Identifying Morphosyntactic Preferences in Collocations

Stefan Evert, Ulrich Heid, Kristina Spranger

Institut für maschinelle Sprachverarbeitung, Universität Stuttgart
Azenbergstr. 12, 70174 Stuttgart, Germany
{evert, heid, spranger}@ims.uni-stuttgart.de

Abstract

In this paper, we describe research that aims to make evidence on the morphosyntactic preferences of collocations available to lexicographers. Our methods for the extraction of appropriate frequency data and its statistical analysis are applied to the *number* and *case* preferences of German adjective+noun combinations in a small case study.

1. Introduction

1.1. Collocations in Computational Lexicography

In the fields of computational lexicography and natural-language processing, a large amount of work has been done, over the past decade, on the development of computational tools for the identification of collocations. Such tools are usually based on a statistical analysis of cooccurrence data from text corpora, as described in (Manning and Schütze, 1999, Ch. 5) and (Evert and Krenn, 2003).

Especially for a language with a richer inflectional morphology than English, lemmatised cooccurrence data are not sufficient as a basis for the description of collocations in a dictionary, whether it is intended for human users or for computational tools. Instead, morphosyntactic preferences need to be taken into account. A simple English example is the adjective+noun combination *open book*, which appears only as a singular in the *British National Corpus* in its collocational sense.[1] In German, the collocation *sich Hoffnung machen* "to have hopes" is much more often found with a plural noun form (*sich Hoffnungen machen*) than with a singular. In addition, the noun does not take an article here, while other collocations prefer a definite or indefinite article and singular number (e.g. *die Hoffnung trügt* "there is false hope", *ein Ende finden* "come to an end"). Similarly, there are case preferences (e.g. *guter Hoffnung sein* "be pregnant", where the noun always has genitive case). Such morphosyntactic preferences often go hand in hand with semantic differences in the collocational base and with the degree of idiomaticity. Many lexicographers would see two different readings of the noun *Grenze* in *seine (eigenen) Grenzen kennen* "know one's (own) limits" and *die untere/obere Grenze* "the lower/upper bound". Very rigid combinations are often idiomatic (*guter Hoffnung sein* "be pregnant") or are best classified as multi-word lexemes (*einige Zeit* "some time" as a multi-word adverb)

More sophisticated tools in computational lexicography should thus provide information about the distribution of collocations with respect to morphosyntactic features such as *number*, *case*, and *definiteness* (and possibly other properties as well). The present paper describes procedures to identify such morphosyntactic preferences, using frequency data from text corpora. In this case study, we look at the features *number* (Sg, Pl) and *case* (Nom, Gen, Dat, Acc) for German adjective+noun combinations. The methods and tools can equally well be applied to other features and other types of collocations, though.

1.2. Approaches to morpho-syntactic preferences

English collocations sometimes show a preference for singular or plural number (as in the *open book* example). To take such phenomena into account, some tools simply obtain frequency counts for word forms rather than lemmatised data. In the *British National Corpus*, *open book* shows a preference for singular (25 × *open book*, 3 × *open books*), while *red rose* does not (60 × *red rose*, 75 × *red roses*). For a language with a more morphological variation, this approach leads to two problems: (i) Unless the morphosyntactic distribution of a collocation is highly restricted, its "frequency mass" will be spread over several different combinations of word forms. For instance, we find 5 × *offenes Buch*, 1 × *offene Buch*, 3 × *offenen Buch*, and 1 × *offenen Buchs* in the *Frankfurter Rundschau* corpus,[2] even though all instances of this collocation are in singular. The smaller cooccurrence frequencies (compared to 10 × the lemmatised combination *offen* + *Buch*) may no longer provide significant evidence for a statistical association, so that the word form combinations are not indentified as collocation candidates. (ii) There is no one-to-one mapping between surface forms and (the values of) morphosyntactic features (→ syncretism). For instance, the noun form *Rose* provides no evidence at all about the *case* feature. (Evert, 2004) reports that only 10.2% of the noun forms (types) found in the *Negra* treebank corpus (Skut et al., 1998) uniquely identify *case* (and Acc is never unique), while 57.80% of the types provide no *case* information at all.

For these reasons, we conclude that both the statistical association and the morphosyntactic preferences of word combinations need to be analysed at the level of lemmas rather than word forms, and the two analyses are independent from each other.[3] By separating the lexical and the morphosyntactic distribution, we can also draw on additional sources of knowledge, such as agreement (of *case*,

[1] There are three instances of *open books*, all of which are are compositional.

[2] The *Frankfurter Rundschau* corpus is a German newspaper corpus of ca. 40 million words. It is part of the ECI Multilingual Corpus 1 distributed by ELSNET.

[3] This does not imply that morphosyntactic rigidity is not useful as an indicator of collocativity. A collocation extraction tool can draw on results from both analyses to make its suggestions.

gender, and *number*) within noun phrases, to reduce the amount of ambiguity in the morphosyntactic analyses. For instance, the noun phrase *der roten Rosen* is uniquely identified as a genitive plural, which cannot be seen by looking only at the noun and the adjective. Table 1 shows ambiguity patterns of the *case* feature for German noun phrases in the *Negra* corpus. More than 20% of all occurrences of common nouns can now unambiguously be classified, and almost 60% of the remainder provide at least partial information (Evert, 2004).

tokens	prop. (%)	value combination			
3664	5.67%	Nom			
971	1.50%		Gen		
7012	10.85%			Dat	
2592	4.01%				Acc
453	0.70%	Nom	Gen		
1	0.00%	Nom		Dat	
20025	31.00%	Nom			Acc
4856	7.52%		Gen	Dat	
1002	1.55%			Dat	Acc
448	0.69%	Nom	Gen	Dat	
916	1.42%	Nom	Gen		Acc
8819	13.65%	Nom		Dat	Acc
18	0.03%		Gen	Dat	Acc
13828	21.40%	Nom	Gen	Dat	Acc

Table 1: Case ambiguity of German common nouns, using agreement within noun phrases for partial disambiguation.

2. Implementation

2.1. Data extraction

The adjective+noun pairs for our case study were extracted from German newspaper corpora comprising a total of approx. 300 million words. The corpora are tokenised, part-of-speech tagged, lemmatised and chunk-parsed. For tokenisation and PoS-tagging we used the TreeTagger (Schmid, 1994). For the partial syntactic analysis, we used YAC (Kermes, 2003), a fully automatic recursive chunker for unrestricted German text. YAC is based on a symbolic regular expression grammar written in the CQP query language which is part of the IMS Corpus Workbench.[4] The German grammar additionally requires morphosyntactic information at the token level, which is annotated using the IMSLex morphology (Lezius et al., 2000). The chunker annotates noun phrases (NP), adverbial, adjectival, and prepositional phrases (PP), as well as verbal complexes. In addition, the chunks also carry partially disambiguated morphosyntactic information.

We extracted cooccurrences of prenominal adjectives and nouns using the annotated chunk boundaries and partially disambiguated morphosyntactic information at the chunk level. In this case study, we looked at the nouns *Tag*, *Zeit*, *Schritt*, and *Kraft*. As a heuristic filtering rule, we excluded NPs that are part of a PP, since a number of adjective+noun pairs occur primarily in adverbial PPs (e.g. *mit letzter Kraft* "with ultimate force").

[4]For more information, see `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`.

For each pair of lexemes we extracted the total cooccurrence frequency as well as unique positive and negative evidence (Evert, 2004) for each value of the features *number* (Sg, Pl) and *case* (Nom, Gen, Dat, Acc).

2.2. Statistical analysis

Our goal in this section is to develop a quantitative measure of morphosyntactic preference that is statistically sound, takes the available amount of corpus evidence into account, and has a meaningful interpretation when it is presented to lexicographers. For the binary feature *number*, which assumes only two different values (Sg and Pl), the procedure is fairly straightforward. Ignoring the issue of ambiguity for the moment, a useful criterion is provided by the proportion of singular (or plural) instances of a given word combination. For example, there are $f = 11$ instances of *wunderschöner Tag* "beautiful day" in our data, $f_{\text{Sg}} = 9$ of which are in the singular. The corresponding proportion of 81.8% seems to indicate a strong preference for singular *number*, but it may just as well be a coincidence (considering the low overall frequency of the combination). Therefore, we apply a statistical hypothesis test in order to obtain an estimate for the *average proportion* of singular occurrences in the language (or sub-language) from which our source corpus is taken. Under the standard random sample assumption, the appropriate estimate is a binomial confidence interval (Lehmann, 1986, 89ff), which can easily be computed with a software package for statistical analysis such as R (R Development Core Team, 2003). In the example above, the R command `binom.test(9,11)` ($9 \times$ Sg out of 11 instances) yields a range of 48.2% – 97.7% for the true average proportion of singulars (at a confidence level of 95%). The lower bound of this interval represents a conservative estimate, which is shown as "prop. of Sg" in the presentation of the results (cf. Table 2). Hence, we cannot even be sure that, on average, there are more singular than plural occurrences of *wunderschöner Tag*.

When some of the occurrences are ambiguous with respect to *number*,[5] f_{Sg} is defined as the number of instances that are uniquely identified as Sg and is thus a lower bound for the true number of singulars in our corpus.[6] Since we are interested in a *conservative* estimate, we can still use the binomial confidence interval (for f_{Sg} out of f instances) to obtain a lower bound on the true average proportion of singulars. As a concrete example, the combination *sonniger Tag* "sunny day" occurs $f = 38$ times in our corpus. Out of these, $f_{\text{Sg}} = 22$ are uniquely identified as Sg, $f_{\text{Pl}} = 13$ are uniquely identified as Pl, and the remaining $f - f_{\text{Sg}} - f_{\text{Pl}} = 3$ instances are ambiguous. The (unknown)

[5]Ambiguity is to be understood here with respect to the information provided by our extraction tools, of course. A human reader will usually be able to classify the ambiguous instances as Sg or Pl, using additional syntactic, semantic, or pragmatic cues from their context.

[6]As (Sch¨onenberger and Evert, 2002) and (Evert, 2004) argue, we cannot simply discard ambiguous data and base our estimate on the proportion of singulars among the unambiguous instances. Therefore, we still have to compare f_{Sg} against the total frequency f (including the ambiguous instances) rather than against $f_{\text{Sg}} + f_{\text{Pl}}$ (where f_{Pl} is the number of unambiguous plurals).

true number of singulars in the corpus must be somewhere between 22 and 25, and the estimated range for the average proportion of singulars is 40.8% – 80.4% (at the 95% confidence level). The lower bound of this range, which is our conservative estimate, is calculated from f_{Sg} and f, so it does not depend on the number of ambiguous instances.

We can use the same procedure for the *case* feature, applying it independently to each of the possible values. Thus, comparing the number f_{Nom} of unambiguous nominatives with the total frequency f of a given word combination, we obtain an indicator of its preference for nominative *case*; from f_{Gen} we obtain an indicator of its preference for genitive *case*; etc. It is also possible to extend these conservative estimates to full ranges for the true average proportion (as in the second example above) by making use of unambiguous negative evidence (see (Evert, 2004) for details).

Unfortunately, this simple approach only makes use of some 20% of the corpus data that provide unambiguous evidence for a unqiue *case* value (cf. Table 1). Systematic ambiguities between two values (Nom|Acc, Gen|Dat, and Dat|Acc) account for another 40% of the data, from which we can obtain at least partial information about morphosyntactic preferences. To this end, we define *ambiguity classes* as sets of feature values, e.g. $A = \text{Nom|Acc}$. The corresponding frequency f_A includes all instances that are either uniquely identified as Nom or Acc, or ambiguous between the two (but all other *case* values can be excluded). The conservative estimate calculated from f_A and f is a lower boundary for the average proportion of occurrences with nominative or accusative *case*.

In our case study, the extracted word combinations were sorted according to the estimated proportion and then manually inspected. This was done separately for each feature value (Sg and Pl for *number*; Nom, Gen, Dat, Acc for *case*) and each ambiguity class (Nom|Acc, Gen|Dat, and Nom|Dat|Acc). Some results of this case study are presented in the following section.

3. Results and applications

Tables 2 and 3 contain examples for *number* preferences of adjective+noun combinations. For each adjective+noun pair, the total corpus frequency is shown together with the number of unambiguous singulars and plurals. A conservative estimate for the average proportion of singulars or plurals, obtained according to the procedure described in Section 2.2., is shown in the fifth column (labelled "prop. of *x*"), followed by an English translation. The mode of presentation is similar to that used in the case study for manual inspection.

The data for *number* preferences often point to collocations that show a tendency towards idiomatisation. The following combinations are examples:

(eine) reife Leistung "good work"	85.41% Sg
(die) treibende Kraft "driving force"	85.23% Sg
(ein) gangbarer Weg "practicable plan"	85.12% Sg

Similar cases can be observed with combinations that prefer a specific *case*, in particular genitive:

gemessenen Schritts "measured"	83.53% Gen
strammen Schritts "briskly"	65.18% Gen

Not only do certain nouns have specific collocations along with their different senses, but these collocations in turn have morphosyntactic preferences. The reading of *Schritt* in the juridic sense (*rechtliche Schritte* "measures", cf. Table 3) is only available with collocations in the plural. Similarly, *Hilfe* has a reading that prefers singular, as in *ärztliche Hilfe* "medical assistance", *medizinische Hilfe* (97.76% and 96.47% Sg, respectively).

In certain cases, the readings tend towards specialised or terminological use: *finanzielle Leistungen* ("benefits", 89.43% Pl) and *vermögenswirksame Leistungen* ("capital-forming payment", 95.28% Pl) are cases in point. The term *installierte Leistung* ("power output of a (number of) power station(s)", 91.80% Sg) illustrates the same phenomenon.

As the data produced by our tools are relevant for lexicography (regarding both paper and NLP dictionaries), we envisage the use of a lexicographic viewing tool for manual selection of lexicographically relevant candidates (Heid et al., 2004). Furthermore, as collocations with strong morphosyntactic preferences tend towards idiomatisation, the tools could also serve as an element of an idiom finder for corpus technology.

4. Further developments

The procedures described in this paper were used to identify adjective+noun combinations with morphosyntactic preferences in a case study, regardless of their collocational status. They can now be combined with standard methods for the identification of collocations (Evert and Krenn, 2003) in one of two ways: (i) Collocation candidates could be identified in a separate first step, based on lemmatised cooccurrence data. Within the set of candidates produced by such a tool, those which seem to have morphosyntactic preferences could be marked. (ii) Alternatively, morphosyntactic preferences could be identified first and used as an additional criterion for the identification of collocations (e.g. by increasing the association scores of morphosyntactically restricted candidates).

In either case, it is necessary to translate the proportion estimate introduced as a quantitative measure in Section 2.2., which was used in our case study to sort the result tables, into a categorical annotation. A possible classification uses the categories *strong*, *weak*, or *no association* and is applied to each feature and each value (or ambiguity set) individually.

Currently, the procedures have only been applied to *number* and *case* preferences. In a similar way, other morphosyntactic features can be accounted for, in particular the *definiteness* of noun phrases. Preferences for the definite or the indefinite article often go hand in hand with *number* preferences: *treibend* + *Kraft* typically occurs in the singular with the definite article (*die treibende Kraft*, "the driving force"). Similarly *der rechte/richtige/falsche Weg* "the right/wrong way", but *ein gangbarer Weg* "a possible way". The extraction of such data is relatively easy on the basis of chunked-parsed text corpora as described in Section 2.1.. In a next version of our tools, we will explore the multi-parametric approach suggested by (Spranger, 2004) for this purpose.

adjective + noun	frequency	#Sg	#Pl	prop. of Sg	translation
kurz + Zeit	2050	2042	8	99.29%	*a short time*
geraum + Zeit	278	278	0	98.92%	*considerable time*
kostbar + Zeit	48	48	0	93.94%	*precious time*
heutig + Zeit	89	87	2	93.09%	*(in) our day*
unbegrenzt + Zeit	7	7	0	65.18%	*unlimited time*
…	…	…	…	…	…
neu + Tag	64	64	0	95.42%	*new day*
historisch + Tag	98	95	3	92.27%	*historic day*
schwarz + Tag	134	120	13	84.14%	*black day*
sch¨on + Tag	215	153	60	65.64%	*beautiful day*
…	…	…	…	…	…

Table 2: Candidates for collocations of the German nouns *Tag* and *Zeit* with a preference for singular *number*.

adjective + noun	frequency	#Sg	#Pl	prop. of Pl	translation
gerichtlich + Schritt	214	2	212	97.09%	*legal measures*
rechtlich + Schritt	561	5	535	93.63%	*legal measures*
…	…	…	…	…	…
gem¨aßigt + Kraft	69	0	69	95.75%	*moderate circles*
…	…	…	…	…	…
sozial + Leistung	406	18	388	93.50%	*fringe benefi ts*
…	…	…	…	…	…

Table 3: Candidates for collocations of the German nouns *Schritt*, *Kraft*, and *Leistung* with a preference for plural.

5. Acknowledgements

Part of the research presented in this paper was carried out within the project TFB-32, a cooperation with the publishing houses *Langenscheidt* and *Duden*, funded at the university by the *Deutsche Forschungsgemeinschaft*.

6. References

Evert, Stefan, 2004. The statistical analysis of morphosyntactic distributions. In *Proceedings of LREC 2004*. Lisbon, Portugal.

Evert, Stefan and Brigitte Krenn, 2003. Computational approaches to collocations. Introductory Course at the European Summer School on Logic, Language, and Information (ESSLLI 2003), Vienna. Slides can be downloaded from `http://www.collocations.de/`.

Heid, Ulrich, Bettina Säuberlich, Esther Debus-Gregor, and Werner Scholze-Stubenrecht, 2004. Tools for upgrading printed dictionaries by means of corpus-based lexical acquisition. In *Proceedings of LREC 2004*. Lisbon, Portugal.

Kermes, Hannah, 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 9, number 3.

Lehmann, E. L., 1986. *Testing Statistical Hypotheses*. New York: Wiley, 2nd edition.

Lezius, Wolfgang, 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 (eds.), *Proceedings of the 9th EURALEX International Congress*. Stuttgart, Germany.

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

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/`.

Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*.

Schönenberger, Manuela and Stefan Evert, 2002. The benefit of doubt. Presentation at the Workshop on Quantitative Investigations in Theoretical Linguistics (QITL), Osnabrück, Germany, October 2002. Slides can be downloaded from `http://www.cogsci.uni-osnabrueck.de/qitl/`.

Skut, Wojciech, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit, 1998. A linguistically interpreted corpus of German newspaper texts. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*. Saarbrücken, Germany. See also `http://www.coli.uni-sb.de/sfb378/negra-corpus/`.

Spranger, Kristina, 2004. Beyond subcategorization acquisition – multi-parameter extraction from german text corpora. In *Proceedings of EURALEX '04*.