A Lexicographic Evaluation of German Adjective-Noun Collocations

Stefan Evert

Institute of Cognitive Science, University of Osnabrück
49069 Osnabrück, Germany
stefan.evert@uos.de

Abstract

This paper describes a small database of 1,252 German adjective-noun combinations, which have been annotated by professional lexicographers with respect to their collocational status and their usefulness for the compilation of a bilingual dictionary. The database is a random sample taken from the most frequent ($f \geq 20$) adjective-noun pairs in a standard newspaper corpus (*Frankfurter Rundschau*). It is particularly useful for the evaluation and development of ranking techniques for multiword candidates. Suitable corpus frequency data (instances of adjective-noun cooccurrences from the same corpus) are made available together with the database.

1. Introduction and background

The work presented here was motivated by two comparative studies that evaluated the usefulness of different association measures for the identification of German adjective-noun collocations (Lezius, 1999; Evert et al., 2000).[1] Both studies seemed to confirm results from previous comparative evaluations carried out for other languages and other types of collocations, e.g. Daille (1994) and Krenn (2000). In particular, the following observations were made:

1. The most useful measure for collocation identification is *log-likelihood* (Dunning, 1993), justifying its well-established role as a default association measure in computational linguistics.
2. Log-likelihood is significantly better than the *chi-squared* measure (even if Yates' continuity correction is applied), as has been claimed by Dunning (1993).
3. A simple ranking of candidates by their cooccurrence *frequency* achieves surprisingly good results, although precision is significantly lower than for log-likelihood.
4. Contrary to the claims of Church and Hanks (1990), *Mutual Information* (MI) is very poorly suited for collocation identification.
5. Many other association measures (including *t-score*) are very close to log-likelihood, but none of them achieves significantly better results for any n-best list of candidates. This observation led to the hypothesis that log-likelihood represents an upper limit for collocation identification methods based on cooccurrence frequency data (the "sonic barrier" hypothesis).

However, both studies also had considerable shortcomings, so that these findings have to be qualified. The most serious problems, which motivated the follow-up study described in this paper, are the following:

- Lezius (1999) only looked at short 100-best lists of candidates, and many of the observed differences are not significant.[2] It is also not clear whether the results can be generalised to practically relevant 1000-best or 2000-best lists.
- Evert et al. (2000) aimed at a complete manual annotation of all recurrent adjective-noun combinations (with $f \geq 2$) in a given corpus, so that recall and baseline precision can be computed. For practical reasons, they chose an unrealistically small corpus of German law texts (approx. 800,000 running words). Real-life applications are likely to use much larger corpora and higher frequency thresholds, which may favour association measures like chi-squared and MI that are oversensitive to low-frequency data.
- Both studies failed to give a precise definition of collocations and did not supply clear guidelines to annotators. As a consequence, inter-annotator agreement was very low (though not reported in the original publications) and it was often impossible to resolve differences by discussion. This raises considerable doubt as to which aspects of the interplay between collocativity and statistical association have been evaluated, and whether a comparison with other studies is meaningful at all.

For these reasons, a follow-up study was designed in order to verify the findings of Lezius (1999) and Evert et al. (2000). The new study was based on a 40-million-word newspaper corpus (*Frankfurter Rundschau*), and candidate collocations were examined by professional lexicographers. This approach ensures a consistent and practically relevant definition of collocations and enables a direct comparison with other studies based on lexicographic (Smadja, 1993) or terminological (Daille, 1994) expert judgements.
The remainder of this paper is organised as follows. Section 2. summarises the initial results of this follow-up study, which have not been published before and provide an important reference point for future experiments with the database. The new adjective-noun database is described in Section 3., while Section 4. documents the file format and availability of the resource.

[1] Following the terminology of the cited studies, we understand *collocations* as a fuzzy concept that encompasses lexicalised, partly lexicalised and other "habitual" word combinations. It is similar in meaning to the current usage of the term *multiword expressions*, but may also include conventionalised word combinations even if they do not show the typical linguistic hallmarks of lexicalisation, i.e. non-compositionality, non-substitutability and non-modifiability (Manning and Schütze, 1999, 184).

[2] In the original study, no significance tests were carried out.

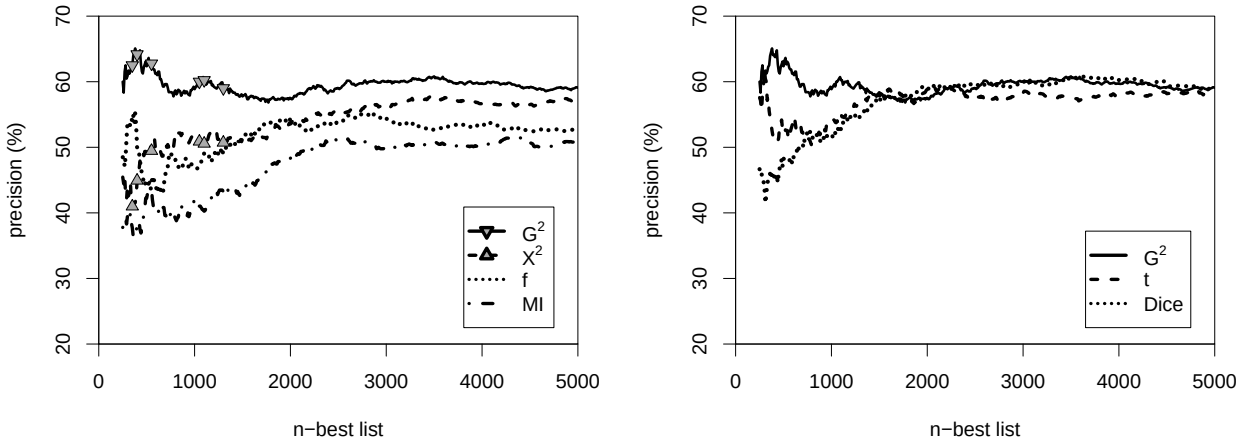

Figure 1: Evaluation results for the identification of adjective-noun collocations for lexicographic purposes. True positives are all word pairs that are considered useful for the compilation of a bilingual dictionary. The following association measures have been evaluated: log-likelihood (G^2), chi-squared with Yates' correction (X^2), Mutual Information (MI), t-score (t), Dice coefficient (Dice) and frequency ranking (f). Grey triangles indicate significant differences between log-likelihood and chi-squared ($\alpha = .05$).

2. The original experiment

For the follow-up experiment, German adjective-noun combinations were extracted from the *Frankfurter Rundschau* corpus (a detailed description of the corpus and extraction procedure is given in Section 3.). After application of a frequency threshold ($f \geq 5$), 5,000-best lists of collocation candidates were prepared according to 7 standard association measures. These measures were selected in order to verify observations made by previous studies. They include log-likelihood, chi-squared, t-score, MI, as well as the Dice coefficient. See Evert (2004, Ch. 3) or `http://www.collocations.de/` for full descriptions of all relevant association measures.

The ranked collocation candidates were manually evaluated by professional lexicographers with respect to their usefulness for the compilation of a bilingual (German-English) dictionary. Since annotation of all 13,533 candidates in the pooled n-best lists would have been prohibitively time-consuming, evaluation was based on a 15% random sample, using the RSE methodology of Evert and Krenn (2005).

The initial results were in accordance with previous studies, as the precision graphs in Figure 1 show (see Evert and Krenn (2005) or Evert (2004) for a detailed explanation of such evaluation graphs). Log-likelihood is the best association measure for this task (left panel). It is significantly better than chi-squared, at least for n-best lists up to $n = 1500$. Frequency ranking performs surprisingly well, but has significantly lower precision than log-likelihood, and MI is worse than frequency ranking. The right panel shows a clear "sonic barrier" effect: for $n \geq 1500$, log-likelihood, t-score and Dice have virtually indistinguishable performance, despite their entirely different mathematical properties.

In summary, this experiment seemed to confirm the results of Lezius (1999) and Evert et al. (2000). The only surprising observation was that log-likelihood achieves almost constant precision ($\approx 60\%$) for all n-best lists. While it is apparently very useful for selecting a large set of 5,000 promising candidates (on par with t-score and Dice), it does not seem to be able to make any further distinctions between these candidates. At the time, this was interpreted as supporting evidence for the "sonic barrier" hypothesis.

One possible explanation for the nearly constant precision of log-likelihood was the fact that the evaluation criterion of "usefulness for dictionary compilation" mixes entirely different types of collocations, ranging from non-compositional multiword expressions to regularly formed, but frequent combinations (which might provide good material for usage examples in the dictionary). In a second evaluation, true positives were therefore restricted to "true" collocations, which are at least partly lexicalised and need to be listed in the dictionary (if only for contrastive reasons). The results were entirely surprising, as the precision curves in Figure 2 show.

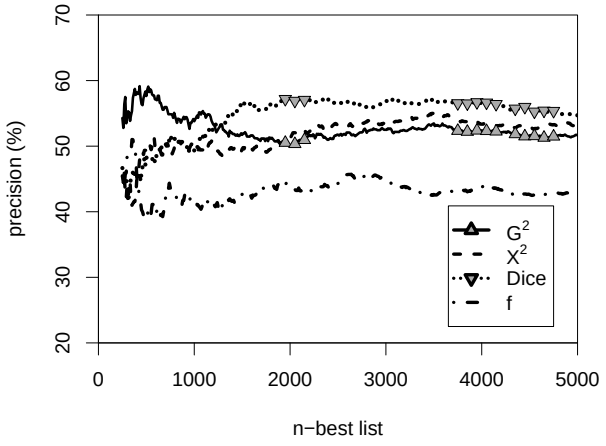

Figure 2: Evaluation results for the identification of "true" adjective-noun collocations, which need to be listed in a bilingual (German-English) dictionary. Grey triangles indicate significant differences between Dice and log-likelihood ($\alpha = .05$).

The precision achieved by log-likelihood is somewhat lower than before, but still almost constant across all n-best lists. Chi-squared is less affected by the modified evaluation criterion and is even slightly better than log-likelihood for $n \geq 2000$, contradicting the argument of Dunning (1993). Most unexpectedly, however, the Dice coefficient (which has never figured prominently as an association measure) obtains significantly higher precision than log-likelihood. With this experiment, the "sonic barrier" hypothesis was falsified: there is indeed room for improvement over log-likelihood.

The choice of association measures for the lexicographic evaluation had been based on the literature on collocation extraction. It seemed quite plausible that other, previously neglected association measures might give even better results than Dice. In order to support experiments with a wide range of different association measures, the manually annotated database was extended to cover (a random sample of) all frequent adjective-noun combinations ($f \geq 20$) in the *Frankfurter Rundschau* corpus. Since the full data set would be biased towards the measures considered in the original experiment, only this high-frequency subset has been publically released and is described in the following sections.

3. Data preparation and manual annotation

The German adjective-noun database (codenamed `La11t`) has been derived from the *Frankfurter Rundschau* corpus, containing approx. 40 million tokens (words and punctuation) of text from German newspaper articles published in the years 1992–1993.[3] The corpus was part-of-speech tagged with TreeTagger (Schmid, 1995) and lemmatised with IMSLex (Lezius et al., 2000). Adjective-noun combinations – consisting of the head of a noun phrase and a prenominal modifying adjective – were extracted using the part-of-speech patterns described and evaluated by Evert and Kermes (2003). Only 8,546 adjective-noun pairs with cooccurrence frequency $f \geq 20$ were retained as collocation candidates.

A random sample of 1,252 candidates ($\approx 15\%$) was manually annotated by four professional lexicographers of Langenscheidt KG, Munich. The main criterion was usefulness for the compilation of a large bilingual (German-English) dictionary, but finer distinctions were also made by the annotators. Each candidate was classified into one of the following 6 categories:

1. *True collocations:* these candidates are at least partly lexicalised and need to be listed in a dictionary. They can be equated with the notion of multiword expression in computational linguistics. (Ex.: *autofreie Zone* 'zone in which no cars are allowed', *böses Blut* 'bad blood', *das gelbe Trikot* 'the yellow jersey')

2. *Habitual combinations:* these candidates have some idiosyncratic properties (often semi-compositional), but usually allow limited substitution of components with semantically related words. Only some items from such a series need to be listed in the dictionary. Habitual combinations fall into the grey area between multiword expressions and free combinations. (Ex.: *brütende Hitze* 'stifling heat', *neuer Anlauf* 'another go', *technische Daten*, 'technical specification')

3. *Familiar combinations:* mostly free, but frequent combinations without contrastive relevance. They often provide good examples to illustrate the usage of a headword. (Ex.: *ehemaliger Schüler* 'former pupil', *günstiges Angebot* 'bargain, good offer', *unbekanntes Ziel* 'unknown destination')

4. *Candidates with unclear status:* these items may assist lexicographers in the compilation process, but are probably not directly relevant for a bilingual dictionary (Ex.: *neuer Meister* 'new champion', *übrige Zeit* 'remaining time')

5. *Non-collocational:* recurrent combinations that are clearly not relevant for a bilingual dictionary, although they might help lexicographers and translators understand the usage of a headword. (Ex.: *Deutsche Bundesbank* 'Central Bank of Germany', *erstes Semester* 'first term at university', *heißer Sommer* 'hot summer')

6. *Trash:* mostly tagging and lemmatisation errors, as well as some combinations that are idiosyncratic for the corpus used. (Ex.: *[unter] anderem Werke [von]*: adverbial misinterpreted as adjective, *Höchster Stadtpark*: district *Höchst* misinterpreted as superlative of adjective *hoch* 'high', *[Die] verliebte Wolke* 'cloud in love': name of a stage play)

For each candidate, the annotators were given up to 10 randomly selected corpus examples. Due to time constraints, an evaluation of inter-annotator agreement could not be carried out, but the four lexicographers discussed all decisions among themselves. In some cases, lemmatisation errors or incomplete extraction of a larger multiword expression were considered as true positives if the correct form could easily be reconstructed from the corpus examples. Table 1 shows the number and percentage of candidates for each of the six categories. The baseline precision of the entire database ranges from 29.3% (if only true collocations in category 1 are accepted as true positives) to 50.9% (if all useful candidates in categories 1–3 are accepted).

1	2	3	4	5	6
367	153	117	45	537	33
29.3%	12.2%	9.4%	3.6%	42.9%	2.6%

Table 1: Number of candidates and corresponding percentage for each annotation category in the `La11t` database.

4. Availability and use

The `La11t` database is made available as a TAB-delimited text file with a single header row specifying variable names

[3] The *Frankfurter Rundschau* corpus is part of the ECI Multilingual Corpus I distributed by ELSNET. See `http://www.elsnet.org/eci.html` for more information and licensing conditions.

for the columns. This is the native format of the UCS toolkit (Evert, 2004); it also works well with statistical software such as R (R Development Core Team, 2008) and spreadsheet programs like Microsoft Excel. The table columns are:

1. `l1` = adjective (lemma)
2. `l2` = noun (lemma)
3. `n.cat` = collocational status (category assigned by lexicographers, cf. Section 3.)

Since the German words contain non-ASCII characters, versions in Unicode (UTF-8) and Latin1 (ISO-8859-1) encoding are provided. The database can be downloaded from the Resources section of `http://multiword.sf.net/`. It may be used freely for academic research and all non-commercial purposes under the terms of the Creative Commons Attribution-Noncommercial (CC-BY-NC) license, version 3.0 unported.

The `La11t` database is primarily useful for the evaluation of association measures and other ranking methods for collocation and multiword candidates. It also supports the optimisation of association measures with machine learning techniques, which can either take the form of a two-way classification task (with true positives belonging to category 1, categories 1–2, or categories 1–3) or of a multi-way classification task that distinguishes between all six categories. As a simplified problem, three-way classification into category groups 1, 2–4 and 5–6 is suggested.

In order to facilitate such experiments, cooccurrence frequency data from the same *Frankfurter Rundschau* corpus are provided together with the database in two formats: (a) a list of *cooccurrence tokens* with adjective and noun lemma, partially disambiguated morphosyntactic information, and the surface realisation of the expression; and (b) a table of *pair types* with their frequency signatures[4] in the UCS data set format (Evert, 2004). It has to be noted that these resources contain no data for some of the adjective-noun candidates, or indicate a cooccurrence frequency far below the threshold of $f \geq 20$. The reason is that the frequency data were obtained from a re-annotated version of the corpus in which some tagging and lemmatisation errors have been corrected (by using improved releases of the tagger and morphology).

5. Acknowledgements

Our thanks are due to the lexicographers at the Redaktion Wörterbücher, Langenscheidt KG, Munich for their enthusiastic support of this project, the annotation work they carried out and their willingness to release the data without further restrictions.

[4] A frequency signature consists of the cooccurrence frequency of a pair type, the marginal frequencies of its two components, and the sample size. It provides the same information as a 2×2 contingency table, and automatic translation between the two data structures is possible.

6. References

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

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

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

Stefan Evert and Hannah Kermes. 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*, pages 83–86.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.

Stefan Evert, Ulrich Heid, and Wolfgang Lezius. 2000. Methoden zum Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In Werner Zühlke and Ernst G. Schukat-Talamazzini, editors, *KONVENS-2000 Sprachkommunikation*, pages 215 – 220. VDE-Verlag.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. See `http://www.collocations.de/` for software and data sets.

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

Wolfgang Lezius, Stefanie Dipper, and Arne Fitschen. 2000. IMSLex – representing morphological and syntactical information in a relational database. In Ulrich Heid, Stefan Evert, Egbert Lehmann, and Christian Rohrer, editors, *Proceedings of the 9th EURALEX International Congress*, pages 133–139, Stuttgart, Germany.

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

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

R Development Core Team, 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. See also `http://www.r-project.org/`.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, March.

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