Acquisition of Phraseological Units from Linguistically Interpreted Corpora
A Case Study on German PP-Verb Collocations

Brigitte Krenn
Universität des Saarlandes
Computerlinguistik
Im Stadtwald
66041 Saarbrücken
e-mail: krenn@coli.uni-sb.de

Abstract

In this paper, we show that accessibility of syntactic information eases collocation extraction from corpora, and supports identification of lexical and structural restrictions related to collocations. For collocation identification we use a corpus that is automatically annotated applying a part-of-speech tagger and a phrase chunker.

1 Introduction

In this paper, we present corpus-based investigations into PP-verb collocations. These are verb-object constructions in which the preposition, the dependent noun and the verb are collocational. We distinguish collocations from arbitrary word combinations by means of

Lexicalization Collocations either consist of collocative (lexically fixed) elements only or comprise lexically fixed (the collocates) and lexically underspecified elements.

Lexical Cohesion Strong lexical cohesion holds between the collocates, i.e. if one collocate appears, it is highly probable that the according partner collocate(s) appear as well.

Recurrence Lexically cohesive word combinations are more likely to cooccur in language data than arbitrary word combinations. Due to this feature, corpus-based techniques are well suited for identification of collocations.

A particular property of a large number of collocations is that they show a variety of grammatical restrictions. Thus collocations are a problem for prevalent grammar theories. On the other hand, access to grammatical information is particularly useful for identification of collocations from corpora, as will be shown in this paper. We use grammatical knowledge to identify a preposition and its dependent noun, and to specify the units within which a verb and a PP may cooccur; and we use corpora to identify the actual occurrences.

2 Retrieval of Collocations from Syntactically Interpreted Corpora

2.1 Syntactically Rich Data

For retrieval of collocations, we use a syntactically annotated corpus. This contrasts the majority of current approaches to collocation extraction which are based on plain word forms or part-of-speech (pos) annotated forms, e.g. [Breidt, 1993], [Frantzi and Ananiadou, 1996], [Haruno *et al.*, 1996], [Ikehara *et al.*, 1996], [Shimohata *et al.*, 1997], [Docherty *et al.*, 1997]. In [Smadja, 1993] it has already been shown that syntactic analysis of phrases containing candidate collocations helps to discard false candidates.

The major benefits of collocation extraction from structurally annotated corpora are that (1) the accuracy of frequency counts increases, i.e. more syntactically plausible collocation candidates are found, and (2) grammatical restrictions related to collocations can be, for the most part, automatically derived from the corpus.

A precondition for such an approach is automatic annotation without hand correction because of the following two reasons. (1) Large amounts of data are required to automatically distinguish collocations from arbitrary word combinations, which makes hand correction quite infeasible. (2) Occurrence frequencies derived from corpora reflect both general language collocations and domain-specific collocations, thus automatic annotation tools are required to allow for flexibility with respect to corpus size and the domains represented.

To tackle the problem, we use a part-of-speech (pos) tagger [Brants, 1996] and a phrase chunker [Brants and Skut, 1998]. The chunker is trained on structurally annotated sentences like the one presented in figure 1. Treebank and structural annotation scheme are described in [Skut *et al.*, 1997; Brants *et al.*, 1997]. The pos-tagset is described in [Thielen and Schiller, 1995]. For training, the structural information is translated into word level tags representing the position of a word within hierarchical structure, word level syntactic category, case, and with respect to lexical elements of NPs, PPs, ADJPs and ADVPs, the category of the mother and grandmother if existing. Information on grammatical functions (HD, SB, NK, AC, ...) is omitted until more training data are available.

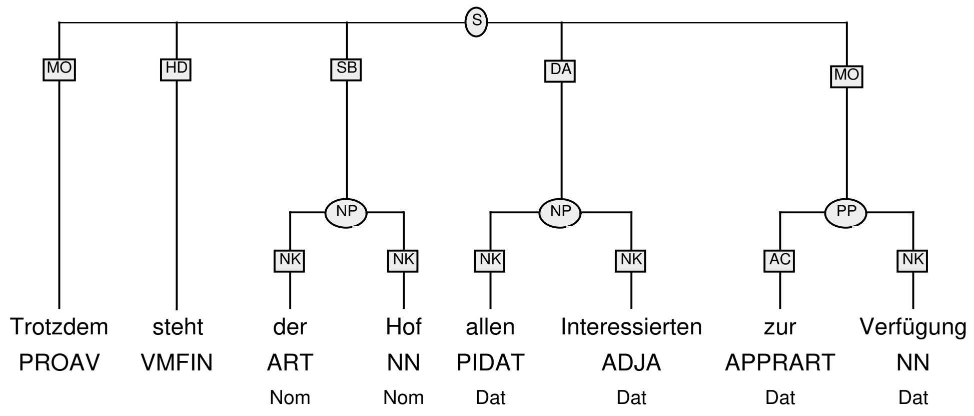

Figure 1: Syntactic information as annotated in the treebank

In our approach, the numerical notion of span[1], i.e. the number of words within which to look for a partner collocate of a word, is replaced by the notion of syntactic span, i.e. the particular syntactic structures within which the collocates of a collocation are to be found. Thus retrieval of syntactically implausible candidates

[1]Cf. [Smadja, 1993; Breidt, 1993; Frantzi and Ananiadou, 1996; Ikehara *et al.*, 1996; Shimohata *et al.*, 1997].

is avoided from the beginning. The approach also allows for more flexibility in accessing collocations, and gives the possibility to search within particular syntactic structures which, in addition, leads to a clear reduction of computation load.

2.2 Corpora

For the investigations presented in this paper we use three different types of data from the Frankfurter Rundschau corpus, a German newspaper corpus (cf. ECI Multilingual CD1).

Part-of-Speech Tagged Corpus: A sequence of 136 000 words (including punctuation) pos-tagged text has been used for training the tagger.

Structurally Annotated Corpus: 10 500 sentences annotated with syntactic structure as shown in figure 1 have been used for training the chunker.

Extraction Corpus: 18 million words from randomly selected parts of the corpus have been automatically annotated with parts-of-speech, phrasal category and syntactic structure covering NP-, PP-, ADJP- and ADVP-chunks. Collocation candidates are identified on the basis of n-gram frequencies over word-pos pairs within particular syntactic positions in a sentence.

The training corpora are partially hand corrected. The extraction corpus is left uncorrected. All tools used for annotation are based on standard techniques for Hidden Markov Model taggers. (For a tutorial on HMMs see [Rabiner, 1989], for basics on pos-tagging see [Krenn and Samuelsson, 1996].)

3 Acquisition of Collocations Containing PP-collocates

Computation of n-grams over word forms only leads to fairly poor results as function words and punctuation occur more often than content words. As a consequence, a large amount of n-grams retrieved need to be discarded. This can be avoided by using syntactic information.

Syntactic information is useful to distinguish potential collocates from arbitrary word combinations. In the case of PP-verb collocations, we have lexical cooccurrence between the PP and the verb, and within the PP itself, i.e. between the preposition and the dependent noun. In addition, there may be lexical and syntactic restrictions with respect to modification. Applying a standard n-gram approach, we would count preposition-noun bigrams within numerical spans. This method is sufficient in the case of completely fixed or highly restricted, frequently occurring PPs. If the PP-collocate is flexible, it is problematic to identify the appropriate preposition-noun combination. The same is true for PP-verb combinations.

3.1 N-Gram Frequencies

Retrieving preposition-noun bigrams from the extraction corpus, we find the following data:

- Arbitrary preposition-noun cooccurrences like *am Samstag, am Wochenende, für Kinder, für Frauen, im Monat.*
- Fixed PPs like *zum Beispiel.*
- PPs with a strong tendency for particular continuation like *auf Seite, in Sachen, im Alter, im Vergleich.*
- PP-collocates of verb-object collocations like *zur Verfügung, in Frage, in Anspruch.*

In order to identify these different types, two strategies were tested. Strategy 1: Trigram frequencies are calculated from the preposition, the noun and the word to the right. Strategy 2: Cooccurrence frequencies are calculated using structural information.

3.1.1 Trigrams over Word-Tag Strings

Applying strategy 1, we gain the following information: *auf Seite* is followed by a cardinal in 834 of 886 coocurrences, *in Sachen* is followed by a common noun in 379 of 497 coocurrences, *im Alter* is followed by a preposition in 705 of 789 cases whereby the most frequent preposition is *von* (518 occurrences) and the second most frequent is *zwischen* (164). *Zur Verfügung* is followed by a verb in 1010 of 2180 cases and more importantly the verbs reduce to forms of *stehen* (491) and *stellen* (519). Thus we have identified two verbal collocates. Similarly, we find overproportional cooccurrences of particular verbs with *in Frage* and *in Anspruch.*

3.1.2 N-grams over Selected Lexical Elements of Particular Phrases

From the data in the previous section, we learn that there is a number of occurrences where PP and verb are adjacent in the surface string. However, syntactic information is necessary to distinguish collocations from arbitrary word sequences. Both phrase boundary and part-of-speech information are needed for the identification of the preposition and its dependent noun, and to identify constructions containing a verb and its PP argument. In German, the latter consists of clauses and phrases with deverbal adjectives, see for example the following occurrences of *zur Verfügung* {*stehen, stellen*} found in the extraction corpus.

> *80 Wohnplätze stehen den Beschäftigten zur Verfügung.* (main clause, verb second)
>
> *Für jede Mannschaft werden mindestens ein Trainer und ein Betreuer zur Verfügung stehen.* (main clause, verb final)
>
> *... , die Räume auch über den regulären Kündigungstermin hinaus zur Verfügung zu stellen , ...* (infinitive clause)
>
> *So sucht die RSG einen Sponsor, der das Material kostenlos zur Verfügung stellt.* (relative clause)
>
> *... zur Verfügung gestellten Gelder seien für den Ankauf wichtiger Materialien eingesetzt worden.* (adjective phrase)

We can observe PP-verb cooccurrences even more clearly when calculating the cooccurrence frequencies of the PP and the verb forms within a phrase or clause. Looking at the 20 most frequent forms we have 1035 occurrences of *stehen* and 960 of *stellen* within 2180 sentences total containing the sequence *zur Verfügung*, see table 1. Looking at the 20 most frequent verbs cooccurring with *in Frage* (650) we have 320 instances of *stellen* and 288 of *kommen*, and for *in Anspruch* (533) we have 520 occurrences of *nehmen*.

PP-Collocate	V-Collocate	Right Neighbour	Cooccurring Main Verb
zur Verfügung (2180)	stehen	491	1035
	stellen	519	960

Table 1: *zur Verfügung* and verbal partner collocates

3.2 Lexical Cooccurrence and Grammatical Restrictions

While the sequence *zur Verfügung* occurs 2180 times in the corpus, the word form *Verfügung* occurs 2400 times. When looking at phrasal chunks containing *Verfügung*, we can distinguish the following classes.

PPs consisting of 'zur Verfügung' only. This constitutes the majority of cases. There is also a small number of occurrences of *zu Verfügung*.

Modified 'zur Verfügung'-PPs. We distinguish two kinds of modification:
(1) modification of the noun like *zur* {*eigenen, freien*} *Verfügung*.
(2) modification of the PP, like {*zeitweilig, schon, nicht, zunächst, weiter, vorübergehend, voll, sofort, nämlich durchaus, nur noch, nochmals, noch, nicht mehr, kostenlos, jedenfalls genügend, flächendeckend, bislang, bereits, bedingt, wieder*} *zur Verfügung*.

NPs containing 'zur Verfügung'-verb collocations as prenominal modifier. See table 2 for a list of occurrencies found in the extraction corpus. From these data we obtain additional evidence that *stehen* and *stellen* are verbal collocates of *zur Verfügung*.

NPs containing 'Verfügung' and PPs with alternative prepositions like *per, mit, auf, gegen, durch, über, wegen, mittels, in, für, bei, aufgrund, als, nach, laut, entgegen*. In both cases, determination and modification varies. The noun cooccurs with adjectives such as *einstweilig, vorläufig, schriftlich, entsprechend, neu, gerichtlich*. The most frequent modifier is *einstweilig*.

While the first three classes relate to the "dispositon" reading of *Verfügung* including the PP-verb collocation *zur Verfügung* {*stehen, stellen*}, the fourth class refers to the meaning "order" or "decree".

4 Relevance of the Data

Collocation data derived from a syntactically annotated corpus can be exploited for grammar theory as well as for applications. In the case of grammar theory, the data support investigations into the relation between generativity of grammar and lexical aspects of language usage. With respect to a theory of collocations, the data can be used for classification of collocations according to their syntactic and semantic distribution in particular corpora, and according to the grammatical restrictions related to the individual collocations.

Knowledge of lexical restrictions to grammar is also useful in computational linguistics. It is useful in parsing where it reduces search space and resolves attachment ambiguities. In generation, it improves the adequacy of utterances. In addition, knowledge on domain-specific collocations is useful for applications like message extraction and data mining.

In the field of lexicography, the possibility of automatic exploitation of arbitrary text corpora supports the construction and update of domain-specific and general language dictionaries.

5 Conclusion

We have shown that syntactically rich data are well suited for collocation extraction and identification of lexical and structural restrictions related to collocations. Structural information allows to identify with high accuracy whether a noun tends to occur within PPs or NPs, and which lexical and structural restrictions hold with respect to modification. In the case of verb-object collocations, syntactic information is useful for identification of the verbal collocate. We get more reliable frequency counts especially if the verb is in sentence initial or in second position. The evidence that a particular verb is a collocate in a verb-object collocation is strengthened by the occurrence of the verb-PP/NP combination as prenominal modifier.

The approach supports identification of a variety of collocations as both lexical recurrence and modification restrictions are general indicators for collocations. The advantage of collocation extraction from a structurally annotated corpus is that calculation of lexical coocurrence is based on frequency counts of words within particular syntactic positions and not just within arbitrary numerical spans. Access to syntactic structure is even more important for identification of modification restrictions.

Current shallow approaches to text processing like part-of-speech tagging and chunk parsing are already reliable enough to function as tools for automatic annotation, even though they produce incomplete annotations. Using these data, we get clearer extraction results than with prevalent approaches operating on n-grams over words or word-tag sequences within numerical spans.

References

[Brants and Skut, 1998] Thorsten Brants and Wojciech Skut. Automation of Treebank Annotation. In *Proceedings of NeMLaP-98*, Sydney, Australia, 1998.

[Brants *et al.*, 1997] Thorsten Brants, Wojciech Skut and Brigitte Krenn. Tagging Grammatical Functions. In *Proceedings of EMNLP-97*, Providence, 1997.

[Brants, 1996] Thorsten Brants. TnT – A Statistical Part-of-Speech Tagger. Technical Report, Universität des Saarlandes, Computational Linguistics, 1996.

[Breidt, 1993] Elisabeth Breidt. Extraction of N-V-Collocations from Text Corpora: A Feasibility Study for German. In *cmp-lg/9603006, revised version of paper presented at 1st ACL-Workshop on Very Large Corpora, 1993*, 1993.

[Docherty *et al.*, 1997] Vincent J. Docherty, Ulrich Heid and Judith Eckle-Kohler. Computational linguistic support for the revision of a commercial dictionary – corpus-based updates of entries and of collocation information. In *Beiträge zur 6. Fachtagung der Sektion Computerlinguistik DGfS-CL*, Heidelberg, Deutschland, Oktober 1997.

[Frantzi and Ananiadou, 1996] Katerina T. Frantzi and Sophia Ananiadou. Extracting Nested Collocations. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 41 – 46, Copenhagen, Denmark, 1996.

[Haruno *et al.*, 1996] Masahiko Haruno, Satoru Ikehara and Takefumi Yamazaki. Learning Bilingual Collocations by Word-Level Sorting. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 525 – 530, Copenhagen, Denmark, 1996.

[Ikehara *et al.*, 1996] Saturo Ikehara, Satoshi Shirai and Hajime Uchino. A Statistical Method for Extracting Uniterrupted and Interrupted Collocations from Very Large Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 41 – 46, Copenhagen, Denmark, 1996.

[Krenn and Samuelsson, 1996] Brigitte Krenn and Christer Samuelsson. A Linguist's Guide to Statistics. Technical Report, University of the Saarland, Dept. of Computational Linguistics, 1996.

[Rabiner, 1989] Lawrence R. Rabiner. Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE 77(2)*, pages 257–295, 1989.

[Shimohata *et al.*, 1997] Sayori Shimohata, Toshiyuki sugio and Nunji Nagata. Retrieving Collocations by Co-Occurrences and Word Order Constraints. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, 1997.

[Skut *et al.*, 1997] Wojciech Skut, Brigitte Krenn, Thorsten Brants and Hans Uszkoreit. An Annotation Scheme for Free Word Order Languages. In *Proceedings of ANLP-97*, Washington, 1997.

[Smadja, 1993] Frank Smadja. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19:143 – 177, 1993.

[Thielen and Schiller, 1995] Christine Thielen and Anne Schiller. Ein kleines und erweitertes Tagset fürs Deutsche. In *Tagungsberichte des Arbeitstreffens Lexikon + Text 17./18. Februar 1994, SchloßHohentübingen. Lexicographica Series Maior*, Tübingen, 1995. Niemeyer.

Table 2: *zur Verfügung*-verb collocations as prenominal modifiers

Tokens	Occurrence Frequency
zur verfügung stehenden mittel	14
zur verfügung stehenden möglichkeiten	4
zur verfügung stehenden mittel	3
zur verfügung stehenden plätze für erstaufnahme	2
zur verfügung stehenden plätze	2
zur verfügung stehenden möglichkeiten und mittel	2
zur verfügung stehenden möglichkeiten des managements	2
zur verfügung stehenden auswechselmöglichkeit	2
zur verfügung gestellten beträge	2
zur verfügung gestellte summe	2
zur verfügung gestellter lkw der fürstlichen brauerei wächtersbach	1
zur verfügung gestellten zimmer	1
zur verfügung gestellten raum	1
zur verfügung gestellten lastwagens	1
zur verfügung gestellten jährlichen mittel von 150 millionen mark	1
zur verfügung gestellten gelder	1
zur verfügung gestellten finanzmittel	1
zur verfügung gestellte manuskripte	1
zur verfügung gestellte geld	1
zur verfügung stehender verbreitungsmittel	1
zur verfügung stehenden zwischenstopp-zeit	1
zur verfügung stehenden zeit	1
zur verfügung stehenden wassermenge der kinzig	1
zur verfügung stehenden truppen	1
zur verfügung stehenden trainingszeiten	1
zur verfügung stehenden straßen und rohstoffe	1
zur verfügung stehenden statistiken	1
zur verfügung stehenden politischen und rechtsstaatlichen mitteln	1
zur verfügung stehenden mittln	1
zur verfügung stehenden mitteln	1
zur verfügung stehenden kader	1
zur verfügung stehenden hortplätze in obertshausen	1
zur verfügung stehenden fz-mittel	1
zur verfügung stehenden bundesmittel	1
zur verfügung stehenden baurechtlichen maßnahmen	1
zur verfügung stehenden anschriften der älteren bürger	1
zur verfügung stehenden 10 000 mark	1
zur verfügung stehende waffe	1
zur verfügung stehende unterlage	1
zur verfügung stehende prominente personal	1
zur verfügung stehende neue vierspurige autobahn	1
zur verfügung stehende mittel	1
zur verfügung stehende kontingent	1
zur verfügung stehende etat	1