The Statistical Analysis of Morphosyntactic Distributions

Stefan Evert

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

Abstract

This paper describes methods for the statistical analysis of quantitative data on the distribution of morphosyntactic features. A key problem is the large amount of ambiguity in automatically extracted data. In the paper, I argue for a conservative approach that treats ambiguous instances as counter-evidence. It is nonetheless possible to obtain detailed morphosyntactic information from the corpus data with the help of partial disambiguation and by exploiting systematic ambiguity classes.

1. Introduction

1.1. Morphosyntactic preferences

Text corpora are a valuable source of raw material for lexical resources in the fields of lexicography (e.g. paper dictionaries), terminology (e.g. terminological databases), and natural-language processing (machine-readable dictionaries etc.). So far, most tools for the extraction of candidate data are based on frequency counts for individual (lemmatised) words and word combinations (collocations). These are combined with a statistical analysis in terms of a random sample model to correct for the effects of chance: only significant results can be generalised beyond the particular source corpus from which they were extracted.

For all three kinds of applications, it is also quite often important to obtain information about the distribution of morphosyntactic features such as *number* (Sg, Pl), *case* (Nom, Gen, Dat, Acc in German), *definiteness* of noun phrases (Def, Ind, Nil = no determiner), etc. Such frequency distributions provide cues for usage preferences both at the level of individual lexemes and at the level of collocations (where they are often even more conspicuous). The arguments in this paper are illustrated on the example of German nouns, but the results apply equally well to other parts of speech, to collocations, and to other languages with a sufficiently rich inflectional morphology (though not necessarily to English, where e.g. most plural nouns can easily be identified by their *-s* suffix).

In German, the surface forms of nouns provide some information about their morphosyntactic features. However, because of syncretism in the inflectional paradigm, the values of features such as *number* and *case* often cannot be uniquely determined from the word form.[1] The amount of ambiguity can be reduced when the syntactic context is taken into account, but some uncertainty will always remain, as is shown by the quantitative data in Section 1.2.. In general, there are three approaches for dealing with such incomplete information: (i) discard the ambiguous data altogether; (ii) apply statistical methods, e.g. a probabilistic parser, to guess the true feature values; or (iii) the principle of maximum entropy, which assumes the most uniform distribution compatible with the observed data. In Section 2.1., I argue that all three methods are liable to produce entirely misleading results under certain circumstances. Therefore, ambiguities have to be fully represented in the base data extracted from a corpus. Following well-established statistical methodology, they are always interpreted as counter-evidence in the quantitative analysis.

The main goal of the statistical analysis described in Section 2. is to detect the tendency of a word type (or a word combination) to occur with a specific value v for a morphosyntactic feature F. I use the notation $[F = v]$ for this condition, e.g. $[\textit{number} = \text{Pl}]$ for plural *number* and $[\textit{case} = \text{Gen}]$ for genitive *case*. The resulting measure of morphosyntactic preference is a conservative estimate for the average proportion $p^*[F = v]$ of tokens with the specified feature-value combination. It is also possible to measure *negative* preference, i.e. a tendency towards the condition $[F \neq v]$. For instance, some words or word combinations may prefer not to occur in genitive *case*, which is indicated by a high proportion $p^*[\textit{case} \neq \text{Gen}]$.

1.2. Syncretism and ambiguity

In this section, we will take a closer look at the amount of morphosyntactic ambiguity in corpus data, using the *case* feature of German nouns as an example. The source of this ambiguity is syncretism, where different feature-value combinations are expressed by the same surface form in an inflectional paradigm. For instance, the noun form *Spiel* "game" can have nominative, dative, or accusative *case*, but not genitive. Quantitative data for this phenomenon was obtained from the Negra corpus (Skut et al., 1998), which consists of 355,096 tokens of German newspaper text with manually corrected part-of-speech tagging, morphosyntactic annotation, and parse trees. For the research presented here, only the tokenisaton and part-of-speech tagging were used, while the annotation of morphosyntactic features (as well as lemmatisation) was performed by the IMSLex morphological analyser (Lezius et al., 2000).

Table 1 shows how many of the word form types of common nouns in Negra fall into each ambiguity pattern for the *case* feature. More than half of all different surface forms provide no information about the *case* of the noun at all. For the statistical analysis, however, the amount of cor-

[1] This problem is exacerbated by homographs, where different forms of two different lexemes share the same surface string. An example is the German noun form *Zwecke*, which can either be the singular of *Zwecke* "tack" or the plural of *Zweck* "purpose". Such ambiguities are comparatively rare and diffi cult to resolve by fully automatic means, so I will ignore them in this paper.

types	prop. (%)	value combination
219	0.92%	Nom
1343	5.65%	Gen
869	3.65%	Dat
8	0.03%	Nom Gen
307	1.29%	Nom Acc
74	0.31%	Nom Gen Dat
840	3.53%	Nom Gen Acc
6379	26.81%	Nom Dat Acc
13750	57.80%	Nom Gen Dat Acc

Table 1: Case ambiguity of German common nouns. This table shows how much information different word form *types* provide about the *case* feature.

pus evidence as measured by the number of tokens is a more important indicator than the number of different types. Table 2 shows that the ambiguity problem is even a little worse on this scale: a unique *case* value can be determined for less than 7% of all noun tokens.

tokens	prop. (%)	value combination
513	0.72%	Nom
2252	3.16%	Gen
2150	3.02%	Dat
12	0.02%	Nom Gen
592	0.83%	Nom Acc
150	0.21%	Nom Gen Dat
2471	3.47%	Nom Gen Acc
21253	29.81%	Nom Dat Acc
41911	58.78%	Nom Gen Dat Acc

Table 2: Case ambiguity of German common nouns. This table shows the amount of corpus evidence (= number of *tokens*) for each *case* pattern.

From the noun forms alone, little can be said about *case* preferences, except for the condition $[case \neq \text{Gen}]$ (i.e. a tendency to avoid the genitive). Fortunately, further disambiguation is possible when the (syntactic) context is taken into account, especially the agreement of determiner, adjectives, and the head noun in a noun phrase (with respect to *case*, *number*, *gender*, and *definiteness*). To this end, the YAC chunk parser (Kermes, 2003) was used to identify noun phrases (including center embedding but without adjuncts) in the Negra corpus, which are automatically annotated with partially disambiguated morphosyntactic information. YAC reaches excellent precision when applied to a corpus with "perfect" part-of-speech tagging, especially for the detection of prenominal adjectives and determiners that are relevant for the disambiguation step (Kermes, 2003, 141ff). Table 3 shows the remaining ambiguity after partial disambiguation using the automatically annotated noun phrases. Now more than 20% of the tokens identify *case* uniquely, another 40% are ambiguous between two *case* values, and only 21.4% still provide no information at all.

So far, we have only looked at ambiguity patterns for a single feature, and the statistical analysis in Section 2. will also be applied to each feature independently. Syncretism cuts across different morphosyntactic features, though, as in the case of the German noun form *Hunde* "dog", which

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

can either be singular and dative (Dat.Sg) or plural but not dative (Nom.Pl|Gen.Pl|Acc.Pl). This form is fully ambiguous with respect to both *number* and *case* taken individually, so that the partial morphosyntactic information it provides is entirely lost. Table 4 shows the combined *case*+*number* ambiguity patterns and their relative frequencies in the Negra corpus after partial disambiguation (based on agreement within noun phrases).[2] The ambiguity pattern of *Hunde* (D1|N2|G2|A2) accounts for 1.64% of the corpus data only. Most patterns decompose orthogonally into separate ambiguity patterns for *case* and *number*, without loss of information (marked ◁ in the table).[3] These patterns account for 57,500 out of the 64,105 tokens found in the corpus (88.9%). The remaining 7,105 tokens (marked –) correspond to little more than 10% of the full data. Therefore, the loss of information entailed by the independent analysis of individual features is relatively modest and does not outweigh its advantages (such as the reduced amount of data that has to be stored and processed, as well as an easier interpretation of the results).

2. Statistical analysis

2.1. The benefit of doubt

In this section, I argue for a conservative approach to incomplete information. Following (Schönenberger and Evert, 2002), ambiguous instances are always treated as evidence against any morphosyntactic preference that may be inferred from the corpus data. Let us consider the *number* feature as an example. Given a type X with corpus frequency f, its instances can be divided into three sets: f_{Sg} unique singulars, f_{Pl} unique plurals, and the remaining $f_{\approx}$

[2] In the feature value combinations, Gen.Sg is abbreviated as G1, Nom.Pl as N2, etc. in order to save space.

[3] For instance, the pattern N1|A1|N2|A2 reduces to the ambiguities Nom|Acc for *case* and Sg|Pl for *number*. This decomposition is orthogonal because any combination of the feature values appears in the original pattern. The pattern N1|G2, on the other hand, is not orthogonal: the individual ambiguities Nom|Gen and Sg|Pl can be combined into G1 and N2, which are not part of the original pattern. Note that all patterns where either *case* or *number* is uniquely defi ned necessarily have an orthogonal decomposition.

tokens	prop. (%)	value combination							
3664	5.67%	N1							◁
456	0.71%		G1						◁
3766	5.83%			D1					◁
2592	4.01%				A1				◁
499	0.77%					G2			◁
3091	4.78%						D2		◁
1	0.00%	N1		D1					◁
12744	19.73%	N1			A1				◁
453	0.70%	N1				G2			–
4787	7.41%		G1	D1					◁
16	0.02%		G1			G2			◁
558	0.86%			D1	A1				◁
155	0.24%			D1			D2		◁
442	0.68%				A1		D2		–
6107	9.45%					N2		A2	◁
67	0.10%	N1	G1	D1					◁
7	0.01%	N1	G1		A1				◁
8425	13.04%	N1		D1	A1				◁
224	0.35%	N1				N2		A2	–
11	0.02%		G1	D1	A1				◁
69	0.11%		G1	D1		G2			–
7	0.01%		G1		A1		D2		–
2	0.00%			D1	A1		D2		–
898	1.39%					N2	G2	A2	◁

tokens	prop. (%)	value combination							
6055	9.37%	N1	G1	D1	A1				◁
381	0.59%	N1	G1	D1		G2			–
352	0.54%	N1		D1	A1		D2		–
950	1.47%	N1			A1	N2		A2	◁
1	0.00%	N1				N2	G2	A2	–
1059	1.64%			D1		N2	G2	A2	–
2514	3.89%					N2	G2	D2 A2	◁
42	0.07%	N1		D1	A1	N2		A2	–
7	0.01%	N1		D1		N2	G2	A2	–
10	0.02%	N1			A1	N2	G2	A2	–
2	0.00%		G1	D1		N2	G2	A2	–
213	0.33%		G1			N2	G2	D2 A2	–
37	0.06%			D1	A1	N2	G2	A2	–
121	0.19%			D1		N2	G2	D2 A2	–
51	0.08%	N1	G1	D1	A1	N2		A2	–
2	0.00%	N1	G1	D1		N2	G2	A2	–
1792	2.77%	N1		D1	A1	N2	G2	A2	–
1	0.00%		G1	D1		N2	G2	D2 A2	–
55	0.09%			D1	A1	N2	G2	D2 A2	–
133	0.21%	N1	G1	D1	A1	N2	G2	A2	–
1027	1.59%	N1		D1	A1	N2	G2	D2 A2	–
622	0.96%		G1	D1	A1	N2	G2	D2 A2	–
137	0.21%	N1	G1	D1	A1	N2	G2	D2 A2	◁

Table 4: Combined *case+number* ambiguity of German common nouns (after partial disambiguation). Patterns that decompose orthogonally into separate ambiguities for *case* and *number* are marked "◁", other patterns are marked "–".

ambiguous tokens. These three numbers add up to the total frequency: $f_{\text{Sg}} + f_{\approx} + f_{\text{Pl}} = f$ (schematised in Figure 1 with $f = 100$, $f_{\text{Sg}} = 10$ and $f_{\text{Pl}} = 30$). Ideally, the statistical analysis should be based on the true number of singulars in the corpus f^*_{Sg} (indicated by a vertical line in the diagram, where there is a slight preference for singulars with $f^*_{\text{Sg}} = 60$). From our incomplete information, we know that f^*_{Sg} must be somewhere in the range $f_{\text{Sg}} \ldots f - f_{\text{Pl}}$ ($10 \ldots 70$ in the diagram), but we cannot narrow down that range.

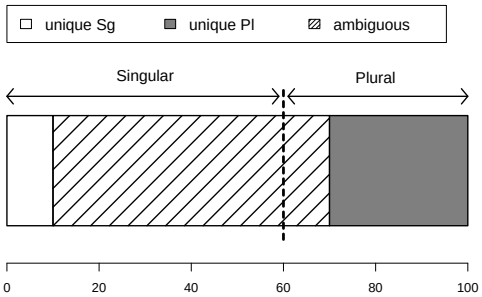

Figure 1: An illustration of how discounting ambiguous cases can produce misleading results.

If the ambiguous data were to be discarded, the remaining tokens might well give a picture that is entirely different from the true situation: in the example shown in Figure 1, $f_{\text{Pl}} = 30$ unique plurals out of $f_{\text{Sg}} + f_{\text{Pl}} = 40$ unambiguous instances indicate a clear preference for $[\textit{case} = \text{Pl}]$ in contrast to the true proportion. Other approaches to incomplete information would also distort the results in an uncontrolled way (in fact, the predicted value of f^*_{Sg} might fall anywhere within the possible range). This can only be avoided when the ambiguity is fully represented and the analysis is based on the whole range of possible values for f^*_{Sg}.

For the binary feature *number* (which assumes only two values, Sg and Pl), the entire morphosyntactic frequency information is represented by the tripartite division shown in Figure 1.[4] For the *case* feature with four different values, much more complex ambiguity patterns exist. In order to simplify the statistical analysis, I reduce these patterns to positive and negative evidence for each of the "simple" conditions $[F = v]$. Again, the set of instances of a type X is divided into f_+ unambiguous positives, f_- unambiguous negatives, and the remaining $f_{\approx}$ ambiguous tokens, with $f_+ + f_{\approx} + f_- = f$. For example, positive evidence (f_+) for $[\textit{case} = \text{Gen}]$ is provided only by unambiguous genitives, while negative evidence (f_-) comes from all ambiguity patterns that do not include Gen (Nom, Dat, Acc, Nom|Dat, Nom|Acc, Dat|Acc, and Nom|Dat|Acc in Table 3). All other patterns contribute to the number of ambiguous tokens ($f_{\approx}$). The true number f^* of instances with $[F = v]$ must be somewhere in the range $f_+ \ldots f - f_-$. The lower bound of this range (f_+) represents the amount of evidence supporting a positive preference $[F = v]$, while the upper bound ($f_+ + f_{\approx} = f - f_-$) represents the evidence supporting a negative preference $[F \neq v]$.

2.2. Confidence intervals

We cannot use the proportion of tokens with $[F = v]$ (i.e. $\frac{f^*}{f}$, which must fall into the range $\frac{f_+}{f} \ldots \frac{f-f_-}{f}$) directly as a quantitative measure for the strength of morphosyntactic preferences. For example, when there are $f_+ = 8$ unambiguous positives out of $f = 10$ tokens, the corresponding proportion of 80% suggests a strong preference for $[F = v]$, but it may equally well be a mere coincidence. In the standard random sample model for corpus frequency data, the

[4] The true number f^*_{Pl} of plurals is given by $f - f^*_{\text{Sg}}$, so that the range of its possible values is a mirror image of the range for f^*_{Sg}.

(true) observed frequency f^*, conditioned on the total frequency f of the type X, has a binomial distribution whose success probability p^* can be interpreted as the average proportion of occurrences of X with $[F = v]$ in the language (or sub-language) from which the corpus was sampled.[5] The maximum-likelihood estimate $p^* \approx \frac{f^*}{f}$ is subject to considerable sampling variation, especially when f is small. A more reliable estimate is provided by the confidence interval $[p_l, p_u]$ of a binomial test (Lehmann, 1986, 89ff). At the commonly-used confidence level of 95%, this estimate is correct (i.e. $p^* \in [p_l, p_u]$) for 19 out of 20 samples. Binomial confidence intervals can easily be computed with a software package for statistical analysis such as R (R Development Core Team, 2003).[6] In the example above, the confidence interval for $f_+ = 8$ out of $f = 10$ tokens is $p^* \in [44.4\%, 97.5\%]$, so there is no significant evidence for a tendency towards $[F = v]$.

In the presence of ambiguous data, binomial confidence intervals have to be computed for every possible value of f^*. Their union defines the *extended confidence interval* $[p_+, p_-]$ for p^*. Note that it is sufficient to compute two intervals: p_+ is the lower bound p_l for $f^* = f_+$, and p_- is the upper bound p_u for $f^* = f - f_-$. The extended confidence interval provides a conservative quantitative measure for the strength of morphosyntactic preferences. A large value of p_+ indicates a tendency towards $[F = v]$, while a small value of p_- indicates a tendency towards $[F \neq v]$.

An implementation of the extended confidence interval in the R language is shown below.[7] This code segment defines an R function `bcf()` which is invoked with `bcf(`f_+`,`f_-`,`f`)$plus` to compute p_+, and `bcf(`f_+`,`f_-`,`f`)$minus` to compute p_-. The confidence level defaults to 95% and can be changed with the optional `conf` parameter. All three arguments may be vectors (of the same length), so that p_+ and p_- can be computed efficiently for an entire set of candidate types.

```
bcf <- function (fP, fM, f, conf=.95) {
  alpha <- (1 - conf) / 2
  lower <- qbeta(alpha, fP, f-fP+1)
  upper <- qbeta(1-alpha, f-fM+1, fM)
  list(plus=lower, minus=upper)
}
```

(Evert et al., 2004) present an application of the statistical methods developed here to the morphosyntactic preferences of German adjective+noun collocations. They use the lower bound $p_+[F = v]$ as an indicator for the strength of preferences, shown in the result tables as "prop. of v".

2.3. Ambiguity classes

By limiting ourselves to "simple" conditions $[F = v]$ and $[F \neq v]$, we have been able to simplify the representation and analysis of complex morphosyntactic ambiguity patterns, obtaining a reliable and easily interpretable measure from the well-understood binomial test. There is a price to pay, though, in the form of a considerable loss of information. For instance, the pattern Nom|Acc, which accounts for 31% of all tokens in Table 3, does not provide unambiguous evidence either for or against any of the two *case* values. The ambiguity inherent in this pattern may be due to our inability to extract complete information from the corpus, but it may also reflect a genuine "underspecified" tendency towards a set of feature values. The fact that the occurrences of type X often have Nom or Acc *case* (rather than Gen or Dat) does not necessarily imply that there is a single preference for $[case = \text{Nom}]$ or $[case = \text{Acc}]$.

In order to account for such underspecified tendencies, I define an *ambiguity class* A as a set of feature values, e.g. $A_1 := \{\text{Nom}, \text{Acc}\}$ for a tendency to occur in the nominative or accusative. The instances of a type X are divided into three subsets, yielding f_+ tokens that satisfy the condition $[F \in A]$ unambiguously, f_- tokens that satisfy $[F \notin A]$ unambiguously, and the remaining $f_\approx$ ambiguous tokens. An extended confidence interval for the ambiguity class A can then be computed as described in Section 2.2.. In the example above, f_+ is obtained from the patterns Nom, Acc, and Nom|Acc, while f_- is obtained from the patterns Gen, Dat, and Gen|Dat, which rule out any of the values in A_1.

3. References

Evert, Stefan, Ulrich Heid, and Kristina Spranger, 2004. Identifying morphosyntactic preferences in collocations. 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.

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/`.

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/`.

[5] Mathematically speaking, $f^* \sim B(f, p^*)$.

[6] The R command is `binom.test(`f^*`,`f`)` for the standard 95% confidence level, and `binom.test(`f^*`,`f`,conf=.99)` for a 99% confidence level.

[7] This implementation uses the incomplete Beta function (which is the distribution function of the Beta distribution) instead of `binom.test()` for better efficiency.