Extraction of V-N-Collocations from Text Corpora: A Feasibility Study for German

Elisabeth Breidt*
Seminar für Sprachwissenschaft
University of Tübingen
Kleine Wilhelmstr. 113, D-72074 Tübingen
breidt@arbuckle.sns.neuphilologie.uni-tuebingen.de

Abstract

The usefulness of a statistical approach suggested by Church and Hanks (1989) is evaluated for the extraction of verb-noun (V-N) collocations from German text corpora. Some motivations for the extraction of V-N collocations from corpora are given and a couple of differences concerning the German language are mentioned that have implications on the applicability of extraction methods developed for English. We present precision and recall results for V-N collocations with support verbs and discuss the consequences for further work on the extraction of collocations from German corpora. Depending on the goal to be achieved, emphasis can be put on a high recall for lexicographic purposes or on high precision for automatic lexical acquisition, in each case leading to a decrease of the corresponding other variable. Low recall can still be acceptable if very large corpora (i.e. 50 - 100 million words) are available or if corpora are used for special domains in addition to the data found in machine readable (collocation) dictionaries.

1 Introduction

Collocations present an area that is important both for lexicography to improve their coverage in modern dictionaries as well as for lexical acquisition in computational linguistics, where the goal is to build either large reusable lexical databases (LDBs) or specific lexica for specialized NLP-applications. We have tested the statistical approach Mutual Information (MI), brought up by Church and Hanks (1989) for linguistics, for a (semi-)automatic extraction of verb-noun (V-N) collocations from untagged German text corpora. We try to answer the question how much can be done with an untagged corpus and what might be gained by lemmatizing, POS-tagging or even superficial parsing.

Choueka (1988) describes how to automatically extract word combinations from English corpora as a preselection of collocation candidates to ease a lexicographer's search for collocations. He only uses quantitative selection criteria, no statistical ones, his main extraction criterion being frequency with a lower threshold of at least one occurrence of the collocation in one million words. He mentions plans to define a

*My thanks go to the IdS that made the two corpora available for research purposes, to Angelika Storrer for her steady encouragement and many fruitful discussions, and to Mats Rooth and Matthias Heyn who introduced me to the corpora tools. I am also greatful to the anonymous reviewers for their helpful comments and constructive criticism.

'binding degree' on how strong the words of a collocation attract each other, which would be similar in spirit to what is calculated with MI. The work described in Smadja and McKeown (1990) and Smadja (1991a.b) is along the same lines as ours, though he uses a different statistical calculation, a z-score, and tagged, lemmatized corpora. Some properties specific to German, however, lead to a type of problem that needs different treatment (section 3.3). Calzolari, Bindi (1990) use MI to extract compounds, fixed expressions and collocations from an Italian corpus, but to our knowledge have not evaluated their results so far.

2 Domain of Investigation

2.1 What Do We Mean by 'Collocation'?

Collocations in the sense of 'frequently cooccurring words' can quite easily be extracted from corpora by statistic means. From a linguistic point of view, however, a more restricted use of the term is preferable which takes into account the difference between what Sinclair (1966) called *casual* vs. *significant* collocations. Casual word combinations show a normal, free syntagmatic behaviour. In this paper, collocations shall refer only to word combinations that have a certain affinity to each other in that they follow combinatory restrictions not explainable with syntactic and semantic principles (e.g. *hammer a nail into sth.* rather than **beat a nail into sth.*).

For collocations that are based on a verb and a noun (preferably an object argument, sometimes however the subject of an intransitive verb), three types of V-N combinations are distinguished for German in the literature: verbal phrasemes (idioms) (e.g. Brundage et al. 1992), support verb constructions (SVCs) (v.Polenz 1989 or Danlos 1992) and collocations in the narrower sense (Hausmann 1989). As Brundage et al. (1992:7) and Barkema (1989:24) point out, the differences between these three types are gradual and "it is hard to find criteria of good selectivity to distinguish collocations from phrasemes". Although our main interest lies in SVCs we will in the following not distinguish between i) SVCs (e.g. *to take into consideration*), ii) lexicalized combinations with support verbs where the noun has lost its original meaning and which belong to phrasemes (e.g. *to take a fancy*), and iii) collocational combinations of support verbs with concrete or non-predicative nouns (e.g. *to take a seat*); we will refer to all these cases as V-N collocations.

2.2 Why V-N Collocations?

Collocations are well suited for statistical corpora studies. The semantics of a collocation in the narrower sense according to Fleischer (1982:63f) is "given by the individual semantics of its components, its meaning differs however in an unpredictable way from the pure sum of its parts. A substantial cause for this unpredictable difference is the frequency of occurrence and the probability with which the occurrence of one component determines the occurrence of the other" (our translation). The unpredictability of a collocation is thus partly caused by the high cooccurrence frequency of its components compared to the relative frequency of the single words. This holds even more for SVCs and phrasemes due to their (partly) non-compositional semantics.

In German, common nouns, proper names and abbreviations of names start with an uppercase letter (sentence beginnings are changed to lowercase in the corpus). So the verb-noun pattern was chosen for our study instead of possible others, because the uppercase makes it possible to extract V-N collocations even from untagged corpora if the verb is used as the key-word. The results of extracting V-N collocations give good indications how promising the retrieval of collocations would be with POS-tagged corpora. Besides, N-N combinations in German are mainly restricted to proper names, and Adj-N collocations are not as extensive in our corpus due to the small number of frequent and interesting adjectives.

3 Resources and Methods Used in the Study

Two untagged corpora were used for our study. kindly supplied by the 'Institut für deutsche Sprache' (IdS), Mannheim: the 2.7 million words 'Mannheimer Korpus I' (MK1) which contains approx. 73% fiction and scientific/philosophical literature and about 27% newspaper texts, and the 'Bonner Zeitungskorpus' (BZK), a 3.7 million words newspaper corpus. Except for the test how results could differ for larger corpora described in section 4.5, where the MK1 was combined with the BZK, the investigation was based on the MK1 on its own, for technical reasons and also because verbs occur more often on average in the MK1 than in the BZK (cf. Breidt 1993).

3.1 Statistical Method and Tools

MI is a function well suited for the statistical characterization of collocations because it compares the joint probability p(*w1*,*w2*) that two words occur together within a predefined distance with the independent probabilities p(*w1*) and p(*w2*) that the two words occur at all in the corpus (for a more detailed description see Church et al. (1991:120) or Breidt (1993:18)):

$$\mathrm{MI}(x,y) = \log_2 \frac{\mathrm{p}(x.y)}{\mathrm{p}(x)\mathrm{p}(y)}$$

Several methods are possible for the calculation of probabilities (cf. Gale and Church 1990); for our purposes we use the simplest one. where the frequency of occurrence in the corpus is divided by the size N of the corpus, $\mathrm{p}(x) = \mathrm{f}(x)/\mathrm{N}$. Distance will be defined as a window-size in which bigrams are calculated.

MI does not give realistic figures for very low frequencies. If a relatively unfrequent word occurs only once in a certain combination, the resulting very high MI value suggests a strong link between the words where it might well be simply by chance. So a lower bound of at least 3 occurrences of a word pair is necessary to calculate MI. The t-test used to check whether the difference between the probability for a collocational occurrence and the probability for an independent occurrence of the two words is significant, is a standard significance test in statistics (e.g. Hatch and Farhady 1982). The statistical calculations were done as described in Church et al. (1991), and were performed together with KWIC queries and the creation of bigrams using tools available at the 'Institut für Maschinelle Sprachverarbeitung', University of Stuttgart[1].

3.2 The 'Standard' Method

Verbs that can occur in SVCs are in the centre of our study because they provide examples for all three types of V-N collocations; besides, the chosen 'potential' support verbs belong to the most frequent verbs in the corpus anyway. V-N collocations were extracted for the following 16 verbs (no translations are given because they differ depending on the N argument): *bleiben*, *bringen*, *erfahren*, *finden*, *geben*, *gehen*, *gelangen*, *geraten*, *halten*, *kommen*, *nehmen*, *setzen*, *stehen*, *stellen*, *treten*, *ziehen*.

Bigram tables of all words that occur within a certain distance of these verbs, together with their cooccurrence frequencies, form the basis for the calculation of MI. Bigram calculations were restricted to words occurring within a 6-word window to the *left* (cf. next section), inclusive of the verb, a span which captures 95% of significant collocations in English (Martin et al. 1983). We will refer to these with B16. For combinations that occur at least 3 times, MI was calculated together with a t-score. From these, candidates for V-N collocations were automatically extracted, sorted by MI. All of these were checked by means of KWIC-listings and classified w.r.t. their collocational status by the author. The classification was in most cases very obvious. If a combination potentially formed a collocation but was not used as such in the corpus it was not counted; a couple of times, where some of the usages were indeed collocations and others not, the decision was made in favour of the predominant case.

3.3 Application for German Corpora: Some Problems

Some properties of the German language make the task of extracting V-N collocations from German text corpora more difficult than for English corpora. A minor difference concerns the strong inflection of German verbs. Whereas in English a verb lexeme appears in 3 or 4 different forms plus one for the present participle, German verbs have 7 to 10 verb forms (without subjunctive forms) for one lexeme and additional 4 for the present participle. This has to be considered for the evaluation of queries based on single inflection forms, because in English more usages are covered with one verb form than in German.

Another point concerns the variable word order in German (see Uszkoreit 1987) which makes it more difficult to locate the parts of a V-N collocation. In a main clause (verb-second order), a noun preceding a finite verb usually is the subject, but it can also be a topicalized complement; in sentences where the main verb occurs at the end (nonfinite verb or subordinate clause) the preceding noun is mostly a direct object or other complement, or an adjunct. A noun to the right of a finite verb can be any of subject, object or other argument due to topicalization or scrambling. We restrict our search to V-N combinations where the noun precedes the verb either directly or within two to five words, because this at least definitely captures complements of main verbs

[1] We greatfully acknowledge that the work reported here would not have been possible without the supplied tools and corpora.

in verb-final position. To find the correct argument to the right of the verb is difficult in an unparsed corpus because of the variable number of intervening constituents.

As illustrated in the last paragraph the assumption that a "semantic agent [...] is principally used before the verb" and a "semantic object [...] is used after it" as described in Smadja (1991a:180) does not hold for German. Therefore, complicated parsing is necessary to distinguish subject-verb from object-verb combinations. The results of V-N extractions reflect this problem. In many if not in most of the uninteresting combinations extracted, the noun to the left of the verb is the subject rather than a complement of the verb (cf. section 4.6).

4 Evaluation of the Results

Below, the top bigrams with *kommen (come)* are shown, and some of the nonsignificant ones (t $\leq$ 1.65), to illustrate MI and t-scores. Bigrams with the infinitive form give best results compared to other inflection forms, possibly because this form covers 1st/3rd pers. pl. present tense, the infinitive and the nonfinite main verb of complex tenses (modals, conditional, future) at the same time. Also, the latter two always occur in verb-final position.

N + *kommen*	Translation	f(x,y)	f(y)	MI	t-score	V-N-Coll.
(zur) Geltung k.	show to advantage	27	96	9.86	5.19	+
(in) Betracht k.	to be considered	9	42	9.47	2.99	+
(in) Berührung k.	come into contact	4	41	8.33	1.99	+
(zur) Anwendung k.	to be used	4	126	6.71	1.97	+
(zu) Tränen k.	come to tears	3	107	6.53	1.70	+
(zur) Ruhe k.	get some peace	4	216	5.93	1.95	+
(auf den) Gedanken k.	get the idea	7	403	5.84	2.58	+
(in den) Himmel k.	go to heaven	3	270	5.20	1.66	+
(zu) Hilfe k.	come to aid	4	477	4.79	1.89	+
...						
(zu) Wort k.	get a chance to speak	3	647	3.94	1.57	+
Vernunft	reason	3	736	3.75	1.55	-
(in) Frage k.	to be possible	4	1054	3.65	1.77	+
(zur) Welt k.	to be born	4	1900	2.80	1.60	+
Sie	You	3	2414	2.04	1.17	-

4.1 Precision and Recall

The question how much is extractable fully automatically can be answered by an evaluation of precision and 'recall' of the described method as it is done for memory tests. Following Smadja (1991a) we define *precision* as the number of correctly found collocations divided by the number of V-N combinations found at all. *Recall* reflects the ratio of the number of correctly found collocations and the maximal number of collocations that could possibly have been found. The latter is slightly difficult to determine, because in principle this means to know the total number of collocations occurring in the whole corpus. Another possibility, to take all collocations that are mentioned in a dictionary as the maximal number of valid collocations, had to be discarded: a comparison with Agricola (1970) or Drosdowski (1970) is not really possible because the collocations found in the corpus are not a subset of those mentioned in the dictionaries. Only 22 of the 43 collocations found with the lemma *bring-* in the MK1 (BI6) belong to the 135 combinations mentioned in the lexical entry for *bringen* in Agricola (1970). Of the remaining 21 in the MK1, 9 can be found in the corresponding noun entries, and 12 do not appear at all though they are 'significant' collocations, e.g. *Klarheit bringen (clarify). zur Entfaltung br. (develop), zur Wirkung br. (bring the effect), in Schwierigkeiten br. (create difficulties), ins Gespräch br. (bring into discussion).* Thus, we decided to use instead the number of collocations with the infinitive as determined by the standard method (BI6) as the basis for recall comparisons, i.e. 100% recall is set to this number.

4.2 Results of the Standard Method

Frequencies for the infinitives of the 16 verbs range from 832 (*kommen*) to 117 (*gelangen*). The number of V-N combinations varies from 46 (*bringen*) to 6 (*erfahren, gelangen, geraten, treten*), precision from 100% (*geraten, ziehen*) to 33% (*erfahren*). Average figures are presented in table 1 below, labeled BI6 Inf. If non-significant combinations are omitted with a t-test (BI6/t Inf), the average of collocations among the extracted V-N combinations is only 95.8% of those found without a significance boundary, but precision rises slightly. With a threshold of MI $\geq$ 6, precision would go up to 82.1% with a still acceptable loss in recall of approximately 10%.

4.3 Experiment 1: Variation of Window-Size

To see whether the collocational nouns could be better located directly to the left of the verb rather than within a couple of words, we reduced window-size to 3 words *including* the verb (this allows one word in between, e.g. 'zu' *(to)* in infinitival constructions). As shown in table 1 for BI3 Inf, precision rises about 10%, but with a recall of 72.1%, because those collocations where other arguments or post modifiers occur between N and V are no longer captured. Taking again only significant combinations (BI3/t Inf) precision rises again slightly. This leads to the conclusion that for German, unless syntactic relations can be determined, a smaller window is preferable to improve a correct detection of preceding object arguments and to exclude unrelated nouns.

Table 1: Average figures for varying window-size and lemmatizing

Bigrams	Ø V-N	Ø Collocations	Precision %	Recall %
BI6 Inf	21.5	13.5	66.3	100 (def.)
BI6/t Inf	18.25	12.9	71.6	95.8
BI3 Inf	12.4	9.5	81.2	72.1
BI3/t Inf	11.5	9.1	83.1	70.0
BI3 Lemma	29.9	16.1	59.8	114.7

4.4 Experiment 2: Simulating Lemmatizing

Because no lemmatizing program was available we used an additional program on top of the bigram calculations for the inflected forms. In order to keep the amount of V-N combinations within a magnitude that could still be checked manually for correctness, we restricted search to a 3-word window to the left. V-N combinations that occurred less than two times with a single inflection form of the verb were sorted out. The inflection forms for the infinitive (also 1st/3rd pers. pl.), 3rd pers. sg. present and past tense, 1st/3rd pers. pl. past and past participle were added up; 1st pers. sg. and 2nd pers. sg./pl. were so rare that they could be ignored. The average results are again presented in table 1 (BI3 lemma); the number of extracted collocations is maximal, but precision is the lowest of all. Precision ranges from 33.3% (*gehen*) to 88.2% (*setzen*), recall from 50% (*erfahren*) to 166.7% (*setzen*). Recall figures are above 100% because the absolute number of collocations found is higher than for BI6 Inf, the basis for the recall calculations. Regarding lemmatization our study shows that one gets more collocations, but at the expense of more uninteresting combinations as well. One explanation for this is that 3rd pers. sg. present/past and 1st/3rd pers. pl. past only occur to the right of their noun argument in subordinate clauses, whereas 1st/3rd pers. pl. present are identical with the nonfinite form which additionally occurs in verb-final position in main clauses with a finite auxiliary or modal verb and in infinitive clauses.

4.5 Experiment 3: Varying Corpus Size

For infinitive *bringen* and lexeme *bring-*, V-N combinations were also calculated with BI6 for a larger corpus consisting of the MK1 and BZK together. For MK1 alone, 31 of 46 combinations are collocations, a precision of 67.4% (recall is set to 100%). With the larger corpus the number of found V-N collocations is more than twice as big, with only a slightly lower precision[2]. Thus, larger corpora would improve results considerably. Results for the lexeme with the highest number of collocations at all (73) are along the same lines; however almost every second V-N combination is no V-N collocation in the sense defined in section 2, i.e. results are much better overall for the infinitive separately. The complete data for *bringen* are listed below.

Table 2: Variations for the verb *bringen*

Bigrams	f(V)	V-N	Coll.	Precision %	Recall %
BI6 Inf	550	46	31	67.4	100 (def.)
BI6/t Inf	550	44	31	70.5	100
BI6 MK1+BZK Inf.	1065	97	63	65.0	203
BI3 Inf	550	33	28	84.9	90.3
BI3/t Inf	550	31	27	87.1	87.1
BI6 Lemma	1508	74	43	58.0	138.7
BI6 MK1+BZK Lemma	3145	142	73	51.4	235.5
BI3 Lemma	1508	46	37	80.4	119.4

4.6 Experiment 4: Simulating Syntactic Tagging

In order to see how much the precision could possibly be improved by determining syntactic relations as done by Smadja (1991a,b) for English, we conducted another test with *bringen*, where we manually excluded those uninteresting extracted combinations in which the nouns were in fact used in subject position of the verb. The results for

[2] The latest runs with the combined corpus showed that for the infinitives precision even rises slightly on average (82.1%) while recall is almost doubled (134,9%); compared to BI3 Inf in table 1.

the two window-sizes, infinitive and lexeme, are shown in table 3. Precision would rise up to 100% with still a good recall of 87.1% if one could consider syntactic relations for the extraction of V-N collocations. The best recall of 43 collocations within 5 words to the left of the lexeme would then still correspond to 78.2% precision as compared to 58% if subjects cannot be detected. These results point in the same direction as Smadja's who reports an improvement from 40 to 80% precision if syntactic relations are considered, with a 94% recall of all collocations that had been found regardless of syntactic relations. However, this cannot as easily be achieved in a large scale for German due to the complicated parsing techniques necessary for the varying word order.

Table 3: Results for *bringen* if subject nouns are excluded manually

Bigrams	V-N	Coll.	Precision %	Recall %
BI3/t Inf (no subj)	27	27	100	87.1
BI6/t Inf (no subj)	39	31	79.5	100 (def.)
BI3 Lemma (no subj)	40	37	92.5	119.4
BI6 Lemma (no subj)	55	43	78.2	138.7

5 Conclusions and Outlook

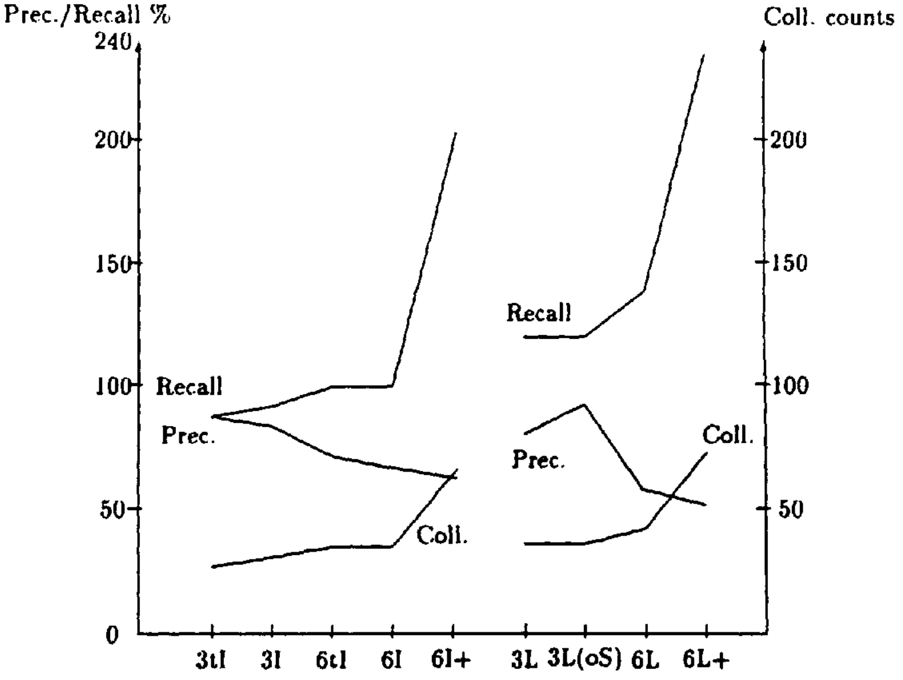

Figure 1: Results for *bringen*

The graphics in figure 1 visualize the results of the experiments for the verb *bringen*; the left y-axis shows recall and precision in per cent, the one to the right the number of counted V-N collocations. The left graph compares the results for the infinitive, the right one those for the lexeme. From left to right are shown: 3-word window with t-threshold (3tI), 3-word window without t (3I), 6-word window with t (6tI) and without (6I), 6-word window for the enlarged corpus (6I+). 3L stands for '3-word window, lexeme', 3L(oS) means the exclusion of subject nouns; 6L and 6L+ are analogous to the infinitive version.

The result for '6I+' implies that larger corpora will improve recall without a serious decline of precision compared to the same method used with the smaller corpus (6I; see also footnote 2). Whether the recall number should at the cost of a bad precision be pushed even higher by calculating MI for lexemes (6L vs. 6L+) can be decided in view of the application the data are extracted for. Once the number of V-N collocations is generally big enough, higher significance and MI thresholds can be used in order to improve precision again. MI sorts the extracted combinations in such a way that the collocations are the better the higher the MI-score is (with a few exceptions which often reflect highly significant, but linguistically uninteresting word combinations from one of the texts; this could hopefully be avoided with a more balanced corpus).

In general, a trade-off has to be found between the number of extracted collocations (recall) and the number of uninteresting items in between (precision), depending on the application. The described approach seems to be a good method for corpora with texts from restricted domains, where a special terminology is used which will thus show up strongly against 'normal' combinations.

Very high precision rates, which are an indispensible requirement for lexical acquisition, can only realistically be envisaged for German with parsed corpora (3L(oS) has the best recall–precision ratio in figure 1); otherwise the main advantage lies in a better lexicographical support, which should not be underestimated both for manually built NLP lexica and for printed dictionaries. Lemmatizing does not seem to be always useful, as a comparison of 6I+ and 3L shows. Possibly the data are blurred because as mentioned on p. 6 the various inflection forms are distributed differently in verb-final and verb-second clauses, at least in the investigated corpus. Restricted lemmatizing with infinitive (1st/3rd pers. pl.) and past participle for a search to the left, and with 3rd pers. sg. pres./past and 1st/3rd pers. pl. past for a search to the right (which is problematic, though) promises to give more precise results, as long as search strategies cannot take into account the syntactic structure of a sentence.

Work is currently in progress to calculate trigrams to check for prepositions in SVCs or for specific (or no) determiners for phrasemes. This will give indications to distinguish SVCs and lexicalized, phraseological SVCs from other collocations. In addition, we plan to consider the variation in span position of the noun within the searched window in order to distinguish fixed phrasemes from flexible ones.

References

Agricola, E., H. Görner, R. Küfner (eds.) (1962/1970). *Wörter und Wendungen. Wörterbuch zum deutschen Sprachgebrauch.* Leipzig: Verlag Enzyklopädie; München.

Barkema, H. (1989). Morphosyntactic flexibility: the other side of the idiomaticity coin. In: Everaert, M., E. van der Linden (eds.). *Proc. of the 1st Tilburg Workshop on Idioms.* 23-40.

Breidt, E. (1993). *Extraktion von Verb-Nomen-Verbindungen aus dem Mannheimer Korpus I.* SfS-Report 03-93. University of Tübingen.

Brundage, J., M. Kresse, U. Schwall, A. Storrer (1992). *Multiword lexemes: a monolingual and contrastive typology for NLP and MT.* IWBS-Report 232, September 1992. IBM Germany, Scientific Centre Heidelberg.

Calzolari, N., R. Bindi (1990). Acquisition of lexical information from a large textual italian corpus. *13th COLING 1990*, Helsinki. 54-59.

Choueka, Y. (1988). Looking for needles in a haystack, or: locating interesting collocational expressions in large textual databases. *Proceedings of the RIAO.* 609-623.

Church, K. W., P. Hanks (1989). Word Association Norms, Mutual Information and Lexicography. *27th ACL*, Vancouver. 76-83.

Church, K. W., W. A. Gale, P. Hanks, D. M. Hindle (1991). Using statistics in lexical analysis. In: Zernik, U. (ed.). *Lexical acquisition: exploring on-line resources to build a lexicon.* Hillsdale, NJ.

Danlos, L. (1992). Support verb constructions. Linguistic properties, representation, translation. *Journal of French Linguistic Study, Vol. 2, No. 1.* CUP.

Drosdowski, G. et al. (eds.) (1970). sl Duden Stilwörterbuch der deutschen Sprache: Die Verwendung der Wörter im Satz. 6th completely revised and extended edition. Mannheim.

Fleischer, W. (1982). *Phraseologie der deutschen Gegenwartssprache.* Leipzig.

Gale, W., K. W. Church (1990). Whats wrong with adding one? *IEEE Transactions on Acoustics, Speech and Signal Processing.*

Hatch, E., H. Farhady (1982). *Research design and statistics for applied linguistics.* Rowley.

Hausmann, F. J. (1989). Le dictionnaire de collocations. In: Hausmann, F. J. et al. (eds.). *Dictionaries: an international handbook for lexicography. Part I.* HSK 5.1. 1010-1019.

Martin, W., B. Al, P. van Sterkenburg (1983). On the processing of a text corpus. In: Hartmann, R. R. K. (ed.). *Lexicography: principles and practice.* London. 77-87.

v.Polenz, P. (1989). Funktionsverbgefüge im allgemeinen einsprachigen Wörterbuch. In: Hausmann, F. J. et al. (eds.). *Dictionaries: an international handbook for lexicography. Part I.* HSK 5.1. 882-887.

Sinclair, J. M. (1966). Beginning the study of lexis. In: Bazell, C. E. et al. (eds.) (1966). *In memory of J. R. Firth.* London. 410-430.

Smadja, F. A., K. R. McKeown (1990). Automatically extracting and representing collocations for language generation. *28th ACL 1990.* 252-259.

Smadja, F. A. (1991a). Macrocoding the lexicon with co-occurrence knowledge. In: Zernik, U. (ed.). *Lexical acquisition: exploring on-line resources to build a lexicon.* Hillsdale, NJ.

Smadja, F. A. (1991b). From n-grams to collocations: an evaluation of Xtract. *29th ACL, Berkeley, CA.* 279-284.

Uszkoreit, H. (1987). *Word order and constituent structure.* CSLI Lecture Notes 8.