58. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 15. bis 17. März 2022

# KORPORA IN DER GERMANISTISCHEN SPRACHWISSENSCHAFT – MÜNDLICH, SCHRIFTLICH, MULTIMEDIAL

Dienstag, 15. März 2022, 15:20 Uhr

## Czech National Corpus: long-term language mapping at your service

Michal Křen, (Universität Prag)

### Introduction

The mission of the Czech National Corpus (CNC) is a long-term, extensive and continuous mapping of the Czech language. For this purpose, CNC systematically collects, processes and annotates language data to produce large general-purpose corpora of Czech with the aim to offer a diverse and high-quality language resources for empirical research. In addition to corpus compilation, CNC is also very active in the development of specialized analytical tools in the form of web-based user applications that enable effective and intuitive work with the language data.

This contribution aims to give an overview of the CNC project (LM2018137; funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures) and to present the main achievements and challenges in the domains of corpus compilation and application development.

### Corpus compilation

Most of the CNC corpora can be characterized as traditional, with emphasis on well-defined composition and reliable metadata. The main data collection areas:

- Contemporary printed Czech: annually updated SYN-series corpora (current overall size 4.7G words, i.e. tokens not including punctuation) that include five 100M-word representative corpora published every five years that cover consecutive time periods;
- Contemporary spoken Czech: spontaneous informal conversations covered by the ORTOFON corpus (2.1M words) and semi-formal monologues covered by the ORATOR corpus (1.2M words);
- Contemporary web Czech: ONLINE monitor corpus of Czech internet media (more than 6G words) updated on a daily basis and complemented by the NET corpus of semi-formal internet communication (discussion forums and blogs; 176M words).

Other data collection areas include multilingual InterCorp parallel corpus (1.5G words), DIAKORP corpus of historical Czech (3.5M words) and DIALEKT corpus of Czech dialects (223k words).

**Application development**

Currently, there are thirteen CNC-developed web applications that view corpus data from various perspectives. All the applications are integrated into the CNC research portal at http://www.korpus.cz/ together with many-faceted user support and related services. This contribution concentrates on selected applications that can be attractive also for linguists who work with languages other than Czech:

- KonText: general-purpose concordancer that supports various corpus types, including spoken and parallel corpora;
- Word at a glance: user-friendly word profiles based entirely on corpus data;
- Calc: corpus calculator divided into modules that cover typical statistical tasks commonly encountered in corpus research.

CNC has an established user community of more than 8,000 registered active users from the Czech Republic and abroad. In 2021, the average number of user queries exceeded 4,000 per day.