# The influence of linguistic pre-processing on candidate data

**Stefan Evert** and **Hannah Kermes**

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Azenbergstr.12
D-70174 Stuttgart, Germany
Stefan.Evert@ims.uni-stuttgart.de

**Abstract**

This paper describes ongoing work on the evaluation of methods for extracting collocation candidates from large text corpora. Our research is based on a German treebank corpus used as gold standard. We present first results for adjective+noun collocations, and plan to extend our work to other types of collocations (e.g. PP+verb pairs).

## 1 Introduction

The extraction of collocations from text corpora is usually performed in a three-stage process (cf. Krenn (2000)):

1. The source corpus is annotated with varying amounts of linguistic information (ranging from part-of-speech tags to full parse trees), depending on the tools available. Then a list of candidates is extracted (often based on part-of-speech patterns). In most cases, candidates are restricted to a grammatically homogeneous set, e.g. adjective+noun pairs where the adjective is a modifier of the noun. This first list will contain both collocational and non-collocational candidates.

2. Linguistic and/or heuristic filters may be applied to reduce the size of the candidate set. For instance, certain "generic" adjectives as well as those derived from verb participles are rarely found in collocations.

3. The remaining candidates are ranked by statistical measures based on their frequency "profiles". Typically, candidates are considered likely to be collocations if their components co-occur much more frequently than expected by chance.

Comparative evaluations such as Krenn (2000) and Evert and Krenn (2001) concentrate on the quality of the statistical measures and the corresponding ranking of the candidates, and to a lesser extent on the performance of linguistic filters. Although Evert and Krenn (2001) are aware of the influence that the first extraction step has on their results, they fail to give a quantitative evaluation of different preprocessing and extraction methods.

Our research aims to fill this gap. Currently, we are evaluating methods for the extraction of adjective+noun pairs from German newspaper text. It is planned to extend our work to other types of collocations, including PP+verb and noun+verb pairs.

## 2 Evaluation procedure

The statistical ranking of candidates is usually evaluated against a manually annotated list of true positives (cf. Krenn

(2000) and Evert *et al.* (2000)). This approach is of little use for the evaluation of candidate extraction methods, though, for several reasons:

- The accuracy of the extraction step influences the final results in two quite different ways: (a) by changing the set of candidates; (b) by changing the frequency profiles of candidates.
- The influence of changes in the frequency profiles depends crucially on which statistical measure is applied in the third stage.
- In many cases, different extraction methods will produce only minor changes in the set of candidates, especially when frequency thresholds are applied. These subtle effects will be masked by the much greater impact of the statisical ranking.
- Simple, window-based extraction methods may find many spurious candidates. Even though some of those might be true positives per se[1], they are not a part of the source corpus and thus should not be included in the list of candidates.

Hence, it is necessary to evaluate the extraction step independently, and to find an appropriate definition of the expected goal of the first processing stage, i.e. what results should ideally be produced.

Clearly, one cannot expect the extraction step to distinguish collocations from non-collocations without access to frequency information. The frequency profiles of candidates should accurately report the number of co-occurrences in the source corpus, and spurious matches should be avoided. This leads to the following goal definition:

> Find all instances of lexeme pairs (or $n$-tuples) that occur in a specific syntactic relationship in the source corpus.

As a consequence, our evaluation is based on instances of candidate pairs, i.e. *tokens* rather than *types*. In our terminology, a *pair type* is a combination of two lexemes, and the corresponding *pair tokens* are individual occurrences of this pair at specific positions in the corpus. Statistical ranking methods are usually applied to and evaluated on pair types.

The experiments reported here investigate the extraction of German adjective+noun pairs, where the noun is the head of a noun phrase (NP) and the adjective appears as a modifier in the NP.

## 3 A gold standard

It is theoretically easy to obtain a gold standard for our evaluation, since the purely syntactic relationships that have to be annotated are less ambiguous than the distinction between collocations and non-collocational candidates. However, the annotation of *tokens* rather than *types* is a prohibitively laborious task.

Fortunately, a German treebank corpus is available, from which the gold standard data can be extracted by automatic means. The Negra corpus (Skut *et al.*, 1998)[2] consists of 355 096 tokens of German newspaper text with manually corrected part-of-speech tagging, morpho-syntactical annotations, and parse trees.

We used XSLT stylesheets to extract a reference list of 19 771 instances of adjective+noun pairs from a version of the Negra corpus encoded in the TigerXML format (Mengel and Lezius, 2000). Unfortunately, the syntactic annotation scheme of the Negra treebank (Skut *et al.*, 1997), which omits all projections that are not

[1] i.e. they would be accepted as collocations by a human annotator

[2] see also `http://www.coli.uni-sb.de/sfb378/negra-corpus/`

strictly necessary to determine the constituent structure of a sentence, is not very well suited for automatic extraction tasks.

So far, we have only been able to extract adjective+noun pairs. We plan to use the TIGERSearch tool[3] in combination with XSLT stylesheets to obtain gold standard data for PP+verb and noun+verb pairs.

## 4 Pre-processing and extraction methods

In addition to the hand-corrected part-of-speech tags in the Negra corpus, we used the IMS TreeTagger (Schmid, 1994) for automatic tagging. With its standard training corpus, a tagging accuracy of 94.82% was achieved. A substantial part of the errors are due to proper nouns missing from the tagger lexicon.

In the next step YAC, a recursive symbolic chunk parser (Kermes and Evert, 2001), was applied to identify adjective phrases (APs), noun phrases (NPs), and prepositional phrases (PPs). An independent evaluation of YAC against NPs extracted from Negra is reported in Kermes and Evert (2002). Due to idiosyncrasies of the Negra annotations, we do not have final results yet. The latest values show a precision of $P = 84.66\%$ and a recall of $R = 87.25\%$ based on perfect tagging. With automatic tagging, $P = 77.66\%$ and $R = 81.90\%$ are achieved.

YAC was specifically designed for automatic extraction, so all AP projections are explicit, and both APs and NPs are annotated with head lemmas, which simplifies the extraction of candidates with XSLT stylesheets tremendously. We created two versions of the chunk annotations, based on the hand-corrected and the automatic tagging, respectively.

Finally, we used three different extraction methods to identify candidate pairs: (a) adjacent adjectives and nouns (based on part-of-speech tagging); (b) adjectives preceding nouns within a given window (we allowed a maximal distance of 10 tokens); (c) the lexical heads of APs and NPs in the chunk annotations, where the AP node is a child of the NP node.[4]

We have evaluated all six combinations of pre-processing and extraction methods. In further experiments, we plan to study the quantitative effects of linguistic filters (excluding adjectives derived from verb participles and/or proper nouns) and lemmatisation (wrt. candidate *types*).

## 5 First results

The reference data extracted from Negra comprises 19 771 instances of adjective+noun pairs. The numbers for automatic extraction range from 17 694 (adjacent pairs based on automatic tagging) to 19 726 (YAC chunks on hand-corrected tagging).

Table 5 lists *precision*[5] and *recall*[6] for all combinations of pre-processing and extraction methods. On the hand-corrected tagging, adjacent pairs yield the highest precision, but recall is much better for extraction from windows or YAC chunks.

The 5% error rate of the automatic tagging reduces the extraction accuracy by approximately the same amount. The chunk-based extraction is slightly less sensitive to tagging errors and achieves both best precision and best recall in this realistic scenario.

Our gold standard contains 16 112 different pair *types*, whereas numbers for automatic extraction range from 14 782 to 16 056. Not surprisingly – considering the large number of hapaxes – there are also

[3] see http://www.ims.uni-stuttgart.de/projekte/TIGER/

[4] These candidates were extracted from the XML output of the chunker with a simple XSLT stylesheet.

[5] *precision* = proportion of correct pair tokens among the automatically extracted data

[6] *recall* = proportion of pair tokens in the reference data that were correctly identified by the automatic extraction

| *candidates from* | perfect tagging | | TreeTagger tagging | |
|---|---|---|---|---|
| | precision | recall | precision | recall |
| adjacent pairs | 98.47% | 90.58% | 94.81% | 84.85% |
| window-based | 97.14% | 96.74% | 93.85% | 90.44% |
| YAC chunks | 98.16% | 97.94% | 95.51% | 91.67% |

Table 1: Results for Adj+N extraction task

considerable differences between the automatically extracted sets of pair types and the gold standard. The best results for YAC chunks on perfect tagging include 660 pair types that are not found in the reference data, while 716 pair types were missed by the automatic extraction.

We have not yet looked at frequency differences. They would be difficult to interpret for such a small sample, and they are only relevant for the very few higher-frequency pairs.

## 6 Conclusion

The extraction of adjective+noun pairs has proven to be a comparatively easy task. Depending on tagging quality, almost perfect results can be obtained. Moreover, even with a straightforward stochastic tagger and naive window-based extraction precision and recall values above 90% provide an excellent starting point for statistical techniques.

## References


Evert, Stefan and Krenn, Brigitte (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.

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

Kermes, Hannah and Evert, Stefan (2001). Exploiting large corpora: A circular process of partial syntactic analysis, corpus query and extraction of lexikographic information. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja, editors, *Proceedings of the Corpus Linguistics 2001 conference*, pages 332 – 340, Lancaster. UCREL.

Kermes, Hannah and Evert, Stefan (2002). YAC – a recursive chunker for unrestricted german text. In *Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain. to appear.

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

Mengel, Andreas and Lezius, Wolfgang (2000). An XML-based representation format for syntactically annotated corpora. In *Proceedings of the Second International Conference on Language Resources and Engineering (LREC)*, volume 1, pages 121–126, Athens, Greece.

Schmid, Helmut (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Skut, Wojciech, Krenn, Brigitte, Brants, Thorsten, and Uszkoreit, Hans (1997). An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, Washington, DC.

Skut, W., Brants, T., Krenn, B., and Uszkoreit, H. (1998). A linguistically interpreted corpus of german newspaper texts. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany.