Multiword Expression Identification for German

Markus Egg <markus.egg@anglistik.hu-berlin.de>
Will Roberts <will.roberts@anglistik.hu-berlin.de>
Valia Kordoni <evangelia.kordoni@anglistik.hu-berlin.de>

WG 1: Lexicon-Grammar Interface

We present ongoing work in a project to automatically identify, describe, and analyse German multiword expressions (MWEs). Specifically, we are interested in verbal constructions with PPs, they combine a particular verb with a PP consisting of a preposition and its nominal argument (hence, such expressions are sometimes termed PNVs, for preposition-noun-verb):

(1) jemandem *zur Verfügung stehen* (‘*to be at* someone’s *disposal*)’

(2) jemandem etwas *zur Verfügung stellen* (‘*to put* something *at* someone’s *disposal*)’

(3) etwas *in Anspruch nehmen* (‘*to make use of* something’)

PNVs may be compositional to some degree: They may allow adverbial modification of the verb, adjectival or genitive modification of the prepositional argument, or there may be other variation (such as the presence or absence of an article to the prepositional argument). E.g., the noun *Verfügung* ‘disposal’ is accompanied by the definite article (fused with the preposition in *zur*), and could be modified by *frei* ‘free’. Modification is obligatory for some PNVs (e.g., *auf* jmds. *Kosten kommen*, ‘to get one’s money’s worth’); here, the noun *Kosten* ‘expense’ needs a genitive noun phrase or a personal pronoun to express the beneficient.

The PNV constructions themselves are verbal in nature and have predictable subcategorisation. For example, *zur Verfügung stehen*, above, is intransitive, while *zur Verfügung stellen* is ditransitive; this pair also illustrates another interesting property of PNVs, which is that they can be found in pairs representing various kinds of verbal alternations. The pair (2) and (3) are a case of a causative alternation, since *stellen* (‘to put’) is the causative of *stehen* (‘to stand’).

We collect our data from SdeWaC (Faaß and Eckart, 2013), a 880-million word corpus of German text assembled from Web search results, which we have automatically parsed using an unlexicalised statistical parser (Petrov et al., 2006). We employ a simple rule-based system to extract instances of verbs with prepositional phrase complements.

To identify MWEs, we represent PNV candidates as triples consisting of the verb, the preposition, and the nominal head of the prepositional argument. We automatically lemmatise verbs, but retain the unlemmatised forms of nouns (indicating case and number) and prepositions (fused with article or not). Another important piece of information is the kind of article that accompanies the noun (definite, indefinite, or none). These features are recorded because most PNVs allow only a very specific combination of them, in fact, such restrictions are good indicators of MWE status. E.g., in (3), neither the number of the noun nor the absence of any article can be varied.[1]

We count the frequency of each PNV candidate in the corpus, along with the frequency of each verb in the corpus, and the frequency of each PP (represented as a combination of preposition + noun) complement to a verb in the corpus. This results in counts for 2.9 million PNV instances. Using the observed counts for verbs, PPs, and PNVs, we are able to rank the PNV candidates

[1] Sometimes case information is needed to distinguish quite systematic ambiguous preposition readings as in *auf* ‘on’ with dative and ‘onto’ with accusative.

using standard MWE association measures. In our work to date, we have been most successful with the Piatetsky-Shapiro association measure $P(A,B) - P(A)P(B)$ (Piatetsky-Shapiro, 1991).

For development, we use the `German_Krenn_PNV` data set (Krenn, 2000)[2], a list of 21,000 German verb-PP combinations extracted from a corpus of newspaper text. These expressions have been manually annotated as being either lexical collocations or not; lexical collocations (which number 1,149) are further classified as either idiomatic (*im Mittelpunkt stehen*, 'to take centre stage') or as support verb constructions (such as *zur Verfügung stellen*, above). Inter-annotator agreement on this data set was calculated to be 75% using linguistically trained annotators. This data set allows us, for example, to evaluate the relative efficacy of different association measures for ranking our PNV candidate list.

Our preliminary findings to date are:

1. The presence or absence of an article to the prepositional argument is a binary feature of PNVs: PNVs tend to either occur always with an article, or never with one. The presence of an article is a weak indicator of MWE status.
2. The majority of PNVs disallow adjectival and genitive modification of the prepositional argument.
3. Most PNVs take a prepositional argument which has a noun head (as opposed to a pronominal form or named entity).
4. Precision is a challenge for the automatic detection of PNV. E.g., our top 500 list of PNV candidates has a precision of 75.2%. False positives are downright wrong or compositional. But we have found that entropy (with how many verbs does the PP co-occur) is a good measure for the likelihood of being compositional.

In future work, we will continue to develop further statistical description of the compositional parameters and syntactic behaviour of PNVs. We then plan to perform semantic analysis of PNVs using paraphrases.

References

Gertrud Faaß and Kerstin Eckart. SdeWaC - A corpus of parsable sentences from the Web. In *Language processing and knowledge in the Web*, pages 61–68. Springer, Berlin, Heidelberg, 2013.

Brigitte Krenn. *The usual suspects: Data-oriented models for identification and representation of lexical collocations*. PhD thesis, Saarland University, 2000.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics, 2006.

Gregory Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In Gregory Piatetsky-Shapiro and William Frawley, editors, *Knowledge Discovery in Databases*, chapter 13, pages 229–238. MIT Press, Cambridge, MA, 1991.

[2] `http://sourceforge.net/apps/trac/multiword/wiki/German_PNV_Krenn`