Can we do better than frequency? A case study on extracting PP-verb collocations

Brigitte Krenn
Austrian Research Institute
for Artificial Intelligence (ÖFAI)
Schottengasse 3
A-1010 Vienna, Austria
brigitte@ai.univie.ac.at

Stefan Evert
IMS, University of Stuttgart
Azenbergstr. 12
D-70174 Stuttgart, Germany
evert@ims.uni-stuttgart.de

Abstract

We argue that lexical association measures (AMs) should be evaluated against a reference set of collocations manually extracted from the full candidate data, and that the notion of collocation needs to be precisely defined so that human collocativity judgments and experimental results are reproducible. We show that identification results achieved by particular AMs do not crucially depend on text type, but that some AMs are much better suited for identifying some classes of collocations than others.

1 Introduction

In the past, a series of measures have been introduced for identifying lexical associations between pairs of words. We refer to these as *association measures* (AMs). Methods range from pure frequency counts to information theoretic measures and statistical significance tests.

AMs are often applied to the task of collocation extraction using the following procedure: A set of candidate lexeme pairs is extracted from a corpus (e.g. adjective-noun pairs, noun-verb pairs, adjacent words, or pairs of words co-occurring within a given window). For each pair, its number of occurrences (the *joint frequency*) and the total frequency of each component (the *marginal frequencies*) are determined. The AM assigns to each pair an association score computed from its joint and marginal frequencies. Depending on the measure, either a high or a low score indicates a strong connection between the two components. The list of candidate pairs is re-ordered according to the AM scores (from best to worst). This re-ordered list is referred to as *significance list* (SL). The top part of the SL (i.e., a small set of n best-scoring candidates) constitutes the basis for manual extraction of true collocations, and is also used for evaluation of the AMs, i.e., the precision of an AM is calculated using the manually identified true positives (TPs).

Many comparative evaluations of the goodness of AMs are based on a vague concept of collocation for which no rigorous definition is given, e.g. (Dunning, 1993), (Lezius, 1999), (Evert et al., 2000); see also the examples in chapter 5 of (Manning and Schütze, 1999). In such experiments, TPs are identified by scanning the SLs for pairs that seem "typical" according to the intuition of the annotator. However, this approach suffers from two serious drawbacks: (i) There is considerable disagreement between annotators on whether a given pair of lexemes is typical or not (cf. Cabré and Estopà (1999)). Hence, the results of experiments may not be reproducible. (ii) The notion of "typicality" overlaps with a number of widely different concepts such as technical terms, proper names, idioms, etc. There is no a priori reason to assume that a given AM is equally well suited for the extraction of all classes of collocations.

In the present study, we have employed only support verb constructions (German: Funktionsverbgefüge, FVG) and figurative expressions (*figur*) for evaluating a variety of AMs. As

bases for collocation extraction, we have used two largely differing corpora, a newspaper and a newsgroup corpus. A brief overview of the AMs considered is given in section 2. The linguistic criteria applied for differentiating FVG and *figur* are presented in section 3. In section 4, we examine the precision obtained by different AMs in identifying PP-verb collocations. We show that applying a particular AM to both corpora leads to similar results (section 5), but that some AMs are much better suited for identifying FVG than *figur*, and vice versa (section 7).

2 Evaluated measures

We have evaluated the following measures:

Mutual information MI (Church and Hanks, 1989), and a heuristically motivated variant of MI where the numerator is squared.

The Dice coefficient, a measure which computes the harmonic mean between the conditional probabilities of the components of a word combination, see (Smadja et al., 1996).

The χ^2 measure which is based on the well-understood χ^2-test, either applied as a test of homogeneity or as a test of independence; the two variants are mathematically and numerically equivalent.

Notationally different but mathematically and numerically equivalent versions of the log-likelihood measure originally presented in (Dunning, 1993).

The t-test measure or t-score which is based on another standard test statistics. The common-birthday measure (Läuter and Quasthoff, 1999) using a crude approximation of the Poisson distribution which yields slightly better results than the exact Poisson distribution, cf. (Evert et al., 2000).

The χ^2 and log-likelihood measures are derived from *two-tailed* statistical tests (where the null hypothesis stipulates that the components of a candidate pair occur independently of each other). Hence, pairs that occur *less* frequently than one would expect from their marginal frequencies will obtain high association scores as well. Since such pairs are assumed to be non-collocational, we have constructed *one-tailed* versions of the two AMs. In our experiments, there was no visible difference in accuracy between the one-tailed and the two-tailed measures.

We compare the AMs listed above against two reference measures: the *baseline precision* obtained by listing all candidates in random order in the SL, and the co-occurrence frequency (referred to as the *frequency* measure). The latter is particularly interesting because it shows how much can be achieved without employing complex statistical methods.

3 Evaluation by linguistic criteria

We describe the linguistic criteria by which collocations were distinguished from arbitrary PP-verb combinations (3.1), and the lists of candidate pairs that we used (3.2).

3.1 PP-verb collocations

Criteria for the distinction of PP-verb collocations from arbitrary combinations: There is a grammatical relation between verb and PP, and the pair can be interpreted as support verb construction and/or a metaphoric or idiomatic reading is available, e.g.: *zur Verfügung stellen* (at_the availability put, 'make available'), *am Herzen liegen* (at the heart lie, 'have at heart').[1]

Criteria for the distinction of FVG and *figur*: Figurative expressions in our interpretation cover idioms (i.e., uninterpretable PP-verb-constructions) as well as a broad range of word combinations that require figurative or metaphoric interpretation. FVG are identified according to the following characteristics: The constructions function as predicates. The noun in the FVG is abstract, it is typically deverbal or deadjectival, and thus has its own argument structure. The noun is also the semantic core of the FVG. It usually combines with more than one verb enabling variation in thematic structure and Aktionsart, see for instance *in Betrieb* {*gehen, nehmen*} ('{go, put} into operation').

These characteristics are valid for a broad range of PP-verb-combinations. However a number of PP-verb combinations exist that show characteristics of FVG but also have figurative aspects. In order to cope with the fuzzy borders between FVG and *figur*, the following decisions have been made

[1] For definitions of and literature on idioms, metaphors and FVG see for instance (Bußmann, 1990).

for classification of the reference data: Semantically opaque word combinations are classified as *figur*. In the case of semantically transparent word combinations, it is distinguished whether the nouns are abstract or concrete, and whether they contribute the main part of the semantics of the predicate. If the noun is concrete, the collocation is classified as *figur*. If the noun is deverbal, deadjectival or another kind of abstract noun, and contributes the major part of the meaning, the collocation is classified as FVG. Otherwise, the collocation is classified as *figur*.

3.2 Candidate lists

The following corpora have been used to identify potential PP-verb collocations: an 8 million word portion of the Frankfurter Rundschau Corpus[2], and a 10 million word sample of a newsgroup corpus developed in the FLAG project at DFKI Saarbrücken.[3]

After part-of-speech tagging and rudimentary syntactic analysis,[4] preposition-noun-verb-triples (PNV) have been automatically selected from each corpus such that P and N are constituents of the same PP, and PP and V co-occur in a sentence. The verbs in the current study are constrained to main verbs. The extracted data are partially inhomogeneous and not fully grammatically correct, because they include combinations with no grammatical relation between the PP and the verb. See Table 1 for a summary of the cadidate data: for example, 453,861 instances (PNV-full-forms) have been automatically extracted from the newspaper corpus. This amounts to 372,121 candidates (types).[5] The number reduces to 10,396 when only data with occurrence frequency $f \geq 3$ are considered, and so forth. The total number of true collocations in the sample where $f \geq 3$ is 1,280 or 12.31%; 5.47% in terms of *figur* and 6.84% in terms of FVG.

corpus	PP-verb pairs	
	instances	candidates
newspaper	453,861	372,121
newsgroup	912,287	631,140

corpus	PP-verb pairs with		
	$f \geq 3$	$f \geq 5$	$f \geq 10$
newspaper	10,396	2,853	743
newsgroup	—	4,795	1,029

corpus	baseline precision		
	total	figur	FVG
newspaper	12.31% (1,280)	5.47% (569)	6.84% (711)
newsgroup ($f \geq 5$)	12.45% (597)	5.38% (258)	7.07% (339)

Table 1: Frequency distributions in the newspaper and the newsgroup corpus

4 Precision curves

In a first experiment, we compare the AMs with respect to their usefulness for extracting collocations from the newspaper corpus, considering only candidates with $f \geq 3$. For each AM, we evaluate the corresponding SL by stepwise taking the first n candidate pairs and comparing them to the list of collocations which have been manually extracted from the complete SL. We compute the proportion of TPs (the *precision*) for each possible value of n. Plotting these precision values against the proportion of the entire SL constituted by the n-best candidates, we obtain a *precision curve* (see Figures 1 and 2).[6]

The three vertical lines correspond to the n-best candidate pairs for $n = 500$, $n = 1000$, and $n = 2000$. Looking at the middle line, we find for instance that among the 1000 best-scoring candidate pairs according to the t-test measure, approximately 30% are true collocations. A similar result can be found for the frequency measure.

Two reference curves are shown in the plots. A solid line, which is the precision curve for SLs

[2]The Frankfurter Rundschau Corpus is part of the European Corpus Initiative Multilingual Corpus I.

[3]http://flag.dfki.de/.
The complete FLAG corpus has been jointly developed at the University of Tübingen and at the DFKI, Saarbrücken.

[4]See (Skut and Brants, 1998) for a description of the tools employed.

[5]In contrast to common practice, we use full form data in the current experiments, as the baseline precision is much higher for full form data than for base forms. In experiments where we have reduced the verbs to their bases, we have found similar differences between the AMs as for full form data.

[6]Colour versions of all plots in this article will be available from `http://www.collocations.de/EK/`.

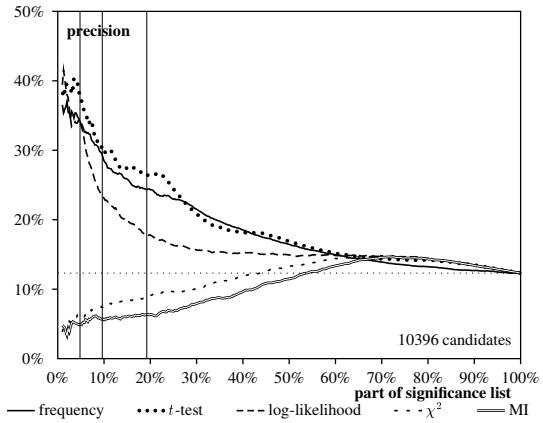

Figure 1: Newspaper data: precision curves 1

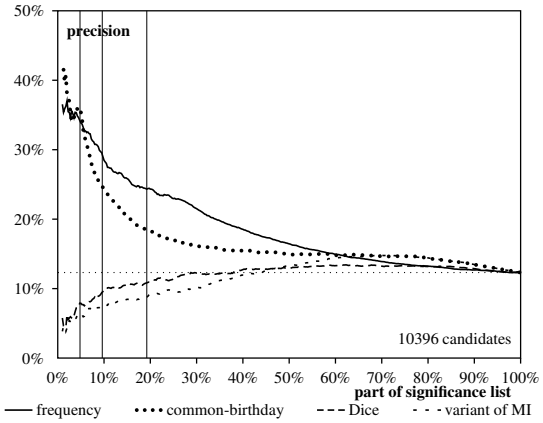

Figure 2: Newspaper data: precision curves 2

ordered by co-occurrence frequency, and a dotted horizontal line, the *baseline precision*, which shows the proportion of true collocations in the candidate set. This line is the expected precision curve for a randomly ordered SL. Hence, for an AM to be useful, its precision curve must be substantially *above* the baseline. In particular, we are interested in AMs that perform significantly better than the frequency reference curve.

Further observations are: Log-likelihood and common-birthday clearly lead to worse results for the PP-verb data than it would have been expected from the results on identification of adjective-noun collocations presented in (Lezius, 1999) and (Evert et al., 2000). MI, Dice and χ^2 perform even worse than random selection. The suitability of AMs for identifying particular classes of collocations will be further explored in section 7. T-score turns out to be the best measure for identifying PP-verb collocations from our sample of the Frankfurter Rundschau Corpus. The precision curve for t-score is slightly above the frequency curve. The significance of these differences will be discussed in section 6.

5 Comparison of the newspaper and the newsgroup data

We compare the newspaper data with $f \geq 3$ to newsgroup data with $f \geq 5$. This seems justified because the number of candidates extracted from the newsgroup corpus exceeds that of the newspaper data (cf. Table 1), and for the given frequency thresholds, we obtain similar baseline precision values. By contrast, the subset of candidates with $f \geq 5$ from the newspaper corpus has a much higher baseline precision of 22.57% (644 true collocations among 2,853 candidates).

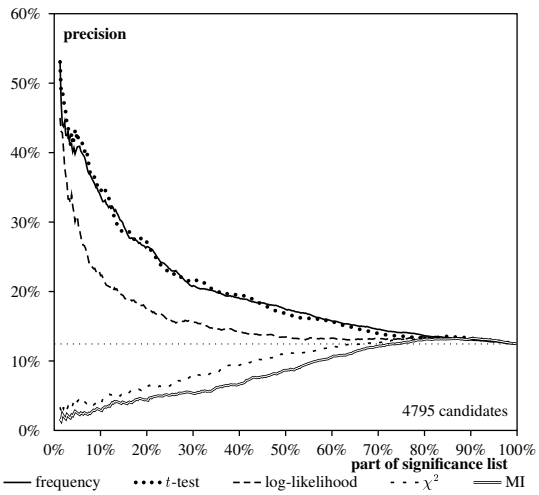

Figure 3: Newsgroup $f \geq 5$: prec. curves 1

Examining the newsgroup data (Figures 3 and 4), we find precision results comparable to those of the newspaper data (see previous section), i.e., t-test and frequency clearly outperform the other AMs, and MI, Dice and χ^2 are below random selection. This result provides strong evidence that identical classes of collocations are similarly distributed in different types of corpora.

Considering Figures 1 to 4, further general observations can be made. The precision curves of the AMs are instable for the first few percent of the SLs, and tend to converge in the second halves of the SLs. Consequently, experimental results

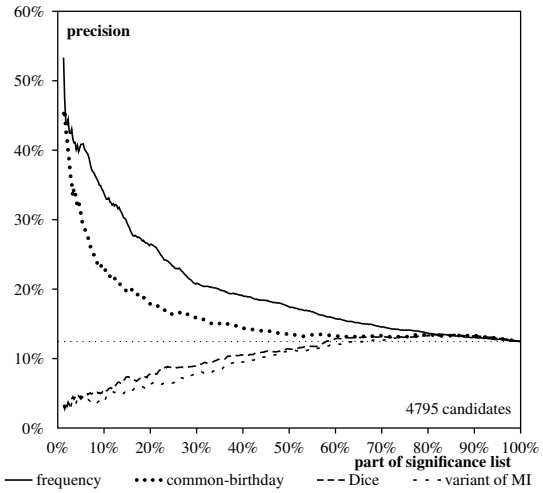

Figure 4: Newsgroup $f \geq 5$: prec. curves 2

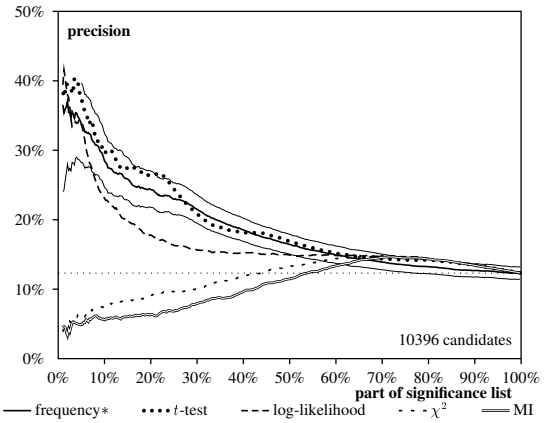

Figure 5: Newspaper: significance of differences

based either on a small number of best-scoring candidates or on more than the first 50% of the SLs are unreliable.

6 Significance of the differences

In order to test the significance of the differences observed in section 4, we apply the χ^2-test for two independent samples as described in (Krenn, 2000, p. 50). For given n and a pair of AMs, the null hypothesis that the two measures will on average identify the same proportion of true collocations is tested against the n best-scoring candidates from the measures' SLs. We use this test to compare the AMs described in section 2 to the frequency measure as a point of reference.

The thin lines above and below the precision curve of the frequency measure in Figure 5 delimit a 95% confidence region for the χ^2-test described above. The null hypothesis will only be rejected (at the 95% level) when the precision curve of an AM is outside this confidence region.

We have found in section 4 that the frequency measure outperforms all other AMs with the exception of t-score, which has better precision throughout the first 30% of the SLs. However, Figure 5 shows that the difference between frequency and t-score is significant only when approx. the first 22% of the SLs are considered (i.e., $n \approx 2,300$). Returning to the question in the title of this paper, we must conclude that none of the AMs is significantly better suited for the extraction of PP-verb collocations than mere co-occurrence frequency.

It has to be noted, though, that the χ^2-test may not be adequate for the comparison of precision curves, for the following two reasons: (i) It is a test designed for independent samples, whereas in our case the two samples (i.e. the n best-scoring pairs from each SL) are dependent, being taken from the same base set of candidates. (ii) The statistical model behind the χ^2-test assumes that the TPs are distributed uniformly within each sample. However, Figures 1 and 2 show that, for many AMs, there are substantially more TPs in the top part of the SL, and hence in the "top part" of the corresponding sample.

Because of (i), the χ^2-test tends to *underestimate* the significance of differences between the precision curves. It is not clear whether (and how) a test of significance can be designed that is appropriate for this task and takes the interdependence of the different SLs into account.[7]

7 Differences between *figur* and FVG

While in sections 4 and 5 no difference between FVG and *figur* has been made in the evaluation of AMs, we now focus on differences in the identification accuracy when FVG and *figur* are evaluated separately. The baseline precision is slightly higher for the FVG, but comparable to that of the *figur* (cf. Table 1). Based on the differences in identification accuracy found for PP-verb data

[7] A possiblity which requires further exploration is to use the McNemar test or the Cochran Q test (Siegel, 1956).

(see section 4) and for adjective-noun data (cf. Lezius (1999) and Evert et al. (2000)), we expect the AMs to differ in their ability to identify FVG and *figur*. This assumption is supported by the results shown in Figures 6 and 7.

While t-test and frequency lead to the best precision results for FVG and figurative expressions, the precision results achieved by χ^2 and, in particular, MI are far *below* the baseline for FVG, but almost identical to the baseline precision for *figur*. In other words, the values of these AMs are (almost) independent from whether or not a candidate pair is a *figur*. Thus we conclude that *figur* cannot at all be characterised in terms of χ^2 and MI values, but that candidates which obtain high MI scores are *less likely* to be FVG than other candidates, even though the MI measure considers them to be strongly correlated.

Figure 7 also reveals that log-likelihood, which is often used as a "general-purpose" AM for a wide variety of extraction tasks, is not very well suited to identify FVG. However its performance is considerably better for *figur*.

Further particularities in the results for *figur* compared to FVG are: (i) The precision curves converge earlier for *figur* than for FVG (in the case of figurative expressions, the AMs converge when approx. 40 to 50% of the data in the respective SLs have been examined; in the case of FVG, they converge when approx. 60% of the SLs have been considered). (ii) In the first $\approx$ 12% of the SLs, frequency and t-score reach much better precision for FVG than for *figur*.

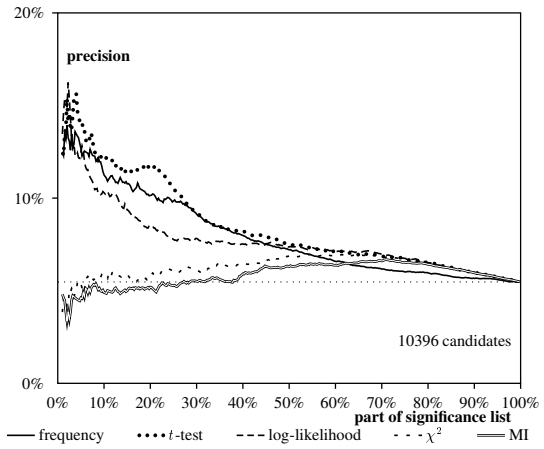

Figure 6: Precision curves for *figur*

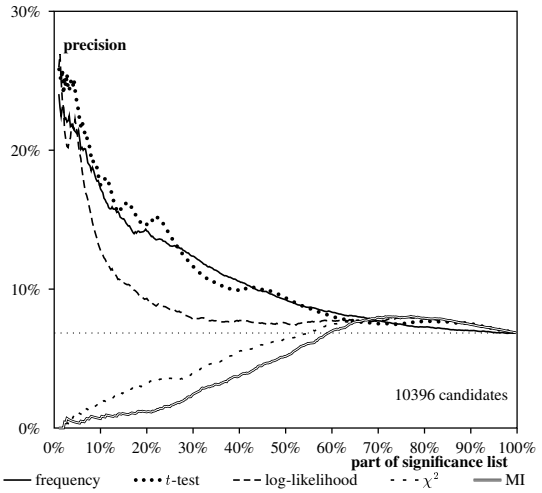

Figure 7: Precision curves for FVG

8 Recall and local precision

In section 7, we observed a *negative* correlation between FVG and the MI measure. Since there are unusually few FVG in the top part of the SL ordered according to the MI measure, some other part of the SL must contain a greater proportion of TPs than a randomly ordered list. We refer to the density of TPs in an arbitrary part of an SL (rather than among the n-best candidates) as *local precision.* Although it is difficult to compute exact local precision values, we can easily obtain approximate results by visual inspection of the recall curves shown in Figure 8.

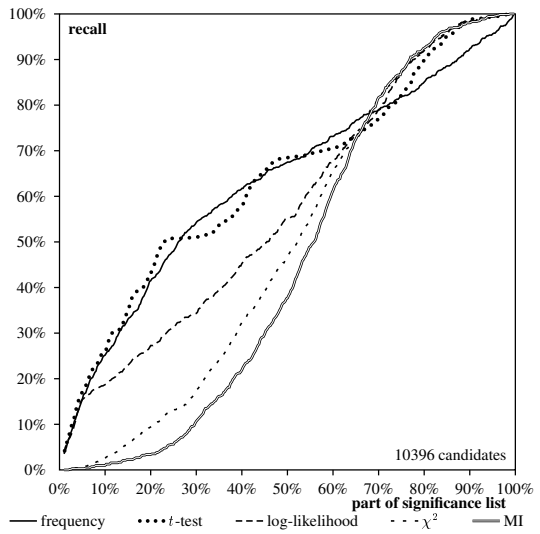

Figure 8: Recall curves for FVG

Recall, in this case, measures what proportion of all FVG is identified by a particular AM (i.e. can be found among the n best-scoring candidates according to that measure). The slope of an AM's recall curve corresponds to the number of FVG contained in that part of the SL, i.e., to local precision. For the MI measure, local precision is extremely low for the 3000 best-scoring candidates. The highest local precision values are achieved for the region of the SL extending (roughly) from the 4000th to the 7000th candidate pair (i.e., x-coordinates between 40% and 70%). This part of the MI recall curve is almost as steep as the initial part of the t-score recall curve. Thus, the best local precision values achieved by MI should be close to the best overall precision values of the t-score measure.

9 The MI mystery

Looking at the MI values computed for the candidate pairs, we find that the region of high local precision corresponds to MI scores ranging approximately from 4.0 to 7.5. Although candidate pairs with MI values above 7.5 are much more frequent than expected under the independence assumption, there are very few FVG among them.

The explanation usually given for the poor performance of the MI measure is that the highest MI scores are almost exclusively assigned to candidate pairs with low joint *and* marginal frequencies. However, for our data such a clear-cut distinction cannot be made.

For some reason that is yet unknown, a large proportion of FVG (around 50% for the newspaper data) have MI values between 4.0 and 7.5. We can exploit this property to construct an optimised AM for the extraction of FVG (but also fine-tuned to the evaluation data). We use the formula $-|MI - 5.75|$, which assigns scores between -1.75 and 0.0 to candidates whose MI score is between 4.0 and 7.5. All other candidates obtain values below -1.75. Figure 9 shows the precision curve for this fine-tuned AM compared to t-score and frequency. The 95% confidence region for the frequency measure is again delimited by two thin lines.

When more than 30% of the SLs are considered, $-|MI-5.75|$ achieves higher precision values than any of the other AMs. In the region between 40% and 80%, the new AM performs *significantly* better than co-occurrence frequency.

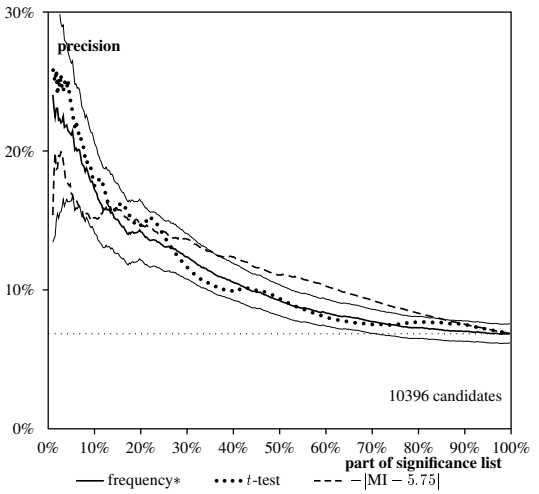

Figure 9: Precision of $-|MI - 5.75|$ for FVG

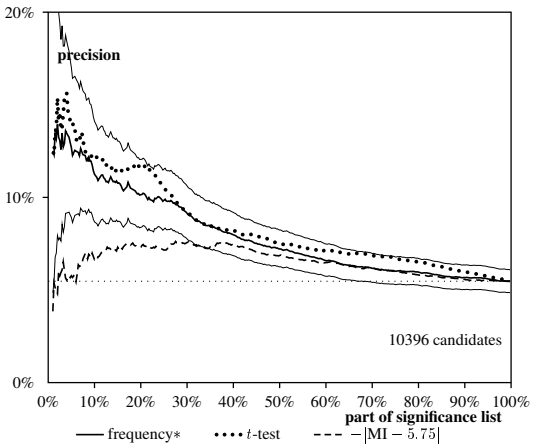

Figure 10: Precision of $-|MI - 5.75|$ for *figur*

Figure 10 shows that our new AM is not the best-performing measure for *figur*, and has significantly lower precision than frequency throughout the first 30% of the SL. However, it is still better than MI or its "squared" variant.

We have thus devised an AM optimised for one particular class of collocation. For this task, it is the only measure that has been able so far to produce significantly better results than mere co-occurrence frequency.

10 Conclusion

The performance of AMs in the identification of PP-verb collocations as described in the present paper is different from the results found in other studies that used less precisely defined definitions of collocativity. Overall, t-score achieves best precision values, but none of the AMs is significantly better than mere co-occurrence frequency. The widely-used log-likelihood measure is *significantly worse* than frequency (Figure 5), and the same holds for the common-birthday measure that Lezius (1999) found to have best performance.

Results do not seem to depend critically on the text type, as the comparison of newsgroup and newspaper data shows. But there are considerable differences between *figur* and FVG. In particular, FVG extracted from the newspaper corpus are surprisingly well characterised by an MI value between 4 and 7.5. Using this information, we can construct an AM adjusted to this class of collocation (and to the newspaper corpus used). This "tuned" AM is the only measure that can identify FVG with significantly better precision than co-occurrence frequency.

The results presented in this paper further suggest that in order to gain deeper insights into the performance of AMs, it is necessary to study the distributional differences between various precisely and narrowly defined classes of collocations. The key question then is which of these classes can be characterised (and thus extracted) in terms of joint and marginal frequencies. With this knowledge at hand, it should be possible to devise fine-tuned AMs.

Acknowledgement

The work of B. Krenn has been sponsored by the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)*, Grant No. P12920. Financial support for ÖFAI is provided by the Austrian Federal Ministry of Education, Science and Culture. The authors also would like to thank Anke Lüdeling for her undaunted insistence that AMs should be evaluated against reference sets defined by narrow linguistic criteria.

References

Hadumod Bußmann. 1990. *Lexikon der Sprachwissenschaft*. Kröner, 2nd edition.

Maria Teresa Cabré and Rosa Estopà. 1999. On the units of specialized meaning used in professional communication. Manuscript, 10 pp., IULA, Barcelona.

K.W. Church and P. Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pages 76 – 83.

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

Stefan Evert, Ulrich Heid, and Wolfgang Lezius. 2000. Methoden zum Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In *Proceedings of KONVENS 2000,* VDE-Verlag, Germany, pages 215 – 220.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations.* DFKI & Universität des Saarlandes, Saarbrücken.

Martin Läuter and Uwe Quasthoff. 1999. Kollokationen und semantisches Clustering. In *11. Jahrestagung der GLDV,* Enigma Corporation, Prag.

Wolfgang Lezius. 1999. Automatische Extrahierung idiomatischer Bigramme aus Textkorpora. In *Tagungsband des 34. Linguistischen Kolloquiums*, Germersheim.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Sidney Siegel. 1956. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, Tokyo.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. In *Computational Linguistics* 22(1), 3 – 38.

Wojciech Skut and Thorsten Brants. 1998. Chunk Tagger. Stochastic Recognition of Noun Phrases. In *ESSLI Workshop on Automated Acquisition of Syntax and Parsing*, Saarbrücken, Germany, August.