Coping with unruly language: non-standard usage and phraseology in a corpus

Alexandr Rosen (Charles University, Prague)

In most cases, corpus annotation does not distinguish regular, expected, standard or compositionally interpretable expressions on the one hand from less predictable evidence of language use on the other. There are some exceptions: individual word forms (colloquial, dialectal or non-words) in mainstream corpora and error annotation in learner corpora, but a principled approach is not a common sight.

Non-standard usage and phraseology defy general rules of grammar: non-standard language may involve performance errors, creative coinages, emerging phenomena; multi-word expressions may have irregular (non-compositional) semantics, syntax, morphology, even pragmatics or phonology.

We start with the assumption that the text in a corpus and its linguistic annotation is where the two Saussurean sides of a single coin converge: the empirical evidence (language use, parole, performance, corpus) and the theory (language as a system, langue, competence, grammar). Moreover, the annotation is also where multiple levels of analysis and linguistic theories may meet and be explicit about any, even irregular phenomena. An annotation scheme defined as a formal grammar can help to identify the difference between the regular and irregular, between the language as a system and the use of language.

All of the above was the motivation behind a briefly introduced project of a Czech parsebank, a corpus annotated by standard stochastic tools and checked by a grammar and valency lexicon. This allows for tasks such as inferring additional linguistic information, checking both the data and the grammar and helping to formulate efficient queries.

The project is still under way. I will show the current coverage of the grammar and the lexicon in terms of linguistic phenomena, and also in terms of statistic results, based on the share of expressions that satisfy the definition of “correctness” as stated by the grammar and the lexicon.

In the final part of the talk I want to focus on two phenomena as treated in the project: analytical morphology as seen from the syntagmatic and paradigmatic perspective, and multi-word expression as a phenomenon at the border between grammar and lexicon. These two contrasts complement the inherent system/use opposition.