

# Collocations of Complex Nouns: Evidence for Lexicalisation

Heike Zinsmeister and Ulrich Heid  
Institute for Natural Language Processing  
University of Stuttgart  
Azenbergstr. 12  
D-70174 Stuttgart  
Germany  
zinsmeis,heid@ims.uni-stuttgart.de

## Abstract

This paper combines a corpus-based study of noun+verb collocations with an attempt to distinguish compositional, regularly formed compounds from lexicalised ones. The identification of lexicalised compounds is relevant for both further NLP-applications as well as for traditional lexicography. We claim that morphologically regular, compositional compounds share most of their collocational preferences with their compound heads, whereas lexicalised compounds have their own collocational preferences, distinct or only marginally overlapping with those of their heads.

We test this claim on corpus data for noun+verb collocations of 85 German simple nouns and their compounds (1,058 types). The extraction relies on stochastic parsing, since this procedure provides syntactically homogeneous data and frequency counts, which are input to the log-likelihood association measure, to compute significance scores for the combinations. The experiments seem to confirm our hypothesis: a collocational analysis of this kind can serve to identify lexicalised compounds.

## 1 Introduction

This paper combines a corpus-based study of noun+verb collocations with an attempt to distinguish compositional, regularly formed compounds from lexicalised ones. The study focuses on collocations of noun+verb pairs where the noun is the syntactic head of the verb's direct object. We compare the collocational preferences of noun+noun compounds with those of their heads. We take the term collocation to denote binary word combinations where one element cannot be easily substituted. This includes a whole range of phenomena such as support verb constructions like (*jemandem*

*Gesellschaft leisten* ('keep sb's company'), fully idiomatic chunks like (*jemandem das Handwerk legen* ('put a stop to sb's activities'), as well as collocations in a narrower sense which are habitual combinations like *eine Kette anlegen* 'put on a necklace').<sup>1</sup> Many such combinations occur significantly often in text. Even though many of these combinations have no word by word translations, we do not include semantic non-compositionality as a defining criterion. This definition is roughly equivalent to those of Hausmann (1989) and (2004) and Benson et al. (1986).

The work described here is prompted by the observation that certain morphologically complex nouns, especially compounds, share most or all of their collocates with the respective compound heads (e.g. *die Kleiderfrage klären* (lit. 'settle the dressing question') and *die Frage klären* ('settle the question')).

These facts may be more regularly distributed than it may seem at first sight. It could be that collocational preferences are governed by the same general (i.e. not lexeme-specific) preferences of a semantic nature, as productivity in idiom variation. Lüdeling (2002) discusses collocational and idiomatic examples, where next to *Zähne putzen* ('brush one's teeth') also *seine Beißerchen putzen* ('brush one's little pearlies') is found in texts, as well as collocations with compounds, such as *seine Schneidezähne putzen* ('brush one's incisors'). This phenomenon seems to be related with the productive modifiability of idiomatic expressions, as found in *da stehen dem Schwarzkittel die Borsten zu Berge* (lit. 'that makes the boar's bristles stand on end'),

---

<sup>1</sup>We also include pairs which are in fact parts of a larger idiomatic expression such as *Nase verdienen* as part of the idiom *sich eine goldene Nase verdienen*, see (Zinsmeister and Heid, 2003).

where *Borsten* replaces *Haare* (and the wild boar the person).

The automatic detection of any kind of broader productive modifiability would require extensive ontological knowledge sources as well as complex reasoning tools. In the case of compound nouns it is much easier to relate the modified phrase - the compound - to the original expression - the nominal head. It only requires a morphological analyser and a mapping tool. We therefore concentrate on compounds although the phenomenon might well be part of a more general effect. The focus on compound nouns is also independently justified (cf. Section 2).

We argue that it is relevant to classify productive uses in the domain of collocations of complex nouns, as well as to find a correlation between the collocational behaviour of compounds and their semantic compositionality. For this correlation, we have a working hypothesis; we suggest that there is a correlation between collocational preferences and the lexicalisation of a compound:

- (i) non-lexicalised, productively built (fully compositional) compounds would then share a large number of collocations with their compound heads;
- (ii) whereas lexicalised compounds (which often but not necessarily are non-compositional) would have their own collocations, distinct or only marginally overlapping with those of the compound head.

The collocational behaviour of compounds would then be usable as an indicator of their lexicalisation. A similar approach was pursued by (Lin, 1999) to automatically identify non-compositional expressions. "[T]he metaphorical usage of a non-compositional expression causes it to have a different distributional characteristic than expressions that are similar to its literal meaning".<sup>2</sup>

We set out to empirically test our hypotheses, by using a large German corpus and an automatic tool to extract verb+object collocations. We search for collocations of nouns which have large numbers of determinative compounds, and

---

<sup>2</sup>See also recent work by (Pearce, 2001) and (Baldwin et al., 2003).

where the compounds are frequent enough to throw up enough collocational examples themselves. This will provide us with collocations and frequency data for both compounds and their compound heads; we will compare the respective figures and attempt to interpret them. Our assessment of the lexicalisation of compounds will have to be an intuitive one.

Section 2 motivates the identification of lexicalised compounds by sketching applications in traditional lexicography as well as in an NLP-system. In Section 3, we describe the preparatory steps for the extraction of the raw data from corpus text; Section 4 reports about a case study carried out on 85 German nouns and their determinative compounds, and Section 5 contains a discussion of the results from a lexicographic point of view, and we propose further work.

## 2 Motivation

If the above hypotheses turn out to be confirmed by the empirical data, this has several implications for lexicography:

- dictionaries may mark non-lexicalised compound nouns and possibly indicate that their collocates are likely to be shared with those of the compound head; alternatively, only non-shared collocations could be marked;
- when it comes to space saving in a printed dictionary, non-lexicalised compounds may be left out from the nomenclature more easily than lexicalised ones (or their micro-structural indications may be reduced); collocational preference data might be useful, then, for lexicographers to decide about the inclusion into the dictionary.

For NLP-applications, in particular for statistical approaches that take lexical information into account (such as lexicalised statistical parsing) a distinction between lexicalised and non-lexicalised compounds could help increase efficiency:

- sparse data problems could be reduced by adding frequency counts of non-lexicalised compounds to the frequency counts of their nominal heads. Only lexicalised compounds would then count as independent

types. This is desirable since low frequency counts often result in suboptimal parameter estimations. Higher frequency counts consequently improve the model.

### 3 Preparing the Extraction: Creating Homogeneous Material

The identification of noun+verb-collocations is a difficult task for automatic extraction from German corpora. In contrast to many other types of collocations, they do not necessarily occur in adjacent word sequences. They are not even restricted to a window of  $n$  adjacent items, which poses problems for classical  $n$ -gram approaches and for many flat, chunking-based approaches. We overcome this problem by making use of a full-fledged clausal analysis as a preprocessing step to the collocation extraction.

The problem of non-adjacency holds especially for languages like German that allow for a relatively free word order of nominal arguments. There is no fixed order related to grammatical functions, instead word order and constituent order depend on various factors like information structure, animacy, definiteness, etc. The grammatical relation between a noun and a verb can not be read off the linearisation. Linguistic knowledge like case morphology and subcategorisation information are needed to determine the relation. The nominal arguments do not occur in fixed positions, and even verbs take part in dislocations: German particle verbs split in Verb Second contexts. The finite verbal part occurs in second position whereas the particle remains in the clause-final base position of the verb, such as in:

Eine Pause<sub>ACC</sub> legte<sub>V</sub> er<sub>NOM</sub> heute nicht  
ein<sub>PARTICLE</sub>  
'He did not take a break today'.

We used a statistical grammar (Schulte im Walde, 2003) that covers the phenomena described above: it recognises (split) particle verbs and identifies verbal arguments independently from linearisation. A manually established context-free grammar with feature constraint annotations functions as backbone. It is trained iteratively by a statistical parser (LoPar, Schmid (2000)) on a newspaper corpus of approximately 35 million words. During training

the grammar rules are enriched with information about their lexical heads. This lexicalisation allows to read off from the trained grammar model the co-occurrence frequencies of lexical heads that are related by grammatical structure (such as object+verb). From an abstract point of view, this lexical co-occurrence data represents a syntactically homogeneous data set of direct object+verb pairs (in the terms of Evert and Krenn (2001)). It can easily be fed into a lexical association measure algorithm.

We extracted pairs of verbs and their internal arguments, i.e. their accusative objects in active clauses or their nominative subjects in passive clauses, thereby generalising over syntactic variations such as linearisation of verb and object, or voice (active/passive alternation) and finiteness of the verb. This ensures exploiting the broad coverage of the grammar, thereby increasing recall and precision.

Adding up all occurrences of a given noun independently of the verbal head results in the estimated frequency count<sup>3</sup> of the noun in general. The same holds for a given verb, respectively. Counting all pair frequencies by generalising over the lexical heads gives the total number of noun+verb-pairs that constitute the background on which the collocations are to be identified.

We ordered the extracted noun+verb-pairs in two ways. Firstly, with respect to their estimated frequency<sup>4</sup>. Secondly, with respect to their log-likelihood score to identify more reliably<sup>5</sup> such pairs that have a high relative association, i.e. potential collocations (cf. Dunning (1993)). The score expresses a degree of confidence with which we can reject the assumption that the co-occurrence of a noun+verb-pair is mere coincidence.

---

<sup>3</sup>Estimated frequencies are fractions instead of discrete number counts. This is due to the fact that estimation takes into account the probability of the parse tree that includes the target structure which normally competes with alternative analyses of the same sentence.

<sup>4</sup>A related approach, a combination of estimated frequency and modelled probability in an EM-based classification model, was used by Prescher (2002) to extract collocation candidates.

<sup>5</sup>Cf. (Evert et al., 2000).

## 4 Experiment

Our experiments are organised as follows. For 85 heads that are related to a total number of 7,518 compound types, we (i) extracted compound+verb-pairs that came with a token frequency higher than 5.0 (1,058 types) and the verb of which did not co-occur with the respective compound head. This turned out to be a quite reliable method for identifying lexicalised compounds. Due to the relatively high frequency threshold, this method cannot deal with sparse data. (ii) To reduce the sparse data problem, we lowered the frequency threshold for compound+verb-pairs to 2.5 (2,024 types) and allowed co-occurrence of the verb and the respective head but added restrictions on the log-likelihood scores (Dunning, 1993) of the word pairs.<sup>6</sup> This method improved the recall.

## 5 Results and Discussion

In our experiments, we manually inspected data for most of the 85 nouns (and their compounds) in our test sample. In this section, we first describe a few clear cases: in Section 5.1.1, we show nouns which share most of their collocates with their head nouns, whereas in Section 5.1.2, we give examples of compounds which have collocations distinct from those of the respective compound heads. To provide a first evaluation, we picked 40 candidates by frequency and analysed the respective data manually (Section 5.2.1). Finally, we used log-likelihood figures to compare collocation preferences of compounds and of their heads; the results are described in Section 5.2.2 and illustrated in Table 1 and Table 2.

### 5.1 Qualitative Evaluation

#### 5.1.1 Shared Collocations

As a sample, the nouns *Fest* and *Kraft* are analysed in more detail, because many of their compounds occur frequently enough to provide interpretable collocational data. We mainly analysed the collocational preferences of noun+noun compounds, but the results seem to

---

<sup>6</sup>The log-likelihood score is not fully adequate for our data since it calculates discrete frequency counts whereas our data consists of continuous estimated frequencies. As the error is expected to be small (see (Evert, to appear) for an extensive discussion on association measures) we used the standard implementation here (see e.g. [www.collocations.de](http://www.collocations.de)).

carry over to verb+noun compounds as well (cf. *Führungskraft<sub>N</sub>* vs. *Schreibkraft<sub>N</sub>*).

The most prominent verbal collocates of *Fest* are *feiern* (estimated frequency of 88.24), *eröffnen* (20.45), *planen* (12.79), *veranstalten* (11.75), and *machen* (10.87). Considering all 94 compounds of *Fest* (types), such as *Abschiedsfest*, *Brezelfest*, *Fußballfest* found in the corpus, 38.21 % of all observed collocations (tokens) of these compounds contain the verb *feiern*; *feiern* was observed in collocations of 49 of the 94 (52.13 % of the) compound types. The next important collocates with compounds of *Fest* are *eröffnen*, *planen*, *veranstalten*, *machen*, *organisieren*, *besuchen*; these verbs account for another 24.87 % of the analysed occurrences.

Another example of the same type is the noun *Kampf*. The collocates shared by many of its compounds are *führen* (13.82 total occurrences), *eröffnen* (3.65 %), *verlieren*, *gewinnen*, *fortsetzen*, *beenden*, *liefern*, *entbrennen*, *ausfechten*, *austragen*, *entscheiden* (together 10.58 % of the occurrences). Interestingly, the second most frequent combination with *Kampf* is *jmdm den Kampf ansagen* ('to challenge sb'). This collocation is lexicalised (or idiomatised?) and it seems to be restricted to the noun *Kampf*, as a combination with *ansagen* seems impossible with any compound of *Kampf*.

The above example illustrates a case where the collocational behaviour of compounds is partly shared with from that of the head. The nouns analysed are mainly monosemous. A polysemous case is the noun *Kraft*. Its compounds fall into two groups: (a) 'power, strength, force': *Triebkraft*, *Symbolkraft*, *Ausdruckskraft*, *Durchsetzungskraft*, ... and (b) 'employee, personnel': *Nachwuchskraft*, *Honorarkraft*, *Führungskraft*, *Halbtageskraft*, ...

Along with the two distinct semantic groups, collocations also group together. With group (a), prominent verbs are *haben*, *stärken*, *bündeln*, *verlieren*, *verleihen*, *beweisen*, *entfalten*, whereas group (b) has *einsetzen*, *einstellen*, *freisetzen*, *suchen*, *anstellen*. Very few - unspecific and likely not collocationally relevant - verbs show up with compounds of both groups: *brauchen*, *geben*, *entwickeln*.

#### 5.1.2 Lexicalised Cases

From the estimated frequency figures for collocations, separately for heads and for com-

pounds, it is easy to extract those cases where a given compound has a highly frequent collocation with a verb and where this verb does not collocate with the respective head at all. This case is the inverse of den *Kampf ansagen*. A few prominent examples are: *Autobahn, Fahrbahn + sperren*, but not *\*Bahn + sperren; Bußgeld verhängen*, but not *\*Geld + verhängen; Hilfestellung + leisten*, but not *\*Stellung + leisten*.

These examples all contain lexicalised compounds which are morphologically regular, but not (e.g. *Hilfestellung*) or only partially (e.g. *Bußgeld*) semantically compositional. Among the heads concerned are mainly very general ones (e.g. *Art, Werk, Wert, Punkt*) which give rise to semantically opaque compounds (like *Handwerk, Kunstwerk, Feuerwerk, Standpunkt, Sportart* etc.).<sup>7</sup>

## 5.2 Quantitative Evaluation

### 5.2.1 A Frequency-Based Evaluation

To evaluate our procedures against the hypothesis that lexicalised compounds do not share (many of their) collocates with their compound heads, we inspected the collocations found with 40 candidates classified as lexicalised compounds by our tools.

The results are depicted in Table 1. 29 of the candidates occurred mainly in non-shared, idiosyncratic collocations such as *Alarmanlage* (‘alarm system’) and *Anhaltspunkt* (‘clue’). Nine candidates such as *Arbeitskampf* (‘industrial action’) and *Arbeitskraft* (‘capacity for work, worker’) showed a mixed behaviour: they have an overlap with the collocations of their head but are rather idiomatic. Only two out of the 40 candidates mainly share the collocation preferences of their head. We take this result as a confirmation of the hypothesis that the analysis of collocational behaviour can be used for identifying candidates of lexicalised compounds.

### 5.3 Log-Likelihood Scores

To evaluate the rest of the extraction experiments, we manually determined lexical compounds from the compound list without taking the noun+verb-collocations into consideration

<sup>7</sup>Interestingly, we have however the following collocations: *Buße verhängen* (‘fine’, lit. ‘impose a fine’) and *Hilfe leisten* (‘provide assistance’). Indeed the semantics of *Bußgeld* is close to that of *Buße*, and so is that of *Hilfestellung* to its non-head *Hilfe*.

<b>Forming idiomatic collocations</b>
Alarmanlage, Anhaltspunkt, Autobahn, Besatzungsmitglied, Bußgeld, Eigentumsverhältnis(-se), Fahrbahn, Feindbild, Feuerwerk, Gangart, Größenordnung, Handwerk, Hilfestellung, Höhepunkt, Meinungsbildung, Notdienst, Spieleabend, Sportangebot, Sportart, Standpunkt, Stellenwert, Streckenführung, Streitwert, Umweltschutz, Urstand, Verkehrsführung, Verwarnungsgeld, Waffenstillstand, Zeitpunkt
<b>Mixed but rather idiomatic</b>
Arbeitskampf, Arbeitskraft, Autofahrer, Grenzwert, Kopfgeld, Motorradfahrer, Ozonwert, Sozialhilfe, Wahlkampf
<b>Sharing of collocation</b>
Mißtrauensantrag, Pressekonferenz

Table 1: Candidates for lexicalised compounds

and compared them with the collocational results. The test used in the manual selection exercise was whether a compound can be replaced by its head without a change in meaning that goes beyond hyperonymy: for example *Verteidigerstellung* (‘position of defender’) can be replaced by its hypernym *Stellung*. We conclude from this behaviour that *Verteidigerstellung* is not a lexicalised compound, whereas *Problemstellung* (‘way of looking at a problem’) is indeed lexicalised, as it is not a specific type of *Stellung*. More generally, compounds of *Stellung* tend to be non-lexicalised as long as the first part of them are common nouns.<sup>8</sup>

Compounds with the head *Art* (‘kind’) deviate from the morphological right-hand head rule in that the semantic head of the compound is its first part, for instance *Baumart* ‘tree type’ is not a kind of *Art* ‘kind’ but a kind of *Baum* (‘tree’). There is a whole class of nouns that behave the same. We expect therefore that there is no significant match in the collocational behaviour of these compounds and their head. This is in fact born out. Among 25 pairs there are only four which have a positive count for a related com-

<sup>8</sup>Due to ambiguities in the morphological analysis the test items include words like *Feststellung* or *Klarstellung*. They are in fact not compounds but nominalisations of complex verbs like *feststellen*, *klarstellen*, and *zufriedenstellen*. We expect them to be opaque in meaning, and also to show individual collocation preferences. They are indeed identified as lexicalised items by our tools.

pound. None of the nine most prominent collocations of *Art*+verb co-occur with any compound of *Art*. Table 2 shows sample data from our experiments. Each line of the table starts with a verb(al collocate) and has, in its second column, a noun, in its fifth column a compound of that noun. Columns 3 and 4 contain figures for the log-likelihood ratio of the collocation between the simple noun and the verb, and for its absolute frequency, respectively. Similarly, columns 6 and 7 contain log-likelihood and frequency figures for the collocation of the verb and the compound.

We consider *Hilfestellung*, for example, as lexicalised: as mentioned above, *Stellung+leisten* does not occur in our corpus (thus "0.00" in columns 3 and 4), whereas *Hilfestellung leisten* occurs with an estimated frequency of 32.55 (column 7) and has a high log-likelihood value (column 6: 306.30). Similarly, *Marktstellung* ('market position') is considered as non-lexicalised, with respect to *Stellung* ('position'): the collocation with *ausbauen* has considerable log-likelihood values with each (18.92 with *Stellung* and 39.30, with *Marktstellung*).

#### 5.4 Residuals

To improve the recall on lexicalised compounds, we tested additional simple grouping heuristics. We extracted sets of compounds with five or more members that co-occur with a verb that has zero-frequency with the corresponding base. This method lead to 77 additional candidates, among them e.g. *Mundwerk* and *Gottesdienst*. Another heuristic was extracting compounds that co-occur with a set of four or more verbs that have all zero-frequency with the corresponding base. We found eleven additional candidates this way, including *Badenwerk* (which is a proper name) and *Dauerwelle*.

Compounds involving proper names or hyphenated compounds were evaluated independently. The latter often feature a proper name or an abbreviation as first compound part. Both types tend to be non-lexicalised and inherit their collocational preferences from their bases.

## 6 Summary

For the extraction of noun+verb-collocations from a German corpus, we used a stochastic grammar; it produced syntactically homogeneous material that was subsequently sorted

by means of an association measure.

We ran an experiment on the collocational preferences of compound nouns, as compared to those of their compound heads. The results suggest that collocate selection is mostly shared between heads and those compounds which are built according to productive morphological rules, and which are not only morphologically, but also semantically compositional. Lexicalised compounds, however, tend to have their own collocations which do not (or only partially) overlap with those of their heads.

This study is part of work on dictionary updating tools, which compare, for example, the head word list of an existing dictionary (that needs to be updated) with the most frequent lemmas from corpora.<sup>9</sup> Many corpus words proposed for inclusion into the dictionary are compounds, and often frequency is not the only criterion for a lexicographer to really include them. If an automatic tool such as the one sketched here could provide hints as to the lexicalisation status of the compounds, another criterion for inclusion into or removal from the dictionary would be operationalised: most lexicographers would want lexicalised compounds to be captured in their dictionary, whereas not all non-lexicalised, productively formed ones are seen as indispensable.

---

<sup>9</sup>This work was carried out in part under the Transferbereich TFB-32, Automatische Exzerption. We acknowledge gratefully the support given to Universität Stuttgart, in this project, by the Deutsche Forschungsgemeinschaft, DFG. For details on TFB-32, see Evert et al. (2004).

verb v	head h	ll(h,v)	f(h,v)	compound c	ll(c,v)	f(c,v)
haben	Abend	0.43	15.65	Feierabend	23.54	6.66
verbringen	Abend	168.73	22.07	Lebensabend	527.35	39.84
haben	Art	8.33	60.47	Eigenart	12.40	5.07
einschlagen	Art	0.00	0.00	Gangart	143.74	11.94
ankündigen	Art	0.16	1.00	Gangart	65.07	9.13
einlegen	Art	0.00	0.00	Gangart	54.18	6.58
trainieren	Art	0.00	0.00	Kampfsportart	50.74	2.86
haben	Art	8.33	60.47	Spielart	10.84	2.85
kennenlernen	Art	0.95	1.00	Sportart	78.07	7.00
ausüben	Art	1.32	1.09	Sportar	27.27	2.91
organisieren	Fest	20.54	6.00	Sommerfest	32.18	4.64
organisieren	Fest	20.54	6.00	Straßenfest	32.71	3.90
betreiben	Kampf	0.00	0.00	Wahlkampf	62.68	12.19
entscheiden	Kampf	36.74	11.17	Machtkampf	59.82	7.47
finanzieren	Kampf	0.00	0.00	Wahlkampf	24.02	5.92
einläuten	Kampf	0.00	0.00	Wahlkampf	27.76	3.50
dominieren	Kampf	0.00	0.00	Wahlkampf	17.56	2.91
treffen	Stellung	0.00	0.00	Feststellung	20.07	3.93
leisten	Stellung	0.00	0.00	Hilfestellung	306.30	32.55
erhoffen	Stellung	0.00	0.00	Hilfestellung	20.47	2.78
bieten	Stellung	0.00	0.00	Hilfestellung	25.36	5.41
verlangen	Stellung	0.00	0.00	Klarstellung	27.73	4.48
fordern	Stellung	0.00	0.00	Klarstellung	14.63	3.80
zerstören	Stellung	0.00	0.00	Luftabwehrstellung	39.55	3.00
ausbauen	Stellung	18.92	6.53	Marktstellung	39.30	4.00
verlieren	Stellung	17.68	12.75	Monopolstellung	18.15	3.00
verbessern	Stellung	1.52	2.01	Rechtsstellung	34.07	3.00
einnehmen	Stellung	52.26	12.20	Spitzenstellung	96.02	8.52
vornehmen	Stellung	0.00	0.00	Weichenstellung	47.68	4.99
erwarten	Stellung	0.00	0.00	Weichenstellung	14.86	2.96
abschalten	Werk	1.88	0.98	Atomkraftwerk	291.55	23.13
bauen	Werk	64.22	22.22	Atomkraftwerk	74.23	12.37
instandsetzen	Werk	0.00	0.00	Bauwerk	41.84	2.98
zerstören	Werk	0.05	1.00	Bauwerk	17.06	2.99
abbrennen	Werk	0.00	0.00	Feuerwerk	166.53	10.57
entfachen	Werk	0.00	0.00	Feuerwerk	64.64	4.87
veranstalten	Werk	0.00	0.00	Feuerwerk	23.42	3.00
legen	Werk	0.25	1.07	Handwerk	502.89	57.32
lernen	Werk	0.00	0.00	Handwerk	135.75	18.40
erlernen	Werk	0.00	0.00	Handwerk	109.22	10.00
beherrschen	Werk	0.00	0.00	Handwerk	80.84	10.94
verstehen	Werk	0.08	1.66	Handwerk	96.54	16.99
erschweren	Werk	0.00	0.00	Handwerk	17.31	3.00
betreiben	Werk	1.63	3.02	Kernkraftwerk	19.68	3.00
liefern	Werk	2.17	3.69	Kernkraftwerk	16.12	2.68
abschalten	Werk	1.88	0.98	Kraftwerk	74.44	7.00
stilllegen	Werk	22.25	5.00	Kraftwerk	58.82	6.00
betreiben	Werk	1.63	3.02	Kraftwerk	36.89	6.00
besetzen	Werk	12.88	6.77	Kraftwerk	14.98	2.99
schaffen	Werk	0.02	4.12	Kunstwerk	29.08	7.48
zerstören	Werk	0.05	1.00	Kunstwerk	13.47	2.99
zerstören	Werk	0.05	1.00	Lebenswerk	31.16	4.00
aufbauen	Werk	4.77	3.96	Netzwerk	77.23	8.99
tragen	Werk	4.63	0.51	Schuhwerk	27.31	3.91
verlieren	Werk	2.20	0.94	Triebwerk	23.22	3.96
billigen	Werk	0.49	1.00	Vertragswerk	31.27	3.21

Table 2: Sample extraction results

## References

- Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An Empirical Model of Multiword Expression Decomposability. In F. Bond, A. Korhonen, D. McCarthy, and A. Villavicencio, editors, *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96.
- M. Benson, E. Benson, and R. Ilson. 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins, Amsterdam and Philadelphia.
- Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19:1:61–74.
- Stefan Evert and Brigitte Krenn. 2001. Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.
- Stefan Evert, Ulrich Heid, and Wolfgang Lezius. 2000. Methoden zum Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In Ernst G. Schukat-Talamazzini Werner Zühle, editor, *KONVENS-2000 Sprachkommunikation*, pages 215–220. VDE-Verlag.
- Stefan Evert, Ulrich Heid, Bettina Säuberlich, Esther Debus-Gregor, and Werner Scholze-Stubenrecht. 2004. Supporting corpus-based dictionary updating. In *Proceedings of Euralex 2004*.
- Stefan Evert. to appear. *The Statistics of Word Cooccurrences: Bigrams and Collocations*. Ph.D. thesis, University of Stuttgart.
- Franz-Josef Hausmann. 1989. Kollokationen in deutschen Wörterbüchern. Ein Beitrag zur Theorie des lexikographischen Beispiels. *Lexikographie und Grammatik*.
- Franz-Josef Hausmann. 2004. Was sind eigentlich Kollokationen? In Kathrin Steyer, editor, *Wortverbindungen – mehr oder weniger fest*, pages 309–334. de Gruyter, Berlin/New York. [= IDS, Jahrbuch 3002].
- Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 317–324. College Park, MD.
- Anke Lüdeling. 2002. Special Topic Session: The productivity of collocations. Presentation at Colloc 2002, Vienna.
- Darren Pearce. 2001. Synonymy in Collocation Extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*, Pittsburgh. Carnegie Mellon University.
- Detlef Prescher. 2002. *EM-basierte maschinelle Lernverfahren für natürliche Sprachen*. Ph.D. thesis, University of Stuttgart. Published as AIMS Vol.8, No.2.
- Helmut Schmid. 2000. Lopar: Design and Implementation. Arbeitspapiere des Sonderforschungsbereichs 340 *Linguistic Theory and the Foundations of Computational Linguistics* 149, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
- Sabine Schulte im Walde. 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph.D. thesis, University of Stuttgart. Published as AIMS Report 9(2).
- Heike Zinsmeister and Ulrich Heid. 2003. Significant Triples: Adjective+Noun+Verb Combinations. In *Proceedings of Complex 2003*, Budapest, Hungary. Complex 2003.