Automatic Construction of Morpho-Syntactic Descriptions of Verbal MWEs from Crowdsourceable Input

Simon Clematide and Marc Luder

Vision

Let humans, NLP tools, corpora and automated corpus queries do what they can do best.

Problem A

Manual creation of exact and detailed morpho-syntactic descriptions of MWEs is tedious.

Solution

Use automatic linguistic analyzers such as morphological analyzer, PoS tagger, and parsers.

Remarks

Dependency parsing is particularly suitable for these needs because it restricts the syntactic structure to dependencies (syntactic relations). There is no need to introduce phrasal categories or to rely on theory-specific phrase structures.[4]

Problem B

Traditional lexicographic entries typically use an infinitive form to represent a construction.

For instance, verbal MWE entry from [1]:

German: *ins kalte Wasser geworfen werden*
English: *to be thrown in at the deep end*

Remarks

This does not represent the typical inflected use of a construction in running text. Especially, subjects are not explicitly represented.
NLP tools trained on running text may fail to correctly analyze this type of language.

Solution

Create simple, inflected sentences with **special tokens** representing fully variable parts of a construction .

German: *Weil jmd von jmdm ins kalte Wasser geworfen wird.*
('because sb was thrown in at the deep end by sb')

Remarks

For German we adapted a PoS tagger and morphological analyzer to provide case-disambiguated analyses for (artificial) tokens like "jmdn" (someone, accusative), "etwasm" (something, dative), etc.
Language-specific canonical forms are needed; German sentences starting with "weil" (*because*) have nice properties concerning the separable verb prefixes.

Real-world sentences contain more than just the construction.

German: *Dennoch wurde sie gleich bei der ersten Kundin ins kalte Wasser geworfen.*
('Nevertheless, she was thrown in at the deep end at the first customer.')
(Source: NZZ, 2013/12/11)

Solution

Non-experts are able to simplify the real-world sentence to a canonical form. This task can be crowd-sourced.

Problem C

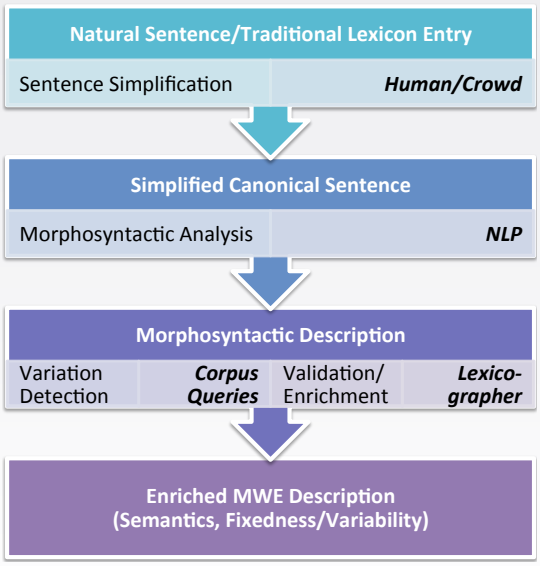
The variability of a construction should be recorded on the basis of empirical corpus-linguistic data.

Solution

A natural view: "Constructions are morpho-syntactic corpus queries."

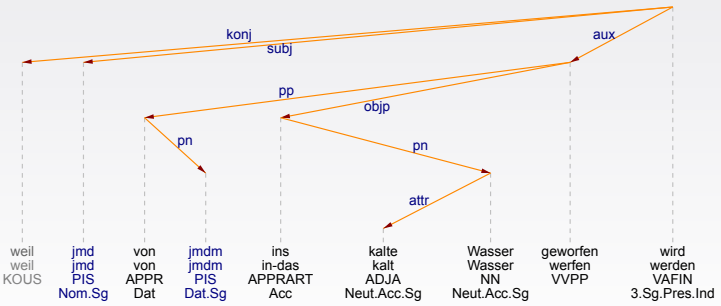
Remarks From the automatically created morpho-syntactic descriptions, we can automatically build corpus queries for the constructions that test selected constructional properties: modifiability, negation, passivization, optionality, slot filler collocations.
Appropriately preprocessed (lemmatization, PoS tagging, parsing) and indexed large corpora are necessary as well as an expressive and efficient query language.

System Architecture



The German psycho-conceptual lexicon "Jakob" [3, 2] used for psychological text analysis contains about 1,300 verbal MWEs built according this schema.
See `http://www.jakoblexikon.ch/lexikon`

Example Analysis



The ParZu parser [5] patched for special tokens can be tested via `http://kitt.cl.uzh.ch/kitt/jakob/lex.html`.

References

[1] Deutsch-Englisches Wörterbuch. http://www.dict.cc, 2013.

[2] Marc Luder. German verb patterns and their implementation in an electronic dictionary. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[3] Marc Luder and Simon Clematide. Constructing a constructional MWE lexicon for psycho-conceptual annotation: Evaluation of CPA and DuELME for lexicographic description. In *Proceedings of the XIV Euralex International Congress*, pages 402–410, Leeuwarden, NL, 2010.

[4] Joachim Nivre. Theory-supporting treebanks. In *Proceedings of The 2nd Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 117–128, Växjö, Schweden, 2003.

[5] Rico Sennrich, Martin Volk, and Gerold Schneider. Exploiting synergies between open resources for German dependency parsing, POS-tagging, and morphological analysis. In *Recent Advances in Natural Language Processing (RANLP 2013)*, pages 601–609, September 2013.

PARSEME, Athens, 10-11 March 2014, WG 1

Contact

University of Zurich[UZH]

Institute of Computational Linguistics
`http://www.cl.uzh.ch`

Simon Clematide
Binzmühlestrasse 14, CH-8050 Zurich
simon.clematide@cl.uzh.ch