Collocation Extraction Based on Modifiability Statistics

Joachim Wermter **Udo Hahn**

Computerlinguistik, Friedrich-Schiller-Universität Jena
Fürstengraben 30, D-07743 Jena, Germany
wermter@coling.uni-freiburg.de

Abstract

We introduce a new, linguistically grounded measure of collocativity based on the property of limited modifiability and test it on German PP-verb combinations. We show that our measure not only significantly outperforms the standard lexical association measures typically employed for collocation extraction, but also yields a valuable by-product for the creation of collocation databases, viz. possible structural and lexical attributes. Our approach is language-, structure-, and domain-independent because it only requires some shallow syntactic analysis (e.g., a POS-tagger and a phrase chunker).

1 Introduction

Natural language is an open and very flexible communication system. Syntax, of course, imposes constraints, e.g., on word order or the occurrence of particular phrasal types such as PPs or NPs, and lexical semantics imposes, e.g., selectional constraints on conceptually permitted sorts or types within the context of specific verbs or nouns. Nevertheless, natural language speakers usually enjoy an enormous degree of freedom to express the content they want to convey in a great variety of linguistic forms.

There is, however, a significant subset of expressions which do not share this rather free combinability, so-called *collocations*. From a linguistic perspective, they can be characterized by at least three recurrent and prominent properties (Manning and Schütze, 1999):

- Non-(or limited) *compositionality*. The meaning of a collocation is not a straightforward composition of the meanings of its parts. For example, the meaning of *'red tape'* is completely different from the meaning of its components.
- Non-(or limited) *substitutability*. The parts of a collocation cannot be substituted by semantically similar words. Thus, *'gut'* in *'to spill gut'* cannot be substituted by *'intestine'* (see also Lin (1999)).
- Non-(or limited) *modifiability*. Many collocations cannot be supplemented by additional lexical material. For example, the noun in *'to kick the bucket'* cannot be modified as *'to kick the* {*holey/plastic/water*} *bucket'*.

Considering these observations, from a natural language processing perspective, collocations should not enter, e.g., the standard syntax-semantics pipeline so as to prevent compositional semantic readings of expressions for which this is absolutely not desired. Hence, collocations need to be identified as such and subsequently be blocked, e.g., from compositional semantic interpretation.

In computational linguistics, a wide variety of lexical association measures have been employed for the task of (semi-)automatic collocation identification and extraction. Almost all of these measures can be grouped into one of the following three categories:

- frequency-based measures (e.g., based on absolute and relative co-occurrence frequencies)
- information-theoretic measures (e.g., mutual information, entropy)
- statistical measures (e.g., chi-square, t-test, log-likelihood, Dice's coefficient)

The corresponding metrics have been extensively discussed in the literature both in terms of their mathematical properties (Dunning, 1993; Manning and Schütze, 1999) and their suitability for the task of collocation extraction (see Evert and Krenn (2001) and Krenn and Evert (2001) for recent evaluations). Typically, they are applied to a set of candidate lexeme pairs which were obtained from preprocessors varying in linguistic sophistication.[1] The selected measure then assigns an association score

[1] On the low end, this may just be a preset numeric window span. In order to reduce the noise among the candidates, however, more elaborate linguistic processing, such as POS tagging, chunking, or even parsing, is increasingly being applied.

to each candidate pair, which is computed from its joint and marginal frequencies, thus expressing the strength of the hypothesis stating whether it constitutes a collocation or not.

While these association measures have their statistical merits in collocation identification, it is interesting to note that they have relatively little to do with the *linguistic* properties (such as those mentioned at the beginning) which are typically associated with the notion of collocativity. Therefore, it may be interesting to investigate whether there is a way to implement a measure which directly incorporates linguistic criteria in the collocation identification task, and even more important, whether such a linguistically rooted approach would fare better in comparison to some of the standard lexical association measures.

In the following study, we will introduce such a linguistic measure for identifying PP-verb collocations in German, which is based on the property of non- or limited modifiability. To the best of our knowledge, this is the first work to use this kind of linguistic measure to acquire collocations automatically. By contrasting our method to previous studies which use the standard lexical association measures, we intend to emphasize a more linguistically inspired use of statistics in collocation mining. Section 2 motivates our definition of the notion of collocation and Section 3 describes our methods, in particular the linguistically grounded collocation extraction algorithm, and the experimental setup derived from it. In Section 4 we present and discuss the results of our experiments.

2 Kinds of Collocations

There have been various approaches to define the notion of 'collocation'. This is by no means an easy task, especially when it comes to defining the demarcation line between collocations and free word combinations (modulo general syntactic and semantic semantic constraints). We favor an approach which draws this line on the semantic layer, *viz.* the compositionality between the components of a linguistic expression.

For this purpose, we distinguish between three classes of collocations based on varying degrees of semantic compositionality of the basic lexical entities involved:

1. *Idiomatic Phrases*. In this case, *none* of the lexical components involved contribute to the overall meaning in a semantically transparent way. The meaning of the expression is metaphorical or figurative. For example, the literal meaning of the German PP-verb combination '[jemanden] auf die Schippe nehmen' is 'to take [someone] onto the shovel'. Its figurative meaning is 'to lampoon somebody'.

2. *Support Verb Constructions/Narrow Collocations*. This second class contains expressions in which *at least one* component contributes to the overall meaning in a semantically transparent way and thus constitutes its semantic core. For example, in the support verb construction 'zur Verfügung stellen' (literal: 'to put to availabilty'; actual: 'to make available'), the noun 'Verfügung' is the semantic core of the expression, whereas the verb only has a support function with some impact on argument structure, causativity or aktionsart. There are, however, also narrow collocations in which the basic lexical meaning of the verb is the semantic core: For example, in 'aus eigener Tasche bezahlen' ('to pay out of one's own pocket') the verb 'bezahlen' is the semantic core. What unifies these two types is the fact that they function as predicates.

3. *Fixed Phrases*. Here, *all* basic lexical meanings of the components involved contribute to the overall meaning in a semantically much more transparent way. Still, they are *not as completely compositional* as to classify them as free word combinations. For example, all the basic lexical meanings of the different lexical components in 'im Koma liegen' (literal: 'to lie in coma'; actual: 'to be comatose') contribute to the overall meaning of the expression. Still, this is different from a completely compositional free word combination, such as 'auf der Strasse gehen' ('to walk on the street').

Our goal is to consider all three types of collocations as a whole, i.e., we will not distinguish between the three different kinds of collocations. However, in order to focus our experiments, we will concentrate on a particular surface pattern in which they occur, *viz.* PP-verb collocations.

3 Methods and Experiments

3.1 Construction and Statistics of the Testset

We used a 114-million-word German-language newspaper corpus extracted from the Web to acquire candidate PP-verb collocations. The corpus was first processed by means of the TnT part-of-speech tagger (Brants, 2000). Then we ran a sentence/clause recognizer and an NP/PP chunker,

both developed at the Text Knowledge Engineering Lab at Freiburg University, on the POS-tagged corpus. From the XML-marked-up tree output, PP-verb complexes were automatically selected in the following way: Taking a particular PP node as a fixed point, either the preceding or the following sibling V node was taken.[2] From such a PP-verb combination, we extracted and counted both its various *heads*, in terms of *P*reposition-*N*oun-*V*erb (PNV) triples, and all its associated *supplements*, i.e., here in this case any additional lexical material which also occurs in the nominal group of the PP, such as articles, adjectives, adverbs, cardinals, etc.[3] The extraction of the associated supplements is essential to the linguistic measure described in subsection 3.3 below.

In order to reduce the amount of candidates for evaluation and to eliminate low-frequency data, we only considered PNV-triples with frequency $f \geq 10$. This was also motivated by the well-known fact that collocations tend to have a higher co-occurrence frequency than free word combinations.[4] Table 1 contains the data for the corresponding frequency distributions.

frequency	PP-verb combinations	
	candidate tokens	candidate types
all	1,663,296	1,159,133
$f \geq 10$	279,350	8,644

Table 1: Frequency distribution for PP-Verb tokens and types for our 114-million-word newspaper corpus

3.2 Classification of the Testset

Three human judges manually classified the PP-verb candidate types with $f \geq 10$ in regard to whether they were a collocation or not. For this purpose, they used a manual, in which the guidelines included the linguistic properties as described in Section 1 and the three collocation classes identified in Section 2.

Among the 8,644 PP-verb candidate types, 1,180 (13.7%) were identified as true collocations. The inter-annotator agreement was 94.8% (with a standard deviation of 2.1).

[2]The verbs in this study are restricted to main verbs and are reduced to their base form after extraction.

[3]It should be noted that both *heads* and *associated supplements* may of course vary depending on the particular linguistic structure targeted for collocation extraction.

[4]Cf. also Evert and Krenn (2001) for empirical evidence justifying the exclusion of low-frequency data.

3.3 The Linguistic Measure

The linguistic property around which we built our measure for collocativity is the non- or limited modifiabilty of collocations with additional lexical material (i.e., supplements). The underlying assumption is that a PNV triple is less modifiable (and thus more likely to be a collocation) if it has a lexical supplement which, compared to all others, is particularly characteristic. We express this assumption in the following way: Let n be the number of distinct supplements of a particular PNV triple (PNV_{triple}). The probability $\mathcal{P}$ of a particular supplement $Supp_k$, $k = [1, n]$, is described by its frequency scaled by the sum of all supplement frequencies:

$$\mathcal{P}(PNV_{triple,Supp_k}) = \frac{f(PNV_{triple,Supp_k})}{\sum_{i=1}^{n} f(PNV_{triple,Supp_i})} \quad (1)$$

with $\sum_{i=1}^{n} f(PNV_{triple,Supp_i}) = f(PNV_{triple})$.[5] Then the *modifiability* $\mathcal{MOD}$ of a PNV triple can be described by its most probable supplement:

$$\mathcal{MOD}(PNV_{triple}) := arg\,max\ \mathcal{P}(PNV_{triple,Supp_k}),\ k = [1, n] \quad (2)$$

To define a measure of *collocativity* $\mathcal{COLL}$ for a candidate set, some factor regarding frequency has to be taken into account. Thus, besides $\mathcal{MOD}$, we take the relative co-occurrence frequency for a specific PNV triple $\mathcal{P}(PNV_{triple})$ (t being the number of candidate types (here, 8,644))

$$\mathcal{P}(PNV_{triple}) := \frac{f(PNV_{triple})}{\sum_{j=1}^{t} f(PNV_{triple_j})} \quad (3)$$

and incorporate it as a second factor to $\mathcal{COLL}$:

$$\mathcal{COLL}(PNV_{triple}) := \mathcal{MOD}(PNV_{triple}) \times \mathcal{P}(PNV_{triple}) \quad (4)$$

3.4 Methods of Evaluation

Standard procedures for evaluating the goodness of collocativity measures usually involve identifying the true positives among the n-highest ranked candidates returned by a particular measure. Because this is rather labor-intensive, n is usually small, ranging from 50 to several hundred. Evert and Krenn

[5]Note that the zero supplement of the PNV triple, i.e., the one for which no lexical supplements co-occur is also included in this set.

(2001), however, point out the inadequacy of such methods claiming they usually lead to very superficial judgements about the measures to be examined. In contrast, they suggest examining various n-highest ranked samples, which allows plotting standard precision and recall graphs for the whole candidate set.

We evaluate the $\mathcal{COLL}$ measure against two widely used standard statistical tests (t-test and log-likelihood) and against co-occurrence frequency. The comparison to the t-test is especially interesting because it was found to achieve the best overall precision scores in other studies (see Evert and Krenn (2001)). Our baseline is defined by the proportion of true positives (13.7%; see subsection 3.2), which can be described as the likelihood of finding one by blindly picking from the candidate set.

4 Experimental Results and Discussion

4.1 Precision and Recall for Collocation Extraction

In the first experiment, we incrementally examined parts of the n-highest ranked candidate lists returned by the each of the four measures we considered. The precision values for various n were computed such that for each percent point of the list, the proportion of true positives was compared to the overall number of candidate items returned. This yields the precision curves in Figure 1 and its associated values at selected list portions in the upper table from Table 2.

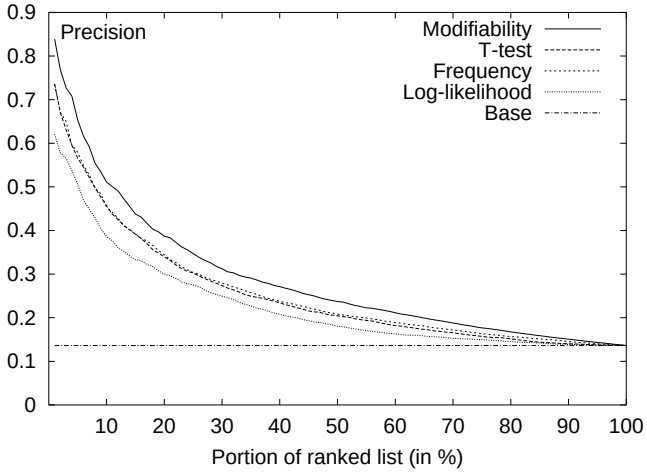

Figure 1: Precision for Collocation Extraction

First, we observe that all measures outperform the baseline by far and, thus, all are potentially useful measures of collocativity. Of the statistical measures, log-likelihood (the most complex one) performs the worst, whereas t-test and frequency, almost indistinguishable, share the middle position, with frequency measurements having a very slight edge at six rank points. This is in contrast to the findings reported by Krenn and Evert (2001), which gave the t-test an edge.[6]

As can be clearly seen, however, our linguistic modifiability measure substantially outperforms all other measures at all points in the ranked list. Considering 1% ($n \approx 86$), its precision value is ten percentage points higher than for t-test and frequency, and even 22 points higher compared to log-likelihood. Until 50% ($n \approx 4322$) of the ranked list is considered, modifiability maintains a three to five percentage point advantage in precision over t-test and frequency. In the second half of the list, all curves and associated values start converging towards the baseline.

We also tested the significance of differences for our precision results, both between modifiability and frequency and between modifiability and t-test. Because in both cases the ranked lists were taken from the same set of candidates, *viz.* the 8,644 PP-verb candidate types, and hence constitute dependent samples, we applied the McNemar test (Sachs, 1984) for statistical testing. We selected 100 measure points in the ranked list, one after each increment of one percent, and then used the two-tailed test for a confidence interval of 99%. Table 3, which lists the number of significant differences for 10, 50 and 100 measure points, shows that almost all of them are significantly different.

# of significance measure points	# of signicant differences comparing **modifiability** with	
	frequency	t-test
10	9	9
50	49	49
100	96	97

Table 3: Significance testing of differences using the two-tailed McNemar test at 99% confidence interval

The recall curves in Figure 2 and their corresponding values in the lower table from Table 2 measure which *proportion of all true positives* is identified by a particular measure at a certain part of the ranked list. In this sense, recall is an even better indicator of a particular measure's performance. Again, the linguistically motivated collocation extraction algorithm outscores all others, even more pronounced than for precision. When examining 20% ($n \approx 1729$), 30% ($n \approx 2593$) and 40%

[6]The reason why frequency performs even slightly better than t-test may very well have to do with the size of our training corpus (114 million words). But this just underlines the fact that large corpora are essential for collocation discovery.

Portion of ranked list considered	**Precision** scores of measure evaluated				
	Modifiablity	T-test	Frequency	Log-likelihood	Baseline
1%	0.84	0.74	0.74	0.62	0.14
10%	0.51	0.46	0.45	0.39	0.14
20%	0.39	0.34	0.34	0.30	0.14
30%	0.31	0.27	0.28	0.25	0.14
40%	0.27	0.23	0.24	0.21	0.14
50%	0.24	0.20	0.21	0.18	0.14
60%	0.21	0.18	0.19	0.16	0.14
70%	0.19	0.17	0.17	0.15	0.14
80%	0.17	0.15	0.16	0.15	0.14
90%	0.15	0.14	0.15	0.14	0.14
($n = 8,644$) 100%	0.14	0.14	0.14	0.14	0.14

Portion of ranked list considered	**Recall** scores of measure evaluated			
	Modifiablity	T-test	Frequency	Log-likelihood
1%	0.06	0.05	0.05	0.05
10%	0.37	0.33	0.33	0.28
20%	0.58	0.50	0.50	0.44
30%	0.69	0.60	0.61	0.55
40%	0.80	0.69	0.70	0.61
50%	0.87	0.75	0.76	0.66
60%	0.93	0.80	0.83	0.72
70%	0.96	0.85	0.88	0.78
80%	0.98	0.89	0.92	0.85
90%	0.99	0.93	0.96	0.92
100%	1.00	1.00	1.00	1.00

Table 2: Precision and Recall Scores for Collocation Extraction at Major Portions of the Ranked List

($n \approx 3458$) of the ranked list, modifiability, respectively, identifies almost 60%, 70% and 80% of all true positives, holding a ten percentage point lead over t-test and frequency at each of these points. When 50% ($n \approx 4322$) are considered, this difference reaches eleven and twelve points (compared to frequency and t-test, respectively).

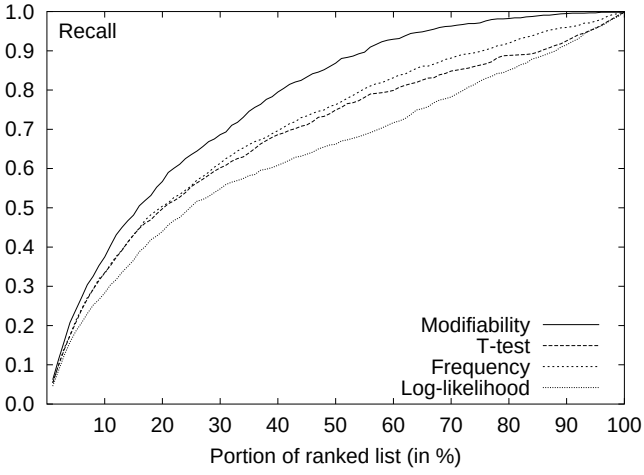

Figure 2: Recall for Collocation Extraction

Even more strikingly, for the identification of 90% of all true positives, modifiability only needs to look at 55% ($n \approx 4754$) of the ranked list. Frequency, on the other hand, needs to examine 75% ($n \approx 6483$) and t-test even 85% ($n \approx 7347$) of the ranked list to reach this high level of recall.

4.2 Modifiability Revisited

The previous subsection showed that a measure for collocation discovery which takes into account the linguistic property of limited modifiability fares significantly better than linguistically not so founded, purely statistical measures. Although the modifiability property constitutes common wisdom about collocations, it has not yet been empirically evaluated. Thus, we ran an experiment which took both the PNV triples classified as collocations and the PNV triples classified as non-collocations and counted the numbers of distinct supplements (referred to as n in Subsection 3.3). From this data, we set up a distribution of collocational and non-collocational PNV triples in which the distributional ranking criterion was the number of distinct supplements (cf. Figure 3).

PNV Triple	NP Supplement	Frequency
'in Griff bekommen'	den/ART Griff/NN	459
'to get under control'	Griff/NN	2
	den/ART gewerkschaftlichen/ADJA Griff/NN	1
	den/ART dramatischen/ADJA Griff/NN	1
	den/ART erzählerischen/ADJA Griff/NN	1
'unter Druck geraten'	Druck/NN	560
'to get under pressure'	politischen/ADJA Druck/NN	6
	erheblichen/ADJA politischen/ADJA Druck/NN	5
	teilweise/ADV lebensgefährlichen/ADJA Druck/NN	1
	wachsenden/ADJA Druck/NN	1
	noch/ADV stärkeren/ADJA Druck/NN	1
	schweren/ADJA Druck/NN	1

Table 4: Collocational PNV Triples with Associated Noun Phrase Supplements

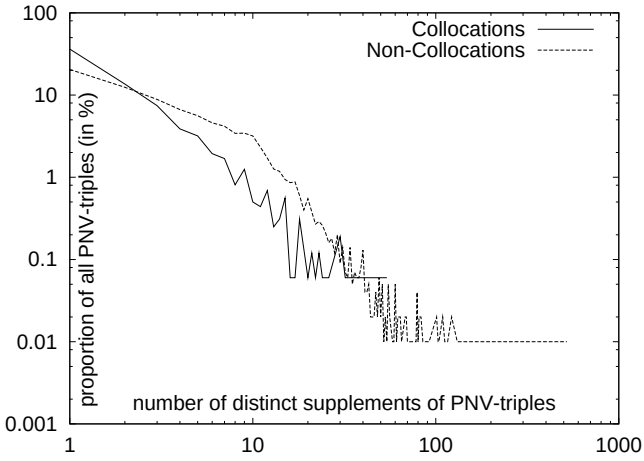

Figure 3: Distribution of Supplements for (Non-) Collocations in PNV Triples. The x- and y-axes are log-scaled.

As Figure 3 reveals, not only is the proportion of collocational PNV triples with only one distinct supplement higher (36%) than the proportion for non-collocational ones (20%), but with each additional supplement, the collocational proportion curve declines more steeply than its non-collocational counterpart. Moreover, the collocational proportion curve already ends with 54 distinct supplements, whereas the non-collocational proportion curve leads up 520 distinct supplements. Thus, we are able to add some empirical grounding to the widespread textbook assumption about the limited modifiablity of collocations.

Another observation (which is also inherent to our linguistic measure) based on this experiment is that some collocations do possess at least limited modifiability. Collocation acquisition is, of course, not a goal by itself, but rather aims at creating collocation lexicons for both language processing and generation (Smadja and McKeown, 1990). From this perspective, our linguistic modifiabilty measure actually yields quite a valuable by-product for the development of lexicons or collocational knowledge bases: A list of possible structural and lexical modifications associated with a particular collocational entry candidate. In our case, these modifications refer to the nominal group of the PP. We illustrate this point in Table 4 with two collocational PNV triples and some of their associated NP supplements plus their frequencies.

As can be seen, both structural and lexical attributes of collocations can thus be obtained. The structural information comes in the form of part-of-speech (POS) tags. From this, possible prenominal POS types and their combinations can be used to describe a collocation's structural make-up. From a lexical viewpoint, the collocation can be described by the lexical semantic word classes used for modification.[7] As can be seen in Table 4 under the PNV triple for 'to get under pressure', the noun *'Druck'* ('pressure') is often modified by a certain semantic class of adjectives, such as *'stark'* ('strong'), *'schwer'* ('heavy'), *'erheblich'* ('considerable', 'grave').

5 Related Work

Although there have been many studies on collocation extraction and mining using only statistical approaches (Church and Hanks, 1990; Ikehara et al., 1996), there has been much less work on collocation acquisition which takes into account the linguistic properties typically associated with collocations.

Smadja (1993), which is the classic work on collocation extraction, uses a two-stage filtering model in which, in the first step, n-gram statistics determine possible collocations and, in the second step, these candidates are submitted to a syntactic valida-

[7] Of course, lexical material is always at least partially dependent on the domain in question. In our case, this is the news domain with all its associated subdomains (politics, economics, finance, culture, etc.).

tion procedure (e.g., determining verb-object collocations) in order to filter out invalid collocations. In a single-judge evaluation of 4,000 collocation candidates, the incorporation of linguistic criteria (via tagging and predicate-argument parsing) boosts precision up to a level of 80% and recall to 94%. These results are, of course, not comparable to ours. First of all, precision and recall are measured at a *fixed* point for a *fixed unranked* candidate list. In order to obtain more reliable evaluation results, we plot these values *continuously* on a *ranked* candidate list. Secondly, our kind of syntactic preprocessing (which is standard nowadays) allows collocation extraction algorithms to better control the *structural types* of collocations.

Lin (1998) acquires a lexical dependency database by assembling dependency relationships from a parsed corpus. An entry in this database is classified as collocation if its log-likelihood value is greater than some threshold. Using an automatically constructed similarity thesaurus, Lin (1999) then separates compositional from non-compositional collocations by taking into account the second linguistic property described in Section 1, *viz.* their non- or limited substitutability. In particular, he checks the existence and mutual information values of phrases obtained by substituting the words with similar ones, which results in the classification of the phrase as being compositional or non-compositional. Although this study offers some promising results, its applicability rather falls into the category of fine-classifying an already acquired set of collocations, e.g., according to the criteria described in Section 2, and thus is not really comparable to our work. Moreover, the linguistic property in his focus is of course a semantic one, whereas ours is purely syntactic in nature.

6 Conclusion

We introduced a new, linguistically motivated measure of collocativity based on the property of limited modifiability and tested it on a large corpus with emphasis on German PP-verb combinations. We showed that our measure not only significantly outperforms the standard lexical association measures typically used for collocation extraction, but also yields a valuable by-product for the creation of collocation databases, viz. possible structural and lexical attributes of a collocation.

Our measure defines the modifiability property in a linguistically simple way, by e.g. ignoring the internal make-up of lexical supplements associated with a collocation candidate. Hence, it may be worthwhile to investigate whether a more sophisticated approach, by e.g. taking into account internal POS types and their distribution etc., would improve our results even more. We may also consider other linguistic criteria (e.g., limited substitutability) to further refine our measure and to categorize already identified collocations.

At the methodological level, our approach, although tested on German newspaper language data, is language-, structure-, and domain-independent. All it requires is some sort of shallow syntactic analysis, e.g., POS tagging and phrase chunking. Thus, in the future we plan to include other syntactic types of collocations, such as verb-object or verb-object-PP combinations, and also apply our methodology to other languages and domains, such as the biomedical field.

Acknowledgements. We would like to thank our students, Sabine Demsar, Kristina Meller, and Konrad Feldmeier, for their excellent work as human collocation classifiers. This work was partly supported by DFG grant KL 640/5-1.

References

T. Brants. 2000. TnT: A statistical part-of-speech tagger. In *Proceedings of the ANLP 2000 Conference*, pages 224–231.

K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

S. Evert and B. Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *ACL'01 – Proceedings of the 39th Meeting of the ACL*, pages 188–195.

S. Ikehara, S. Shirai, and H. Uchino. 1996. A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. In *Proceedings of the COLING'96 Conference*, pages 574–579.

B. Krenn and S. Evert. 2001. Can we do better than frequency? A case study on extracting pp-verb collocations. In *Proceedings of the ACL Workshop on Collocations*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the COLING/ACL'98 Conference*, pages 768–774.

D. Lin. 1999. Automatic identification of non-compositional phrases. In *ACL'99 – Proceedings of the 37th Meeting of the ACL*, pages 317–324.

C. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

L. Sachs. 1984. *Applied Statistics*. Springer.

F. A. Smadja and K. R. McKeown. 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Meeting of the ACL*, pages 252–259.

F. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.