Bootstrapping a database of German multi-word expressions

Alexander Geyken*

*Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstr. 22/23, 10117 Berlin, www.dwds.de
geyken@bbaw.de

Abstract

We pre-classified 32,000 entries from the *Wörterbuch der deutschen Idiomatik* (Schemann, 1993) using an inductive description of POS sequences in conjunction with a Brill Tagger trained on manually tagged idiomatic entries. This process assigned categories to 86% of entries with 88% accuracy. Further manual classification resulted in a database of multi-word expressions where each entry is associated with a sequence of POS-tag/token pairs. The second phase of our project, currently underway, addresses the association of a sequence of POS-tag/token pairs with a corpus example. To this end, we generate a weighted finite state transducer from the sequences for each entry and apply a finite state filter to the corpus. The filter will extract those sequences in the corpus that correspond to the longest match of the multi-word expression.

1. Introduction

For the past decade, statistical methods have been widely used for the extraction of multi-word expressions (Smadja, 1992; Oakes, 1998). These approaches commonly target word groups that frequently appear in the same context. A challenge for all these methods is presented by 'rare collocations', i.e. groups of words with low frequency. Sparsity of data is not just a matter of corpus size, since even very large corpora such as the 1 billion word corpus of the DWDS (www.dwdscorpus.de) do not provide enough occurrences of common multi-word expressions (MWE) for statistical methods. (Geyken et al., 2004) demonstrated this in a small case study with 46 idioms randomly selected from a commonly used dictionary of German idioms (Duden 11). The idioms had been previously linguistically analyzed and described in the context of the research project on idioms at the BBAW (www.bbaw.de/forschung/kollokationen/). The one billion word corpus was searched for all occurrences, including variations, of these idioms. The frequency distribution shows that 9 entries occur between 1 and 10 times, 13 idioms were found between 11 and 25 times, 15 idioms occur between 26 and 100 times, and 15 idioms show up more than 100 times in the corpus.

On the other hand, multi-word expressions (henceforth MWE) have been compiled manually, resulting in a few specialized print-dictionaries with considerable coverage. For example, the "Wörterbuch der deutschen Idiomatik" (Schemann, 1993) consists of more than 32,000 entries and constitutes the largest printed collection of German multi-word expressions. However, print dictionaries suffer generally from the fact that they are competence based and that their relation to attested examples remain sometimes unclear. Schemann freely admits this in the introduction of his dictionary, saying that it is " ... impossible to provide suitable attested examples for such a big number of entries (Schemann 1993:XV) ".

We proceed in two steps. First, dictionary entries are annotated in an appropriate way; second, a knowledge-based formalism that detects occurrences of the multi-words entries in the corpus is developed. This paper describes the first step and briefly sketches the second.

2. Method

2.1. Schemann's dictionary

The starting point for this work were 32,000 idiomatic entries of Schemann's dictionary. The notion of "idiom" is used in a very broad sense by the author; Schemann considers all those expressions idioms where a word is constrained by a larger context. Particular emphasis is placed on the presentation of the syntagmatic contexts for all entries. Hence, an entry is not a linear multi-word expression but rather a set of different syntagmatic contexts around the target word. The selections range from completely frozen expressions to MWE where all components are substitutable. The complexity of form varies from two-word expressions to complex PP-VP constructions.

The following examples illustrate the variety of entry structures found in Schemann's dictionary.

(1) einen zischen
one sizzle
to have a swift one (a drink)

(2) ein absolutes/ das absolute Gehör haben
an absolute/ the absolute ear have
to have perfect pitch

(3) (voll) auf jdn./... abfahren
(full) on sb./... leave
to (really)/... fancy sb.

(4) ein Glas über den Durst trinken
a glass over the thirst drink
to have one (drink) too many

These entries demonstrate the lexicographic shorthand used by Schemann: / to indicate a mutually exclusive choice between elements to the left and right of the symbol, ( ) to indicate an optional phrase that may be omitted, [...] to designate arbitrary clauses, and placeholders such as jdn. (s.b.) to designate flexible components that admit substitution, in this case an accusative noun phrase.

This variety of entry structures means that Schemann's dictionary is not immediately usable for automatic processing.

Therefore, some simple preprocessing steps are applied in order to transform this format into a more suitable machine-readable format. Entries with optional elements are rendered simultaneously as multiple phrases: a phrase with the optional element and a phrase without. Likewise, `[...]` can be thought of as an unconstrained substitution. Hence, it is possible to remove the token altogether without loss of significant information.

Somewhat more difficult is the disambiguation of the lexical elements within the entries. The word *einen* may be interpreted as a determiner, a verb, or a cardinal number. Preprocessing can readily identify such placeholders, whose strict usage within Schemann limits their ambiguity. Significantly more difficult, however, is the resolution of the scope of the `/` operator. It is not clear in example 2 whether the resolution should be *[ein absolutes]/ [das absolute] Gehör haben* or *ein [absolutes/das] absolute Gehör haben*.

These examples illustrate that multi-word entries of a paper dictionary cannot be extracted as simple strings for the purposes of NLP.

2.2. Automatic pre-classification

Given the 32,000 entries and an average length of 4.3 tokens for each entry, manual association of each token to its part-of-speech (POS) tag would not be feasible. On the other hand, parsing the entries of an idiom dictionary is not equivalent to parsing naturally occurring text, since the entries follow a formalized structure that is quite different from that of ordinary language. The tokens are generally lemmatized, alternatives are marked with slashes and many entries themselves do not constitute full sentences. This makes the entries unsuitable for processing with existing taggers.

We defined a small tagset of 10 lexical categories (ADJ, ADV, CONJ, DET, N, NA, PREP, PRON, PTK, V), tagged approximately 6,000 dictionary entries by hand and called this as our training corpus. Next, we defined a model of plausible POS-patterns **S** inductively as the union of the following POS sequences: **S** $:= S_0 \cup S_1 \cup S_2 ... \cup S_7$.

ID	POS-sequence	description
S_0	Pron.Det.V ...	specific sequences
S_1	$NP := (N\|Det.N\|...)$	atomic phrases
$S_{1'}$	$P := (NP\|PP\|AdjP)$	atomic phrases
S_2	$P_n := P.(P)+$	atomic combinations
S_3	$P.V$	atoms and verbs
S_4	Conj.P := $Conj.P_n$	conjunction phrases
S_5	AdvP_n	adverbial phrases
S_6	AdvP.ConjP.P_n	combinations of $S_{4,5}$
S_7	$K_{1,2,4...6}.VP$	complex combinations

Table 1: main construction classes

We then trained the Brill Tagger (Brill, 1994) on these entries and applied the tagger to the remaining 26,000 entries. In the four examples above, the examples (1), (3) and (4) can be tagged successfully.

- example (1): ↦ det.V [1]
- example (3): ↦ Prep NA V
- example (4): ↦ Det N Prep N V

Generally, the tagging process is limited either in cases with lexicon gaps or in cases where a dictionary entry corresponds to more than one sequence. For example, in (2), some additional pre-processing is necessary because of the above mentioned scope problem of the `/` operator. As a result, this process assigned categories to 86% of the entries with 88% accuracy (Geyken and Boyd-Graber, 2003).

It is clear that the regularity of the normalized patterns of multi-word units in the dictionary contributes to this success. Remarkably, our approach worked well without relying on any language-specific linguistic resources other than the training corpus.

2.3. Fine-grained manual classification

The above-mentioned association of the computed POS sequences with predefined construction classes yields a satisfactory recall of 86%. Our decision to extend the pre-classification was motivated by the need to account for the remaining 14% of unrecognized sequences and by the lack of precision for complex entries (such as (2) above), but also by the fact that the initial tag set was sometimes too simple, as the following example illustrates:

(5) sich$/_{pron}$ ein$/_{det}$ paar$/_{pron}$ Tränen$/_N$ abquetschen$/_V$
oneself a few tears squeeze
to squeeze out a few tears

Here, the sequence ´pron.det.pron.N.V´ is not specific enough. It should instead be re-encoded as ´reflexive-pron.det.indefpron.N.V´.

A fine-grained manual classification should take into account these different pronouns, it should also make a difference between definite and indefinite articles or a difference between auxiliary, modal and full verbs. Therefore, we decided to re-encode the entries with a richer tag set. A very commonly used tag set for German is the STTS-tag set (Schiller et al., 1995). This tag set forms the basis for most German POS taggers. Thus, this step is intended to make the dictionary entries matchable with the text data.

The manual re-classification of the entries according to the STTS tag set was carried out by taking advantage of the pre-classification in the following way. Simple entries such as adjective-noun sequences can be re-written via a straightforward translation of our initial simple tag sets into the STTS tag sets. The same procedure applies to the re-encoding of simple noun phrase sequences followed by a simple verb.

Hence, rather than describing the multi-word entries in alphabetic order, we took advantage of the pre-classification step and ordered the entries according to their POS-sequence. Encoding multi-word expressions ordered by POS sequences is not only faster but also less error

[1] In this example the tagger disambiguates the entry. The other possible sequence V.V with 'einen' as the infinitive 'to unite' would be meaningless.

prone. Pre-classification helped us to organize encoding projects too: the validation of simple patterns like adj_N or prep_det_N can be done by novices, whereas the validation resp. encoding of more complex patterns, such as those containing verbs or predicative complements, are left to more experienced linguists.

Compared with complete manual annotation, this method is considerably more efficient. Given an average of 25 encoded entries per hour, manual tagging of the training set required about 80 hours. Tagging all 32,000 entries of the Schemann dictionary at the same rate would take more than 1,280 hours. Our method required training the Brill tagger (60 hours), correction of the imprecise or wrong results (470 hours), plus the time required for the manual tagging of the training set (80 hours). This totals 610 hours, which amounts to a time saving of about 50%.

3. A Database of MWE

Our manual processing yielded a database currently containing 27,000 entries with the following types of information: the citation form of the entry, a POS sequence according to STTS-tagset and one or more sequences of word/POS-tag pairs for each entry.

It is interesting to look at the frequency distribution of the idiom entries according to their POS-sequence. If we extract all distinct POS-sequences we obtain more than 4000 different patterns, of which almost half occur only once. Even though lexical substitutions and syntactic alternations account largely for this striking number of distinct idiom patterns, the result clearly demonstrates that multi-word expressions cannot be reduced to a limited set of syntactic patterns.

An overview of the most frequent patterns is given in the following table where the patterns are sorted by frequency. In order to facilitate the readability, the patterns are expressed in the simple tag set.

pattern	**n. of idioms**
$NA.Det.N.V$	1788
$NA.Prep.N.V$	1433
$NA.Prep.Det.N.V$	1349
$NA.N.V$	994
$NA.Det.Adj.N.V$	963
$NA.Prep.N.V$	879
$NA.Prep.Det.N.V$	849
…	…
$Det.Adj.N$	427
…	…
$Pron.Det.N.N.V$	2

Table 2: frequencies of construction classes

Generally, each of the above mentioned patterns corresponds to more than one POS-pattern expressed in the STTS-tag set. For example, the most frequent pattern $NA.Det.N.V$ (1788 entries) is decomposed in the database into the following subpatterns:

- Subclass $ARTDEF$: die Kontenance verlieren (to lose one's self control)
- Subclass $ARTINDEF$: eine Regel aufstellen (to establish a rule)
- Subclass: $PINEG$: keinen Zug vertragen (not to be able to stand drafts)
- Subclass: $PPOS$: seine Leute kennen (to know your own people)
- Subclass: $PPOS|ARTDEF$: das/sein Maul halten (to shut up)

In these examples the determiner of the POS-pattern $NA.Det.N.V$ is expressed as a definite article (ARTDEF), an indefinite article (ARTINDEF), a negative indefinite pronoun (PINEG), a possessive pronoun (PPOS), or a disjunction of a possessive pronoun and a definite article.

Some POS-patterns of very well known idioms are quite rare. For example, the well known idiom 'das ist des Pudels Kern' (that's the heart of the matter - PDPRON DET N N VK) has only one pattern equivalent: 'das ist des Rätsels Lösung' (that's the solution of the enigma).

Obviously the database itself inherits some of the limitations of Schemann's dictionary, the most important being that the dictionary entries mix core and contextual components without explicit mark-up. This is not necessarily a problem for a print dictionary since the context may facilitate the reader's understanding and usage of the idiom. It is a problem, however, for linguistic processing. For example, the idiom 'über den Durst trinken' is encoded in the dictionary as 'ein Glas über den Durst trinken' (cf. example 4 above). The first noun phrase 'ein Glas' is merely a contextual, not a core component of the idiom: it receives a literal interpretation and can be replaced by semantically similar nouns. The distinction between core and contextual idiom components is not indicated in the dictionary, which presents an obstacle to further linguistic and computational processing, as we will see in the next section.

4. Populating the database with corpus examples

The second phase of our project, currently underway, addresses the association of a sequence of POS-tag/token pairs to a corpus example. To this end, we generate a finite state transducer from the POS/token sequence for each entry and we associate weights to each POS/token pair depending on the POS-pattern. The weights have to be adjusted manually for each POS-pattern whereas the local grammar itself can be generated automatically from the database of multi-word expressions. A finite state filter, e.g. (Karttunen et al., 1996; Senellart, 1998), is then applied to the corpus. This filter is supposed to match the local grammar with those sequences of the corpus that correspond to the longest match of the multi-word expression.

This method is straightforward for multi-word expressions with short patterns such as $adj.noun$ or $prep.N.V$. Here one would generate a grammar of the form $'\$adj.\$noun'$ where the $ stands for the lemma operator, meaning that all word forms belonging to the morphological paradigm of the word are accepted. The second case is more complex since the verb can occur either to the left or to the right of the PP. Hence, one could generate a grammar that detects all '$prep.\#1.\$N$'[2], with the verb at a certain distance. In both cases, however, it is comparatively safe [3] to suppose that the entry consists only of core components. Hence, all the components of the idiom have to be present in the text, even though the local grammar has to take into account the fact that the order of the pattern components can change and that the verb can consist of a separable prefix.

As stated in the previous section, the entries in Schemann's dictionary contain core components as well as contextual elements. Furthermore, multi-word expressions may occur in contexts that differ significantly from the citation form. One has to take these factors into account by admitting shorter sequences than the one given in the citation form as instances of a given multi-word expression. This can be realized by setting pattern-specific thresholds.

For example, we have shown above that the first NP of example (4) can be omitted since it is not a core component of the idiom 'über den Durst trinken'. Also, it is not sure if the definite determiner is always realized in instances of the idiom. We can associate weights to all lexical categories of the pattern: a weight of '2' to the determiners, a weight of '5' to the first noun and the verb, and a weight of '10' to the preposition and the second noun (cf. (6)). This corresponds to the intuition that the most characteristic parts of all multi-word expressions of the pattern $det.N.prep.det.N.V$ are the preposition and the noun in the second NP.

(6) ein/$_{det}$ Glas/$_N$ über/$_{prep}$ den/$_{det}$ Durst/$_N$ trinken/$_N$
2 5 10 2 10 5
to have one (drink) too many

We set a threshold of '22': only those patterns where the sum of the weights exceeds the threshold are considered as instances of the idiom. With this threshold, the following sequences are correctly matched (the sum of the weights are marked in brackets): 'ein Glas über den Durst trinken' $< 34 >$, 'einen über den Durst trinken' $< 27 >$, 'becherte ... über den Durst' $< 22 >$.

On the other hand, the following sequences below the threshold are correctly rejected because they correspond to the literal interpretation: 'über einen Durst' $< 20 >$, 'ein Glas, um den Durst zu löschen' $< 19 >$, 'ein Glas über die Pflanze gießen' $< 17 >$, or 'mehr trinken ... als man Durst hat' $< 15 >$.

Further work is planned along the following lines. First, the weights associated with the POS-patterns in the database of multi-word expressions as well as the corresponding threshold have to be experimentally evaluated and validated. If either the weights or the thresholds are not appropriate, numerous examples of these lists generated by the transducer have to be discarded since they are not examples of the idiomatic expression but strings with literal interpretations. Also, the word distances in finite state transducer generated from the token/POS-sequences have to be assessed. In both cases, evaluation will be done on the basis of a manually annotated corpus.

Second, evidence from the corpora will lead to a modification of the description of the database. Variations of the multi-word expressions found in the corpora will lead to a new mark-up of the database entry where core components and contextual elements are distinguished. This work on the basis of a careful manual annotation of the 1 billion word corpus of the DWDS (www.dwdscorpus.de) is currently carried out in the aforementioned idiom project on the basis of a subset of 2,000 verb-NP and verb-PP entries (www.bbaw.de/forschung/kollokationen/).

5. Acknowledgements

This work is supported by the Wolfgang-Paul Preis from the Alexander-von-Humboldt Foundation to Christiane Fellbaum. The author would also like to thank the linguists who described and annotated many of the complex entries of the competence base, namely R. Finkbeiner C.Hümmer, K. Krell, and E. Stathi, as well as C. Bohn, A. Chichlova and C. Niederhaus for their annotation work.

6. References

Brill, E., 1994. Some Advances in Transformation-Based Part of Speech Tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*.

Geyken, A. and J. Boyd-Graber, 2003. Automatic Classification of Multi-Word Expressions in Print Dictionaries. *Linguisticae Investigationes*, XXVI:2.

Geyken, A., A. Sokirko, I. Rehbein, and C. Fellbaum, 2004. What is the optimal corpus size for the study of idioms? In *DGfS-Jahrestagung, Mainz 25.-27.02.2004*.

Karttunen, L., J.-P. Chanod, G. Grefenstette, and A. Schiller, 1996. Regular Expressions for Language Engineering. *Natural Language Engineering*, 2(4):305–328.

Oakes, M., 1998. *Statistics for Corpus Linguistics*. Cambridge University Press.

Schemann, H., 1993. *Deutsche Idiomatik. Die deutschen Redewendungen im Kontext*. Pons.

Schiller, A., A. Teufel, S. Stöckert, and C. Thielen, 1995. Vorläufige Guidelines für das Taggen deutscher Textcorpora mit STTS. Technical report, IMS, Univ. Stuttgart and SfS, Univ. Tübingen.

Senellart, J., 1998. Reconnaissance automatique des entrées du lexique-grammaire des phrases figées. *Le lexique-grammaire. Travaux de Linguistique*, 37:109–127.

Smadja, F., 1992. Retrieving collocations from text:Xtract. *Computational Linguistics*, 19(1):143–177.

[2] #1 means that there is a distance of 1 between $prep$ and N.

[3] one could weaken this condition here too; we will deal with this problem in the next paragraph.