The Czech National Corpus: research infrastructure for empirical language-oriented inquiry

The Czech National Corpus: research infrastructure for empirical language-oriented inquiry

The Czech National Corpus (CNC) project, established in 1994, strives to continually map the Czech language in all available dimensions (from the time, regional and genre perspective). The CNC builds and makes available large electronic text collections (language corpora) serving as a basis for research on current Czech (both written and spoken) as well as historical Czech and other languages. It also develops the methodology of empirical linguistic research and tools for language corpora exploration.

Since 2012 the CNC has been recognized as a research infrastructure for empirical language-oriented inquiry in many fields of social sciences and humanities (esp. linguistics, psychology, sociology, history, NLP etc.). Thanks to its large and high-quality language resources the CNC is a sought-after partner in many international research projects. Besides these activities, CNC also focuses on consulting, providing analyses for research or popularizing purposes, providing data packages for research on Czech as well as other languages for contrastive research, and automatic text processing.

Key collaborators

- Anna Čermáková
- Michal Křen
- Karel Kučera
- Vladimír Petkevič

Selected outputs

- Čermáková, A., Chlumská, L., Malá, M. (eds): Jazykové paralely. NLN. Praha 2016.
- Petkevič, V.: Morfologická homonymie v současné češtině. NLN. Praha 2014.
- Čermák, F. – Křen, M. (eds): *A Frequency Dictionary of Czech: Core Vocabulary for Learners*. Routledge, London 2011.