Extraction tools for collocations and their morphosyntactic specificities

Julia Ritz*, Ulrich Heid†

* Institut für Linguistik
Universität Potsdam
Postfach 601553
14415 Potsdam
Germany
julia@ling.uni-potsdam.de
† Institut für Maschinelle Sprachverarbeitung (IMS)
Universität Stuttgart
Azenbergstr. 12
70174 Stuttgart
Germany
Ulrich.Heid@ims.uni-stuttgart.de

Abstract

We describe tools for the extraction of collocations not only in the form of word combinations, but also of data about the morphosyntactic properties of collocation candidates. Such data are needed for a detailed lexical description of collocations, and to support both their recognition in text and the generation of collocationally acceptable text. We describe the tool architecture, report on a case study based on noun+verb collocations, and we give a first rough evaluation of the data quality produced.

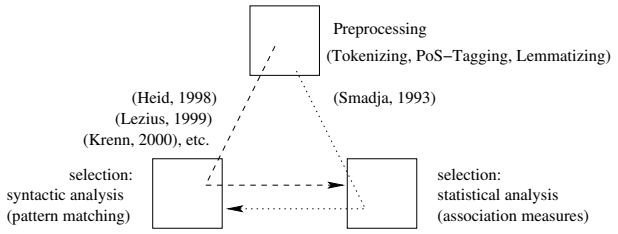

Figure 1: Collocation extraction approaches

1. The Problem

In much of the more recent work on the extraction of collocation candidates from text corpora, an architecture is used that relies on a sequence of steps (see figure 1). Typically, first corpora are preprocessed by means of tokenizing, part-of-speech (=PoS) tagging and lemmatization. As a second step, pattern-based extraction routines are used, which provide syntactically homogeneous sets of candidate word pairs (in a given grammatical relation: 'relational cooccurrences' (Evert, 2005)), for example pairs of nouns and attributive adjectives preceding them. Since obviously not all pairs are collocational, statistical filtering by means of association measures (such as, for example, the log likelihood ratio test (Dunning, 1993)) is used as a third step, to identify pairs with a statistically significant cooccurrence frequency, and to order them according to the strength of their association. This architecture was used, among many others, by (Heid, 1998; Lezius, 1999; Krenn, 2000), etc. It differs from Smadja's (1993) approach: he first determines significant word pairs and then uses their occurrence within a syntactic relation as a filtering criterion.

Either approach will provide pairs of lemmas or pairs of word forms that show sufficient evidence of (relational) cooccurrence. But both can not account for effects of idiomatization, nor for differences in readings of any of the two elements of a collocation. It has often been noticed that many collocations have strong preferences to appear in certain morphosyntactic forms rather than others. Examples are preferences with respect to number (see example 1[1]) or determination (see example 2). There are also combined preferences (for more than one dimension) and there are collocations which show up in two (or more) different morphosyntactic forms (see example 3), but not in all theoretically possible forms.

(1) in Schwierigkeiten$_{pl}$ stecken
'to be in trouble'

(2) zur $_{prep+def}$ Diskussion stehen
'to be under consideration'

(3) im$_{prep+def}$ Dienst$_{sg}$ des X stehen
bei X in Diensten$_{pl,indef}$ stehen
'to be in X's service'

These restrictions on morphosyntactic variability have been interpreted as signs of idiomatization ((Helbig, 1984): 'lexikalisierte Funktionsverbgefüge') or of opacity (Tutin, 2004). Some such restrictions are also valid for idioms.
Moreover, if an extraction tool just provides lemma pairs like *Hoffnung* + *machen*, this result may in fact be due to the presence of several different collocations in the corpus data, some of which can be distinguished in terms of

[1] Abbreviations used in examples and tables:
sg - singular
pl - plural
prep - preposition
def - definite determiner
indef - indefinite determiner
modif - modifier (in a broader sense)

morphosyntactic properties: *jemandem Hoffnung machen* (typically with *Hoffnung* in the singular and with a dative complement) '*to inspire hope to someone*' should be distinguished from *sich Hoffnungen machen* (*'to entertain hope'*), reflexive, with a strong preference for the plural of *Hoffnung*. These two readings would be collapsed into one output by a collocation extractor designed according to the architecture sketched above. We[2] thus expect a double improvement from more sophisticated extraction tools: (i) access to more details about collocations, especially to those morphosyntactic properties that contribute to the idiomatic behaviour of collocations, and (ii) on that basis, more possibilities to (semi-automatically) tell apart different collocations (or readings of collocations) that involve the same lemmas.

2. Acquisition tools

Our collocation candidate extraction tools are designed to do two jobs in one go: (i) identify collocation candidates of the noun+verb type, and (ii) identify their respective morphosyntactic properties and preferences. As an input, we use PoS-tagged and partially parsed German newpaper text. It is parsed with the recursive chunker YAC (Kermes, 2003)[3].

2.1. Contexts

Our first aim was to extract collocation candidates from 'secure contexts', i.e. from contexts where a syntactic relation between the NPs or PPs and the verbs exists. For German noun+verb collocations, prenominal participle constructions are such secure contexts: mostly, the object or prepositional complement precedes the participle immediately (see example 4). Moreover, in many cases, the start of the participle phrase is clearly marked, its end trivially being the participle itself. Exceptions are cases where the beginning of the adjective phrase is not clearly marked, as in example 5. Some of these cases cannot be resolved even with subcategorisation information.

(4) die zur Diskussion stehenden Fragen
(lit.)*'the to (+definite determiner) discussion standing questions'*
'the questions under discussion'

(5) Es handelt sich um ein Konzept, mit dem [[[Banken]$_{NP}$ führende]$_{AP}$ Kunden]$_{NPsubj}$ gewinnen.
Es handelt sich um ein Konzept, mit dem [Banken]$_{NPsubj}$ [[führende]$_{AP}$ Kunden]$_{NPobj}$ gewinnen.

A disadvantage of participle constructions is that they are not used very frequently. Furthermore, the matching of participles to verbs is not a trivial task. The following issues cannot be guaranteed: (i) same reading in verb and participle (e.g. *bekannt - bekennen*), (ii) meaningfulness of the reconstructed verb (e.g. *wiedergewählt - wiederwählen*, *erdölexportierend - erdölexportieren*) and (iii) analogous subcategorisation and/or collocation behaviour (e.g. *geboren - gebären*).
Under these conditions, we opted for additional, less secure, contexts: verb final constructions (= subclauses) and constructions with a modal verb in the left sentence bracket[4]. Although there is a high degree of freedom in constituent ordering in German, the NP or PP immediately preceding the verb (complex) in the right sentence bracket contains most likely the base noun of the collocation. Rarely, adverbs or embedded phrases may intervene, or collocations may be coordinated.

2.2. Extraction procedures

The extraction process includes (i) a pattern matching step, and (ii) a feature determination step. We designed patterns for the syntactic contexts mentioned above[5], based on the chunk annotation of YAC. For each instance matching a pattern, the values of the following features are determined:

- lemma of the noun (=potential base)
- lemma of the verb (=potential collocate)
- number of the noun (singular, plural)
- case of the noun (nominative, accusative, genitive, dative)
- determination of the noun (definite, indefinite, null, demonstrative, possessive, quantifying)
- modification of the noun (adjective, cardinal number, PP, genitive NP, compounding etc.)
- negation[6] (yes/no)
- auxiliaries and modal verbs under which the potential collocate is embedded
- sentence from the corpus (used as an example of the feature set which has been extracted from it)

The resulting feature/value pairs are stored in a relational data base, on which interpretation tools can operate.
Interpretation includes grouping by features, the determination of quantitative preferences (e.g. 91% singular), and, optionally, a word formation analysis of the elements of the collocation.
A morphological analysis of the nouns provides clues as to whether compound nouns (e.g. *Rauchpause* "smoking break") share collocates with their heads (e.g. *Pause einlegen* "have a break"). Table 1 shows a comparison between potential collocates of the simplex *Plan* (*'plan'*) vs. potential collocates of nouns having *Plan* as their morphological

[2]This work has been carried out at the Institute for Natural Language Processing (IMS), at the University of Stuttgart.

[3]The chunker YAC is based on the corpus query language CQP (Christ, 1994); its annotation can, in turn, be queried using this language. See also `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/cqp-tutorial.pdf`.

[4]For topological field theory, see (Wöllstein-Leisten et al., 1997).

[5]More details about these patterns can be found in (Ritz, 2006)

[6]Negation with *kein* is considered as negation + quantifying determiner.

v	prep	occurrences with **Plan** (*simplex*)	(+*compounds*)
vorsehen	in	5	47
bauen	nach	5	5
ausweisen	in	0	11
festschreiben	in	1	5
befassen	mit	0	3

Table 1: Collocational behaviour of *Plan* (simplex vs. compounds)

head (including compounds such as *Bebauungsplan* (*'local plan'*), *Haushaltsplan* (*'budget'*) etc.). Under the assumption that transparent (productively formed, non-lexicalized) compounds tend to share collocates with their heads (cf. (Zinsmeister and Heid, 2004)), compound data may be used to reinforce quantitative tendencies observed for the head nouns; at the same time, differences in the collocational preferences of compounds vs. compound heads may indicate that the respective compounds are lexicalized. We can retrieve and interpret compound data separately as well as together with data about the compound heads.

3. Results and Evaluation

n	v	prep	f	translation, specificities
Denkmalschutz	stehen	unter	67	"to be under monumental protection", sg (95.63%), no determiner (95.63%), present participle (95.63%)
Depression	leiden	unter	13	"to suffer from depression", pl (79.42%), no determiner (79.42%), present participle (79.42%)
Dienst	stehen	in	22	"to be in someone's service", present participle (87.27%)
Dienst	stellen	in	18	"to put into service", sg (84.67%), no determiner (84.67%), past participle (84.67%))
Diskussion	bringen	in	12	"to bring into discussion", sg (77.91%), determiner (77.91%), def (77.91%), no fusion (77.91%), past participle (77.91%)
Diskussion	stehen	zu	31	"to be under discussion", sg (85.59%), determiner (85.59%), def (85.59%), fusion of preposition and determiner (85.59%), present participle (90.79%)
Diskussion	stellen	zu	15	"to put under discussion", sg (81.9%), determiner (81.9%), def (81.9%), fusion of preposition and determiner (81.9%), past participle (81.9%)

Table 2: Sample extraction results (participle constructions)

On the hypothesis that collocations are exclusively characterized by a deviant morphosyntactic behaviour, we used strong preferences for certain feature values to identify collocation candidates. In this experiment, the following feature values were taken into account: singular or plural, existence of a determiner, the sort of the determiner (definite, indefinite, demonstrative, possessive or quantifying), and (in the case of participle constructions) the tense of the participle. For each of these criteria, a threshold of 60% was used.

Sample results are shown in tables 2 (participle constructions) and 3 (constructions with full verbs in the right sentence bracket). Along with the lemma combinations, the tables contain the frequency (f) and the 'typical' morphosyntactic behaviour.

n	v	f	specificities
Abhilfe	schaffen	193	sg (100%), no determiner (96.37%)
Abitur	ablegen	15	sg (100%), determiner (100%), def (86.67%)
Abitur	machen	20	sg (100%), determiner (60%)
Abkommen	abschließen	24	sg (79.17%), determiner (83.33%)
Abkommen	akzeptieren	10	sg (90%), determiner (90%), def (80%)
Abkommen	anschließen	10	sg (100%), determiner (100%), def (90%)
Abkommen	aushandeln	15	sg (100%), determiner (100%), indef (100%)
Abkommen	beitreten	54	sg (94.44%), determiner (94.44%), def (90.74%)
Abkommen	einhalten	21	sg (61.90%), determiner (95.24%), def (95.24%)
Abkommen	erreichen	19	sg (100%), determiner (89.47%), indef (68.42%)
Abkommen	erzielen	34	sg (97.06%), determiner (97.06%), indef (97.06%)
Abkommen	geben	12	sg (83.33%), determiner (100%), indef (66.67%)
Abkommen	kündigen	10	sg (100%), determiner (100%), def (100%)
Abkommen	paraphieren	11	sg (100%), determiner (100%), def (81.82%)
Abkommen	ratifizieren	23	sg (86.96%), determiner (95.65%), def (86.96%)
Abkommen	schließen	70	sg (88.57%), determiner (90%), indef (84.29%)
Abkommen	sein	32	sg (84.375%), determiner (81.25%)
Abkommen	treten	31	sg (83.87%), determiner (83.87%), def (80.65%)
Abkommen	umsetzen	27	sg (77.78%), determiner (81.48%), def (62.96%)
Abkommen	unterschreiben	14	sg (92.8571428571429%), determiner (100%), def (57.14%)
Abkommen	unterzeichnen	274	sg (89.78%), determiner (91.61%)
Abkommen	verlängern	13	sg (100%), determiner (100%), def (84.62%)
Abkommen	zustandekommen	11	sg (100%), determiner (100%), indef (72.73%)
Abkommen	zustimmen	17	sg (94.11%), determiner (100%), def (100%)
Absicht	bestehen	12	sg (75%), determiner (100%), def (66.67%)
Absicht	haben	103	sg (96.12%), determiner (99.03%), def (94.17%)
Absicht	sein	44	sg (86.36%), determiner (68.18%)
Abstand	betragen	10	sg (80%), determiner (100%), def (100%)
Abstand	gewinnen	12	sg (100%), no determiner (83.33%)
Abstand	halten	13	sg (100%), no determiner (84.62%)
Abstand	nehmen	98	sg (98.98%), no determiner (98.98%)
Abstand	sein	19	sg (100%), determiner (57.89%), def (57.89%)
Abstand	verringern	27	sg (92.59%), determiner (100%), def (96.30%)
Abstand	werden	17	sg (100%), determiner (70.59%), def (70.59%)

Table 3: Sample extraction results (constructions with full verbs in the right sentence bracket)

From a corpus of nearly 300 million words, we extracted 96,421 instances (token combinations) of prenominal participles (1,892 lemma pair types with $f > 4$). Therefrom, 573 lemma combinations were identified as collocation candidates. For these constructions, we achieve only a precision of 35%[7]. When extracting constructions with the full verbs in the right sentence bracket from the same corpus, from an extracted 1.3 million instances (over 750,00 lemma pair types; 10,934 with $f \geq 10$), 9,340 were identified as collocation candidates, resulting in a precision of 66%.

[7] As mentioned above, these constructions are very rare: only 0.035% of the text and only 0.157% of the occurrences of the nouns are indeed extracted.

f	n	v	prep
315	Exil	leben	in
237	Tod	verurteilen	zu
192	Leben	rufen	in
184	Verfügung	stehen	zu
132	Auge	fassen	in
122	Leben	kommen	um
117	Parlament	vertreten	in
79	Verfügung	stellen	zu
77	Boden	liegen	an
76	Tod	bedrohen	mit
74	Vertrag	festlegen	in
73	Stocken	geraten/raten	in
72	Kommunist	hervorgehen	aus
65	Nähe	liegen	in
61	Amt	scheiden	aus
56	Ausland	leben	in
54	Gespräch	bringen	in

Table 4: Combinations with a preference for the definite determiner (participle constructions)

Table 3 contains a short list of noun+verb-collocation candidates for the nouns *Abhilfe*, *Abitur*, *Abkommen*, *Absicht* and *Abstand*. Among typical collocations such as *Abkommen schließen* (*'to conclude an agreement'*), *Abkommen einhalten* (*'to fulfill an agreement'*), etc., trivial combinations, such as *Abkommen erzielen*, *Abkommen erreichen* (both *'to arrive at an agreement'*) are found. The combination *Abkommen treten* shows the relevance of collocation combinations, as it is erroneously brought forward by the fact that *in Kraft treten* (*'to enter into force'*) is a typical collocate of *Abkommen* in a N_{Subj}+V collocation.

As mentioned above, restrictions with respect to determination and modification, as well as a restricted set of collocates have been discussed in the literature as signs of idiomatization. Table 4 shows a few examples of combinations with a marked preference for the definite article (from prenominal participles, verb+PP data). These examples include support verb constructions (e.g. *zur Verfügung stehen* (*'to be available'*) and *zur Verfügung stellen* (*'to provide'*)), idioms like *ins Leben rufen* (*'to call into life'*), *ins Auge fassen* (*'to envisage'*), but also prominent trivial combinations such as *im Parlament vertreten* (*'represented in Parliament'*, typically a participle construction), *in der Nähe liegen* (*'to be nearby'*). The combination *am Boden liegen* has two readings: a literal one (*'to lie on the floor'*, a trivial word combination) and an idiomatized one (*'to be devastated'*).

Modification preferences also seem to produce rather idiomatic combinations: table 5 contains a few items which in our corpus data do not occur with any kind of modifier, and table 6 shows cases which recurrently use the same PP modifier. Some of these sequences are combinations of collocations (e.g. with the support verb constructions *zur Verfügung stellen* (*'to provide'*) and *in Kraft treten* (*'to put into force'*)), some are idioms (*jemandem einen Strich durch die Rechnung machen* (*'to upset someone's plans'*), *jemandem den Wind aus den Segeln nehmen* (*'to take the wind out of someone's sails'*), *zwei Fliegen mit einer Klappe schlagen* (*'to kill two birds with one stone'*), *Nägel mit Köpfen machen* (*'to put one's money where one's mouth is'*), *jemandem Steine in den Weg legen* (*'to put obstacles in someone's way'*)).

f	n	v	
19	Pech	haben	
17	Revue	passieren	(lassen)
16	Gehör	finden	
...			
13	Schulbank	drücken	
12	Tanzbein	schwingen	
11	das Weite	suchen	
9	Auftrieb	geben	

Table 5: Combinations never used with any modifier

f	n	v	modif (PP)
255	Polizei	mitteilen	an Montag
137	Grenzwert	überschreiten	nach Smog-verordnung
64	Sprecher	sagen	auf Anfrage
50	Aussicht	haben	auf Erfolg
47	Mensch	kommen	zu Schaden
45	Strich	machen	durch Rechnung
43	Waffenstillstand	treten	in Kraft
43	Amt	stellen	zu Verfügung
41	Mensch	kommen	um Leben
41	Geld	stellen	zu Verfügung
40	Anfang	treten	in Kraft
40	Wind	nehmen	aus Segeln
36	Fliege	schlagen	mit Klappe
32	Nagel	machen	mit Kopf
28	Gesetz	treten	in Kraft
28	Stein	legen	in Weg
28	Berufung	einlegen	gegen Urteil

Table 6: Combinations recurrently used with the same PP

Even though this has not been much discussed in the literature yet, we think that also strong preferences for possessive and quantifying determiners may be an indicator of idiomatization: in table 7[8], we show a few collocations and the absolute and relative frequencies for each kind of determiner used with them. The idiomatic combination *Hut + nehmen*, e.g., appears to have a strong preference for the possessive determiner, but also occurs with a definite determiner in a considerable amount of data. In contrast, the collocation *Veto + einlegen* (also with a preference for the possessive determiner), is never used with a definite, but in some cases with an indefinite determiner.

[8] Abbreviation in table 7: n.r. - not relevant.
In the table, frequencies of the null determiner are split according to number (sg/pl) because, in German, in the singular, null determination only occurs with mass nouns and in idioms, whereas in the plural, it is the regular realisation of indefiniteness.

f	n	v	poss	pl, null	sg, null	def	indef	quant
214	Stimme	abgeben	179 (84%)	20 (9%)	-	7 (3%)	4 (2%)	3 (1%)
212	Veto	einlegen	147 (69%)	n.r.	11 (5%)	-	25 (12%)	29 (16%)
174	Zustimmung	geben	147 (84%)	n.r.	6 (3%)	10 (6%)	3 (2%)	6 (4%)
123	Amt	niederlegen	109 (89%)	1 (1%)	-	10 (8%)	-	-
93	Amt	aufgeben	73 (78%)	3 (3%)	-	15 (14%)	-	-
147	Hut	nehmen	113 (77%)	-	-	34 (23%)	-	-
163	Einfluß	haben	-	n.r.	63 (39%)	1 (1%)	9 (6%)	84 (52%)
40	Auskunft	erteilen	-	7 (18%)	11 (28%)	-	1 (3%)	21 (53%)
29	Schwierigkeit	machen	-	10 (34%)	-	-	1 (3%)	18 (62%)

Table 7: Distribution of determiners

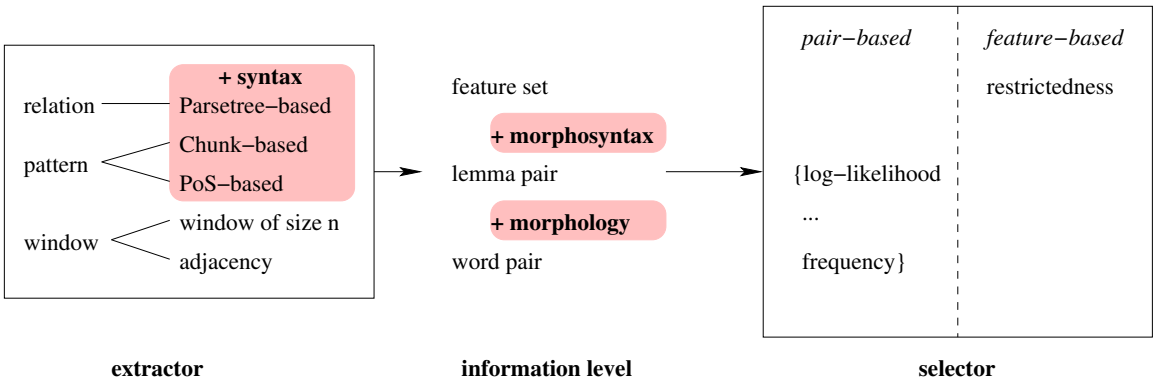

Figure 2: Methods of collocation extraction and selection

4. Conclusions: Towards large-scale extraction of detailed multiword data

On the assumption that restrictions in the morphosyntactic behaviour of multiword sequences and in lexical combinatorics can be correlated with idiomacity (in the sense of partial compositionality or non-compositionality), the tools may be used to identify multiword items that show idiomatization effects. As the distinction between, e.g., support verb constructions, other verb+object collocations and VP idioms can not be drawn on the basis of morphosyntactic data, our tools are obviously only meant to provide data for more detailed (manual) linguistic analysis. However, large-scale descriptive work on a broad basis of data is only possible, if the collection and structuring of sample data is automated.

By combining pattern-based search and the extraction of linguistic features, we are able to provide lexeme cooccurrence data and context parameters in one go. Figure 2 schematically compares different approaches to the extraction and selection of collocation candidates. From window-based to pattern-based to relation-based extraction, an increasing amount of lingistic knowledge is needed. Typically, this knowledge is permanently annotated in the corpus, and a variety of tools can benefit from the available annotations. With the addition of morphological knowledge, the abstraction step from word pairs to lemma pairs becomes feasible; and the addition of morphosyntactic knowledge, as underlying the work presented here, opens up the possibility to describe the extracted data in terms of restrictedness.

This latter type of description is possible on the basis of chunked corpora (as in our case) or of parsed corpora. Some of our results clearly show the lack of information about grammatical relations, since N_{Subj}+V collocations (e.g. *Abkommen + in Kraft treten*) are contained in table 3, next to N_{Obj}+V collocations. At the expense of recall, we can restrict the search in our database of intermediate results to cases with a clearly marked accusative, in which case the false positives disappear (along with all those N_{Obj}+V collocations whose nouns have no unambiguous accusative form). Overall, the use of chunked corpora provides a major information gain over PoS-pattern-based or window-based approaches. Yet, frequency- or significance-based selectors can still be combined with the tools presented here. One integration possibility is to identify collocation candidates by means of association, and retrieve the morphosyntactic preferences of each collocation in a second step.

Future enhancements of the proposed tools include the use of external lexical knowledge to reduce case ambiguity (e.g. *mitteilen* ('*to inform*') in *die Polizei teilt mit* can only be intransitive, the group is thus of the type N_{Subj}+V). A syntactic subcategorisation lexicon would exactly provide such information; this would allow us to stick to the efficiency of chunking (instead of extracting from parsed text) while getting information about (at least some) grammatical relations. Similarly, a full integration of morphology into the interpretation tools is still outstanding: a morphology system would annotate compound nouns in the data base with their compound head, in an automatic interpretation step.

Furthermore, a number of features have been extracted but not yet analyzed in detail with respect to their relevance for collocation and idiom extraction. These include negation, tense preferences, embedding under *lassen* and other modal verbs, modification of the noun (by means of adjectives), or modification of the verb (by means of adverbs). It would also be interesting to identify cases where adjective and adverb seem both possible (*brieflich*$_{ADV}$*/brieflichen*$_{ADJ}$ *Kontakt halten*, *'stay in contact by mail'*).

5. References

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94: 3rd conference on Computational Lexicography and Text Research*, pages 23–32, Budapest. `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`.

Ted E. Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences - Word Pairs and Collocations*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Ulrich Heid. 1998. Towards a corpus-based dictionary of German noun-verb collocations. In *Proceedings of the EURALEX International Congress 1998*, pages 301 – 312, Liège.

Gerhard Helbig. 1984. *Studien zur deutschen Syntax*. Leipzig: VEB Verlag Enzyklopädie.

Hannah Kermes. 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Brigitte Krenn. 2000. Collocation Mining: Exploiting Corpora for Collocation Identification and Representation. In *Proceedings of KONVENS 2000*, pages 209 – 214, Ilmenau, Deutschland.

Wolfgang Lezius. 1999. Automatische Extrahierung idiomatischer Bigramme aus Textkorpora. In R. Rapp, editor, *Tagungsband des Linguistischen Kolloquiums 1999*, Germersheim, Germany.

Julia Ritz. 2006. Collocation Extraction: Needs, Feeds and Results of an Extraction System for German. EACL 2006 Workshop 'Multi-word-expressions in a multilingual context', April 6th, 2006. To appear.

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

Agnès Tutin. 2004. Pour une modélisation dynamique des collocations dans les textes. In Geoffrey Williams and Sandra Vessier, editors, *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France.

Angelika Wöllstein-Leisten, Axel Heilmann, Peter Stepan, and Sten Vikner. 1997. *Deutsche Satzstruktur*. Stauffenburg Verlag, Tübingen, Germany.

Heike Zinsmeister and Ulrich Heid. 2004. Collocations of Complex Nouns: Evidence for Lexicalisation. In Geoffrey Williams and Sandra Vessier, editors, *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France.