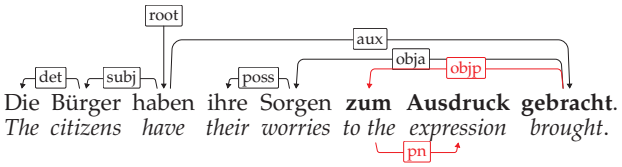
Multiword expression identification for German: Statistical description of PNV compounds

Markus Egg, Will Roberts and Valia Kordoni
Department of English and American Studies, Humboldt-Universität zu Berlin

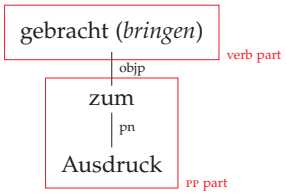
Introduction

- Preposition-Noun-Verb constructions (PNVs) are a class of multiword expressions (MWEs) consisting of a verb dominating a prepositional phrase (PP) with a nominal argument.
- These can be idiomatic (1, 2), or light verb constructions (3):
  - (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 *zum Ausdruck bringen* ('*to express* something')
- PNVs may contain an article on the PP argument (which may be fused with the preposition, as with *zur* and *zum*, above).
- Variable compositionality: PNVs may allow modification (adverb modifying the verb, or adjective or genitive modifying the noun). Modification may be obligatory.
- The free word order and verb-second sentence structure in German tends to render *n*-gram methods inapplicable:



Method

- We use SdeWaC (Faaß and Eckart, 2013), a 880-million word corpus of German text assembled from Web search results.
- Automatic preprocessing:
  - POS tags and lemmas from the TreeTagger
  - parsed using the unlexicalised statistical Berkeley Parser (Petrov et al., 2006)



- We collect statistics on the verbal part (lemmatised) and on the PP part (preposition and PP argument, unlemmatised).
- We identify PNVs using the Piatetsky-Shapiro (1991) association measure: $P(A,B) - P(A)P(B)$.
- 2,886,132 PNV types recorded in corpus:

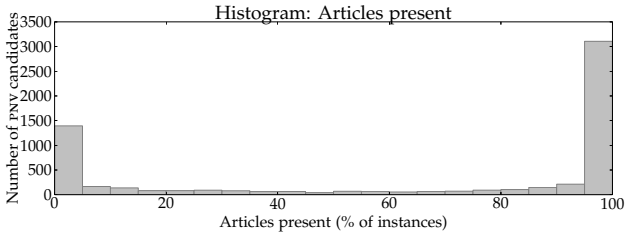
Ranked PNVs			Correct?	(Krenn, 2000)?
zur	Verfügung	stehen	✓	✓
zur	Verfügung	stellen	✓	✓
in	Lage	sein	✓	
in	Anspruch	nehmen	✓	✓
im	Mittelpunkt	stehen	✓	✓
mit	sich	bringen	✓	
von	Bedeutung	sein	✓	
in	Kraft	treten	✓	✓
zum	Ausdruck	bringen	✓	✓
auf	Weg	machen	✓	✓
zur	Folge	haben	✓	
zur	Kenntnis	nehmen	✓	✓
vor	allem	sein		
ins	Leben	rufen	✓	✓
in	Ordnung	sein	✓	
im	Rahmen	werden		
etc.				

Diathesis alternations

- PNVs are verbal in nature, and have subcategorisation preferences:
  - *zur Verfügung stehen* is transitive
  - *zur Verfügung stellen* is ditransitive
- PNVs can have alternations or appear in semi-regular verb "families":
  - *zur Verfügung stehen* (stative)/*stellen* (causative)
  - *zum Ausdruck kommen* (inchoative)/*bringen* (causative)
  - *in Ordnung sein* (stative)/*bringen*/*halten* (causative)

Statistical description of PNVs

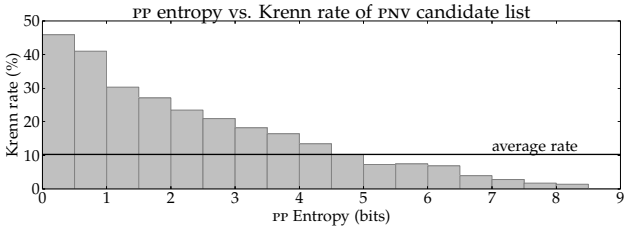
- Question: Do PNVs vary in whether they appear with articles?
  - That is, does a given PNV sometimes appear with an article and sometimes without?
- We take the best 6,474 PNV candidates from the ranked list.
- For each PNV candidate, we collect all instances observed in the corpus (mean = 77).
- We measure what percent of these instances appear with an article, and plot a histogram over all 6,474 candidates.



- Finding: Presence of article is a binary feature
  - PNVs either occur with an article always, or never.

PP entropy

- Problem: Our ranked list of PNVs still includes compositional constructions or idiomatic PPs with common verbs (e.g., *im Rahmen* 'in the context of').
- Intuition: PNVs are likely to feature PPs which co-occur with relatively few verbs.
- We use the 1,149 PNVs manually listed by Krenn (2000) as a gold standard.
- We define the *Krenn rate* as the precision of subsets of the 6,474 candidates.
- PP entropy: the degree of uncertainty in predicting the verb which dominates a given PP.



- Finding: PP entropy is a promising indicator of PNV status.

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 *COLING-ACL*, pages 433–440, 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.