Towards a database of Czech multi-word expressions

Milena Hnátková and Alexandr Rosen

Institute of Theoretical and Computational Linguistics
Faculty of Arts
Charles University in Prague

PARSEME 2016
7th final general meeting
Dubrovnik
26–27 September 2016

Work done

- MWEs annotated in a 3.6 GW representative corpus of Czech
- Non-verbal, verbal, clausal, similes, proverbs, conjunctions
- Rule-based identification, based on a lexicon of 36K entries
- Technical solutions to variability of MWEs

Work in progress

- Discovery of additional MWEs using stochastic methods
- Mapping entries to the PARSEME taxonomy
- Conceptual and formal integration with a syntactic module
- Designing and building a MWE database