

Mining for Preposition-Noun Constructions in German

**Katja
Keßelmeier**

**Tibor
Kiss**

**Antje
Müller**

**Claudia
Roch**

**Tobias
Stadtfeld**

**Jan
Strunk**

Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
D-44801 Bochum, Germany

{kesselmeier, tibor, mueller, roch, stadtfeld,
strunk}@linguistics.rub.de

Abstract

Preposition-noun constructions (PNCs) are problematic in that they allow the realization of singular count nouns without an accompanying determiner. While the construction is empirically productive, it defies intuitive judgments. In this paper, we describe the extraction of PNCs from large annotated corpora as a preliminary step for identifying their characteristic properties. The extraction of the data relies on automatic annotation steps and a classification of noun countability.

1 Introduction

In many languages, the realization of a determiner with a singular count noun is mandatory. Yet, the same languages show apparently exceptional behaviour in allowing the combination of prepositions with *determinerless* nominal projections. Minimally, such a construction consists of a preposition and an unadorned count noun in the singular, as illustrated in (1).

(1) *auf Anfrage* (after being asked), *auf Aufforderung* (on request), *durch Beobachtung* (through observation), *mit Vorbehalt* (with reservations), *unter Androhung* (under threat)

The construction is not restricted to the collocation-like combinations in (1). It can be extended in all possible ways allowed for nominal projections with the characteristic property that the resulting projection does not contain a deter-

miner. More complex constructions are illustrated in (2).

(2) *auf parlamentarische Anfrage* (after being asked in parliament), *bei absolut klarer Zielsetzung* (given a clearly present aim), *mit schwer beladenem Rucksack* (with heavily loaded backpack), *nach mehrfacher Verschiebung der Öffnung* (after postponing the opening several times), *unter sanfter Androhung* (under gentle threat)

Sometimes, the constructions in (1) and (2) have been called *determinerless PPs* (cf. Quirk et al., 1985). Since determiners combine with nominal projections, and not with prepositions, we will refrain from using this terminology and call the phrases in (1) and (2) *preposition-noun constructions* (henceforth: PNCs).

Until recently, PNCs have been considered as exceptions in both theoretical and computational linguistics. A striking example is their treatment in the Duden grammar of German, which considers the realization of a determiner with a singular count noun mandatory and treats PNCs as exceptions that can be listed. But Baldwin et al. (2006) have pointed out that the equivalent construction is productive in English; Dömges et al. (2007) have verified the empirical productivity of the construction in German on the basis of a stochastic model. They remark, however, that the empirical productivity of the construction does not correspond to its intuitive productivity: while speakers of German are able to understand PNCs occurring in newspaper texts, they are reluctant

to coin new PNCs. Hence, the linguist is confronted with a phrasal combination whose properties cannot easily be determined by introspective judgments. Consequently, it still remains unclear which factors allow a singular count noun to appear without an article if embedded under a preposition.

It has also been assumed that PNCs and PPs can be distinguished by the simple fact that PNCs do not show up as ordinary PPs. That is, it should not be possible to transform a PNC into a PP by just adding a determiner. However, this assumption is not correct. In example (3), we can either use a PNC or a PP containing a singular NP or a PP containing a plural NP without any change in its grammaticality (and only the slightest changes in interpretation).

(3) *Milosevic unterschrieb*
Milosevic signed
auch unter Androhung/der
even under threat_{sg}/the
Androhung/Androhungen
threat/threats_{pl}
von NATO-Bombardementen
of NATO-air-raids
nicht.
not
'Milosevic did not even
sign on pain of NATO air
raids.'

As speaker intuition cannot be used to determine the properties of this construction, we are pursuing an alternative strategy. We assume that the constitutive properties of PNCs can be determined by making use of *Annotation Mining* (Chiarcos et al., 2008). To this end, we annotate large corpora both automatically and manually, and extract the pertinent constructions from the annotated corpora, including not only PNCs, but also corresponding PPs (as illustrated in (3)), in order to determine the characteristics that distinguish PNCs from ordinary PPs.

The data are extracted from a large newspaper corpus, the *Neue Zürcher Zeitung corpus* (1993-1999), which contains approximately 200 million words. Carried out as a case study, we have initially opted for an inline XML format, but will move on to a stand-off format. We use standard tools available for the analysis of large corpora for automatic annotation, in particular two part-of-speech taggers, a morphological analyser and a phrasal chunker. In addition, we had to develop a genuine classifier for noun countability, since

we are only interested in those PNCs in which the noun is classified as countable. Noun countability cannot be determined as a lexical property but must be considered a contextual property (cf. Allan, 1980). To this end, we have developed a classification system by chaining together a decision tree and a naïve Bayes classifier.

In section 2, we describe the automatic morphosyntactic and categorial annotation of the corpus provided by two different taggers and present the classification of noun countability. Section 3 describes the indexing and search procedures. We also present a small-scale evaluation of the extraction method. Section 4 briefly describes manual annotation steps that are further required to carry out annotation mining.

2 Corpus processing

2.1 Construction of the corpora

The construction of the corpora started with plain-text files for each volume of the NZZ newspaper from 1993 to 1999. The first step was to identify the document structure and to extract meta-information about genre, date, and author for each article. Since headlines and titles are often formulated in telegraph style with anomalous use of articles, it was very important to determine the membership of a sentence in a title-section or a paragraph. This was done using simple heuristic methods. To further facilitate the preprocessing, the daily issues of the newspaper were stored in 2092 individual files. A daily issue contains approx. 98,000 tokens on average, which turned out to be a size that could be handled well by all tools employed.

2.2 Automatic morphosyntactic analysis

The tokenization and sentence-boundary detection of the corpora was performed using the *Punkt* system (Kiss and Strunk, 2006). After converting the data into the customary format for tagging (one token per line), two taggers were used simultaneously to process the corpora.

The *Regression-Forest Tagger* (Schmid and Laws, 2008) does not only produce POS tags but also performs a morphological analysis of each token based on SMOR (Schmid et al., 2004). It thus provides the lemma and morphosyntactic features of nouns, including their number value and whether we are dealing with a common or proper noun. To maximize the quality of the morphological analysis we trained the morphological component of the RFT on a full lexicon of all word types occurring in our corpora. The

high accuracy of this tagger for identifying the number value of nouns (a preliminary test resulted in over 97% accuracy) was the main argument for using RFT.

The *TreeTagger* (Schmid, 1995) provides POS annotations as well, but in addition determines non-recursive chunks essential for the identification of PNCs and regular PPs.

To aggregate the output of the two taggers in a standard common format, we have not only integrated the annotations of the two taggers for each daily issue into a single valid inline XML data format, but also reorganized and enhanced the previously extracted meta-information, and defined an individual ID for every token, sentence, segment, and article. The user can thus identify sentences or tokens unambiguously even in huge corpora and across different preprocessing and annotation tools. Table 1 exemplifies the token ID `NZZ_1994_04_27_a32_seg5_s13_t4` and Figure 1 shows a small example of the constructed inline XML data format.

Name of newspaper	NZZ
Year	1994
Month	04
Day	27
Number of article (in daily issue)	32
Number of segment (in article)	5
Number of sentence (in segment)	13
Number of token (in sentence)	4

Table 1. Structure of the global IDs.

2.3 Countability classification

Allan (1980) suggests that countability is not a lexical property, but determined by the formal context of a noun. Nevertheless, his classification system accounts for the fact that most nouns show a preference for a countability class.

In the present system, we employ the idea of a countability preference particularly in those cases where the context is neutral with regard to countability.

The first step therefore was to determine the countability preferences. We annotated 10,000 German lemmas for their most probable countability class (e.g. *Auto* (car) countable, *Wasser* (water) uncountable). Four trained linguists annotated each noun. Nouns that did not receive a unique annotation were discarded. We furthermore dismissed all nouns that did not show a *class-plausible* ratio of singular and plural occurrences, using the information provided by the RFT. The remaining 4,267 nouns (74% countable, 26% uncountable) were used as prototypical members of their countability class. For these nouns, we counted the co-occurring contexts in the corpora and stored them in the form of a 3-tupel (RFT-POS, TT-POS, lemma) (cf. Table 2).

Context (C)	+count	-count	P(C +count)
PIAT PRO viel	0	1765	0.0005
KOKOM CONJ wie	327	1200	0.2145
VMFIN VFIN sollen	37	237	0.1376
ART ART einen	246	15	0.9391
PIAT PRO keine	4287	2969	0.5907
...			

Table 2. Example context tuples used by the countability classifier.

We used the *m-estimate* variant of a naïve Bayes classifier (Mitchel, 1997) to determine the probability of a noun being countable given the context (cf. the posterior probabilities given in the last column of Table 2).

For each unseen noun, we calculate a score for being either +COUNT or -COUNT by multiplying the calculated probabilities of occurring contexts, weighted with their frequency. If the normalized score for countability exceeds a defined threshold, the noun is classified as countable.

```
<art source="Neue Zürcher Zeitung" genre="WIRTSCHAFT" date="27.04.1994"
  misc="Nr. 97 31" id="NZZ_1994_04_27_a32"> [...]
<para>
<s id="NZZ_1994_04_27_a32_seg5_s13"> [...]
<tt_chunk type="PC">
<tok tt_pos="APPR" rft_pos="APPR" rft_lemma="auf" rft_morph="Auf"
  tok_id="NZZ_1994_04_27_a32_seg5_s13_t4">auf</tok>
<tok tt_pos="NN" rft_pos="N" rft_lemma="Anfrage" rft_morph="Reg.Acc.Sg.Fem"
  tok_id="NZZ_1994_04_27_a32_seg5_s13_t5">Anfrage</tok>
</tt_chunk> [...]
</s> [...]
</para> [...]
</art>
```

Figure 1. Abbreviated example of the inline XML format used for the annotation.

If the score is below a second threshold it will be classified as uncountable. A score between those two values results in a classification as *unknown*.

The second classifier bases its classification on the calculated singular/total-ratio of the noun. We trained a decision-tree classifier on all annotated nouns using cross-validation. A singular/total-ratio above 0.997 results in a classification as -COUNT, while a value below 0.98 as +COUNT. Nouns with a value between these two thresholds are classified as *unknown*.

A noun is considered as countable or uncountable if both classifiers reach the same conclusion. Otherwise it is marked as *unknown*.

A first evaluation based on 100 nouns classified as countable and 100 classified as uncountable showed an accuracy of the classifier of 93% in case of countable and 88% in case of uncountable nouns. A more detailed description of the process can be found in Stadtfeld et al. (2009).

3 Indexing and search

3.1 Conversion and indexing

The automatic annotation of our corpora with morphosyntactic features and non-recursive chunks and the training of an accurate countability classifier provide us with all the information necessary to identify and extract PNCs (and also regular PPs). Since the corpora we are currently using already comprise more than 208 million tokens and we are planning to at least double the size of our data base by adding further corpora, we require a search tool that is able to deal with this huge amount of data efficiently.

The (Open) Corpus Workbench (CWB) developed at IMS Stuttgart¹ (Evert, 2005) is well suited to index and query shallow linguistic annotation and has been designed to cope with corpora of more than 100 million words.² Moreover, it is also able to index token spans (such as chunks) delimited by XML tags. Therefore, the only minor conversion step necessary to index our inline XML corpus files with CWB consisted in converting the XML annotation of individual tokens into a tab-delimited column format while leaving the XML tags for higher units such as

¹ <http://cwb.sourceforge.net/>

² We are also currently looking into the possibility of adopting the search and visualization tool ANNIS2 developed at Humboldt University Berlin and the University of Potsdam (<http://www.sfb632.uni-potsdam.de/~d1/annis/>). This would be especially useful for the manual inspection of individual examples in later stages of the project. We are currently testing whether ANNIS2 will scale up to very large corpora.

chunks and sentences intact. The information about tokens was encoded by positional attributes, while the information about larger units was encoded using structural attributes. Most importantly, the detailed global IDs defined for all units during the aggregation step were also indexed in CWB in order to enable the unambiguous identification of the extracted constructions for the subsequent manual annotation steps.

We originally planned to create just one big index of all our corpora and to query them all at once in order to make searching less laborious. However, this turned out to be impossible because of RAM limitations. We therefore backed off to indexing whole year volumes of our newspaper corpora separately.

3.2 Searching and extracting PNCs

After indexing the corpora with CWB, we formulated the query shown in (4) to search for PNCs. This query expresses the fact that PNCs form a preposition chunk (PC) which consists of a specific preposition, here exemplified with *an* (on, to), followed by any number of words which are not determiners (i.e., not articles, demonstratives, possessive pronouns, etc.) and finally a regular noun that is both countable and singular.

```
(4) <tt_chunk_type = "PC">
    [(word="an" %cd) &
     (rft_pos="APPR")]

    [(rft_pos!="(ART|...)" ) &
     (tt_pos!="(ART|...)" )]*

    [(tt_pos="NN") &
     (rft_morph!=".*\Pl\..*")
     & (countability="count")]
</tt_chunk_type>
```

The list of 23 prepositions that we examine in our study is given in (5). It includes all simple prepositions that typically take an NP complement and also assign case to it.

```
(5) an, auf, bei, binnen,
    dank, durch, für, gegen,
    gemäß, hinter, in, mit,
    mittels, nach, neben,
    ohne, seit, über, um,
    unter, vor, während,
    wegen
```

Examples of prepositions that were excluded are *ab* (from) and *bis* (until), which often occur with a PP or adverbial complement, and the preposi-

tion *zwischen* (between), which demands a coordinated NP. In general, all prepositions that deviate significantly from the pattern PP = P + NP were excluded.

The query results are exported from CWB as a list of the IDs of all sentences containing at least one PNC. From these lists, reasonably sized working packages can be created, the relevant sentences can be extracted from the inline XML format based on their IDs and can be converted to the format of the annotation tool used for manual annotation (see section 4).

3.3 Evaluation

We performed a small-scale evaluation of our strategy for extracting PNCs in order to determine its effectiveness and quality in terms of precision and recall. For this evaluation, we chose one daily issue of the NZZ randomly: April 3rd, 1997. This issue contains 6,081 sentences and 91,357 tokens. We constructed a gold standard list of PNCs by searching for all occurrences of the prepositions in (5) based only on their word form. All true examples of PNCs were then manually extracted from this large list of 5,304 hits. This yielded a much smaller list of true positives comprising 161 PNCs.³

We then used the query expression in (4) to extract all putative PNCs with one of the 23 prepositions from the same NZZ issue based on the automatic morphosyntactic and countability annotation. This resulted in an even shorter list of 56 putative PNCs.

A comparison of the manually and automatically extracted lists of PNCs yielded 27 true positives, 29 false positives, and 134 false negatives. This corresponds to a precision of 48.21% and a recall of 16.77%. The precision of our PNC extraction strategy is satisfactory for our purposes, since irrelevant constructions can still be excluded during the manual annotation phase. The false positives mostly consisted of determinerless nominal complements of prepositions in headlines and coordinations. Since the use of articles follows special rules in these contexts, such examples were excluded in the manual extraction. The low recall is more problematic. It is due to the fact that the countability classifier only classifies nouns for which it has gathered enough contextual information (cf. section 2.3). As discussed in section 1, PNCs are a productive construction and therefore occur with a large number

of nouns that the countability classifier has never encountered before. The low recall thus comes from the notorious problem of data sparseness.

This can be shown by extracting a second list of putative PNCs based on the automatic annotation that includes not only nouns classified as countable but also all nouns that were not classified because of a lack of evidence. A comparison of this automatically extracted list with the gold standard results in 143 true positives, 467 false positives, and 18 false negatives, corresponding to a precision of 23.44% and a recall of 88.82%. Recall can thus be increased fivefold, while only halving precision. It might therefore be more sensible to use a classification as uncountable as a knockout criterion rather than to search positively for countable nouns. It is also clear that the coverage of the countability classifier should be improved by training it on larger corpora.

4 Manual annotation

While the automatic annotation steps described in sections 2 and 3 suffice to extract PNCs from corpora, this is only a preliminary task. We are interested in the characteristic properties which distinguish PNCs from PPs, and hence have to annotate further features of PNCs (and corresponding PPs) manually. This step is performed in small batches, since the annotation tool we use cannot deal with large amounts of data and small working packages are also more convenient for the human annotators.

We annotate the relevant constructions with various features such as valency, morphological complexity and etymological status (native vs. borrowed) of the noun and furthermore the semantic interpretation of the respective preposition and noun.

MMAX2 (Müller and Strube, 2006) is employed for manual annotation.⁴ It features stand-off annotation, which enables us to keep the original corpus and the added annotations separate. Although the annotation tool makes a conversion and preprocessing of the data and the definition of an annotation scheme inevitable, the user has a maximum degree of flexibility in making the tool fit his purposes. Another advantage of MMAX2 is the possibility to create an arbitrary number of independent annotation levels. The annotator is able to add both markables, i.e. spans of tokens, at different levels and pointer relations between the markables.

³ This small number of PNCs shows that huge corpora are indeed required to study such more peripheral constructions.

⁴ <http://mmax2.sourceforge.net/>

As a preparatory step, it is necessary to create an MMAX2 project for every batch of sentences, based on the IDs extracted using CWB. Each project consists of several tiers containing the information annotated automatically in the preceding steps, e.g. the information provided by the sentence boundary detection system (level: *sentences*), the TreeTagger (*tt_pos* and *chunks*), the RF-Tagger (*rft*) with the attributes (*rft_pos*, *rft_lemma*, *rft_morph*) and finally the information from the countability classifier (*countability*). New levels for the manual annotation have to be created for the interpretation of the prepositions and nouns (*prep-meaning*, *noun-meaning*), as well as two levels for the valency of the noun, in order to be able to create pointer relations between the noun and its dependents (*noun-valency*, *noun-dependents*).

Last but not least, we also define a level at which metadata about the annotation process will be inserted. This will be important to assure completeness of annotation, in particular after reintegrating the manually annotated sentences into the entire corpus. Once the annotation of the PNCs has been completed, we will restart extraction and annotation with ordinary PPs, corresponding to the PNCs we have identified in the first cycle.

5 Conclusion and outlook

The extraction of PNCs is an important yet preliminary step in the determination of the characteristic properties of PNCs.

In this paper, we have shown how automatically annotated data can be used as a basis for extracting the pertinent construction from large corpora. Since we are preparing the data for annotation mining (particularly for clustering and classification), reaching a high recall is as necessary as reaching a high degree of accuracy. Our evaluation has shown some shortcomings of the extraction process in this respect, but a variety of alternative strategies can be considered.

In the current state of affairs, where PNCs have mostly been investigated by looking at individual examples, even an extraction with a relatively low recall facilitates further investigation and will thus be useful to eventually determine the constituting factors of this construction.

References

Keith Allan. 1980. Nouns and countability. *Language* 56(3): 541-567.

Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. 2006. In search of a systematic treatment of determinerless PPs. In Patrick Saint-Dizier, editor, *Syntax and Semantics of Prepositions*. Springer, Dordrecht, pages 163-179.

Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. 2008. A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues*. Special Issue Platforms for Natural Language Processing. ATALA, 49 (2).

Florian Dömges, Tibor Kiss, Antje Müller and Claudia Roch. 2007. Measuring the productivity of determinerless PPs. In *Proceedings of the ACL 2007 Workshop on Prepositions*, pages 31-37, Prague, Czech Republic.

Stefan Evert. 2005. *The CQP Query Language Tutorial* (CWB version 2.2.b90). Institut für Maschinelle Sprachverarbeitung, University of Stuttgart.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4): 485-525.

Tom Mitchell. 1997. *Machine Learning*. McGraw-Hill, New York.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. (English Corpus Linguistics Vol. 3)*. Peter Lang, Frankfurt, pages 197-214.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

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

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of LREC 2004*, pages 1263-1266, Lisbon, Portugal.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING 2008*, Manchester, UK.

Tobias Stadtfeld, Tibor Kiss, Antje Müller, and Claudia Roch. 2009. Chaining classifiers to determine noun countability. Submitted to *ACL 2009*, Singapore.