Ulrich HEID, Institut für maschinelle Sprachverarbeitung – Computerlinguistik, Universität Stuttgart

Towards a corpus-based dictionary of German noun-verb collocations

Abstract

We[1] describe our attempts to automatically extract raw material for a dictionary of German noun-verb collocations from large corpora of newspaper text. Such a dictionary should be *about collocations* and it should include a description of their linguistic properties, rather than listing the mere lexical cooccurrence.
Since most statistical collocation finding tools do not provide other than lexical cooccurrence information, we first use symbolic extraction tools, based on a regular grammar over part-of-speech tagged and lemmatized text, and we use statistical filters thereafter.
We first list the types of information which should be contained in a collocational dictionary for Natural Language Processing, then sketch our extraction methods and finally discuss and illustrate our initial results.

Keywords: Collocations, text corpora, semi-automatic lexical acquisition.

1. Introduction: motivation and objectives

Other than for English (and, to some extent, for French[2]), there is no major collocational dictionary for German. The only available ones are either rather small (such as (Duden 1988)), or they cover different types of lexical combinatory phenomena, without distinguishing them (cf. (Agricola *et al.* 1988)).

There is a need for information about German collocations, not only in printed dictionaries, but even more so in Natural Language Processing (NLP): any broad coverage grammar needs a detailed lexicon, which must contain a realistic number of collocations and the appropriate linguistic descriptions.[3] Such a dictionary is needed, because it helps to rule out spurious syntactic analyses which are only due to rule interaction (for examples, see (Heid 1998)).

Constructing such a dictionary manually is a difficult and potentially error-prone task; we thus aim at a procedure, where raw material ("collocation candidates") is extracted automatically from large text corpora (typically a total of 300 million words), and the lexicographer then manually selects those candidates which should go into the targeted dictionary.[4] The entries for the selected items are then produced automatically, by means of reformatting.[5] The use of corpora is in line with the assumption that collocations are a phenomenon of language use: where else could we capture collocations in use with relatively simple electronic means, if not in corpora[6]?

In our view, a collocational dictionary – at least for learners and for NLP – should not only be a *list of* collocations, but also a source of information *about* collocations: concentrating on noun-verb-collocations[7], we will list the types of information needed for the intended applications, and we will show how the (morpho-)syntactic part of this information can be provided by corpus exploration. For light verb constructions (*Funktionsverbgefüge*), more syntactic information on collocations can be extracted from corpora than we can find in most existing dictionaries.

2. Information types in a dictionary of German noun-verb collocations

A collocational dictionary should be not only a dictionary *of collocations*, but also a dictionary *about collocations*. Listing relevant collocations is important to identify the lexical aspects of collocations ("which words go together?"), but this is not sufficient to describe these combinations for a foreign language learner to use them appropriately, or for an NLP program to analyze or generate collocational text. More information is necessary, at the lexical, (morpho-)syntactic, semantic and pragmatic level.

2.1. The starting point

The view of collocations underlying this work is in the tradition of HAUSMANN, BERGENHOLTZ and MEL'CUK. Our starting point is constituted by the following elements of a definition of noun-verb collocations in general and light verb constructions in particular:

- Collocations involve two lexemes ((Hausmann 1989)), plus "grammatical words", such as determiners, prepositions, etc. (cf. (Bergenholtz/Tarp 1994:407)). For ease of reference, we adopt HAUSMANN's distinction between base and collocate (now more generally seen by (Hausmann 1997) as an instance of the distinction between Autosemantikon and Synsemantikon; the difference is also reflected by MEL'CUK's distinction between key word and value of the lexicalfunction).
- The two items participate in a well-formed grammatical construction; with respect to the categories of the items involved, we distinguish *noun*-noun, *noun*-verb, *noun*-adjective, *verb*-adverb, *adjective*-adverb combinations (cf. (Hausmann 1989), *bases* (in HAUSMANN's terms) in italics).
- In *noun*-verb collocations, the noun may be a subject, a complement, or sometimes an adjunct of the verb (cf. (Bergenholtz/Tarp 1994:407f, Helbig 1984); in the literature on German light verb constructions, often, only verb-complement cases are discussed (noun as object, prepositional object), whereas we follow MEL'CUK, whose Lexical Function FUNC typically used for subject-verb collocations is on a par with OPER and LABOR, used to describe light verb constructions of the verb-complement type.
- HAUSMANN and MEL'CUK use semantic and pragmatic criteria (semantic polarity and determination relation; regularity/schematicity of the collocation's semantics; disponibility "en bloc" in native speakers, "déjà-vu-effect", etc.) to distinguish typical (i.e. dictionary-relevant) collocations from trivial lexical combinations (cf. *pay attention*, collocational, vs. *pay the amount*, trivial).

  This distiction is sometimes hard to apply in the lexicographer's daily task of selecting dictionary-relevant items; we tend, here, to follow (Bergenholtz/Tarp 1994:407), who suggest that frequency in a corpus is usable as an important piece of additional information in this decision process. In addition we are experimenting with statistical relevance measures which may better capture the intuitions about "frequent combinations".
- MEL'CUK has provided, with the Lexical Functions (LFs) of the Meaning↔Text-Model an approach to the semantic description of collocations. (Fontenelle 1997) has annotated LFs in his dictionary. We are, due to the nature of this – exclusively syntax-based – extraction

exercise, unable to provide this information, but we see it as an important next step in the construction of a collocational dictionary.

- (Helbig 1984 and others classified the different types of light verb constructions, roughly according to idiomaticity (HELBIG: "lexikalisierte" vs. "nicht-lexikalisierte" Funktionsverbgefüge). The distinction is partly correlated with referential availability of the noun and with its determination properties (cf. (Bausewein 1990)). These properties are identifiable in corpora.

2.2. An outline of a descriptive model

If we accept that collocations are combinations of two lexemes, and that they together form a linguistic object which needs to be described in a dictionary, it is consequent to also accept that in total there are three kinds of objects to be described in a collocational dictionary: the base (in our case the noun), the collocate (the (light) verb) and the collocation as a whole; if there are no space restrictions, it makes sense to ensure that collocations have lemma status in a dictionary (cf. (Heid 1994:241)). Moreover, a multi-level description, involving the levels of morphosyntax, syntax, semantics and pragmatics, in addition to the merely lexical level, is necessary. In figure 1, we summarize the main types of information needed at each level, and we illustrate each information type with an example, from German light verb constructions.

#	Obj.	Level	Property	Example
1	Col.	lexical	lexeme cooccurrence	*Rede + halten*
2		syntactic	subcat. of collocation	*In der Lage sein zu + INF*
3		semantic	aktisonsart, aspect	"INCEP $OPER_1$ (*Ansicht*)"
4			synonymy	*ins Schwitzen kommen | geraten*
5		pragmatic	diasystematic marks	*Sorge tragen zu* (swiss)
6			frequency	f (*Vorschlag machen*) >> f (*Vorschlag unterbreiten*)
7	N.	morphosynt.	number	(jmdm.) *Grenzen setzen* (plur.)
8		syntactic	determination	*zu der Ansicht gelangen, dass...* definite, non-fused article (**zur Ansicht gelangen, dass*)
9			modification potential	*eine ? [] | gute Entwicklung nehmen*
10			subcategorization	*zu der Ansicht kommen, dass...*
11	V.	morphosynt.	active/passive	*Eile ist geboten* (passive)
12			tense preference	
13		syntactic	subcategorization	*zu der Ansicht kommen*

Figure 1: Information types to be included in a dictionary of German noun-verb collocations

Not all of the information types may be evident. Collocations may as such have subcategorized complements (i.e. act as complex predicates, cf. (Bausewein 1990), etc.): *in der Lage sein* subcategorizes for a *zu*-Infinitive, whereas none of its components does so; this information must be kept separate from the description of the subcategorization of the verb or

of the noun. Collocations may as such need to be labelled diasystematically: *zu etw. Sorge tragen* ("care about") is very frequent in Switzerland, but quite uncommon in Germany. Some collocations come with number restrictions on the noun: *in die Gänge kommen* requires the plural, whereas *in Gang kommen, Protokoll schreiben* requires the singular noun; the latter without a possible mass noun reading. We accept the need for keeping track of the nounform, but we still suppose that collocation refers to lemmata, only preferring certain forms over others.

2.3. Information types available in other dictionaries and through corpus extraction methods

The above list of types of information for a dictionary of light verb constructions is by no means new. Many researchers (cf. (Helbig 1984), (Persson 1975) etc.) have indicated the need to keep track of some or all of these. But dictionaries, so far, mostly give only very little information on collocations[8], in many cases nothing but the lexical cooccurrence statement itself (number 1 in figure 1 above), the category of base and collocate, and the grammatical function of the noun (13: e.g.: ROBERT/COLLINS, (Cohen 1986)). MEL'CUK's ECDs are prominent examples of dictionaries that contain a semantic description of the collocation (3, 4: via Lexical Functions) and determination information for the noun (8). Dictionaries which give collocational examples contain this information implicitly (cf. (Heid 1998, forthcoming)). Most of the more detailed information is not available in larger quantities: the ECDs are limited in coverage, and only (Ilgenfritz *et al.* 1989) is a sizeable collocational dictionary.

Published corpus-based extraction methods for collocations also often only address a subset of the information types needed. Since the work of (Church *et al.* 1991), statistical measures like Mutual Information (MI) and t-score are generally accepted as simple to use collocation extractors. This is true for the *lexical description* of the collocation only (number 1 in our table above). MI and t-score do not themselves lead to any information about the subcategorization of the light verb (with/without preposition, subject/complement, cf. number 13 of figure 1.). And moreover, (Breidt 1993) has shown that German separable verb prefixes are a major stumbling block for adjacency- and window-based extraction, as used in MI and t-score implementations.

Only (Smadja 1993)'s work, and, in particular, the extraction work based on low-level parsing described by (Grefenstette/Fontenelle/Heid 1996) has the potential to provide morpho-syntactic and syntactic information on nouns (number 10 in our table) and verbs (number 11, 13): our work is thus based on a symbolic approach which simulates low-level parsing. It grew out of the work described in (Grefenstette/Fontenelle/Heid 1996). An integrated approach, which brings together robust parsing and statistical measures, has been advocated by (Krenn 1998).

3. Extraction procedures for collocations

3.1. Corpus pre-processing

The analysis of the German corpora relies on standard tools and methods for low-level processing. The corpora are tokenized (word and sentence boundaries) and part-of-speech

tagged with the STTS tagset[9] using SCHMID's decision tree tagger.[10] The tagging process includes lemmatization, based on morphological and part-of-speech information.

3.2. Corpus queries for collocations

We mainly use query templates to extract evidence for noun-verb combinations. These are similar to queries used in a concordancing tool. A simple frequency filter, operating on the output of the symbolic extraction procedures, ensures that the listing of results contains only those lexeme combinations which appear at least 2, 5, 10, 20... times (threshold set by the user) in the analyzed corpus. Among the candidates thereby retained, the more significant (and indeed more collocational) ones can be separated from the trivial ones by use of statistical relvance measures.

We use the CQP/XKWIC corpus query tools (see (Christ 1994)), a toolbox which supports regular expressions over word forms and annotations of any type as well as set operations on the extraction results. The extraction templates (i.e. possibly complex queries with variables) make use of information about sentence boundaries, sequencing and adjacency of word forms, lists of lemmas (e.g. for function words), and boolean expressions over word forms, lemmas and/or part-of-speech shapes.

The query templates extract contexts where a verb (a potential collocate) appears at the right sentence boundary: this includes all subordinate clauses with verb-final word order *(weil ... eine Rede gehalten hat)* and verb-second cases with finite auxiliaries *(... kann ... eine Rede halten)*. To capture all relevant cases (all tenses, all relevant combinations with auxiliaries) several partial part-of-speech-shape descriptions need to be combined. The extraction templates furthermore make use of the empirical fact that mostly the base noun is found to the immediate left of the verb (complex): only adverbs and noun complements can intervene.[11] Although verb-first and some verb-second contexts are not exploited, in this setup, it is useful not to enlarge the basis of raw material: inmain clauses, full parsing would be needed to identify verb-complement pairs, whereas a "local grammar" is fully sufficient in our setup. The verb (collocate) and the nominal head of the preceding noun (or prepositional) group (i.e. of the base) are identified in the sentences extracted, and their cooccurrence is counted.

3.3. Automating the extraction

The query templates are organized in a hierarchy, thus partitioning the corpus into subsets of sentences which all display a homogeneous behaviour with respect to the following distinctions:

- reflexive vs. non-reflexive verbs[12];
- "prepositional" vs. "accusative/dative" constructions (*im Vordergrund stehen* vs. *eine Frage stellen*); property no. 13 of figure 1);
- details of the noun group, at the levels of determiners (no determiner vs. definite vs. indefinite; property no. 8 of figure 1), of adjectival modification of the noun (*ein jähes Ende finden*, property no. 9), and of the presence or absence of genitives and/or prepositional phrases to the right of the noun (*im Zusammenhang mit x stehen*, properties 2, 10).

The hierarchy corresponds to a sequence of corpus queries, which lead to the extraction, merging and complement building of subsets of corpus sentences. This sequence can be called automatically, on any text pre-processed as described in section 3.1. To this end, we use the macroprocessor for the CQP language by (Schulze 1996).

The subclassification into reflexive/non-reflexive is necessary for the lexical characterization (property 1 in figure 1) of the collocate verb. The (morpho-)syntactic verb properties (numbered 11 and 12, above) are in principle accessible through the modelling of the verb complex, but have not yet been exploited. As mentioned above, no semantic information (property 3) can be gathered with our means from the corpora. We use frequency counts to identify the most frequent collocations, and list, for each lexical combination, the most frequent morphosyntactic forms (properties 7, 8, 9).

4. Linguistic and lexicographic uses of the extracted raw material

The procedures sketched out above are the first step of three in a more detailed scenario. It involves (1) collocation candidate identification and broad classification; (2) more detailed morphosyntactic and syntactic analysis of collocation candidates, and (3) an analysis of the correlation between lexical classes, collocational behaviour and syntactic properties.
Our extraction procedures provide evidence for all kinds of phenomena on the cline from free combinations, selectionally restricted combinations, more or less idiomatic collocations through to variable and fixed idioms. Heuristically, we assume that, the more restricted the (morpho-)syntactic potential, the more likely the linguistic object in question is idiomatic; we are aware to miss out a whole range of idioms here, but the extraction is anyway not supposed to produce an idiom collcetion. Rather, we need to accept the fact that there is no automatic way of cutting up the lists of candidates into "trivial" vs. "collocational", except frequency and statistical relevance measures. See the examples in 2, where the singular/plural distinction matters for that purpose. The last step is the correlation of different types of data and the use as raw material for a broad semantic classification.

Bringen	Gang in	vlast_p_n	334	HGC
halten	Gang in	vlast_p_n	80	HGC
kommen	Gang in	vlast_p_n	210	HGC
setzen	Gang in	vlast_p_n	515	HGC
bringen	Gänge in	vlast_p_det_n	5	HGC
kommen	Gänge in	vlast_p_det_n	26	HGC

Figure 2: Collocations vs. idiomatic expression: *In Gang kommen | setzen* ... ("get | put ... in motion") vs. *in die Gänge kommen* ("get organized"); HGC in the listing is the name of the 200 million word corpus used.

4.1. Using the data as raw material for lexicography

Interestingly, light verb constructions come out with highest frequency, if the verbs cooccurring with a given noun are sorted by frequency (cf. Grefenstette/Fontenelle/Heid 1996). 3 shows a few examples with the base *Risiko* which illustrate this fact (this table lists the verb lemma, the noun form, the syntactic subcategorization of the verb ("nop" stands for

"no preposition", indeed for direct/indirect objects), the morphosyntax of the noun group, the frequency and the corpus name).

ein#gehen	Risiko	nop	vlast_det_n	66	HGC
ein#gehen	Risiko	nop	vlast_det_adj_n	23	HGC
tragen	Risiko	nop	vlast_det_adj_n	11	HGC
ein#gehen	Risiko	nop	vlast_ein_n	11	HGC
tragen	Risiko	nop	vlast_det_n	9	HGC
begrenzen	Risiko	nop	vlast_det_n	7	HGC
dar#stellen	Risiko	nop	vlast_det_adj_n	6	HGC
Über#nehmen	Risiko	nop	vlast_det_n	4	HGC
vermeiden	Risiko	nop	vlast_det_n	4	HGC
dar#stellen	Risiko	nop	vlast_ein_n	3	HGC

Figure 3: Frequency sorting of a few collocations with the base *Risiko*: frequency figures refer to determination types in verb-last sentences

The list of nouns appearing in the same type of collocational construction, and with the same collocate verb is interesting for the lexicographer who wishes to compare the collocational behaviour of semantically related nouns (cf. the work done, without corpus basis, by (Mel'cuk/Wanner 1994)). A simple example are nouns which cooccur with the verb *einschlagen*: the nouns in the upper line illustrate the light verb use found in the corpus, whereas the lower line illustrates the literal use (manual sorting):

ein#schlagen: Gangart, Kurs, Laufbahn, Richtung, Weg, Wege
ein#schlagen: Fenster, Scheibe

More generally, the patterning of nouns with certain collocate verbs is an interesting piece of information in view of the construction of semantically oriented (collocational) dictionaries. So far, we have no automatic tools for that purpose.

4.2. Genericity of the tools – use of the data for linguistic research

The procedures used in our extraction toolbox are completely generic. We are thus able to use them on any text preprocessed according to the procedures described in section 3.1. We have, for example, compared the results obtained on newspaper text (204 million words, minimal occurrence of the collocations: 10 times) for the nouns *Termin* and *Treffen* ('date', 'meeting') with those obtained on the VERBMOBIL dialogues (ca. 1 million, minimal occurrence: 2): since the VERBMOBIL dialogues all deal with agreeing on a date for a meeting, clearly VERBMOBIL offers more diversity and better coverage of this collocational area; but the newspaper has also e.g. *der Termin platzt* ('is cancelled'), a situation not allowed in the VERBMOBIL scenario.

Provided corpora are available, a collocational exploration of texts of different kinds becomes possible with the tools described here.

The collection of more material on collocations is also an important step towards a better empirical foundation of linguistic work on collocations: some of it had to be rather

anecdotical in the past, and with the availability of larger amounts of data, we expect that it will become easier to describe German collocations and to verify descriptive hypotheses.

5. Further work

The raw material at hand allows for more detailed analyses: combinations of noun-verb and noun-adjective collocations (*ein biblisches Alter erreichen, in ADJ Verhältnissen leben, eine ADJ Entwicklung nehmen*, etc.), the use of possessives in collocations (*sein Veto einlegen*), and other questions can be approached on the basis of the available material.

Moreover, a quantitative assessment of the quality of our extraction routines is still outstanding. In particular the question needs to be addressed, whether the exclusion of verb-first contexts has an impact on the proportions of material extracted. We assume that it does not have one, but this needs to be verified.

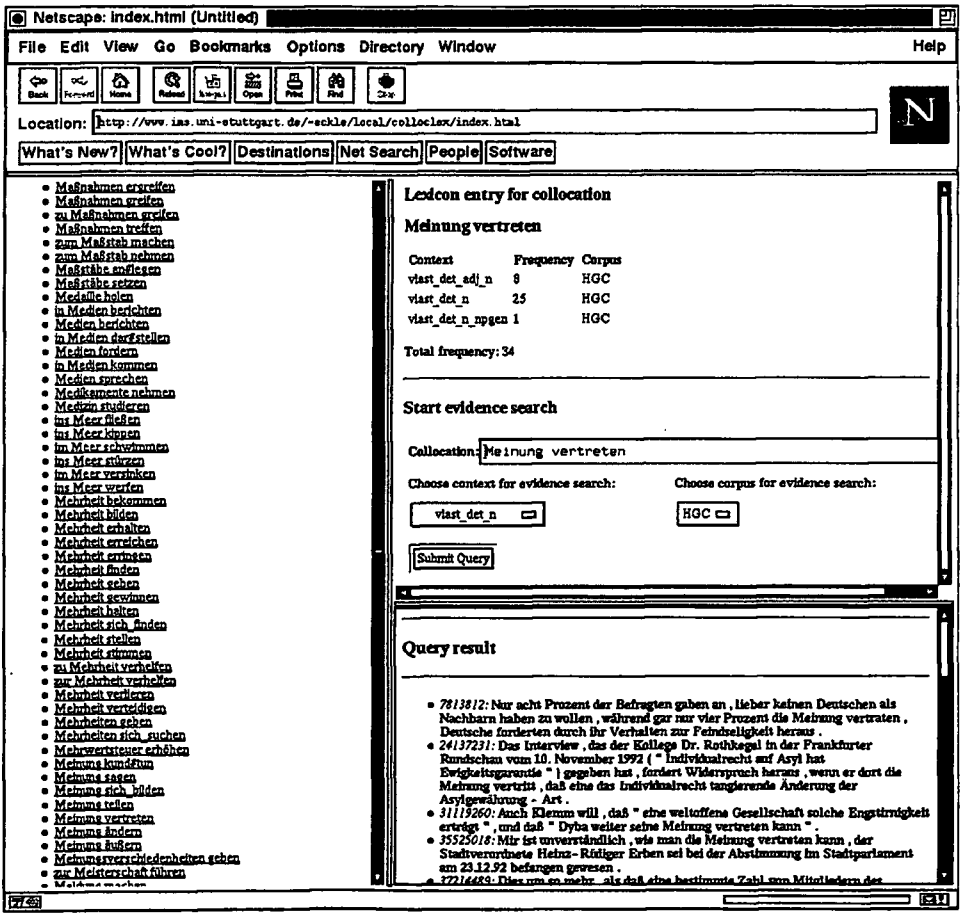

Figure 4: GUI for interactive work with the corpus-based collection of collocation candidates

Finally, raw material for other collocation types is currently also being extracted: in particular, we are working on noun-adjective and adjective-adverb collocations.

6. Notes

1 The work described here was done in collaboration with Judith ECKLE-KOHLER, who designed and implemented the extraction routines. In addition, Jonas KUHN and Carme COLOMINAS participated in discussions about earlier versions of this paper. Many thanks to all of them.

2 For English, the COBUILD collocations dictionary on CD-ROM, the BBI dictionary, (Benson *et al.* 1986), and (Kozlowska/Dzierzanowska 1993) are most well-known. For French, a collocational dictionary for German learners has been published by (Ilgenfritz *et al.* 1989. The Translation Bureau of the Canadian Public Works and Government Services is a major producer of (bilingual French/English) collocational dictionaries for sublanguage, where the main part of the dictionary is French, and an English index is provided; examples are (Lainé 1993) and (Pavel/Boileau 1994).

3 In our context, a broad coverage grammar of German in the framework of *Lexical Functional Grammar* (LFG, cf. (Dalrymple/Kaplan/Maxwell III/Zaenen 1995), (Butt/Fortmann/Rohrer 1996)) is the intended application for which the collocational component of a lexicon has to be provided. (Fontenelle 1997) has been, as far as we can see, the first researcher to develop a substantial collocational dictionary which can also be used by computational tools. His dictionary – derived from the ROBERT/COLLINS bilingual dictionary – has the big advantage of including semantic descriptions and of being bilingual.

4 To this end, a WWW-based interface is offered, where corpus examples for each identified collocation are extracted and displayed on demand; a screenshot is reproduced in figure 4.

5 There may be several more steps between the corpus-based extraction and the formatting of the entries. In particular, the lexicographers' selection work also has to include measures for quality control. This aspect (of "lexicon engineering") is only recently starting to receive attention. See now (Eckle-Kohler 1998), for quality control in a corpus-derived syntactic dictionary.

6 As our corpora come exclusively from news stories, clearly, the resulting raw material will reflect journalistic use; the extraction tools are however generic, such that they can be used on other corpora as well, if these become available.

7 We are currently also working on A+Adv and N+A collocations; N+V collocations are however more challenging because they are usually not found adjacently and thus require more sophisticated extraction routines.

8 See also the results of an analysis of English dictionaries by (Bahns 1996).

9 STTS stands for Stuttgart-Tübingen TagSet. STTS is compatible with and trivially mappable onto the EAGLES morphosyntax specifications ELM-DE (cf. (Teufel/Stöckert 1996)). It contains 54 tags with categorial, distributional and lexical distinctions (see (Schiller/Teufel/Thielen 95)). Tagging accuracy is around 97% with the decision tree tagger.

10 See http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html.

11 HERINGER, PERSSON and others pointed out that this is not a criterion to distinguish light verb constructions from "trivial" combinations; however, we are not aware of many exceptions. Thus we can use this modelling in our discovery procedures, knowing that non-collocational noise must be separated out by other means. Example of an adverb:..., *ob Kovacs seine Rede überhaupt halten würde* (1 against 350 examples of *Rede* + *halten* without adverb, in a 100 M word corpus).

12 *Haben* and *sein* may also appear in collocations (*Angst haben*); these cases are captured by a separate template set.

7. References

Erhard Agricola (Ed.): *Wörter und Wendungen*; Wörterbuch zum deutschen Sprachgebrauch; VEB Bibliographisches Institut Leipzig, 1988.

Jens Bahns: *Kollokationen als lexikographisches Problem. Eine analyse allgemeiner und spezieller Lernerwörterbücher des Englischen.*; Lexicographica Series Maior 74; Niemeyer, 1997.

Karin Bausewein: *Akkusativobjekt, Akkusativobjektsätze und Objektsprädikate im Deutschen. Untersuchungen zu ihrer Syntax und Semantik*, Tübingen: Niemeyer, 1990.

Morton Benson, Evelyn Benson, Robert Ilson: *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*, Amsterdam, Philadelphia 1986.

Henning Bergenholtz, Sven Tarp: *Mehrworttermini und Kollokationen in Fachwörterbüchern*, in (Schaeder/Bergenholtz (Eds.) 1994): 385 – 419.

Elisabeth Breidt: "Extraction of v–n–Collocations from Text–Corpora: A Feasibility Study for German". In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives. 2.6.1993, Ohio State University, Columbo.* Association for Computational Linguistics, 1993.

Miriam Butt, Christian Fortmann, Christian Rohrer: "Syntactic Analyses for Parallel Grammars: Auxiliaries and Genitive NPs", in: *Proceedings of COLING 1996*, Copenhagen, 1996.

Oliver Christ: "The XKwic User Manual", internal report, Stuttgart: IMS, 1994.

Kenneth Ward Church et al.: "Using Statistics in Lexical Analysis", in: Uri Zernik (ed.): *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 115–163.

Jeremy Clear: "I can't see the sense in a large corpus", in: (Kiefer/Kiss/Pajzs (Eds.) 1994): pp. 33–48.

Betty Cohen: *Lexique de cooccurrents; Bourse – conjoncture économique*; Montréal, Linguatech, 1986.

Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell, Annie Zaenen (eds.). *Formal Issues in Lexical-Functional Grammar*, CSLI Publications, Stanford, CA, 1995.

Dudenredaktion (Eds.): *Duden; Stilwörterbuch der deutschen Sprache; die Verwendung der Wörter im Satz*, Duden Band 2; Dudenverlag Mannheim, 1988.

Judith Eckle-Kohler: "Methods for quality assurance in semi-automatic lexicon acquisition from corpora"; in: *Proceedings of the EURALEX 1998 International Congress (Liège)*, 1998.

Thierry Fontenelle: *Turning a Bilingual Dictionary into a Lexical-Semantic Database*; Lexicographica Series Maior 79; Niemeyer, 1997.

Gregory Grefenstette, Thierry Fontenelle, Ulrich Heid: "The DECIDE Project: Multilingual Collocation Extraction", in: *Proceedings of the 7th Euralex International Congress* 1996, Göteborg, 1996.

Franz Josef Hausmann: "Le dictionnaire de collocations", in: *(Hausmann et al. (Ed.) 1989): Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch*, (Berlin: de Gruyter): 1010-1019, 1989.

Franz Josef Hausmann: "Semiotaxis und Wörterbuch"; in: (Konerding/Lehr 1997), pp. 171-179.

Ulrich Heid: "Décrire les collocations – deux approches lexicographiques et leur application dans un outil informatisé", in: *Terminologie et Traduction*, 2-3, 1992.

Ulrich Heid: "On Ways Words Work Together – Topics in Lexical Combinatorics", in: W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, P. Vossen (Eds.), *Euralex 1994, Proceedings*, Amsterdam, 1994, pp. 226-257.

Ulrich Heid: "Building a dictionary of German support verb constructions from text corpora"; in: *Proceedings of the First International Conference on Language Resources and Evaluation*; Granada, 28-30 May 1998.

Ulrich Heid: "Finding hidden collocations – a computational analysis of a computational dictionary", to appear in: *Proceedings of the IXth International Symposium on Lexicography at Copenhagen University*, (Tübingen: Niemeyer).

Gerhard Helbig: "Probleme der Beschreibung von Funktionsverbgefügen im Deutschen", in: Gerhard Helbig: *Studien zur deutschen Syntax*, Bd.2, (Leipzig) 1984.

P. Ilgenfritz, N. Stephan-Gabinel, G. Schneider: *Langenscheidts Kontextwörterbuch Französisch – Deutsch*, Langenscheidt, Berlin, München, 1989.

Ferenc Kiefer, Gabor Kiss, Julia Pajzs (Eds.), *Papers in Computational Lexicography – COMPLEX '94*, Proceedings of the 3rd International Conference on Computational Lexicography, Budapest, Hungary, (Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences), 1994.

Klaus-Peter Konerding, Andrea Lehr: *Linguistische Theorie und lexikographische Praxis*; Symposiumsvorträge, Heidelberg 1996; Lexicographica Series Maior 82; Niemeyer 1997.

Christian Douglas Kozlowska, Halina Dzierzanowska: *Selected English Collocations*; Wydawnictwo Naukowe Pwn, Warszawa, 1993.

Brigitte Krenn: Acquisition of Phraseological Units from Linguistically Interpreted Corpora – A Case Study on German PP-Verb Collocations"; in: *Proceedings of the First International Conference on Language Resources and Evaluation*; Granada, 28-30 May 1998.

Claude Lainé: *Vocabulaire combinatoire de la CFAO mécanique/Combinatory Vocabulary of CAD/CAM in Mechanical Engineering*, Bulletin de terminologie 219; Ministre des Approvisionnements et Services Canada, 1993.

Igor A. Mel'cuk et al.: *Dictionnaire explicatif et combinatoire du français contemporain. Recherches Lexico-Sémantiques*, I, Montréal, Presses Universitaires de Montréal, 1984.

Igor Mel'cuk, Leo Wanner: *Towards an Efficient Representation of Restricted Lexical Cooccurrence*, in: *Proceedings of EURALEX-94 International Congress*, 1994.

Silvia Pavel, Monique Boileau: *Vocabulaire des systèmes dynamiques et de l'imagerie fractale/Vacabulary of Dynamical Systems and Fractal Imagery*; Ministère des Approvisionnemens et Services Canada, 1994.

Ingemar Persson: *Das System der kausativen Funktionsverbgefüge. Eine semantisch-syntaktische Analyse einiger verwandter Konstruktionen*; Lunder germanistische Forschungen 42, Kristianstad 1975.

Burkhard Schaeder/Henning Bergenholtz (Eds.) *Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern*, Tübingen: Narr, 1994.

Anne Schiller, Simone Teufel, Christine Stöckert, Christine Thielen: "Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS", Stuttgart/Tübingen, 1995.

Bruno Schulze: "MP user manual", in: *Proceedings of the First International Conference on Language Resources and Evaluation*; Granada, 28-30 May 1998.

Frank Smadja: "Retrieving Collocations from Text: Xtract", in: *Computational Linguistics*, Vol. 19, Nr.1, 1993, pp. 143-177 [= Special Issue on Using Large Corpora I].

Simone Teufel, Christine Stöckert: "EAGLES specifications for German morphosyntax", Stuttgart, 1996.