Separating the wheat from the chaff – Corpus-driven evaluation of statistical association measures for collocation extraction

Brigitte Krenn[1], Stefan Evert

In our contribution, we present a three step procedure for the evaluation of statistical association measures (AM) for lexical collocation extraction. The evaluation strategy provides a general framework for the evaluation and direct comparison of different AMs. The presented approach is corpus-driven and task-oriented, i.e. the collocation candidates identified by individual AMs are compared to a list of true positives manually extracted from the corpus data according to certain predefined, task-specific annotation criteria. For practicability a random sample approach to evaluation is defined, and methods of the assessment of intercoder agreement are discussed.

In diesem Beitrtag wird ein dreistufiges Modell für die Evaluierung von statistischen Assoziationsmassen (AM) zur Kollokationsextraktion aus Korpusdaten vorgestellt. Der Ansatz zeichnet sich dadurch aus, dass er eine generelle Methode zur Evaluierung und zum direkten Vergleich unterschiedlicher Assoziationsmasse darstellt. Der Ansatz ist datenorientiert in dem Sinn, dass die mittels AM identifizierten Kollokationskandidaten anhand einer manuell aus den Korpusdaten extrahierten Liste von tatsächlichen Kollokationen evaluiert werden. Die manuelle Annotierung der „echten“ Kollokationen erfolgt auf Basis klar definierter Annotationsrichtlinen. Um die Brauchbarkeit der Annotationsrichtlinien abzuschätzen, werden Methoden zum Testen der Übereinstimmung zwischen mehreren Annotierern vorgestellt. Aus Gründen der Praktikabilität wird eine Methode eingeführt, bei der die Evaluierung auf Basis einer Zufallsstichprobe anstelle der vollen Datenmenge erfolgt.

[1] The Austrian Research Institute for Artificial Intelligence (ÖFAI) is supported by the Austrian Federal Ministry for Education, Science and Culture, and by the Austrian Federal Ministry for Transport, Innovation and Technology.

2. Introduction

Over the recent years, a variety of methods for the extraction of multiword units, terms, phraseological units, etc. – which we henceforth subsume under the term lexical collocation – have been proposed. To assess the practical usefulness of the different methods, a number of questions need to be addressed such as the following: What types of collocations are to be extracted automatically? What are the domain and size of the extraction corpus? How appropriate is a certain method for the treatment of high frequency versus low frequency data? Only the last question can be satisfactorily answered by a mathematical discussion of the statistical association measures (AMs) employed. In general, one has to be aware that evaluation results are valid only for data from specific corpora, extracted with specific methods, and for a particular type of collocations. This situation demands corpus- and task-driven empirical evaluation experiments.

In our contribution, we present an approach to the evaluation of methods and tools for extraction of lexical collocations from corpora which (i) is independent of the candidate extraction strategy employed (positional or relational n-tuples, adjacent or nonadjacent n-tuples); (ii) allows for a direct and meaningful comparison of different AMs; (iii) is corpus-driven, i.e., the collocation candidates identified by the individual AMs are compared to a list of manually identified true positives (TPs) in the extraction corpus; and (iv) can be applied to random samples from the candidate set to reduce the amount of manual annotation work.

Our approach implements a three step procedure where: in Step 1, lexical tuples are extracted from a source corpus and the frequency data for each candidate type (i.e., distinct tuple) are represented in the form of a statistical contingency table; in Step 2, AMs are applied to the contingency tables, resulting in a candidate list of n-tuples ranked according to their associated AM scores; in Step 3, the usefulness of the individual AMs for collocation extraction is assessed by comparing the candidate lists to a full list of TPs in the candidate set that have been manually identified by expert annotators. The more TPs there are in a given n-best list, the better the performance of the measure. This performance is quantified by the n-best precision and recall of the AM. As manual inspection of all candidate data extracted from a corpus is very time consuming and thus not feasible in daily practice, we argue for an evaluation based on random samples, an approach which we have shown to be feasible through practical experimentation.

In Section 2, we describe the general strategy for evaluating AMs. The random sample approach is presented in Section 3. In Section 4, we address criteria for the distinction of collocations and non-collocations (Section 4.1), and methods for assessing intercoder agreement (Section 4.2).

2. General Evaluation Strategy

In the following, we present a general procedure for evaluating the usefulness of different measures for the identification of lexical collocations from text corpora.

Step 1 – Extraction of lexical tuples: Lexical tuples are extracted from a source corpus, and the cooccurrence frequency data for each candidate type are represented in the form of a contingency table. It is important to bear in mind that the types of lexical tuples identified and their frequencies are strongly influenced by the extraction strategies employed. For instance, depending on the extent to which the characteristic linguistic properties of a class of collocations are taken into account by the extraction process, there will be a larger or smaller amount of noise in the data. Extracting clearly defined linguistic units such as verb-object combinations from German text (as opposed to arbitrary cooccurrences of verbs and nouns within sentences) results in a larger number of *Funktionsverbgefüge*[2] among the lexical tuples. An extraction strategy specifically geared towards Funktionsverbgefüge is described in (Breidt 93). The choice between using full or lemmatized forms of words also influences the frequency counts in the contingency tables.

In contrast to the relational approach, where the grammatical relations of word combinations are taken into account, the positional approach extracts lexical tuples from numerical spans[3] without making use of any explicit linguistic knowledge. As a result, the extracted tuples belong to a broad mix of linguistic units and include a large proportion of noise such as: combinations that cross phrase boundaries, or lexical realizations of grammatically prominent but non-lexicalized word combinations, e.g. article-noun or other closed class –

[2] Funktionsverbgefüge are a specific type of verb-object collocations typically constituted by a main verb deprived of its major lexical semantics and a predicative noun. For a discussion of Funktionsverbgefüge see (Krenn 2000, p. 74ff.).

[3] The numerical span defines the number of words to right and/or left of a *keyword* w that are considered as potential *collocates* for the keyword.

open class combinations. To avoid this, lists of function words (or other words that appear to be detrimental to the extraction quality) have been introduced as stop lists. Radical elimination of closed class words, however, is not desirable in all cases: for some lexical collocations the closed class element has a distinctive function. For instance, in the two variants of the German Funktionsverbgefüge *in Betrieb gehen* (to start operating) versus *ausser Betrieb gehen* (to stop operating) the prepositions *in* and *ausser* convey inchoative and terminative *Aktionsart*, respecitvely.

Whichever strategy has been employed for extracting the lexical tuples, the AMs under investigation are applied to the cooccurrence frequency data in the contingency table of each candidate type. As most AMs are designed for two-dimensional contingency tables, we currently restrict our example evaluations to 2-tuples, i.e. to pairs.[4] Each lexical tuple is thus a pair of two lexical units which may themselves be simple or complex, and its cooccurrence frequency information is represented by a 2×2 contingency table. For instance, our data in used for the identification of adjective-noun collocations are derived from frequency counts of adjective-noun pairs (where the adjective modifies the nominal head of a NP), whereas the data for the identification of verb-object collocations stem from frequency counts of preposition-noun pairs (the preposition and nominal head of a PP) in combination with main verbs.

In this paper, we present the verb-object data, which consist of (P+N,V) combinations extracted from an 8 million word portion of the Frankfurter Rundschau corpus[5] such that every PP (represented by the combination P+N) in a sentence is combined with every main verb V that occurs in the same sentence. In this way, we obtain 294 534 distinct PNV combinations (lemmatized pair types), 80% of which occur only once in the corpus ($f = 1$). Another 15% occur only twice ($f = 2$), and merely 5% have occurrence frequencies $f \geq 3$. This illustrates the Zipf-like distribution of lexical tuples typical for corpus data. For the evaluation experiments, we use the 14 654 PNV types with $f \geq 3$ as candidates for lexical collocations. We henceforth refer to them as the PNV data set.

[4] In general, lexical n-tuples are represented by n-dimensional contingeny tables with 2^n cells.

[5] The Frankfurter Rundschau (FR) Corpus is a German newspaper corpus, comprising approximately 40 million words of text. It is part of the ECI Multilingual Corpus 1 distributed by ELSNET. See http://www.elsnet.org/resources/ecicorpus.html for details.

In Table 1, we show the frequency information for the Funktionsverbgefüge *in Frage stellen* ("to question"). From this table we can see that the combination *in Frage stellen* (O_{11}) occurs 146 times in the PNV data set, whereas the combination of *in Frage* with any other main verb but *stellen* (O_{12}) occurs 236 times, the combination of *stellen* with any other P+N combination but *in Frage* (O_{21}) occurs 3 901 times, and the number of (P+N,V) combinations where P+N is not *in Frage* and V is not *stellen* (O_{22}) occurs 10 371 times.

	stellen	any main verb but *stellen*
in Frage	O_{11} = 146	O_{12} = 236
any P+N but *in Frage*	O_{21} = 3 901	O_{22} = 10 371

Table 1: contingency table for in Frage stellen

Step 2 – Application of association measures: AMs are applied to the frequency information collected in the contingency tables. Every AM assigns a score to each lexical tuple in the data set. For each individual AM, the candidate list is ordered from highest to lowest score. Thus we obtain as many different rankings of the candidate set as AMs are applied to the data. Since, by the usual convention, higher scores indicate stronger statistical association (which is interpreted as evidence for collocativity) we use the first n candidates from each such ranking for evaluation. When there are ties in the rankings, they need to be resolved in some way in order to select exactly n candidates. To avoid biasing the evaluation results, ties are broken randomly in our experiments.

For the illustration experiment in this paper we compare the following measures: (i) t-score (Church et al., 91) and log-likelihood (Dunning, 93), (ii) Pearson's chi-squared test (with Yates' correction applied) and (iii) plain cooccurrence frequency. Log-likelihood and t-score are two widely-used AMs. While the chi-squared test is considered as the standard test for association in contingency tables, it has not found widespread use in collocation extraction tasks (although it is mentioned by Manning & Schuetze 99).

Step 3 – Evaluation of the candidate lists generated by the AMs against manually annotated data: In order to assess the usefulness of each individual AM for collocation extraction, the (ranked) candidate lists are compared to a manually identified list of tuples where true positives (TP) and false positives (FP) are marked by a human annotator based on (hopefully) clearly defined annotation guidelines. The more TPs there are in a given n-best list, the better the performance of the measure. This performance is quantified by the n-best precision and recall of the AM.

Let t(n) be the number of TPs in a given n-best list and t the total number of TPs in the candidate set. Then the corresponding n-best precision is defined as p = t(n)/n and recall as r = t(n)/t. Precision-by-recall plots are the most intuitive mode of presentation, as detailed in Evert (2004b), Section 5.1. Since they can be understood as mere coordinate transformations of the original precision plots, we restrict our presentation here to precision plots. For a predefined list size n, the main interest of the evaluation lies in a comparison of the precision achieved by different AMs, while recall may help to determine a useful value for n. Evaluation results for many different list sizes can be combined visually into a precision plot as shown in Figure 1 for the full PNV data set and the task of identifying Funktionsverbgefüge and figurative expressions. Detailed annotation guidelines can be found in (Krenn 2004).

Figure 1 shows that the precision of AMs (including frequency sorting) typically decreases for larger n-best lists, indicating that the measures succeed in ranking collocations higher than non-collocations, although the results are far from perfect. Of course, the precision of any AM converges to the baseline for n-best lists that comprise almost the entire candidate set. The baseline precision (6.41% in this example) is the proportion of collocations in the entire candidate set, i.e. the total number of TPs (here, 939) divided by the total number of collocation candidates (here, 14 654). The x-axis covers all possible list sizes, up to n = 14 654. Evaluation results for a specific n-best list can be reconstructed from the plot, as indicated by the thin vertical lines for n = 1 000, n = 2 000 and n = 5 000. In the example, the differences between the AMs vanish for n ≥ 8 000. In our example, larger lists are hardly useful for collocation extraction as all measures have reached a recall of approx. 80% for n = 8 000. From the precision graphs we see that t-score clearly outperforms log-likelihood for n ≤ 6 000, and in the range 2 000 ≤ n ≤ 6 000 even simple frequency sorting is better than log-likelihood. Chi-squared achieves a poor performance on the PNV data and is hardly superior to the baseline, which corresponds to random selection of candidates from the data set. This last observation supports Dunning's claim that the chi-squared measure tends to overestimate the significance of (non-collocational) low-frequency cooccurrences (Dunning, 93).

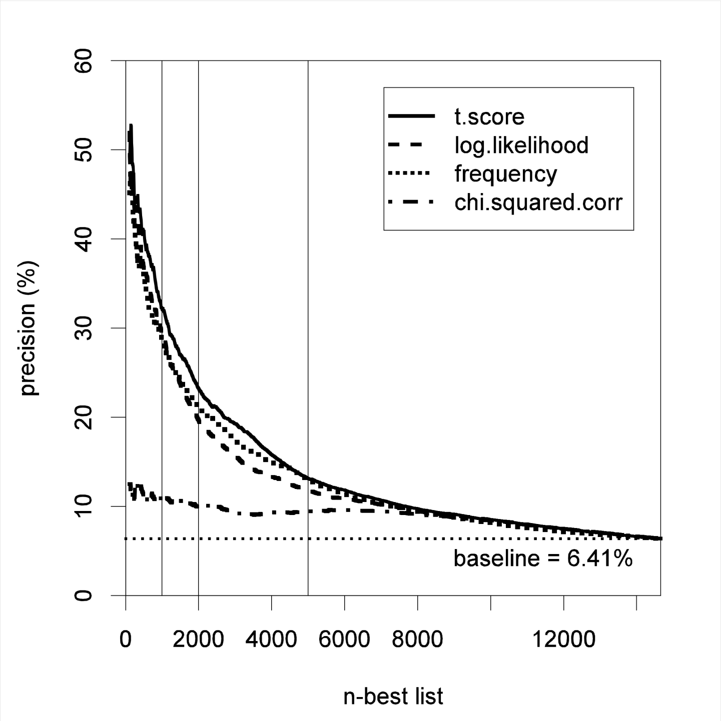

Figure 1: Precision plots for AMs on a figurative expressions / Funktionsverbgefüge extraction task, obtained by annotation of the full PNV data set

On the one hand, extensive experimentation is important, as the quality of individual AMs strongly depends on the properties of the candidate data set (especially the source corpus and the strategies used for the extraction of lexical tuples) and the kind of lexical collocations to be extracted, cf. (Krenn, 2000), (Krenn & Evert, 2001), (Evert & Krenn, 2001). On the other hand, manual annotation of TPs on the full data set is not feasible for practical work in collocation extraction, because it is highly time consuming. In a way, it also makes the use of AMs redundant. As a way out of this dilemma, we propose to evaluate AMs based on a random sample from each data set. Thus Step 3 of our general evaluation strategy is replaced by a random sample evaluation (RSE) procedure which is outlined in the following section.

3. Random Sample Evaluation (RSE)

3.1 Basics Procedure

To achieve a substantial reduction in the amount of manual annotation work, only a random sample R of the full data set C ($R \subseteq C$) is annotated. The ratio |R| / |C| is called the sampling rate, and is usually small (10% to 20% of C). The manual annotation identifies all TPs from T (the TPs in the full data set) that belong to the sample R, i.e. the set $T \cap R$. The baseline precision b is estimated by the proportion of TPs in the random sample, i.e. $b' = |T \cap R| / |R|$. Similarly we can estimate the true precision p(A) of any subset $A \subseteq C$ by the ratio

$$p'(A) = |A \cap T \cap R| / |A \cap R| = k'(A) / n'(A)$$

which is called the sample precision of A. We use the shorthand notation n'(A) for the number of candidates sampled from A, and k'(A) for the number of TPs found among them. Correspondingly, an estimate for the n-best precision $p_{g,n}$ of an AM g is given by

$$p'_{g,n} = p'(C_{g,n}) = k'_{g,n} / n'_{g,n}$$

In addition, a statistical measure for the accuracy of this estimate can be computed in the form of a confidence interval for p'. A special hypothesis test can then be used to determine whether the observed difference between the p'-values of two different AMs is significant or whether it could be a result of the random sampling process (see Evert 2004, Section 5.3 for details).

3.2 Evaluation Examples using RSE

In the following, we present RSE evaluations on two different data sets: (1) the PNV data set introduced in Section 2, and (2) an adjective-noun (AN) data set which will be introduced below.

PNV data: The right panel of Figure 2 shows graphs of $p'_{g,n}$ for $n \leq 5\,000$, estimated from a 10% sample of the PNV data set. Note that the x-coordinate is n, not the the number $n'_{g,n}$ of sampled candidates. The baseline shown in the plot is the sample estimate b'. The thin dotted lines above and below indicate a confidence interval for the true baseline precision. From a comparison with the true precision graphs in the left panel, we see that the overall impression given by the RSE is qualitatively correct: t-score emerges as the best measure, mere frequency sorting outperforms log-likelihood (at least for n>=4 000), and chi-squared is much worse than the other measures, but is still above the baseline. However, the findings are much less clear-cut than for the full evaluation; the precision graphs become unstable and unreliable for n<=1000 where log-likelihood seems to be better than frequency and chi-squared seems to be close to the baseline. This is hardly surprising considering the fact that these estimates are based on fewer than one hundred annotated candidates.

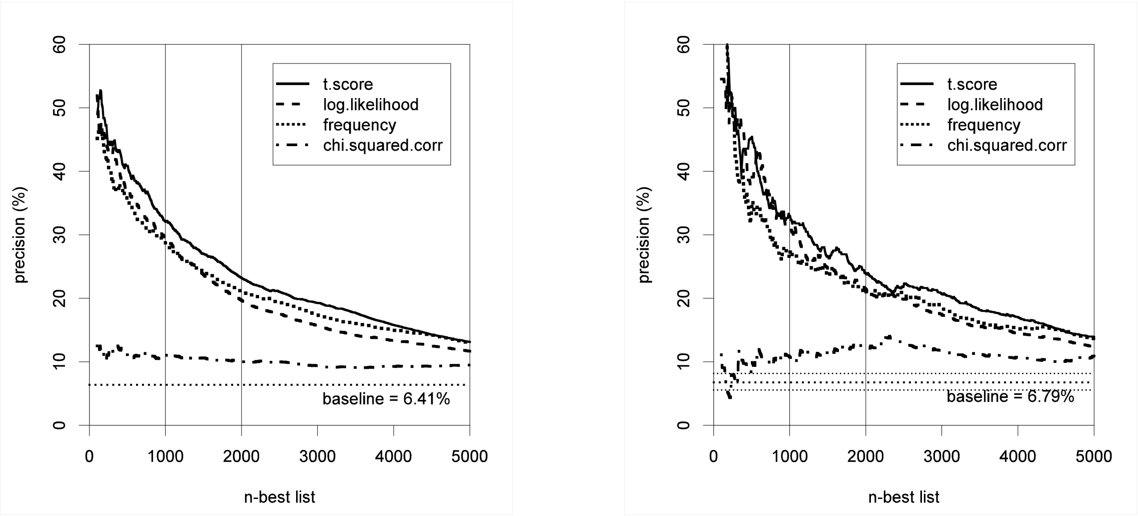

Figure 2: An illustration of the use of random samples for evaluation: precision plots for the 5000-best candidates from the full PNV data set (left) and the corresponding estimates from a 10% sample (right)

AN data: Figure 3 shows another example of an RSE evaluation. Here, German adjective-noun combinations were extracted from the full Frankfurter Rundschau Corpus, using part-of-speech patterns as described in (Evert & Kermes, 2003), and a frequency threshold of f ≥ 20 was applied. From the resulting data set of 8 546 candidates, a 15% sample was manually annotated by professional lexicographers (henceforth called the AN data set). In contrast to the PNV data, which uses a linguistically motivated definition of collocations, the annotators of the AN data set also accepted ``typical'' adjective-noun combinations as true positives when they seemed useful for the compilation of dictionary entries, even if these pairs would not be listed as proper collocations in the dictionary. Such a task-oriented evaluation would have been impossible if, e.g., an existing dictionary had been used as a gold standard. The results of the AN evaluation experiment are quite surprising in view of previous experiments and conventional wisdom. Frequency-based ranking is not significantly better than the baseline, while both t-score and log-likelihood are clearly outperformed by the chi-squared measure, contradicting the arguments of (Dunning, 1993). For 1 000 ≤ n ≤ 3 000, the precision of chi-squared is significantly[6] better than that of log-likelihood.

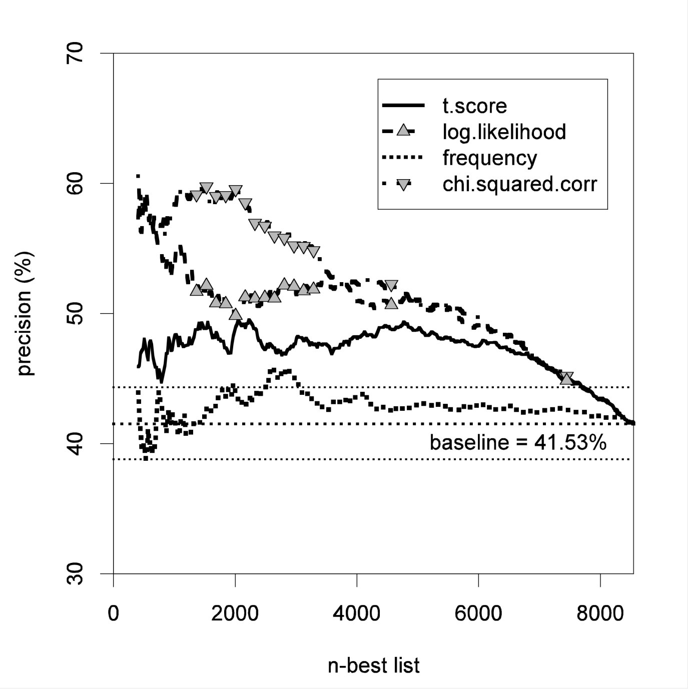

Figure 3: RSE of German adjective+noun combinations

4. Manual Annotation of Candidate Data

4.1 Phenomena

There is no single theory of lexical collocations. The phenomena subsumed under the term are manifold, ranging from lexical proximities in texts (cf. Firth's notion of collocation, Firth 1957) to syntactic and semantic units characterized by semantic opacity as well as syntactic irregularity and rigidity. However, lexical collocations differ from arbitrary word combinations in various ways:

[6] For testing the significance of differences between two AMs see (Krenn & Evert, 2001).

The determining elements of a collocation are lexically selected, i.e., certain words cooccur preferentially, e.g. German *Jacke anziehen* (jacket + put on) versus *Hut aufsetzen* (hat + put on) versus *Kette anlegen* (necklace + put on).

A number of lexical collocations are morphologically or syntactically marked, e.g. *zu Felde ziehen* (‘to campaign’) where *Feld-e* is an ancient inflection form that is unusual outside the collocation; *ins Rollen bringen* (in+ARTICLE rolling bring, ‘to set something rolling’) where the article cannot be separated from the preposition without meaning loss (** in das Rollen bringen*). *Rollen* also cannot be pronominalised: *weil sie die Affäre ins Rollen bringt* (because she the affair into rolling brings) versus ** weil sie die Affäre in es bringt* (because she the affair into it brings) or ** weil sie die Affäre ins Rollen bringt, das endlos ist* (because she the affair into rolling brings, which (= the rolling) endless is).

A number of lexical collocations are semantically restricted: in the case of idioms the meaning of a complex linguistic unit is opaque, e.g. *ins Gras beissen* (‘die’). In other cases, semantic compositionality is restricted or a metaphoric interpretation is required, e.g. *am Herzen liegen* (‘to be important’).

Modification can be restricted, e.g., *ins Gras beissen* in the reading ‘die’ can only be modified as a whole, i.e., only modification of the VP is possible: *weil er irgendwann ins Gras beisst* (because he some day in+the grass bites, ‘because he dies some day’) versus ** weil er ins grüne Gras beisst* (because he in+the green grass bites), similarly *weil die Affäre schnell ins Rollen* kommt (‘because the affaire is set rolling quickly’) versus ** weil die Affäre ins schnelle Rollen kommt* (because the affaire into+the quick rolling comes).

Apart from the variety of definitions and terminology related to lexical collocations, which has been influenced by different linguistic and lexicographic traditions, there is a general problem with lexical collocations: individual instances can be found in a continuum ranging from syntactically and semantically fully flexible linguistic units, which are only marked by lexical selection, to linguistically rigid units such as idioms. Moreover, a variety of word combinations are habitually used in certain contexts. For instance, a frequent tuple in the PNV data set is *um Uhr beginnen* (at o’clock begin). The PP-verb pair does not exhibit any characteristics that would make it qualify as a collocation, nevertheless it is typical for the announcement of events and therefore might be a combination of interest (a TP) in a certain context of use.

In practical work on collocation extraction, we typically find an opportunistic approach to collocativity, i.e., the definition of TPs depends on the intended application rather than being motivated by (linguistic) theory, and it covers a mixture of different phenomena and classes of lexicalized word combinations. This is particularly obvious in lexicography where the definition of TPs strongly depends on their usefulness for the update or compilation of a particular dictionary, as is the case for the AN data presented in Section 3. Accordingly, it is important that a clear picture about the kinds of TPs relevant for a certain extraction task is given beforehand, providing guidelines for the annotation of TPs and FPs in Step 3 of the general evaluation procedure. All this requires a data- and task-driven experimental approach to evaluating the true usefulness of a particular AM or combination of AMs for the extraction of certain kinds of lexical collocations. In order to keep the amount of manual work limited, the random sample evaluation procedure is indispensable.

Despite the availability of well-defined criteria and explicit annotation guidelines, annotators may make different decisions for some of the collocation candidates because of individual mistakes, differences in their intuition, disagreement about the precise interpretation of the guidelines, etc. Thus the assessment of intercoder agreement on a certain annotation task is important. In the following section, we will briefly discuss methods for the assessment of intercoder agreement and report on results for the annotation of Funktionsverbgefüge and figurative expressions on the PNV data set. For the annotation guidelines employed here see (Krenn, 2004).

4.3 Intercoder Agreement

A widely used means for measuring intercoder agreement is the kappa statistic (Cohen, 60), where the observed agreement between coders is compared with chance agreement. Kappa values lie within the interval [0,1]. While the interval boundaries are well defined (kappa = 0 indicates mere chance agreement, and kappa = 1 indicates perfect agreement), the intermediate values are hard to interpret. A widely used interpretation of kappa values for natural language coding tasks was suggested by Krippendorff (1980), especially with regard to dialogue annotation. Krippendorff distinguishes the following values:

kappa $\leq .67$	to be discarded
$.67 \leq$ kappa $\leq .8$	shows tentative agreement

kappa ≥ .8	definite agreement

When using kappa values, it is important to be aware of their inherent uncertainty: the computed values are sample estimates, whose variance has to be taken into account for the comparison with a fixed scale (see Fleiss et al. (1969) for a formula spcifying the variance of kappa estimates). Thus, intercoder agreement needs to be interpreted on the basis of the intervals constituted by observed kappa values and their standard deviations, rather than by looing only at the observed values. Moreover, the kappa values are not very intuitive, because the definition of the kappa statistic starts from the unrealistic assumption that there were only chance agreement between annotators. Of course, from a linguistic point of view we would expect that there is a substantial agreement between annotators, because of common language competence of native speakers and the availability of clearly defined annotation guidelines. An attempt to formulate a more realistic measure of intercoder agreement is presented in (Krenn et al. 2004). This measure estimates the proportion of true agreement between annotators by dividing the observed ("surface") agreement into true and chance agreement. The authors present results for pair-wise intercoder agreement between a reference annotator and 12 other annotators on a Funktionsverbgefüge/figurative expression coding task (using the PNV data set). A comparison of the resulting confidence intervals for the proportion of true agreement with the confidence intervals of the corresponding kappa values shows a striking similarity for each of the 12 pairings. This finding may open up a new and more intuitive interpretation of the kappa values, but further investigation is required to validate our conjecture. The results of the experiment also show that high intercoder agreement is achieved by trained linguists while the agreement of non-expert annotators with the reference is much lower. This provides evidence that reliable annotation of collocation data requires expert annotators.

5. Conclusion

In this paper, we have presented a general procedure for evaluating the usefulness of different association measures (AMs) for extracting lexical collocations from corpus data. We have provided evidence that the assessment of AMs requires extensive experimentation, because the usefulness of individual measures depends on the properties of the source corpus, the candidate extraction strategies that were used, and the type of collocations to be identified. Thus, results obtained in a particular setting cannot easily be generalised to different settings. The enormous differences in extraction quality that the same AM may show in different tasks is illustrated by the examples of PP-verb (PNV) and adjective-noun (AN) collocations.

A full manual inspection of the candidate data extracted from a specific corpus for true positives (TPs) and false positives (FPs) is rarely feasible. To remedy this situation, we argue for an evaluation based on random samples from the full candidate set, so that only a small portion (10 to 20%) of the initial data need to be inspected manually. This is an important precondition for the broad empirical studies that are needed to obtain a better understanding of the general properties of AMs in collocation identification tasks.

We have argued that clear and detailed guidelines are required for the manual annotation TPs among the candidate data. The precise definition of TPs depends both on the task and the data, so special-purpose annotation manuals have to be created for each experiment. This fact is hardly surprising considering that (i) there is no single definition or theory of lexical collocations, and (ii) practical work on collocation extraction typically takes an opportunistic approach to collocativity, where the definition of TPs is largely determined by their intended use. For instance, lexicographers that are updating a dictionary, "typical" word combinations will often be of greater interest than collocations in a narrow linguistic sense.

Finally, we have discussed the necessity for and different ways of measuring intercoder agreement. We have presented a procedure that is more meaningful and intuitive than the widespread practice of using observed kappa values, and have concluded that a reliable manual annotation of lexical collocations requires expert annotators.

References

Breidt, E., June 1993. Extraction of N-V-collocations from text corpora: A feasibility study for German. In: *Proceedings of the 1st ACL Workshop on Very Large Corpora*. Columbus, Ohio.
(A revised version is available from http://arxiv.org/abs/cmp-lg/9603006)

Church, K., Gale, W., Hanks, P., Hindle, D., 1991. Using statistics in lexical analysis. In: *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Lawrence Erlbaum, pp. 115–164.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**: 37–46.

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

Evert, S., 2004a. An on-line repository of association measures.
http://www.collocations.de/AM/

Evert, S., 2004b. *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* PhD dissertation, University of Stuttgart.

Evert, S., Krenn, B., 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, pp. 188–195.

Evert, S., Kermes, H., 2003. Experiments on candidate data for collocation extraction. In: *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics*. pp. 83–86.

Firth, J.R., 1957. *Papers in Linguistics 1934 – 1951.*Oxford University Press, London.

Fleiss, J. L., Cohen, J. and Everitt, B. S., 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin,* **72**(5): 323–327.

Krenn, B., 2000. *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*. Vol. 7 of Saarbrücken Dissertations in Computational Linguistics and Language Technology. DFKI & Universität des Saarlandes, Saarbrücken, Germany.

Krenn, B., 2004. Manual zur Identifikation von Funktionsverbgefügen und figurativen Ausdrücken in PP-Verb-Listen.
http://www.collocations.de/guidelines/

Krenn, B., Evert S. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France, July.

Krenn, B., Evert, S., Zinsmeister H., 2004. Determining intercoder agreement for a collocation identification task. In: *Proceedings of Konvens '04*. Vienna, Austria.

Krippendorff, K., 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

Manning, C. D., Schütze, H., 1999. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA.