Using Small Random Samples for the Manual Evaluation of Statistical Association Measures

Stefan Evert

IMS, University of Stuttgart, Germany

Brigitte Krenn

ÖFAI, Vienna, Austria

Abstract

In this paper, we describe the empirical evaluation of statistical association measures for the extraction of lexical collocations from text corpora. We argue that the results of an evaluation experiment cannot easily be generalized to a different setting. Consequently, such experiments have to be carried out under conditions that are as similar as possible to the intended use of the measures. Finally, we show how an evaluation strategy based on random samples can reduce the amount of manual annotation work significantly, making it possible to perform many more evaluation experiments under specific conditions.

Key words: collocations, cooccurrence statistics, evaluation, association measures

1 Introduction

In this contribution, we propose a three-step procedure for empirically evaluating the usefulness of individual statistical association measures (AMs) for the identification of lexical collocations in text corpora. In order to reduce the manual annotation work required, we propose a random sample evaluation (RSE) where the AM(s) most appropriate for a certain task and a specific extraction corpus are identified on the basis of a random sample extracted from the extraction corpus in question.

Email addresses: evert@ims.uni-stuttgart.de (Stefan Evert), brigitte@oefai.at (Brigitte Krenn).

1.1 Motivation

All statistics-based approaches to natural-language processing require a thorough empirical evaluation. This is also the case for the extraction of collocations from text corpora using statistical association measures (AMs). Common practice in this area, however, is that evaluations have a middlingly ad-hoc character. Authors typically look at small lists of n highest-ranking collocation candidates and decide, most often by rule of thumb, which of the lexical tuples in the candidate list qualify as true positives (TPs), while the actual discussion focuses on the mathematical properties of the proposed measure. [1] This is without dispute an important issue, but not sufficient to get a complete picture of the usefulness of a certain AM in practice.

A common approach to the identification of lexical collocations is their semi-automatic extraction from text corpora. First, n-tuples of syntactically related words are extracted as collocation candidates, which are then annotated with AM scores. Finally, the candidates with the highest scores are inspected by a human expert in order to select the true collocations (= TPs). The extraction step is usually based on a syntactic pre-processing of the corpus, although some researchers define cooccurrence purely in terms of the distance between words (e.g. Sinclair, 1991), if only because the necessary pre-processing tools are not available (cf. Choueka, 1988). Most AMs are designed for word pairs, although first suggestions for an extension to n-tuples have been made (da Silva and Lopes, 1999; Blaheta and Johnson, 2001). [2] Although our example data consist of word pairs that occur in specific syntactic relations, the proposed evaluation procedure is independent of the number of words in a lexical tuple and the extraction method used. In any case, the resulting set of collocation candidates will be huge, most of them occurring just once or twice in the corpus (in accordance with Zipf's law).

The simplest approach to improving the quality of automatically extracted collocation candidates is to rank them by their cooccurrence frequencies, following the intuition that recurrence is a good indicator of collocativity (see e.g. Firth, 1957, Ch. IV). Further improvements are expected from AM scores, since the statistical association between the component words of each candidate is assumed to correlate better with collocativity than mere cooccurrence frequency. Association measures can be applied to a candidate set in three different ways: (a) use a certain AM value as a threshold to distinguish between collocational and non-collocational word combinations; (b) rank the

[1] See Evert (2004b) or Evert (2004a) for a comprehensive listing of known AMs.

[2] Such extensions typically focus on plain sequences of adjacent words called n-grams (e.g. Dias et al., 1999), rather than tuples of syntactically related words that may be quite far apart in the surface form of a sentence (cf. Goldman et al., 2001).

candidates according to their AM scores and select the n highest-ranking candidates for manual annotation (called an n-best list); (c) leave it to the human annotator how many candidates from the ranked list she is willing to inspect. The direct use of threshold values is not very common in practical work, which most often focuses on n-best lists where n is determined *a priori* by external requirements.[3] In the paper, we will therefore concentrate on (b), which is equivalent to (a) for a suitably chosen threshold (ignoring the possibility that there may be ties in the ranking). Procedure (c) can also be seen as equivalent to (b), except that it does not use a pre-determined size for the n-best list (instead, n is determined interactively by the annotator). Some collocation extraction methods apply various filtering techniques to reduce the size of the candidate set (e.g. Smadja, 1993). Although these methods do not result in a ranking of the candidates, they are directly comparable with the n-best lists of AMs, provided that n is chosen to match the number of candidates that remain after filtering. In this way, our evaluation procedure can also be applied to such methods.

From theoretical discussions, log-likelihood (Dunning, 1993) emerged as a statistically sound measure of association. Since it is also convenient in practical work, it has become popular as an all-purpose measure in computational linguistics. Even though most evaluation experiments have confirmed log-likelihood as the most useful AM for collocation extraction so far (specifically Daille (1994), Lezius (1999), and Evert et al. (2000)), sorting by mere cooccurrence frequency (without a sophisticated statistical analysis) has also led to surprisingly good results. However, Krenn (2000) found the t-score measure (Church et al., 1991) to be optimal for the extraction of German PP-verb collocations (which she defined as figurative expressions and *Funktionsverbgefüge*) from newspaper text and Usenet group discussions. On these data, the log-likelihood ranking was significantly worse than simple frequency sorting for n-best lists with $n \geq 2000$ (see Evert and Krenn, 2001). This example shows that log-likelihood may not always be the best choice. On the other hand, measures such as MI and t-score, which are widely used in computational lexicography, will be suboptimal for most other tasks. With a felicitous choice of measure, it is often possible to improve substantially on frequency sorting, log-likelihood and other standard AMs (e.g. Krenn and Evert, 2001). The practical usefulness of individual AMs depends on such different issues as the type of collocation to be extracted, domain and size of the source corpora, the tools used for syntactic pre-processing and candidate extraction, and the amount of low-frequency data excluded by setting a frequency threshold. Therefore, only an empirical evaluation can identify the best-performing AM

[3] One exception is the work of Church and Hanks (1990), who use an empirically determined threshold for the MI measure to select collocation candidates. In a later publication, this procedure is augmented by a theoretically motivated threshold for the t-score measure (Church et al., 1991).

under a given set of conditions.

1.2 Our Approach in a Nutshell

Step 1 of the proposed evaluation procedure is the extraction of lexical tuples from the text corpus. In step 2, the AMs under investigation are applied to the lexical data. In step 3, the candidate data are manually evaluated by a human annotator. Each candidate is marked as a true positive (TP) or false positive (FP). Finally, the AMs are evaluated against this manually annotated data set by computing the precision and recall of the respective n-best lists.

There are two major reasons why a meaningful evaluation of AMs requires manual annotation of the candidate data. (1) No existing lexical resource can be fully adequate for the description of a new corpus (i.e. any corpus that did not serve as a basis for the compilation of the resource). This argument is similar to the case made for lexical tuning by Wilks and Catizone (2002) with respect to word senses. Some researchers have tried to circumvent manual annotation of the candidate data by using a paper dictionary or machine-readable database as a "gold standard". Unfortunately such a gold standard necessarily provides only partial and inadequate coverage of the true collocations that will be found in a corpus. (2) Nowadays dictionaries become increasingly corpus-based. This poses the additional danger of introducing a bias in favour of whichever association measure or other method was used to extract collocation candidates for the dictionary.

Irrespective of the (non)generalizability of AMs, manual annotation of the candidate data is an expensive and time-consuming task. Random sample evaluation helps to reduce the amount of manual annotation work drastically. To do so, in step 3 of our evaluation procedure we use a random sample of the candidate data for manual annotation instead of the full set. The most appropriate AM(s) for the given extraction task and the complete extraction corpus can then be predicted on the basis of this random sample. After describing the mathematical background of the RSE procedure and appropriate tests for the significance of results, we illustrate its utility with an evaluation of German PP-verb pairs (Krenn, 2000). This example shows that the RSE results are comparable to those of a full evaluation. A second example, carried out on German adjective-noun data, provides further evidence for the necessity of repeated evaluation experiments, especially as the results obtained on the adjective-noun data contradict those of the PP-verb data. [4]

[4] The RSE procedure for the evaluation of AMs is implemented as an R library in the UCS toolkit, which can be downloaded from `http://www.collocations.de/`. All evaluation graphs in this paper (including confidence intervals and significance tests) were produced with the UCS implementation. R is a freely available program-

In such a situation – where it is difficult to generalize evaluation results over different tasks and corpora, and where extensive and time-consuming manual inspection of the candidate data is required – RSE is an indispensable means to make many more and more specific evaluation experiments possible.

Section 2 is dedicated to the empirical evaluation of measures for collocation extraction. In Section 2.1, we present a general procedure for manual evaluation, which is then applied to a selection of AMs and the task of extracting collocations from German PP-verb data (Section 2.2). In the following, we argue that only an evaluation based on random samples (RSE) allows us to study the usefulness of AMs in a wide range of situations. Section 3 presents the mathematical details of the evaluation procedure. First, we introduce a formal notation for the evaluation process (Section 3.1), followed by an explanation of the RSE method (Section 3.2). Finally, we address the sampling error introduced by the use of random samples, resulting in confidence regions for n-best precision graphs (Section 3.3) and statistical tests for the significance of performance differences between AMs (Section 3.4).

2 Evaluation

2.1 General Strategy

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. For instance, consider German preposition-noun-verb (PNV) triples, which we use to illustrate the evaluation procedure in Section 2.2. As most AMs are designed for word pairs, we interpret the PNV triples as PP-verb pairs, represented by the combination (P+N,V). [5] For each pair type (p+n,v), we classify the pair tokens (P+N,V) extracted from the corpus into a contingency table with four cells, obtaining the following frequency counts: [6]

$$\begin{array}{ll} O_{11} := f(P = p, N = n, V = v) & O_{12} := f(P = p, N = n, V \neq v) \\ O_{21} := f(P \neq p, N \neq n, V = v) & O_{22} := f(P \neq p, N \neq n, V \neq v) \end{array} \quad (1)$$

ming language and environment for statistical computing (cf. R Development Core Team, 2003).

[5] Note that this pairing (rather than e.g. (N, P+V)) is motivated both by the syntactic structure of the PNV triples and by the properties of support-verb constructions (*Funktionsverbgefüge*), where the verb typically indicates Aktionsart.

[6] The notation O_{ij} for the cell frequencies follows Evert (2004a). Note that we use upper-case letters (P,N,V) as variables for word *tokens* and lower-case letters (p,n,v) as variables for word *types*, again following Evert (2004a).

Step 2: Application of the association measures. AMs are applied to the frequency information collected in the contingency table. The result is a candidate list of pair types and their associated AM scores. For each individual AM, the candidate list is ordered from highest to lowest score. 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. There will often be ties in the rankings, which need to be resolved in some way in order to select exactly n candidates. For the evaluation experiments, we break ties randomly to avoid biasing the results (cf. page 10).

Step 3: Manual annotation. In order to assess the usefulness of each individual AM for collocation extraction, the (ranked) candidate lists are compared to a gold standard. The more true positives (TP) 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. [7] 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, Section 2.3. [8]

2.2 Evaluation Experiment: Data

For illustration of the proposed evaluation strategy, we consider PP-verb combinations extracted from an 8 million word portion of the Frankfurter Rundschau corpus. [9]

Step 1: Extraction of lexical tuples. Every PP, represented by the preposition P and the head noun N, is combined with every main verb V that occurs in the same sentence. For instance, the combination of P=*in*, N=*Frage*, and V=*stellen* occurs in 146 sentences. 80% of the resulting 294 534 PNV combina-

[7] 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 $r := t(n)/t$. Note that precision and recall are closely related: $p = rt/n$ (see also page 11).

[8] Precision-by-recall plots are the most intuitive mode of presentation (see Evert, 2004b, Sec. 5.1). However, they can be understood as mere coordinate transformations of the original precision plots, according to the equation $r = np/t$. It is thus justified to consider only precision plots in this paper.

[9] The Frankfurter Rundschau (FR) Corpus is a German newspaper corpus, comprising ca. 40 million words of text. It is part of the ECI Multilingual Corpus 1 distributed by ELSNET. ECI stands for European Corpus Initiative, and ELSNET for European Network in Language And Speech. See `http://www.elsnet.org/resources/ecicorpus.html` for details.

tions (lemmatized pair types) occur only once in the corpus ($f = 1$), another 15% occur twice ($f = 2$), and only 5% have occurrence frequencies $f \geq 3$. This illustrates the Zipf-like distribution of lexical tuples that was mentioned in Section 1. For the evaluation experiment, we use the 14 654 PNV types with $f \geq 3$ as candidates for lexical collocations. We refer to them as the PNV-FR data set throughout the article. [10] For each (p+n,v) pair type in the PNV-FR data set, the frequency counts for the cells $O_{11}, O_{12}, O_{21}, O_{22}$ of the contingency table are determined according to Eq. (1). In the example above, there are $O_{11} = 146$ cooccurrences of *in Frage stellen*, $O_{12} = 236$ combinations of *in Frage* with a different verb, $O_{21} = 3\,901$ combinations of *stellen* with a different PP, and the total number of pair tokens is $N = O_{11}+O_{12}+O_{21}+O_{22} = 406\,159$.

Step 2: Application of the association measures under investigation to the frequency information in the contingency tables. For the illustration experiment, the measures tested are two widely-used AMs – t-score (Church et al., 1991) and log-likelihood (Dunning, 1993) – as well as Pearson's chi-squared test (with Yates' correction applied) and plain cooccurrence frequency. The chi-squared test is considered as the standard test for association in contingency tables, but has not found widespread use in collocation extraction tasks (although it is mentioned by Manning and Schütze (1999)). Every AM assigns a specific value to each PNV type in the PNV-FR data set. Thus we obtain four different orderings of the candidate set.

Step 3: Manual annotation. In the semi-automatic extraction process, the candidate set is passed on to a human annotator for manual selection of the true collocations. For the purposes of an evaluation experiment, each candidate is marked as a true positive (TP) or false positive (FP). The PNV-FR data set has been annotated according to the guidelines of Krenn (2000). [11]

2.3 Evaluation Experiment: Results

Figure 1 displays precision graphs for n-best lists on the PNV-FR data set, ranked according to t-score, log-likelihood, chi-squared and frequency. The baseline of 6.41% is the proportion of collocations in the entire candidate set, i.e. the total number of TPs (939) divided by the total number of collocation candidates (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

[10] See Krenn (2000) and Evert and Krenn (2001) for a detailed description. Evert (2004b, Ch. 4) gives a theoretical justification for a frequency threshold of $f \geq 3$.

[11] Annotation of true collocations is a tricky task that requires expert annotators, especially as the borderline between collocations and non-collocational word combinations is often fuzzy. See Krenn et al. (2004) for a discussion of intercoder agreement on PP-verb collocations in the PNV-FR data set.

plot, as indicated by thin vertical lines for $n = 1\,000$, $n = 2\,000$ and $n = 5\,000$ (which are also shown in Figure 2). From the precision graphs we see that t-score clearly outperforms log-likelihood for $n \leq 6\,000$. Even simple frequency sorting is better than log-likelihood in the range $2\,000 \leq n \leq 6\,000$. Chi-squared achieves a poor performance on the PNV-FR 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, 1993). Figure 1 also 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. In our example, the differences between the AMs vanish for $n \geq 8\,000$ and larger lists are hardly useful for collocation extraction (all measures have reached a recall of approx. 80% for $n = 8\,000$).

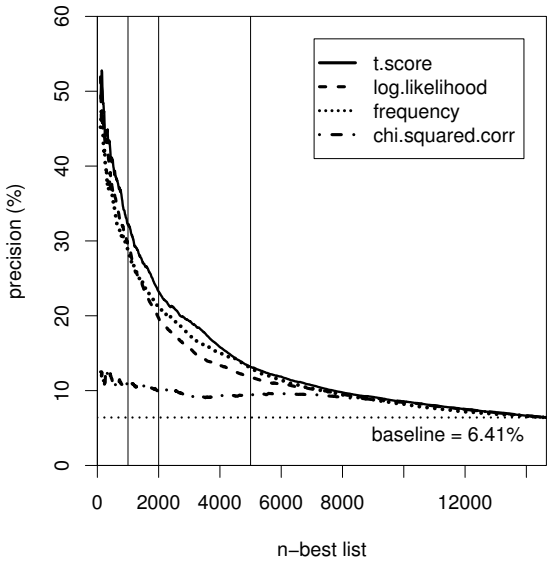

Fig. 1. Evaluation of association measures with precision graphs.

The data in Figure 1 clearly show that log-likelihood, despite its success in other evaluation studies and despite its wide-spread use, is not always the best choice. To make a reliable recommendation for an AM that is suitable for a particular purpose, an empirical evaluation has to be carried out under conditions that are as similar as possible to those of the intended use. The evaluation has to be repeated whenever a novel use case arises, because the performance of a particular AM cannot be predicted from the mathematical theory, and the evaluation results cannot be generalized to a substantially different extraction task.

In most cases, a manual annotation of true positives is necessary, although

some researchers have tried using existing dictionaries as a "gold standard" (e.g. Pearce, 2002). Since manual coding is a time-intensive (and often expensive) task, only a few large-scale evaluations have been carried out so far (Daille, 1994; Krenn, 2000; Evert et al., 2000). In addition, there are some small case studies such as Breidt (1993), Lezius (1999), and several articles where the usefulness of a newly-suggested AM is supported by a short list of extracted collocation candidates (e.g. Dunning, 1993). In order to cover a wide range of settings, a method is needed that reduces the required amount of manual annotation work drastically. This is achieved by annotating only a random sample selected from the full set of candidates, and estimating the true precision graphs from the sampled data. Especially for large-scale extraction tasks, it can also be useful to carry out a preliminary evaluation (based on a very small sample) on the data set that will be used for semi-automatic collocation extraction. We refer to this procedure as *tuning* of AMs.

In the remainder of this paper, we argue that RSE is an appropriate means (a) to predict the n-best precision of a given AM, and (b) to select the best-performing AM from two or more alternatives (typically a range of well-known and tested AMs such as log-likelihood, chi-squared, and t-score). In doing so, we establish RSE as a viable alternative to full evaluation, and we demonstrate its potential for AM tuning. Further research is necessary in order to determine whether the improvements achieved by tuning will outweigh the additional effort of the preliminary RSE step.

Concerning (a), the methods described in Section 3.3 yield a confidence interval for the true precision value, which gives a general indication of whether the results of the extraction procedure will be good enough for the intended use. For instance, lexicographers are interested in candidate lists that contain a fairly large amount of TPs, but the results need not be perfect. Thus it is important to know whether a certain AM can improve on the baseline precision: if the estimated precision is not substantially better than the baseline, there is little point in the application of statistical methods. The RSE estimates for different n-best lists and the corresponding confidence intervals can be combined into a precision graph similar to Figure 1. This graph can also help to determine an appropriate list size n, e.g. where the estimated precision drops below a useful threshold.

Concerning (b), it is obvious that, for a given list size n, the AM that achieves the highest n-best precision in the RSE should be used. However, any other AM that is not significantly different from the best measure may achieve equal or better precision on the full n-best list. Section 3.4 details the necessary significance tests. Significant differences between two AMs can then be marked in the precision graphs. It will rarely be possible to find an AM that is significantly better than its competitors for all n-best lists, though.

3 Random sample evaluation

3.1 Notation

Before describing the use of random samples for evaluation, we need to introduce a formal notation for the evaluation method described in Section 2. Let C be the set of candidates, and $S := |C|$ its size.[12] For the PNV-FR data set, we have $S = 14\,654$ and an example of an element $x \in C$ is the (p+n,v) pair type $x = (\textit{in+Frage}, \textit{stellen})$ representing the German collocation *in Frage stellen* "call into question". For each candidate pair x, an AM g computes a real number from the corresponding contingency table, called an association score. The actual values of the scores are rarely considered, though (see Footnote 3 on page 3 for an exception). Normally, only the ranking of the candidates according to the association scores is of importance. Since there is usually a substantial number of candidates whose contingency tables are identical (and candidates with different tables may occasionally obtain the same scores), the ranking will almost always contain ties. In order to determine n-best lists that include exactly the specified number of candidates (and are thus directly comparable between different measures), the ties need to be broken by random ordering of candidates with identical scores.[13]

Since the actual scores are normally discarded and ties are broken by random selection, we can represent an AM g by a ranking function $g : C \to \{1, \ldots, S\}$ (with respect to the candidate set C). This function assigns a unique number $g(x)$ to each candidate x, corresponding to its rank. An n-best list $C_{g,n}$ for the measure g contains all candidates x with rank $g(x) \leq n$, i.e.,

$$C_{g,n} := \{x \in C \mid g(x) \leq n\} \tag{2}$$

for $n \in \{1, \ldots, S\}$. By definition, $|C_{g,n}| = n$ (since all ties in the rankings have been resolved). Manual annotation of the candidates results in a set $T \subseteq C$ of true positives, which forms the basis of the evaluation procedure. The baseline precision b is the proportion of TPs in the entire candidate set: $b := |T| \,/\, |C|$. For any subset $A \subseteq C$, let $k(A) := |A \cap T|$ denote the number of TPs in A ($A \cap T$ is the set of TPs that belong to A). The true precision $p(A)$ of the set A is given by $p(A) := k(A)/\,|A|$, and the recall is given by $r(A) := k(A)/\,|T|$. We are mainly interested in the true precision of n-best lists (i.e. with $A = C_{g,n}$),

[12] $|C|$ stands for the cardinality of the set C, i.e. the number of candidates that it contains.

[13] A similar strategy, viz. randomization of hypothesis tests, is used in mathematical statistics for the study and comparison of hypothesis tests when the set of achievable p-values is highly discrete (see e.g. Lehmann, 1991, 71–72).

for which we use the shorthand notation $k_{g,n} := k(C_{g,n})$ and

$$p_{g,n} := p(C_{g,n}) = k_{g,n}/n \ . \tag{3}$$

Note that the baseline precision b can be obtained by setting $A = C$, i.e. $b = p(C)$. The plot in the left panel of Figure 2 displays the n-best precision $p_{g,n}$ achieved by four different AMs for n ranging from 100 to 5 000.[14] It is a "zoomed" version of the left third of the precision plot in Figure 1.

As was pointed out in Section 2, the main object of interest for the evaluation of AMs is the true n-best precision $p_{g,n}$. It is used to identify the best-performing measure g^* for given n and to compare its precision $p_{g^*,n}$ with the baseline b. Unless $p_{g^*,n}$ is significantly larger than b, there is no point in the application of AMs to rank the candidate set. Note that the n-best recall $r_{g,n}$ is fully determined by the corresponding precision $p_{g,n}$ and can be computed according to the formula $r_{g,n} = p_{g,n}n/bS$. Consequently, it does not provide any additional information for the evaluation, and neither does the F-score.[15]

Precision graphs visually combine the results obtained for many different n-best lists, but one has to keep in mind that they are mainly a presentational device. It is not the goal of the evaluation to find an AM that achieves optimal results for *all* possible n-best lists (i.e. whose precision graph is "above" all other graphs), and this will rarely be possible (cf. Figure 1).

3.2 Evaluation of a random sample

To achieve a substantial reduction in the amount of manual work, only a random sample $R \subseteq C$ is annotated. The ratio $|R| \,/\, |C|$ is called the sampling rate, and will usually be comparatively small (10% – 20%).[16] Since the manual annotation now identifies only those TPs which happen to belong to the sample R, i.e. the set $T \cap R$, it is necessary to estimate the full set T by statistical inference. As a first result, we obtain a maximum-likelihood estimate $\hat{b}$ for the baseline precision, which is given by the proportion of TPs in the random sample: $\hat{b} := |T \cap R| \,/\, |R|$. In the same manner, we can estimate the

[14] The four measures are g_1=t-score, g_2=log-likelihood, g_3=frequency-based ranking, and g_4=chi-squared. Precision values for $n < 100$ were omitted because of their large random fluctuations, which result in highly unstable graphs.

[15] The F-score is defined as the harmonic mean between precision and recall. It is often used for the evaluation of information retrieval tools, part-of-speech taggers, etc. in order to strike a balance in the tradeoff between high precision and high recall. In our application, however, this tradeoff is pre-empted by the choice of a specific list size n.

[16] Some remarks on how to choose the sampling rate can be found in Section 3.5.

true precision $p(A)$ of any subset $A \subseteq C$ by the ratio

$$\hat{p}(A) := \frac{|A \cap T \cap R|}{|A \cap R|} = \frac{\hat{k}(A)}{\hat{n}(A)}, \tag{4}$$

which is called the sample precision of A. We use the shorthand notation $\hat{n}(A)$ for the number of candidates sampled from A, and $\hat{k}(A)$ 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

$$\hat{p}_{g,n} := \hat{p}(C_{g,n}) = \frac{\hat{k}_{g,n}}{\hat{n}_{g,n}} \tag{5}$$

Note that the number $\hat{n}_{g,n}$ of annotated candidates in $C_{g,n}$ (which appears in the denominator of (5)) does not only depend on n (as in the definition of $p_{g,n}$, cf. (3)), but also on the particular choice of the random sample (the random sample picks a specified number of candidates from the full set C, but the number that falls into $C_{g,n}$ is subject to random variation). Consequently, $\hat{n}_{g_1,n}$ and $\hat{n}_{g_2,n}$ will usually be different for different measures g_1 and g_2. We return to this issue in Section 3.3.

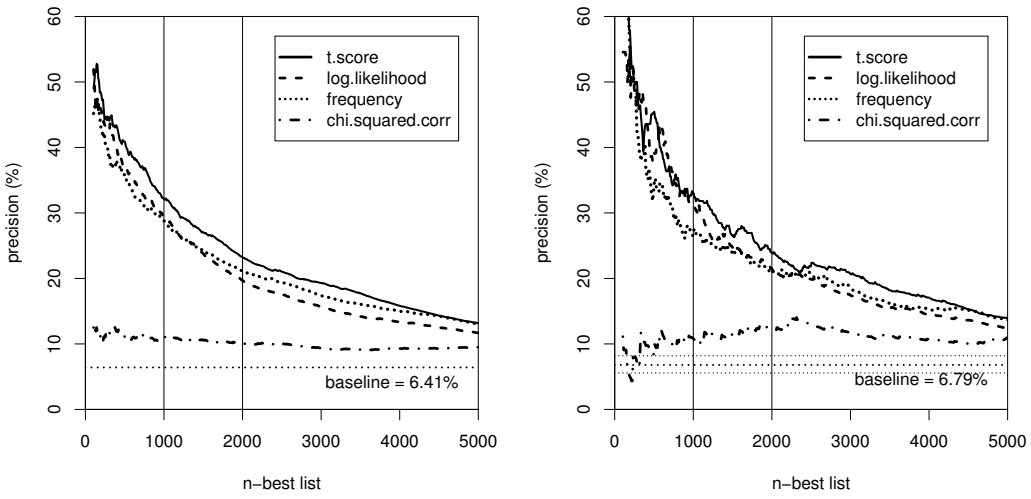

Fig. 2. An illustration of the use of random samples for evaluation: precision graphs for the full PNV-FR data set (left panel) and the corresponding estimates obtained from a 10% sample (right panel).

The right panel of Figure 2 shows graphs of $\hat{p}_{g,n}$ for $n \leq 5\,000$, estimated from a 10% sample of the PNV-FR data set. Note that the x-coordinate is n, not $\hat{n}_{g,n}$. The baseline shown in the plot is the estimate $\hat{b}$. The thin dotted lines above and below indicate a confidence interval for the true baseline precision (cf. Section 3.3). 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 $n \geq 4\,000$), 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 $n \leq 1000$ 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.

3.3 Confidence regions

In the interpretation of the RSE graphs, we use the sample precision $\hat{p}_{g,n}$ as an estimate for the true n-best precision $p_{g,n}$. Generally speaking, $\hat{p}(A)$ serves as an estimate for $p(A)$, for any set $A \subseteq C$ of candidates. The value $\hat{p}(A)$ will differ more or less from $p(A)$, depending on the particular sample R that was selected. The difference $\hat{p}(A) - p(A)$ is called the sampling error of $\hat{p}(A)$. We need to take this sampling error into account by constructing a confidence interval $\hat{\Pi}(A)$ for the true precision $p(A)$, as described e.g. by Lehmann (1991, 89ff). At the customary 95% confidence level, the risk that $p(A) \notin \hat{\Pi}(A)$ (because the selected sample R happens to contain a particularly large or small proportion of the TPs in A) is 5%. In order to define a confidence interval, we need to understand the relation between $p(A)$ and $\hat{p}(A)$, i.e., the sampling distribution of $\hat{p}(A)$. For notational simplification, we omit the parenthesized argument in the following discussion, writing $p := p(A)$, $\hat{p} := \hat{p}(A)$, $\hat{k} := \hat{k}(A)$, etc. In addition, we write $n := |A|$ for the total number of candidates in A.

The sample estimate $\hat{p}$ is based on $\hat{n}$ candidates that are randomly selected from the n candidates in A. In other words, $\hat{p}$ is a random variable whose sampling distribution depends on the true precision p, i.e. p is a parameter of the distribution. Our goal is to make inferences about the parameter p from the observed value of the random variable $\hat{p}$. However, $\hat{p} = \hat{k}/\hat{n}$ also depends on the number of candidates sampled, which is itself a random variable.

In contrast to $\hat{k}$ and $\hat{p}$, $\hat{n}$ is a so-called ancillary statistic, whose sampling distribution is independent from the parameter p.[17] Since the particular value of $\hat{n}$ does not provide any information about p, we will base our inference on the conditional distribution of $\hat{p}$ given the observed value of $\hat{n}$, i.e. on probabilities $P(\hat{p} \,|\, \hat{n})$ rather than $P(\hat{p})$. These conditional probabilities are equivalent to the probabilities $P(\hat{k} \,|\, \hat{n})$ because $\hat{k} = \hat{p} \cdot \hat{n}$. Assuming sampling with replacement,

[17] See Lehmann (1991, 542ff) for a formal definition of ancillary statistics and the merits of conditional inference.

we obtain a binomial distribution with success probability p, i.e.

$$P\Big(\hat{k} = j \;\Big|\; \hat{n}\Big) = \binom{\hat{n}}{j} p^j \, (1-p)^{\hat{n}-j}. \tag{6}$$

From (6), we can compute a confidence interval $\hat{\Pi}$ for the parameter p based on the observed values $\hat{k}$ and $\hat{n}$ (see Lehmann, 1991, 89ff). The size of this interval depends on the number $\hat{n}$ of candidates sampled and the required confidence in the estimate. Binomial confidence intervals can easily be computed with software packages for statistical analysis such as the freely available program R (R Development Core Team, 2003).

We have assumed sampling with replacement above in order to simplify the mathematical analysis, although $R \subseteq C$ is really a sample without replacement (since R is a subset which may not contain duplicates). For a sample without replacement, (6) would have to be replaced by a hypergeometric distribution with parameters k (the total number of TPs in A) and $n-k$ (the total number of TPs in $C \setminus A$). While binomial confidence intervals can be computed efficiently with standard tools, similar confidence sets for $p = k/n$ based on the hypergeometric distribution would require a computationally expensive custom implementation. The binomial distribution provides a good approximation of the hypergeometric, given that the sampling rate ($\hat{n}/n \approx |R| \,/\, |C|$) is sufficiently small. When one is worried about this issue, it is always possible to simulate sampling with replacement on the computer. The resulting sample is a multi-set R' in which some candidates may be repeated. In practice, each candidate will be presented to the human annotators only once, of course.

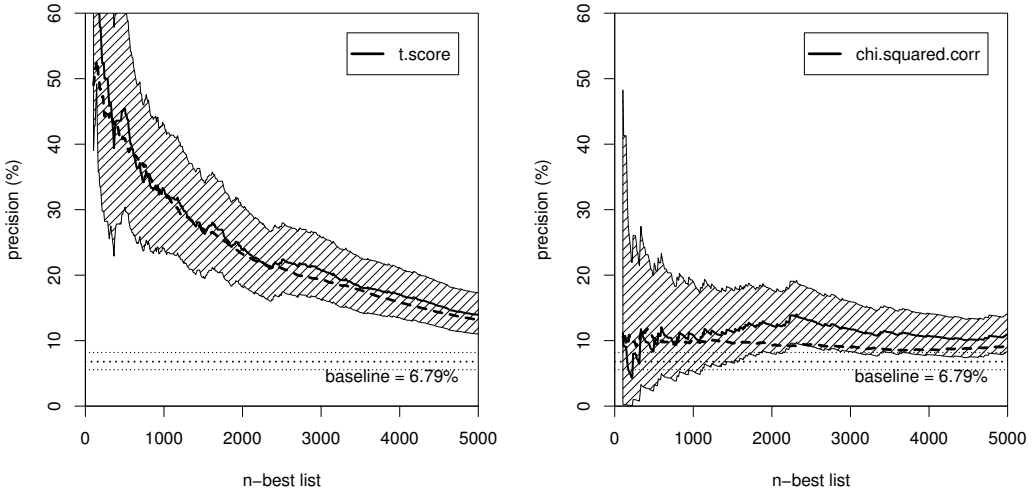

Fig. 3. Confidence intervals for the true precision $p_{g,n}$. The solid lines show the sample estimate $\hat{p}_{g,n}$, and the dashed lines show the true values of $p_{g,n}$ computed from the full candidate set.

Setting $A = C$, we obtain a confidence interval $\hat{\Pi}(C)$ for the baseline precision

b. This interval is indicated in the right panel of Figure 2 (and subsequent RSE graphs) by the thin dotted lines above and below the estimated baseline $\hat{b}$. Setting $A = C_{g,n}$, we obtain a confidence interval $\hat{\Pi}_{g,n} := \hat{\Pi}(C_{g,n})$ for the n-best precision $p_{g,n}$ of an AM g. Such confidence intervals are shown in Figure 3 as shaded regions around the sample-based precision graphs of t-score (left panel) and chi-squared (right panel). By the construction of $\hat{\Pi}_{g,n}$, we are fairly certain that $p_{g,n} \in \hat{\Pi}_{g,n}$ for most values of n, but we do not know where exactly in the interval the true precision lies. In other words, the confidence intervals represent our uncertainty about the true precision $p_{g,n}$. For instance, the RSE shows that t-score is substantially better than the baseline and reaches a precision of at least 20% for n-best lists with $n \leq 2000$. We can also be confident that the true precision is lower than 20% for $n \geq 4000$. However, any more specific conclusions may turn out to be spurious. For the chi-squared measure, we cannot even be sure that its performance is much better than the baseline, although $p_{g,n}$ may be as high as 20% for small n.

For comparison, the true n-best precision is indicated by a dashed line in both graphs. As predicted, it always lies within the confidence regions. For t-score, the difference between $p_{g,n}$ and $\hat{p}_{g,n}$ happens to be much smaller than the confidence intervals imply. On the other hand, the true n-best precision of chi-squared is close to the boundary of the confidence intervals for $n \geq 2000$. This example illustrates that the uncertainty inherent the sample estimates is in fact as large as indicated by the confidence intervals.

3.4 Comparison of association measures

The confidence intervals introduced in Section 3.3 allow us to assess the usefulness of individual AMs by estimating their n-best precision and comparing it with the baseline. However, the main goal of the evaluation procedure is the comparison of different AMs, in order to identify the best-performing measure for the task at hand. As we can see from the left panel of Figure 4, the confidence regions of the t-score and log-likelihood measures overlap almost completely. Taken at face value, this seems to suggest that the RSE does not provide significant evidence for the better performance of t-score on the PNV-FR data set. The true precision may well be the same for both measures. Writing g_1 for the t-score measure and g_2 for log-likelihood, the hypothesis $p_{g_1,n} = p_{g_2,n} =: p$ is consistent with both sample estimates ($\hat{p}_{g_1,n}$ and $\hat{p}_{g_2,n}$) for any value p in the region of overlap, i.e. $p \in \hat{\Pi}_{g_1,n} \cap \hat{\Pi}_{g_2,n}$.

This conclusion would indeed be correct if $\hat{p}_{g_1,n}$ and $\hat{p}_{g_2,n}$ were based on *independent* samples from $C_{g_1,n}$ and $C_{g_2,n}$. However, there is usually considerable overlap between the n-best lists of different measures (for instance, the 2 000-best lists of t-score and log-likelihood share 1 311 candidates). Both

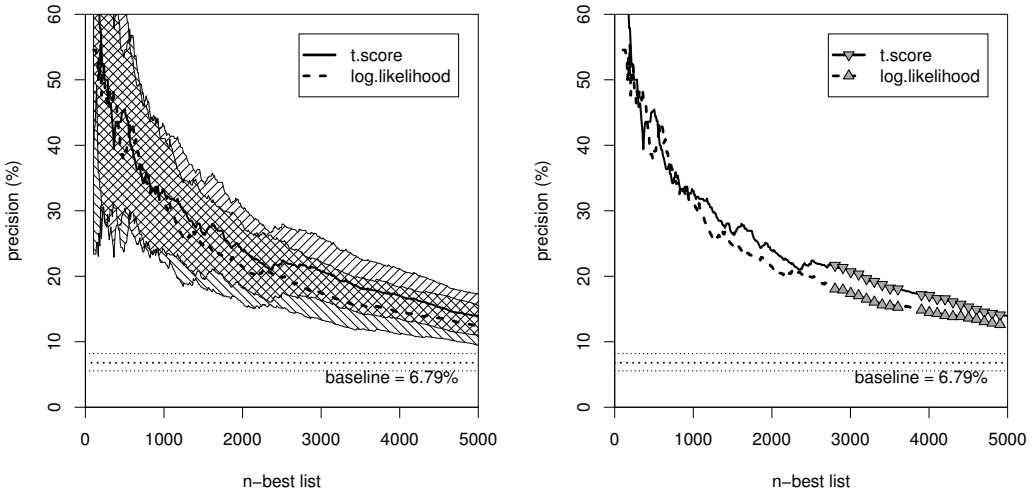

Fig. 4. Comparison of the t-score and log-likelihood measures.

samples select the same candidates from the intersection $C_{g_1,n} \cap C_{g_2,n}$ (namely, $C_{g_1,n} \cap C_{g_2,n} \cap R$), and will consequently find the same number of TPs. Any differences between $\hat{p}_{g_1,n}$ and $\hat{p}_{g_2,n}$ can therefore only arise from the difference sets $C_{g_1,n} \setminus C_{g_2,n} =: D_1$ and $C_{g_2,n} \setminus C_{g_1,n} =: D_2$.

Setting $A = D_i$ for $i = 1, 2$, it follows from the argument in Section 3.3 that the conditional probability $P(\hat{k}(D_i) \,|\, \hat{n}(D_i))$ has a binomial distribution (6) with success probability $p(D_i)$. Since $D_1 \cap D_2 = \emptyset$, the samples from D_1 and D_2 are independent, and so are the two distributions. Furthermore, $p_{g_1,n} > p_{g_2,n}$ iff $p(D_1) > p(D_2)$, and vice versa. Our goal for the comparison of two AMs is thus to find out whether the RSE provides significant evidence for $p(D_1) > p(D_2)$ or $p(D_1) < p(D_2)$. To do so, we have to carry out a two-sided hypothesis test with the null hypothesis $H_0 : p(D_1) = p(D_2)$. Since the sample sizes $\hat{n}(D_1)$ and $\hat{n}(D_2)$ may be extremely small (depending on the amount of overlap), asymptotic tests should not be used. [18] Exact inference for two independent binomial distributions is possible with Fisher's exact test (Fisher, 1970, 96f (§21.02)), which is applied to the following contingency table:

$\hat{k}(D_1)$	$\hat{k}(D_2)$
$\hat{n}(D_1) - \hat{k}(D_1)$	$\hat{n}(D_2) - \hat{k}(D_2)$

Implementations of Fisher's test are available in most statistical software packages, including R. In the right panel of Figure 4, the grey triangles indicate n-best lists where the RSE provides significant evidence that the true precision of t-score is higher than that of log-likelihood (according to a two-sided

[18] A standard test for equal success probabilities of two independent binomial distributions is Pearson's chi-squared test. This application of the test should not be confused with its use as an association measure.

Fisher's test at a 95% confidence level). Despite the enormous overlap between the confidence intervals, the observed differences are (almost) always significant for $n \geq 3000$.

3.5 A second example and some final remarks

Figure 5 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 by Evert and Kermes (2003), and a frequency threshold of $f \geq 20$ was applied. From the resulting data set of 8 546 candidates, a 15% sample was manually annotated by professional lexicographers (henceforth called the AN-FR data set).[19] In contrast to the PNV-FR data, which uses a linguistically motivated definition of collocations, the annotators of the AN-FR 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 an existing dictionary had been used as a gold standard. The results of this 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 $1\,000 \leq n \leq 3\,000$, the precision of chi-squared is *significantly* better than that of log-likelihood.

Summing up, the evaluation examples for the PNV-FR and AN-FR data sets clearly show that the usefulness of individual AMs for collocation extraction has to be determined by empirical evaluation under the specific conditions of the intended use case. Results obtained in a particular setting cannot be generalized to different settings, and theoretical predictions (such as Dunning's) are often not borne out in reality. The RSE approach helps to reduce the amount of work required for the manual annotation of true positives, making evaluation experiments such as the adjective-noun example above possible.

One question that remains is the choice of a suitable sampling rate, which determines the reliability of the RSE results, as given by the width of the binomial confidence intervals $\hat{\Pi}_{g,n}$ for the true n-best precision (Section 3.3). Interestingly, this width does not depend on the sampling rate, but only on the total number $\hat{n}_{g,n}$ of candidates sampled from a given n-best list (and on the

[19] We would like to thank the Wörterbuchredaktion of the publishing house Langenscheidt KG, Munich for annotating this sample. The evaluation reported here emerged from a collaboration within the project TFB-32, funded at the University of Stuttgart by the DFG.

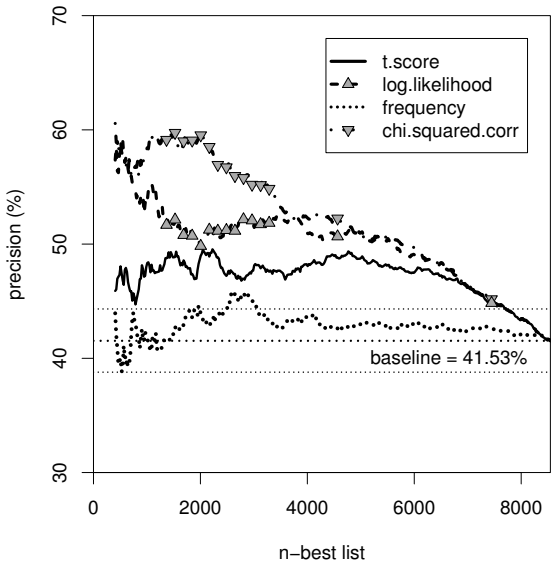

Fig. 5. RSE of German adjective+noun combinations.

observed precision $\hat{p}_{g,n}$). Thus, a 20% sample from a 500-best list achieves the same reliability as a 5% sample from a 2000-best list (since $\hat{n}_{g,n} \approx 100$ in either case). The RSE procedure can therefore also be applied to large n-best lists, provided that they achieve sufficiently high precision. [20] The precise width of the confidence intervals can be predicted with the help of a binomial confidence interval chart (e.g. Porkess, 1991, 47, s.v. *confidence interval*). Unfortunately, it is much more difficult to predict the sampling rate that is necessary for differences between AMs to become significant (Section 3.4). The power of Fisher's test depends crucially on the amount of overlap between the two measures being compared, i.e. on the number of candidates sampled from the difference regions, $\hat{n}(D_1)$ and $\hat{n}(D_2)$. In addition, power calculations for Fisher's test are much more complex than in the binomial case.

4 Conclusion

With the random sample evaluation (RSE) we have presented a procedure that makes the evaluation of association measures (AMs) for a specific type of collocation and for a specific kind of extraction corpus practically feasible. In this way, an appropriate AM can be selected depending on the application setting, which would otherwise not be possible because the results of an evaluation experiment cannot easily be generalized to a different situation. Based

[20] As a rule of thumb, estimates from small samples ($\hat{n} \approx 100$) are of little use when the observed precision $\hat{p}$ drops below 20%. Larger samples ($\hat{n} \approx 500$) extend the useful range down to $\hat{p} \approx 10\%$.

on a data set of German PP-verb combinations, we have shown that the RSE allows us to estimate the precision achieved by individual AMs in this particular application. Using the RSE procedure to evaluate the same AMs on a second data set of German adjective-noun combinations, we have collected further evidence that the evaluation of AMs for collocation extraction is a truly empirical task, obtaining results that contradict both widely-accepted theoretical arguments and the results of previous evaluation experiments. In the light of these findings, the RSE is indispensable as it allows researchers and professional users alike to carry out many more evaluation experiments by reducing the amount of manual annotation work that is required. Our findings also demonstrate the potential for tuning AMs to a specific collocation extraction task, based on the manual annotation of a very small sample from the extracted data set. The RSE procedure for the evaluation of AMs is implemented as an R library in the UCS toolkit, which can be downloaded from `http://www.collocations.de/`. All precision graphs in this paper (including confidence intervals and significance tests) were produced with the UCS implementation.

Acknowledgements

Suggestions by several anonymous reviewers and by Alexandra Klein from ÖFAI have helped make this article much more understandable than it might have been. We would also like to thank the Wörterbuchredaktion of the publishing house Langenscheidt KG, Munich for the manual inspection of German collocation candidates. 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.

References

Blaheta, D., Johnson, M., July 2001. Unsupervised learning of multi-word verbs. In: Proceedings of the ACL Workshop on Collocations. Toulouse, France, pp. 54–60.

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`).

Choueka, Y., 1988. Looking for needles in a haystack. In: Proceedings of RIAO '88. pp. 609–623.

Church, K., Gale, W. A., 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.

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

da Silva, J. F., Lopes, G. P., July 1999. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In: 6th Meeting on the Mathematics of Language. Orlando, FL, pp. 369–381.

Daille, B., 1994. Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. Ph.D. thesis, Université Paris 7.

Dias, G., Guilloré, S., Lopes, J. G. P., 1999. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In: Proceedings of Traitement Automatique des Langues Naturelles (TALN). Cargèse, France.

Dunning, T. E., 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. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, to appear.

Evert, S., Heid, U., Lezius, W., 2000. Methoden zum Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In: Zühlke, W., Schukat-Talamazzini, E. G. (Eds.), KONVENS-2000 Sprachkommunikation. VDE-Verlag, pp. 215 – 220.

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.

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.

Firth, J. R., 1957. A synopsis of linguistic theory 1930–55. In: Studies in linguistic analysis. The Philological Society, Oxford, pp. 1–32.

Fisher, R. A., 1970. Statistical Methods for Research Workers, 14th Edition. Oliver & Boyd, Edinburgh.

Goldman, J.-P., Nerima, L., Wehrli, E., July 2001. Collocation extraction using a syntactic parser. In: Proceedings of the ACL Workshop on Collocations. Toulouse, France, pp. 61–66.

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., Evert, S., July 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In: Proceedings of the ACL Workshop

on Collocations. Toulouse, France, pp. 39–46.

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

Lehmann, E. L., 1991. Testing Statistical Hypotheses, 2nd Edition. Wadsworth.

Lezius, W., 1999. Automatische Extrahierung idiomatischer Bigramme aus Textkorpora. In: Tagungsband des 34. Linguistischen Kolloquiums. Germersheim, Germany.

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

Pearce, D., 2002. A comparative evaluation of collocation extraction techniques. In: Third International Conference on Language Resources and Evaluation (LREC). Las Palmas, Spain, pp. 1530–1536.

Porkess, R., 1991. The HarperCollins Dictionary of Statistics. HarperCollins, New York.

R Development Core Team, 2003. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3. See also `http://www.r-project.org/`.

Sinclair, J., 1991. Corpus, Concordance, Collocation. Oxford University Press, Oxford.

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

Wilks, Y., Catizone, R., 2002. What is lexical tuning? Journal of Semantics 19, 167–190.