Czech National Corpus in 2020: Corpora and Applications

Czech National Corpus (CNC) is a long-term project striving for extensive and continuous mapping of the Czech language by compilation, maintenance and providing public access to a range of various corpora to support empirical linguistic research. CNC is also very active in creating user applications for working with corpora integrated into the CNC web portal at http://www.korpus.cz/. This contribution will discuss the main CNC achievements in corpus compilation and application development. In both domains, it will aim at an overview of the recent development supplemented by a brief outline of future plans.

Overview of the main corpus compilation areas:

- Contemporary written Czech: annually updated SYN-series corpora (4.5 G), including four 100-million representative corpora that cover consecutive time periods.
- Contemporary spoken Czech: spontaneous informal conversations: ORAL v1 corpus (5.4 M) and the new-generation ORTOFON corpus (two-tier transcription, full balance, 1 M), semi-formal monologues: ORATOR corpus (580 k).
- Multilingual parallel corpus InterCorp with Czech texts aligned on sentence level with their translations to or from 40 languages (annual updates, 1.73 G).
- Specialized corpora: DIAKORP corpus of historical Czech, DIALEKT dialectal corpus, Koditex corpus for the analysis of register variation in Czech, NET corpus of semi-formal internet communication etc. NET will soon be supplemented by a corpus of Czech web media (including social networks) updated on a daily basis.

The development of digital humanities and emphasis on empirical methods in linguistics creates significant demands on the development of user-friendly web applications. Until recently, there were five such web applications (KonText, SyD, Morfio, KWords, Treq). In the last year, they were supplemented by another five brand new applications (Word at a glance, Calc, Lists, Mapka, KorpusDB) that will be focused on in the presentation.

References

Cvrček, V., Čermáková, A. and Křen, M. (2016). Nová koncepce synchronních korpusů psané češtiny. *Slovo a slovesnost* 77(2): 83–101.

Cvrček, V. and Vondřička, P. (2011). Výzkum variability v korpusech češtiny. In F. Čermák (Ed.), *Korpusová lingvistika Praha 2011. 2. Výzkum a výstavba korpusů*. Praha: NLN, pp. 184–195.

Cvrček, V. and Vondřička, P. (2012). Nástroj pro slovotvornou analýzu jazykového korpusu. In *Gramatika a korpus 2012*. Hradec Králové: Gaudeamus.

Čermák, F. and Rosen, A. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13 (3): 411–427.

Goláňová, H. and Waclawičová, M. (2019). The DIALEKT corpus and its possibilities. *Jazykovedný časopis* 70(2): 336–344.

Hnátková, M., Křen, M., Procházka, P. and Skoumalová, H. (2014). The SYN-series corpora of written Czech. In *Proceedings of LREC 2014*. Reykjavík: ELRA, pp. 160–164.

Komrsková, Z., Kopřivová, M., Lukeš, D., Poukarová, P. and Goláňová, H. (2017). New Spoken Corpora of Czech: ORTOFON and DIALEKT. *Jazykovedný časopis*, 68(2): 219–228.

Kopřivová, M., Lukeš, D., Komrsková, Z. and Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. *Korpus – Gramatika – Axiologie*, 15: 47–67.

Křen, M. et al. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of LREC 2016*. Portorož: ELRA, pp. 2522–2528.

Kučera, K. and Stluka, M. (2014). Corpus of 19th-century Czech Texts: Problems and Solutions. In *Proceedings of LREC 2014*. Reykjavík: ELRA, pp. 165–168.

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. In *Proceedings of LREC 2020*. Marseille: ELRA. (in press)

Machálek, T. (2020). Word at a Glance: Modular Word Profile Aggregator. In *Proceedings of LREC 2020*. Marseille: ELRA. (in press)

Rosen, A. and Vavřín, M. (2012). Building a multilingual parallel corpus for human users. In *Proceedings of LREC 2012*. İstanbul: ELRA, pp. 2447–2452.

Škrabal, M. and Vavřín, M. (2017). The Translation Equivalents Database (Treq) as a Lexicographer's Aid. In: *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pp. 124–137.

Zasina, A. and Komrsková, Z. (2019). Koditex – korpus diverzifikovaných textů. *Studie z aplikované lingvistiky*, 10(1): 127–132.