Data-driven identification of fixed expressions and their modifiability
Villada Moirón, María Begoña

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2005

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 8
Conclusions and Future directions

The research questions dealt with in this thesis are:

- Can data-driven methods identify which fixed expressions require special lexical mention?

- To what extent can we determine the potential for modification of fixed expressions using a corpus-based approach?

After establishing how the linguistic behavior of fixed expressions differs from the behavior of (syntactic- and semantically) regular and productive expressions, this thesis concentrated on the automatic identification of fixed expressions and the establishment of their potential for morpho-syntactic variation and modification. Concerning the identification problem, the models were evaluated in two different tasks, the first was the acquisition of collocational prepositional phrases and the second, the acquisition of support verb constructions.

Among the various aspects that show a divergence between productive expressions and fixed expressions, there is the lexical affinity between component words. Other generalizations about the linguistic behavior of fixed expressions are difficult to establish for three reasons. First, irregularities vary from expression to expression; second, there is no agreed list of sufficient features that characterize a fixed expression nor an established definition of fixed expression. Third, earlier (more theoretical) studies on the properties of fixed expressions – in concrete, the prototypical expressions, i.e. idioms – disagree with more recent (empirical) corpus-based approaches. Two controversial issues are: the extent to which fixed expressions allow morpho-syntactic
variation and modification and whether they show internal semantic structure. Both issues affect the appropriate representation and description of fixed expressions in a lexicalist constraint-based grammar. Because ultimately a linguistic description of fixed expressions is needed, this work took side with recent corpus-based approaches and designed a corpus-based model to explore the potential of variation in Dutch support verb constructions.

The characterization of the various linguistic aspects of fixed expressions, as well as further constraints specific to Dutch collocational prepositional phrases and support verb constructions, informed the design, development and evaluation of the identification models.

A part-of-speech tagged corpus and a fully parsed corpus were used. Originally, collocational prepositional phrases (CPPs) were thought to be fixed, therefore a part-of-speech tagged corpus was used as extraction data. In contrast, constituent phrases and lexemes in support verb constructions may be non-adjacent; this led us to use a fully parsed corpus. During the extraction of candidate datasets, we tried to avoid potential errors made by the automatic annotation tools. Thus, we tried to anticipate parser errors by ignoring the head-complement dependencies proposed by the parser and instead, we made use of phrasal chunks and word level information.

In the CPP identification task, the candidate representation in the datasets is a pattern \( P \) base NP P as it is found in the corpus. The fact that the NP varies in surface form introduces a data sparseness problem affecting those CPPs that allow limited variation. In the support verb construction identification task, we corrected, to some extent, the data sparseness problem by representing a candidate support verb construction as the pattern abstraction \( v \ prep \ noun \). Nominal morphological information was kept, thus the noun exhibits its surface form; in contrast, tense inflection was discarded by lemmatizing the verb and further, other determiners and modifiers were ignored.

In both identification tasks, candidate patterns were treated as bigrams. An advantage is that common association measures could be applied. However, this introduces limitations given that all candidates need to be formatted and, association scores of expressions that include a highly frequent word are overestimated. Although the log-linear model does not encounter these limitations, the corpus was not large enough for this model.

In the first hybrid model, a frequency cutoff was applied to maximize the reliability of the statistical scores. Although a cutoff brings about an improvement in the precision, the cutoff typically affects the coverage (or recall) of the models. In the context of the second identification task, we argued that the use of a frequency cutoff is undesirable when the validation data is limited. Beforehand, we ignore the impact of the cutoff on the cover-
age of the identification model. Instead, we identified a statistic that is less sensitive to low-frequency counts.

Validation data is typically limited in the identification of fixed expressions, at least for languages other than English. Quantitative and qualitative evaluation of the performance of identification models is recommended when only limited validation data is available. Furthermore, standard precision and recall can only be approximated, thus, in the second identification task, a different evaluation methodology was introduced. In addition to using accuracy graphs (nowadays a standard technique), the novel aspect of our approach is to assess the performance of the tests with the uninterpolated average precision which indirectly measures recall also. This measure avoids the arbitrary decision of setting the size of the nbest list and, it is more informative than standard precision.

Error analysis showed that the restricted syntactic flexibility or the non-compositional meaning of support verb constructions cannot be captured by simple statistical tests. Thus, systematic errors made by the second hybrid model are discarded by applying linguistic diagnostics. This filtering mechanism exploits the idiosyncratic syntax and/or semantics of support verb constructions thus, producing an error rate decrease of 24.7% and consequently, significantly improving the accuracy of the automatic identification model.

To conclude this part, those identification models that capture morphosyntactic irregularities are more efficient than models that only take into account the lexical affinities between component words and the syntagmatic relationship between them. Nevertheless, the identification models need to be further tested with other types of fixed expressions.

Large corpora provide a potentially rich source of fixed expressions and an invaluable collection of their linguistic behavior. In a next step, a corpus query tool retrieves evidence of morpho-syntactic variation and modification in those instances of the patterns found in syntactically annotated corpora. The extracted evidence shows a distinction between various types of support verb constructions that range from totally fixed to flexible expressions. We also established that specifier variation, in particular, certain determiner changes correlate, albeit less than perfectly, with adjectival modification. Finally, we observed that specifier variation and adjectival modification in the svc's we studied do not delimit semantically decomposable expressions.

Evidence of determiner alternation, morphological variation and insertion of modifiers is crucial to improve the description of the mentioned patterns in lexical resources. A few limitations remain. In order to extract reliable evidence of modification and other morpho-syntactic restrictions, the corpus-based method assumes knowledge of the valence patterns of each expression. Furthermore, the output of this corpus exploration method needs to be manu-
ally supervised to ensure that the support verb construction interpretation is present. This means that the method cannot be fully automated at this stage. However, the method is efficient in retrieving very specific linguistic evidence from an automatically annotated corpus. This evidence can be crucial for a lexicographer or a grammar writer to decide the appropriate lexical representation of support verb constructions.

A last drawback affects corpus-based research in general. The fact that no corpus evidence is found does not guarantee that a certain linguistic diagnostic is not satisfied or that variation and/or modification within a support verb construction is not possible. Therefore, one can never be totally certain about making categorical statements such as ‘expression X belongs to the fixed class’.

The significance and achievements of this work can best be appreciated in a larger context. The cyclic investigation process that includes (i) the empirical study of the linguistic behavior of fixed expressions observed in large corpora, (ii) the incorporation of relevant features and irregularities in identification models, (iii) the qualitative assessment and error analysis of the output of the models and (iv), the incorporation of the extracted fixed expressions in a lexicalist constraint-based grammar can provide valuable evidence and insights of the linguistic behavior of these expressions. Other outcomes of this cyclic process are updated and improved lexical resources, tools to extract lexical information from corpora and means to improve the automatic annotation of large corpora. The availability of these computational resources has opened new possibilities in the field of linguistic exploration in corpus linguistics.

### 8.1 Future directions

**Descriptive studies of fixed expressions** In order to advance the linguistic analysis of fixed expressions as well as the theory of idiomatic phenomena, more corpus-based research on other types of fixed expressions is needed.

In the future, we would like to investigate other types of support verb constructions, idioms and verb particle combinations. The general purpose is to extend identification models to handle these fixed expressions types as well as to increase the coverage of existing lexical resources. Furthermore, from a more theoretical perspective we aim at determining what morpho-syntactic properties pinpoint internal semantic structure in fixed expressions.
8.1. Future directions

Identification models  Both hybrid models encounter a few limitations. First, the size of candidates in the datasets varies from 3 to 4 or more words. Although we found an abstract representation of the data that allows us to apply association measures to candidate patterns of length 3, it is desirable to have a method that generalizes to patterns of any length.

Second, the association measures do not reliably estimate the lexical affinity between highly frequent component words. This work has shown that an identification model that captures syntactic and semantic irregularities is more efficient in the task of acquiring a lexicon of support verb constructions. Thus, in the future more expressive models need to incorporate features that take into account lexical, morphological and syntactic idiosyncrasies of fixed expressions. Semantic properties are also important, however, before integrating this type of knowledge in acquisition models other developments in the construction of automatic thesauri and other ontologies are needed.

Machine learning algorithms such as maximum entropy models and latent semantic analysis provide models that set no constraint on the length of candidate patterns and at the same time can combine various sources of linguistic knowledge. This is an avenue for future research.
Chapter 8. Conclusions and Future directions