Tools for collocation extraction: preferences for active vs. passive

Ulrich Heid, Marion Weller

* Institute for Natural Language Processing
University of Stuttgart, Germany
{heid,wellermn}@ims.uni-stuttgart.de

Abstract

We present and partially evaluate procedures for the extraction of noun+verb collocation candidates from German text corpora, along with their morphosyntactic preferences, especially for the active vs. passive voice. We start from tokenized, tagged, lemmatized and chunked text, and we use extraction patterns formulated in the CQP corpus query language. We discuss the results of a precision evaluation, on administrative texts from the European Union: we find a considerable amount of specialized collocations, as well as general ones and complex predicates; overall the precision is considerably higher than that of a statistical extractor used as a baseline.

1. Introduction

1.1. The notion of collocation underlying this study

This paper is concerned with German noun+verb-collocations (NVCs), their extraction from corpus data and the analysis and semi-automatic description of their morphosyntactic properties, in particular their preferences for the active vs. passive voice.

Our view on collocations is a lexicographic one, in line with e.g. the *Oxford Collocations Dictionary for Students of English*, or with the tentative definition given by (Bartsch 2004): 76:

> *[...] collocations are lexically and/or pragmatically constrained recurrent cooccurrences of at least two lexical items which are in a direct syntactic relation with each other.*

Our work on collocation candidate extraction from text corpora is intended as a first step in lexicographic work, i.e. in the creation of collocationally rich dictionary entries for both NLP and human users (cf. e.g. (Heid et al. 2007)). To this end, we extract more than just word cooccurrence data; in our view, to describe a collocation, also an account of its morphosyntactic preferences has to be given, if it has such preferences; otherwise, it has to be shown that the only idiosyncratic element is the lexical selection.

1.2. Choices with respect to corpus preprocessing

There exist many approaches to collocation extraction, some of which are based on statistical measures only, while others rely on morphosyntactic and syntactic annotation of corpus data or on a combination of both, statistical and symbolic devices. Approaches that make use of symbolic extraction procedures (corpus query, analysis of parsing results, etc.) may require more or less deep preprocessing of the corpora. Depending on the degree of detail present in the annotation, more or less linguistic knowledge needs to go into the extraction procedures and/or into later interpretation steps. For example, one may use a full parser to identify and annotate relations between e.g. verbs and their subjects, verbs and their objects, extract these pairs from the corpus and then analyze them further, with respect to their frequency and significance of association.

In our work, we start from much less detailed preprocessing, in fact from a flat annotation of the corpora: our German texts are tokenized, POS-tagged with Schmid's TreeTagger (cf. (Schmid 1994)), using the STTS tagset, and chunked with YAC, a recursive chunker (cf. (Kermes 2003)), which identifies adjective, noun and prepositional chunks grouping all pre-head material together with the head of a given phrase. It does not account for post-head modifiers, nor for attachment, but it provides the start and end points of the German verbal complex (main verb and auxiliaries). We assume that this amount of annotation is sufficient for our task; it can be provided for very large corpora.

1.3. Problems in the extraction of German noun+verb collocations from text

The extraction of noun+verb collocations of a language like English does not pose a particular problem, as regular expressions over parts of speech and possibly start and end points of chunks may suffice to get both an acceptable recall and a good precision.

For German, the situation is somewhat different, due to three types of problems. First, German has three different models of verb placement which need to be taken into account. In addition, German is not a configurational language. Thus, its constituent order is relatively free, at least in the 'Mittelfeld', the topologically central part of the German sentence. Thirdly, German does not fully compensate the lack of configurationality with its morphological case; only 21 % of the noun phrases contained in the Negra treebank and analyzed by (Evert 2004) are indeed unambiguous with respect to case. These facts have an incidence on the architecture of our extraction tools.

The remainder of this paper is structured as follows: we first describe our extraction architecture and the procedures used to identify passives (section 2); we then present some results of the extraction work (section 3) and an evaluation of some of our results (section 4). We conclude in section 5.

1.4. Corpus data considered

We have used different types of corpora for the experiments reported here. The bulk of the texts come from newspapers from Germany[1], Switzerland and Austria[2]. However, we also analyzed the German part of the *Acquis Communautaire* Corpus (cf. (Steinberger et al. 2006), 16 M) and a corpus of texts from juridical journals provided by a cooperation partner (78 M).

2. Extracting German noun+verb collocations and their linguistic properties from text corpora

2.1. Outline architecture

Our tools are conceived as a sequence of corpus processing steps: the input is tokenized, POS-tagged, lemmatized and recursively chunked as described above. After this preprocessing, we use extraction patterns based on the corpus query language CQP[3] to extract noun+verb pairs and their contexts. All verb pair occurrences are stored in a database, along with attribute/value pairs for the morphosyntactic features we are interested in (see below). In a subsequent step, the candidates can be ordered by frequency, or their association strength can be determined by means of an association measure, such as the log likelihood ratio test (cf. (Dunning 1993)). This basic architecture is described in (Ritz/Heid 2006) and (Ritz 2006).

This sequencing, which is inverted with respect to e.g. (Smadja 1993)'s approach, has the advantage of allowing us to work with syntactically homogeneous candidate material (cf. also (Krenn 2000), (Evert 2005) or (Sereţan/Wehrli 2006)); furthermore, as we are interested in separating active from passive occurrences, we simply use two separate queries which extract distinct sets of sentences, and we store along with each collocation candidate instance, whether it comes from an active or a passive clause.

To identify the morphosyntactic properties of the collocation candidates, we use the morphological annotation contained in the corpus data, and we note attribute/value pairs for the following properties of each candidate: number, determiner and possible modifiers of the noun, negation, quantifiers and the presence or absence of a modal auxiliary; in addition, we identify the voice (active/passive), the passive auxiliary (*sein* for stative passive vs. *werden* for the dynamic passive), as well as the verb placement model from which the candidate is extracted. All parameter values of each candidate are stored in the database; to identify preferences, we consider all occurrences of a given verb+noun pair, and we calculate preferences according to the calculus proposed by (Evert 2004).

2.2. Accounting for German word order

As mentioned above, German has three models of verb placement, verb-first (`v-1` in our tables), verb-second (`v-2`) and verb-final (`vlast`). The first one is used in questions and conditionals, the last one in subclauses, and the verb-second model in all other cases. These three models are illustrated in table 1 below[4].

As can be seen in table 1, the verb-first model and often the verb-second model lead to a separation of the verb and its complement noun phrase, whereas these two elements of the collocation candidate are typically adjacent or only separated by adverbial constituents in the verb-final case. Consequently, we mainly extract active clauses from the verb-final model. Verb-final clauses make up for about 25 % of all finite verbs in the German TIGER treebank: clearly, our procedure reduces recall drastically, but it helps considerably to improve precision. We do not think that this choice will influence our data about the distribution of morphosyntactic features; in the active, all collocations are likely evenly represented in the three word order models.

2.3. Extracting passive data

To extract passives, we consider all three verb placement models, as the passive auxiliary and possible modal auxiliaries, together with conjunctions at the start of a subclause and the participle determine quite clearly in which domain the complements of a verb may be found. Passives share this property with active clauses under the verb-final model. Table 2 shows the three models under the passive.

We identify passives with sequence models: figure 1 shows a query for the passive in the verb-final model. The query match begins with a sentence introducing conjunction or a relative pronoun (line 2); in the relevant NP (line 4-9), which is typically immediately left of the verb complex, no measure nouns (`meas`), pronouns, proper nouns (`ne`) or cardinal numbers are allowed, as these don't form relevant collocations. This NP can be followed by an arbitrary number of tokens, e.g. adverbs, (but no further NP, line 10) and an optional tense auxilary (line 11). The past participle (line 12) is the verbal collocate, followed by the passive auxiliaries *sein* or *werden* and possibly by further tense auxiliaries and/or modal auxiliaries. The match may not cross sentence boundaries (line 17).

3. Results

3.1. Frequency of passives

The overall frequency of passives in our texts varies between ca. 5.8 % and 15.3 %. The lowest figure is found in newspapers (ca. 5.8 % on a 76 M word corpus composed of *Frankfurter Rundschau* and *Stuttgarter Zeitung* (1992/93, see above)). The highest amounts of passives (ca. 15.3 % of all verb forms) is found in the administrative language of

[1] We use data from the *Frankfurter Rundschau* (ECI version, 1992/93, 40 M), the *Stuttgarter Zeitung* (1992/93, 36 M), the *Frankfurter Allgemeine Zeitung* (1997-99, 80 M) and *Die Zeit* (1998 - 2005, 50 M).

[2] Parts of the DeReKo corpus jointly created by the Institut für deutsche Sprache, Mannheim, and the universities of Tübingen and Stuttgart.

[3] Cf. (Evert 2005b).

[4] Abbreviations of topological fields in table 1: VF = Vorfeld (first constituent of the sentence); LK = Linke Satzklammer (verbal or conjunction position); MF = Mittelfeld (the typical place of NPs and PPs, which may occur in any order, depending e.g. on information structural constraints); RK = Rechte Satzklammer (second possible verbal position); NF = Nachfeld (position of e.g. extraposed material).

Type	Model	VF	LK	MF		RK	NF
Question	v-1		Löst	der Mitarbeiter	[...] das Problem?		
Conditl.	v-1		Löst	der Mitarbeiter	[...] das Problem,		so ...
Decl. sent.	v-2	Der Mitarb.	löst	[...] das Problem			
Subclause	vlast		weil	der Mitarbeiter	[...] das Problem	löst	

Table 1: Models of verb placement in German: examples

	VF	LK	MF		RK	
v-1	(Es)	wird	die Frage	dann	gestellt	
		Kann	die Frage		gestellt	werden?
v-2	Die Frage	wird		dann	gestellt	
	Die Frage	kann	hier	wohl	gestellt	werden
vlast		weil	die Frage	dann	gestellt	wird
		daß	die Frage	dann	gestellt	werden kann

Table 2: German verb placement models for passives and constructions with modal verbs

Acquis Communautaire. The juridical texts show ca. 7 % of passive occurrences. We thus look at relatively rare occurrences. It is then all the more significant, at least of the administrative style of the *Acquis* corpus, to find combinations like *Beihilfe + zahlen* (pay financial aid), *Präsident + ermächtigen* (entrust + the president) overwhelmingly in the passive.

3.2. Morphosyntactic preferences of collocations

In table 3, we show some results[5] for the noun *Rechnung* (`account`) as used in the *Acquis Communautaire Corpus*: there is a burst with *Rechnung tragen* (`keep track`), which always shows up without article and in the singular. The other word pairs use the meaning of 'bill', i.e. *Rechnung ausstellen*, *erstellen* (`make out`), *bezahlen* (`pay`). The data for *Rechnung ausstellen* show variability with respect to number, determination and voice, which points towards a compositional interpretation, as suggested by e.g. (Fazly/Stevenson 2006) while the figures for *Rechnung tragen* are a clear sign of fixedness and idiomaticity.

3.3. Word order preferences of collocations in the passive

Passives show roughly the same distribution over the three word order models as actives. Verb-second cases make up for roughly half of all occurrences, verb-final for more than one third, and verb-first for about 10-12 %. One would thus expect passive forms of individual collocation candidates to be accordingly distributed over the three word order models. However, a subset of rather frequent candidates (cf. table 4) do not or very rarely appear in the verb-second model.

In the cases listed in table 4, the noun is not the true direct object of the verb, but rather a part of a complex predicate,

[5]Table columns: f = absolute frequency, det_type = type of determiner, num = number, order = word order models.
Cells: det_type: def = definite, indef = indefinite, null = no art icle, dem = demonstrative, poss = possessive, quant = quantifying; num: sg = singular, pl = plural; for word order types, see above.

Candidate	A:V-L	P:V-1	P:V-L	P:V-2
Auffassung vertreten	1321	53	97	48
Bezug nehmen	783	439	492	0
Rechnung tragen	2287	481	492	0
Gebrauch machen	2095	216	430	0
Sorge tragen	241	31	43	0

Table 4: Collocations which do not appear in the verb-second passive: idiomatized collocations (complex predicates)

cf. *Bezug nehmen* (`make reference`). The Vorfeld position, to the left of the finite verb in a `v2` sentence, seems not to accept certain non-topical constituents; the nouns of lexicalized support verb constructinos seem to be equally in that position. In fact, most occurrences in table 4 are idiomatic[6]; *Auffassung vertreten* (voice + opinion) has been added to exemplify the behaviour of non-idiomatized collocations. The instances which lack a v-2 passive are complex predicates. Similarly, they can neither be separated by a verbal element (*er hat auf X Bezug genommen*, but not **Bezug hat er auf X genommen*). The only contexts where a verb second passive can occur are either contrastive ones (e.g. with a negated quantifier under emphasis: *kéin Bezug wird auf X genommen*, or cases where the finite verb is a modal auxiliary. This property can be used to detect these lexicalized SVCs, at least frequent ones.

4. Evaluation

4.1. Work towards a gold standard for noun+verb collocation candidates

As with any linguistic data extraction task, a complete evaluation implies both an assessment of precision and of recall; for collocation extraction, (Evert/Krenn 2001) and (Evert 2005) have shown in detail how a full evaluation of lexical cooccurrence data can be carried out. Such an

[6]*Bezug nehmen* (make reference), *Rechnung tragen* (take into account), *Gebrauch machen* (make use), *Sorge tragen* (care about).

```
     MACRO passive_verb-final(0)
1    (
2    [pos = "(KOU(S|I)|PRELS)"]
3    []*
4    <np>
5    @[!pp & !ap & _.np_f not contains "ne" & _.np_f not contains "pron"
6       & _.np_f not contains "meas" & _.np_h != "@card@"]
7    [!pp & !ap & _.np_f not contains "ne" & _.np_f not contains "pron"
8       & _.np_f not contains "meas" & _.np_h != "@card@"]*
9    </np>
10   [!np & pos != "(\$.|KOUS|VMFIN)"]*
11   [pos = "V.*"]*
12   [pos = "VVPP"]
13   [lemma = "(werden|sein)"]
14   [pos = "V.*"]*
15   [pos = "(\$.|KON)"]
16   )
17   within s
```

Figure 1: Sample query for the extraction of noun+verb collocation candidates in passive subclauses

f	n_lemma	v_lemma	det_type	num	active_passive	order
5	Rechnung	ausstellen	def	Sg	passive	vlast
4	Rechnung	ausstellen	indef	Sg	active	vlast
4	Rechnung	ausstellen	def	Sg	active	vlast
1	Rechnung	ausstellen	def	Pl	active	vlast
1	Rechnung	bezahlen	indef	Sg	passive	vlast
1	Rechnung	erstellen	def	Sg	active	vlast
1	Rechnung	erstellen	def	Sg	passive	vlast
1387	Rechnung	tragen	null	Sg	active	vlast
262	Rechnung	tragen	null	Sg	passive	v-1
136	Rechnung	tragen	null	Sg	passive	vlast
10	Rechnung	tragen	def	Sg	active	vlast

Table 3: Collocation candidates with *Rechnung* (account) and their morphosyntactic preferences

evaluation necessarily has to rely on a gold standard corpus; for (adjacent) pairs of attributive adjectives and nouns, (Evert/Krenn 2001) used manually annotated lists of candidate data. For verb+complement pairs, a detailed recall analysis would require a full (manually corrected) parse.

As we are in addition interested in the morphosyntactic properties of the collocation candidates extracted, a gold standard corpus should also include annotations of this kind. We are working on the creation of a data set of this kind, which should contain the following types of annotations:

- Collocation: yes or no (`coll=+|-`);
- Base lemma, collocate lemma: annotated to ease the comparison with the database of results (`bs=..., ct=...`);
- Active/passive (`ap=a|prespart|ps|pw|pzu|hzu|perfpart`): active, active/present participle, “haben ... zu”; passive/sein, passive/werden, passive/zu, past participle;
- Syntactic function of NP/PP (`na|nd|nn|np|nrefl|nil`): object, indirect object (dat), subject, PP (with prep), reflexive, nil.

Individual annotations are given as features of the XML encoding as shown in (1). We expect a first set of ca. 1000 annotated sentences to be available by mid-2008. Ad interim, we can only evaluate the precision of our tools.

(1) `<s snum=3 vn=na ap=a coll=+ bs=Taetigkeit ct=ausueben >`
Die Agentur übt ihre Tätigkeit ausschließlich im Hinblick auf das Gemeinwohl aus. `</s>`

4.2. Evaluating morphosyntactic property extraction

As the tools consist of several components and are conceived to identify not only significant lexical cooccurrences, but also their morphosyntactic properties, it makes sense to evaluate both these functions, separately.

We have evaluated the following components:

- Identification of word order models (’w.o.’ in table 5)

- Identification of active vs. passive under the verb-second and verb-final word order model ('a/p');
- Identification of the correct chunk size to determine noun+verb collocation candidates ('chu');
- Identification of syntactically well-formed verb+complement groups (with accusative or dative complements, 'v+c.').

For the evaluation we used small samples of three times fifty sentences randomly picked from our results database; the results were created from the 1992/93 issue of the *Frankfurter Rundschau*. The three subsets concern sentences classified as follows by our system:

- verb-second, passive;
- verb-final, active;
- verb-final, passive.

The results are given in percentages, in table 5.

context type	w.o.	a/p.	chu.	v+c.
verb-second, passive	100.0	100.0	96.0	96.0
verb-final, active	56.0	98.0	100.0	88.0
verb-final, passive	100.0	84.0	100.0	80.0
complete set, average	85.3	94.0	98.7	81.3

Table 5: Precision values for the identification of word order, active/passive, chunk size and verb+complement candidates, for selected result subsets

A first observation is that actives seem to cause more problems than passives. With respect to the word order classification, this is due to the fact that the tools mistakenly count many constructions with a verb-second modal or tense auxiliary into the verb final class (example: *er wird nicht nur das Umweltamt übernehmen*, 'he will not only take over the office for environmental affairs'); this does not affect the correctness of the extraction of the verb+complement pairs, but it leads to a misclassification with respect to word order.
The relatively low figures for the identification of verb+complement pairs are due to two types of phenomena which are rather hard to cover in a setup without full parsing and detailed lexical resources, and which show the limitations of our approach based on flat annotations and slim preprocessing:

- complex nominal phrases with embedded prepositional phrases the attachment of which can not be calculated in the tool setup: for example, in the sentence *[...] ob er Funktionen in einer der DDR-Parteien oder Massenorganisationen innegehabt hatte* ('whether he had had a function in one of the DDR parties or mass organizations'), the verb+complement pair *Funktionen innehaben* ('have functions') should be identified. Due to a chunking problem with the coordinated PP, the tool identifies # *Massenorganisation innehaben* as a candidate.
- complex predicates which are part of verb+complement groups: the sentence *das Arbeitsleben wird anhand unzähliger Utensilien in Erinnerung gerufen* ('countless objects remind of workers' life', lit. 'workers' life is with countless objects brought-into-remembrance') should produce the verb+complement pair *Arbeitsleben + in Erinnerung rufen*; as there is no lexical information about the multiword *in Erinnerung rufen*, we get # *Arbeitsleben rufen* as a result. The same phenomenon occurs with predicative constructions like *gerecht werden* ('satisfy'), *höher schrauben* ('increase'), where only the verbal part of the construction is presented as part of the verb+complement group.

Our figures are somewhat lower than those reported by (Ritz 2006) for the identification of morphosyntactic features in prenominal participles. She reports a chunking quality of 96 to 99.5 % for this specific construction, and 99% precision in the identification of singular/plural, determination etc. within noun groups in prenominal participles. Ritz' work concentrated on a construction from which it is possible to extract these data with a very high precision, as much less variation is to be expected than in full sentences.

4.3. Evaluating the extraction of collocation candidates

We also have carried out a precision evaluation of the extraction of collocation candidates. This evaluation was done in the framework of the project *Collocations en Contexte*, on data extracted from the *Acquis Communautaire* corpus.
We analyzed two samples:

1. the 500 and the 774 most frequent verb+complement candidates (by lexical types);
2. the 2338 verb+complement candidate types for the 619 most frequent nouns of the *Acquis Communautaire* corpus, with a cooccurrence frequency of at least 4[7].

The candidates were evaluated according to the following classification:

- true positives:
  - complex predicates (e.g. *Bezug nehmen* ('make reference'));
  - collocations (e.g. *Zeugnis ausstellen* ('make out + certificate')) which are regularly used in general language;
- syntactically valid verb + complement groups with a sublanguage-specific meaning (conceptual collocations): e.g. *pH-Wert einstellen* ('set pH value');
- true negatives: irrelevant ad hoc combinations and misclassified verb + subject cooccurrences.

[7] We took the 1000 most frequent nouns and extracted all cooccurrence data for these; by considering only those collocation candidates which occurred at least 4 times, the set was reduced to 619 nouns.

The results obtained on set 2 are given in percentages in table 6.

Criteria	set 2
True positives + sublang. coll	68.9 %
– True positives	20.5 %
– – Complex predicates	2.1 %
– – Collocations	18.4 %
– Sublanguage collocations	48.5 %
True negatives:	31.0 %
– subject + verb	7.8 %
– other	23.2 %

Table 6: Evaluation results for verb+complement candidates in the *Acquis Communautaire* corpus

Due to the rather restrictive view on true positives adopted in the framework of the evaluation, the overall amount of true positives obtained is rather low; if the sublanguage specific combinations extracted by the tool are however added, the overall performance of the tool is quite acceptable.
As far as the smaller set 1 is concerned, we carried out a comparison of our tools with a baseline constituted by the top 500 candidates by log likelihood extracted by the statistical extraction tool presented in (Todirascu et al. 2008). We arrive at 44.6 % correct candidates in the top 774 candidates, whereas the statistical tool only provides 31.4 % true positives. The latter only relies on pos-tagging, constant distance between noun and verb and on the log likelihood value of the pairs.
The discrepancy between only 20.5 % true positives in set 2 and over 40 % in set 1 can be explained by the high frequency of the complex predicates and of the general language collocations (both predominantly found in the top 774 candidates), and, conversely, the large amount of lower frequency sublanguage-specific combinations, which makes up for almost half of the data of set 2. Thus, as the *Acquis Communautair*e corpus is highly specialized, the present figures should be interpreted with care, as far as their generalizability to other corpora is concerned.

5. Conclusions

We have presented and partially evaluated a set of extraction procedures for collocations and their morphosyntactic preferences, especially for the active vs. passive voice. The tools rely on tokenized, tagged, lemmatized and chunked text, but don't require full parsing. The precision achieved is acceptable, but the use of rather constrained contexts (verb-last active sentences) reduces the recall. As we aim at providing lexicographers with data for dictionary enhancement, emphasis is on precision, as a high precision alleviates their task of removing false positives.
The tools produce useful data about the use of collocations in the passive and clearly signal idiomatized collocations (complex predicates) which do not figure in v-2 passives, thereby providing a partial (low recall) recognizer for such non-compositional constructions.
In the future, we intend to finalize the suggested test set for recall evaluation. We will then experiment with a full parsing based collocation extractor and compare the performance of both approaches. As our tools allow us to extract the morphosyntactic properties of noun+verb-collocations with reasonable quality, we will use the data produced by the tool to further analyze morphosyntactic fixedness phenomena, in order to better understand their correlation with semantic opaqueness and idiomaticity. Furthermore, we will use the tools to learn more about the interaction between collocations and syntactic subcategorization.

6. References

Sabine Bartsch: *Structural and functional properties of collocations in English*, A corpus study of lexical and pragmatic constraints on lexical co-occurence, (Tübingen: Narr), 2004

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

Stefan Evert, Brigitte Krenn: "Methods for the Qualitative Evaluation of Lexical Association Measures", in: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 2004

Stefan Evert: "The Statistical Analysis of Morphosyntactic Distributions", in: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon*, (Lisbon: ELRA), 2004: 1539 – 1542

Stefan Evert: *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, available from `http://www.collocations.de/phd.html`, 2004, published 2005; software: `http://www.collocations.de/software.html`

Stefan Evert: "The CQP Query Language Tutorial (CWB version 2.2.b90)", ms. (Stuttgart: IMS), 2005

Afsaneh Fazly, Suzanne Stevenson: "Automatically constructing a lexicon of verb phrase idiomatic combinations", in: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2006,* (Trento/New Brunswick: ACL) 2006: 337 – 344

Ulrich Heid, Dennis Spohr, Julia Ritz, Christiane Schunk: "Struktur und Interoperabilität lexikalischer Ressourcen am Beispiel eines elektronischen Kollokationswörterbuchs", in: Georg Rehm, Andreas Witt, Lothar Lemnitzer (Hrsg.): *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen - Proceedings der GLDV-Jahrestagung 2007*, (Tübingen: Gunter Narr Verlag), 2007: 313 – 322.

Hannah Kermes: *Offline (and Online) Text Analysis for Computational Lexicography*, Diss., Stuttgart, (Stuttgart: IMS), 2003 [= AIMS, 9:3]

Brigitte Krenn: *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*, (Saarbrücken: German Research Center for Artificial Intelligence and Saarland University), 2000, [= Dissertations in Computational Linguistics and Language Technology, Volume 7]

Jonathan Crowther et al.: *Oxford Collocations Dictionary for Students of English*, (Oxford: Oxford University Press), 2002

Julia Ritz: “Collocation Extraction: Needs, Feeds and Results of an Extraction System for German” , in: *EACL 2006 Workshop on Multi-word-expressions in a multilingual context*, (Trento: EACL) 2006

Julia Ritz, Ulrich Heid: “Extraction tools for collocations and their morphosyntactic specificities”, in: *Proceedings of the Linguistic Resources and Evaluation Conference, LREC-2006*, Genova, Italia, 2006 [CD-ROM]

Helmut Schmid: “Probabilistic Part-of-Speech Tagging Using Decision Trees”. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*. Manchester, UK, 1994

Violeta Sereţan, Eric Wehrli: “Multilingual collocation extraction: Issues and solutions”, in: *Multilingual Language Resources and Interoperability. Proceedings of the COLING/ACL 2006 post-conference workshop.*, (Sydney/New Brunswick: COLING/ACL) 2006: 57 – 66

Frank Smadja: “Retrieving Collocations from Text: Xtract”, in: *Computational Linguistics*, Vol. 19.1 (1993): 143 – 177 [= Special Issue on Using Large Corpora I]

Ralf Steinberger et al.: “The JRC ACquis: A multilingual aligned parallel corpus with 20+ languages”, in: *Proceedings of the 5th LREC Conference*, Genova, 2006: 2142 – 2147

Amalia Todirascu et al.: “A hybrid approach to extracting and classifying verb+noun constructions”, in: *Proceedings of LREC-2008*, Marrakesh (this volume).