declared.

```
stellen   ⟨ acc, pp(in,bewaring) ⟩
```

Concerning the syntactic behavior of LVCs, Alpino currently only declares whether the expressions allow passive or not and the type of passive. The current representation allows intervening adjuncts and other material between the syntactic dependents. No explicit constraints are stated with regards to topicalization, wh-extraction, coordination, clefting, etc.

5 Related work

Automatically annotated corpora have been used before to identify (prepositional) support verb constructions and to asses their variation and modification potential. Led by (Krenn, 2000) and continued by (Spranger, 2004) (among others), most work focused on German support verb constructions and figurative expressions. Our use of fully parsed corpora and the treebank query tool to extract relevant evidence introduces a fundamental difference with the cited work.

Analytic techniques to annotate syntactically flexible (but idiosyncratic) expressions in lexical resources are discussed in (Breidt et al., 1996; Sag et al., 2001) and (Odijk, 2004). Within a similar line of work, (Sag et al., 2001) propose lexical selection, inheritance hierarchies of constructions and the notion of idiomatic construction to formalize the syntax and semantics of truly fixed, semi-fixed and syntactically flexible expressions. Assuming a regular syntactic behavior and having checked that component lexemes satisfy certain predicate-argument relationships, the semantics layer assigns the idiomatic interpretation to syntactically flexible expressions. (Sag et al., 2001) only mention light verb plus noun constructions. Supposedly, the Dutch prepositional LVCs fall into the syntactically flexible group.

6 Conclusion and further improvements

The corpus-based method extracts evidence of variation and modification within support verb constructions. The method is sufficiently efficient in extracting proof of morphological productivity, specifier variation and adjectival modification inside LVCs, but at least one instance of each type of variation needs to be manually assessed to determine whether the LVC interpretation is present. The evidence retrieved allows us to establish the required syntactic structure, lexical restrictions and furthermore, a preliminary classification of LVCs. Our findings form the basis of the lexical annotation of these expressions in Alpino.

A few ideas to enhance the method described in order to improve the quality of the retrieved evidence follow. During compilation of the raw subcorpus, we will adapt the method so that, for each P N V triple, all verb and noun variant forms are retrieved from an existing lexicon. This ensures that the 'subcorpus compiler' collects all possible variants from the TwNC. Given that the parsed data includes dependency relations we are trying different methods to infer the complete subcategorization frame of each LVC. So far, an LVC is represented as a P N V triple, but we need to know other syntactic requirements of the predicate. Access to subcategorization frames ought to improve the extraction of variation evidence. Finally, the experiments described concentrate on support verb constructions. It is sometimes difficult to distinguish a support verb construction from an idiomatic expression. Thus, some of the expressions might perfectly belong to the idioms class, rather than the support verb construction group. A related question is how to distinguish

the literal use of triples from the support verb construction use automatically. This still needs a solution.

Acknowledgements

I would like to thank Gertjan van Noord and the three anonymous reviewers for their invaluable input on this research. This research was supported in part by the NWO PIONIER grant 220-70-001 and also the IRME STEVIN project.

References

Anne Abeill´e. 1995. The Flexibility of French Idioms: a representation with lexicalized tree adjoining grammar. In Martin Everaert, Erik-Jan van der Linden, Andre Schenk, and Rob Schreuder, editors, *Idioms: Structural & Psychological Perspectives*. Lawrence Erlbaum Associates.

T. Baldwin, J. Beavers, L. van der Beek, F. Bond, D. Flickinger, and I.A. Sag, to appear. *In search of a systematic treatment of Determinerless PPs*. Computational Linguistics Dimensions of Syntax and Semantics of Prepositions. Kluwer Academic.

Gosse Bouma and Geert Kloosterman. 2002. Querying dependency treebanks in XML. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume V, pages 1686–1691, Las Palmas de Gran Canaria, Spain.

Elisabeth Breidt, Frederique Segond, and Giuseppen Valetto. 1996. Local grammars for the description of multi-word lexemes and their automatic recognition in texts. In *COMPLEX96*, Budapest.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL conference on applied Natural Language Processing*, pages 356–363, Washington,D.C.

Miriam Butt. 1995. *The structure of complex predicates in Urdu*. Ph.D. thesis, Stanford University.

Bart Hollebrandse. 1993. Dutch light verb constructions. Master's thesis, Tilburg University, the Netherlands.

Brigitte Krenn and Gregor Erbach. 1994. Idioms and support verb constructions. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, pages 365—395. CSLI.

Brigitte Krenn. 2000. *The Usual Suspects: Data Oriented Models for the Identification and Representation of Lexical Collocations*. Ph.D. thesis, DFKI & Universitat des Saarlandes.

Rosamund Moon. 1998. *Fixed expressions and Idioms in English. A corpus-based approach*. Clarendom Press, Oxford.

Tim Nicolas. 1995. Semantics of idiom modifi cation. In Martin Everaert, Erik-Jan van der Linden, Andre Schenk, and Rob Schreuder, editors, *Idioms: Structural & Psychological Perspectives*. Lawrence Erlbaum Associates, New Jersey.

Jan Odijk. 2004. Reusable lexical representation for idioms. In *Proceedings of 4th International Conference on Language Resources and Evaluation 2004*, volume III, pages 903–906, Portugal.

R.J.F. Ordelman. 2002. Twente Nieuws Corpus (TwNC), August. Parlevink Language Techonology Group. University of Twente.

Carl Pollard and Ivan A. Sag. 1994. *Head–Driven Phrase Structure Grammar*. The University of Chicago Press, CSLI: Stanford.

Susanne Riehemann. 2001. *A constructional approach to idioms and word formation*. Ph.D. thesis, Stanford University.

Ivan Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2001. Multiword expressions: a pain in the neck for NLP. LinGO Working Paper No. 2001-03.

Manfred Sailer. 2000. *Combinatorial Semantics & Idiomatic Expressions in Head-Driven Phrase Structure Grammar*. Ph.D. thesis, University of Tuebingen.

Kristina Spranger. 2004. Beyond subcategorization acquisition. Multi-parameter extraction from German text corpora. In *Proceedings of the 11th EURALEX International Congress*, volume I, pages 171–177, France.

Leonoor van der Beek, Gosse Bouma, Jan Daciuk, Tanja Gaustad, Robert Malouf, Gertjan van Noord, Robbert Prins, and Begoña Villada. 2002. Algorithms for Linguistic Processing NWO PIONIER Progress Report. Available electronically at `http://odur.let.rug.nl/~vannoord/alp`., Groningen.

Begoña Villada Moir´on. 2004. Distinguishing prepositional complements from fi xed arguments. In *Proceedings of the 11th EURALEX International Congress*, volume III, pages 935–942, Lorient, France.

Begoña Villada Moir´on. 2005. *Data-driven Identification of fixed expressions and their modifiability*. Ph.D. thesis, University of Groningen.