Czech National Corpus Listed as a CLARIN K-Centre

On December 4, 2018, Czech National Corpus (CNC) has been officially approved as one of the K-centres of CLARIN, a European infrastructure (ESFRI) focused on language resources and tools for Humanities and Social Sciences. The aim of the CNC K-Centre is to provide information, consulting and technical assistance in the area of corpus linguistics with the emphasis on Czech.

This official recognition is also a confirmation of the merits of the CNC project that is being fostered since 1994 by the Institute of the Czech National Corpus (ICNC), Faculty of Arts, Charles University. The main aim of the ICNC is continuous mapping of Czech in its many forms (written and spoken) by the means of building language corpora and making them publicly available. Apart from the data collection, ICNC also focuses on corpus linguistics methodology and develops user interfaces for working with the corpora.

In 2011, CNC was included on the Roadmap of the Large Research Infrastructures of the Czech Republic, which resulted in strengthening its orientation towards user services, namely the launch of the korpus.cz web portal with integrated user support including an on-line helpdesk and knowledge base. As a result, CNC currently has more than 6,500 registered active users from the Czech Republic and abroad, with the average number of user queries exceeding 2,500 per day. There are also many scientific outputs and theses based on the CNC resources, the project registers almost 300 of them only in 2017.

Language corpus is an electronic collection of authentic texts that is designed to represent the language use. Corpora typically feature rich annotation both on the level of texts (bibliographical information and other metadata) and

individual words (lemmatization, morphological and/or syntactic annotation). Corpora typically serve as large bases of data for empirical research, especially in linguistics, but also in other fields (literary science, psychology, sociology, history etc).