Automatic Construction of Morpho-Syntactic Descriptions of Verbal MWEs from Crowdsourcable 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 “jmdm” (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, optionalnality, slot filler collocations. Appropriately preprocessed (lemmatization, PoS tagging, parsing) and indexed large corpora are necessary as well as an expressive and efficient query language.

REFERENCES

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

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