Proceedings of he 8th Workshop on Challenges in the Management of Large Corpora, pages 52–57
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Czech National Corpus in 2020: Recent Developments and Future Outlook

Michal Křen
Institute of the Czech National Corpus
Faculty of Arts, Charles University
Nám. Jana Palacha 2, 116 38 Prague, Czechia
michal.kren@ff.cuni.cz

Abstract

The paper overviews the state of implementation of the Czech National Corpus (CNC) in all the main areas of its operation: corpus compilation, annotation, application development and user services. As the focus is on the recent development, some of the areas are described in more detail than the others. Close attention is paid to the data collection and, in particular, to the description of web application development. This is not only because CNC has recently seen a significant progress in this area, but also because we believe that end-user web applications shape the way linguists and other scholars think about the language data and about the range of possibilities they offer. This consideration is even more important given the variability of the CNC corpora.

Keywords: language infrastructures; national corpora; corpus compilation; application development; user services

1. Introduction

Czech National Corpus (CNC) is a long-term project that strives for extensive mapping of the Czech language. This effort results mostly in compilation, maintenance and providing public access to a range of various corpora with the aim to offer a diverse, representative, and high-quality data for empirical research mainly in linguistics. An important point here is the continuity of the data collection that enables researchers to carry out longitudinal studies of language development or to study changes of public discourse in different time periods.

Apart from the corpus compilation, CNC is also very active in creating web applications for working with corpora, as well as in providing user support and all kinds of related services integrated into the CNC web portal at http://www.korpus.cz/.

CNC has an established and growing user community of more than 8,000 registered active users from the Czech Republic (ca 76 %) and abroad (ca 24 %). In 2019, there were on average 3,164 user interactions per day. An interaction is understood here as entering a query into one of the CNC web applications; any further work with the query results is not counted in this number, as well as any other interaction with the CNC web portal.

This contribution builds on the paper presented at CMLC 3 (Křen, 2015) and discusses the main CNC achievements since then. It gives an overview of recent developments in the given domains supplemented by an outline of future plans.

2. Corpus Compilation

The most of the CNC corpora can be characterized as traditional, with emphasis on well-defined composition, reliable metadata and high-quality data processing. The following gives an overview of the main data collection areas:

• Contemporary written (printed) Czech is covered by the SYN-series corpora (Hnátková et al., 2014). Every year, the series is updated with ca 150 million running words (i.e. tokens not including punctuation) of fresh data, mostly newspapers and magazines, so its overall size now reaches 4.5 billion running words. In addition, a 100-million representative corpus is selected from the SYN-series data every five years. Starting with SYN2000, there are now four such representative corpora, with the fifth one, SYN2020, to be published by the end of 2020. All these representative corpora contain a large variety of fiction, non-fiction, newspapers and magazines, with detailed bibliographic and register annotation (Cvrček et al., 2016; Křen et al., 2016), and thus continuously map the Czech printed production by covering consecutive time periods.

• Contemporary spoken Czech can be divided into several areas. First and foremost, it is the spontaneous informal conversations that can be considered a CNC flagship in this area. These are covered by two corpus series: the recently released ORAL v1 corpus (Kopřivová et al., 2017; 5.4 million running words) that summarizes many years of data collection and is now surpassed by the new-generation ORTOFON corpus. ORTOFON features a two-tier transcription (orthographic and phonetic), it is designed as a representation of contemporary spontaneous spoken Czech and therefore, it is fully balanced in terms of the main sociolinguistic categories of speakers (Komrsková et al., 2017; 1 million running words). These are complemented by a one-tier ORATOR corpus that covers semi-formal monologues (the first version released in 2019; 580,000 running words). The compatible orthographic tiers of ORTOFON and ORATOR constitute a suitable base for further extension of data collection to another spoken language domains.

• Parallel corpora are represented by InterCorp, a multilingual parallel corpus (Čermák and Rosen, 2012; Rosen and Vavřín, 2012) with Czech texts aligned on sentence level with their translations to or from 40 langages (27 of them lemmatized and/or tagged). The core of the InterCorp consists of manually aligned and proofread fiction, and it is supplemented by collections of automatically processed texts from various domains. InterCorp is updated every year, the total size of aligned texts released in the latest version of InterCorp amounts to 1.73 billion running words.

• Historical Czech: DIAKORP with its current size 3.5 million running words includes texts from the 14th century, with a recent focus on the 19th century (Kučera and Stluka, 2014; Kučera et al., 2019). In the long-term perspective, one of the main goals is to compile a representative monitor corpus of written Czech that would cover the period from the 19th century to the present and enable a systematic study of language change.

• Specialized corpora for specific research topics:

◦ DIALEKT dialectal corpus (Goláňová and Waclawičová, 2019; 100,000 words) with two-tier transcriptions of older dialectal recordings (from the 1960s until the 1980s), as well as newer probes (from 1990s until present).

◦ Koditex corpus created for the conducting a multidimensional analysis of register variation in Czech (Cvrček et al., 2018; Zasina and Komrsková, 2019; 9 million running words). For this reason, the corpus was compiled to be as diverse as possible, and therefore, it includes samples also from domains not covered by CNC (e.g. transcripts of TV discussions).

◦ NET corpus of semi-official internet communication, currently discussion forums and blogs (published in 2019; 41 million words). NET is not meant to be "just another web-crawled corpus", so the emphasis is not on size, but rather on rich metadata and high-quality text processing.

◦ ONLINE corpus of Czech web media and social networks that will be published in spring 2020 with an overall size of several billions of running words. Source data for the ONLINE corpus are provided by the Dataweps company. This cooperation will also make it possible to update the corpus on a daily basis, with the update size estimated at ca 4 million running words a day.

Apart from the CNC-compiled corpora mentioned above, there are also a number of hosted corpora available via the CNC web portal (see section 5 for more details).

3. Annotation

There are two main kinds of linguistic annotation being actively maintained by CNC: morphological tagging and syntactic parsing. For this purpose, CNC mostly adapts language-independent software tools to enhance their accuracy on Czech, and in particular, in the individual language domains. This is why CNC also works on the creation of training data from these domains that will be used to train third-party tools like MorphoDiTa (Straková et al., 2014).

Currently, there is an ongoing effort to develop a uniform tagging scheme that would cover the very different language varieties present in CNC: written Czech, informal spoken Czech, Czech used on the internet (discussion forums, social networks etc.) and Czech of the 19th century. This effort is coordinated with the authors of the MorfFlex CZ morphological dictionary,[1] in order to make the resulting tagging scheme as close to it as possible. The scheme will be first used for processing the SYN2020 corpus and it will remain stable for a couple of years to come.

Syntactic level annotation is – similarly to the morphological one – carried out by adapting the existing language-independent third-party tools for syntactic parsing and their enhancement by various methods (Jelínek, 2014; Jelínek, 2019); this sometimes requires also the creation of small treebanks (Jelínek, 2017). Currently, only the newer representative written Czech corpora are available with syntactic annotation: SYN2015 and the forthcoming SYN2020.

[1] http://hdl.handle.net/11234/1-1673

4. Application Development

The emphasis on empirical methods in linguistics and the development of digital humanities highlight the need of quantitative utilization of the variety and volume of the data using statistical methods. Furthermore, we believe that end-user web applications shape the way linguists think about the language data and about the range of possibilities they offer. This creates significant demands on the development of user-friendly web applications aimed at researchers in the humanities and social sciences that would easily use them as powerful sources of reliable information.

Currently, there are eight such web applications in CNC, three of which have been developed in 2019 (Word at a Glance, Lists, Calc).

• KonText (http://kontext.korpus.cz; Machálek, 2020a): continually developed web-based general-purpose corpus concordancer that supports various corpus types including spoken and parallel corpora. It is built above the data retrieval and indexing libraries of NoSketch Engine including its core library manatee-open (Rychlý, 2007). The main distinctive features of KonText can be divided into three main groups (Machálek, 2020a):

◦ query construction: query syntax highlighting; tag builder widget for interactive selection of individual values designed both for Universal Dependencies (UD) and positional tagsets; advanced query history with an easy overview, filtering, and marking for later reuse;

◦ data selection: interactive creation of subcorpora by "zooming into" the selected parts of a corpus down to the document level which enables easy examination of its contents;

◦ result presentation and manipulation: easy control over all operations on the query result (including their reproducibility and editable processing chain); rendering of dependency syntax trees; visual representation of dialogues with a clear indication of speaker turns and overlaps for spoken corpora.

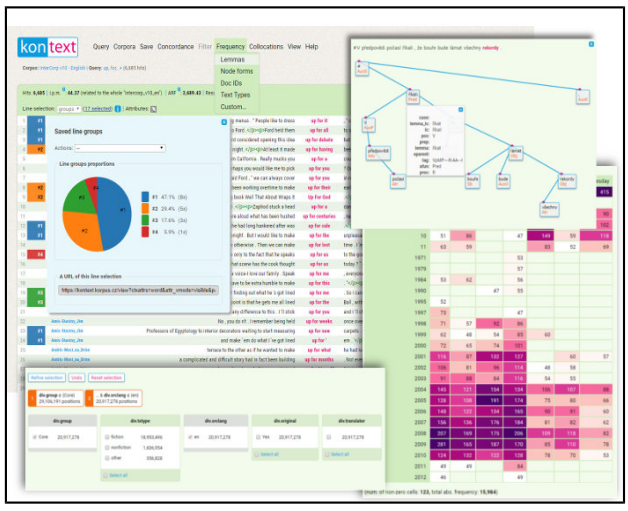

Figure 1: KonText.

In addition, there are many other KonText features not mentioned above, including possible integration with third-party services that may provide additional information about the searched terms. KonText is a mature software developed at GitHub[2] and deployed by some of the CLARIN centres in Europe.

• SyD (http://syd.korpus.cz/; Cvrček and Vondřička, 2011): web application for the corpus-based analysis of language variants. In the synchronic part, frequency distribution and collocations of variants can be compared across different domains of contemporary written and spoken texts, while the diachronic part shows their development over time. SyD provides easily interpretable summarized information with lively visuals and graphics, and it is thus very popular also among the general public.

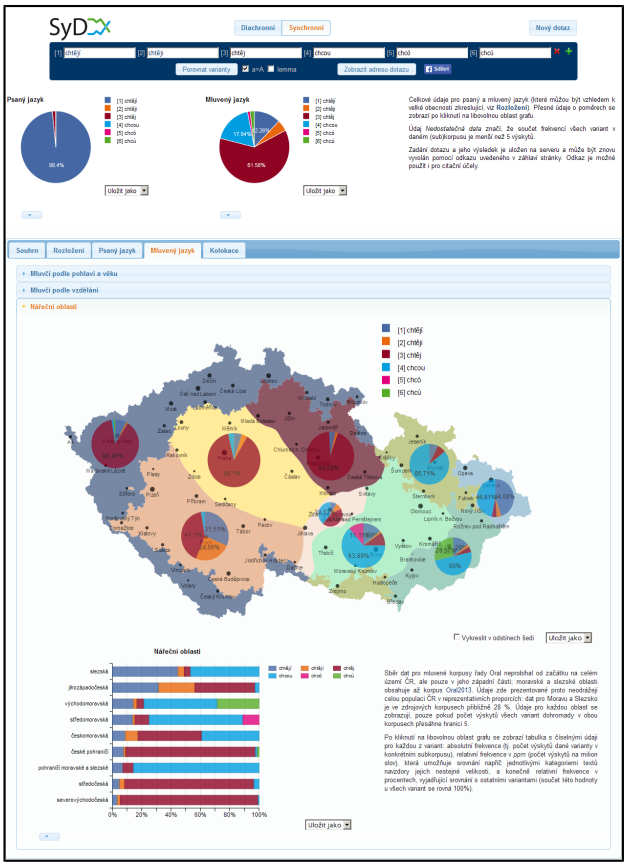

Figure 2: SyD.

• Morfio (http://morfio.korpus.cz/; Cvrček and Vondřička, 2012): web application for the study of word formation and derivational morphology in corpora of contemporary written Czech (extension to other langauges is on the way). Morfio searches the corpus to identify and analyse selected derivational patterns, as specified by prefixes, suffixes or word roots. It can be used to analyse the morphological productivity of affixes and to estimate the accuracy of a selected derivational model. It also includes a list of morphonological alternations.

[2] https://github.com/czcorpus/kontext

Figure 3: Morfio.

• KWords (http://kwords.korpus.cz/): web application for the identification of keywords (i.e. statistically prominent words usually connected with the text topic) in Czech and English texts. It enables users to upload their own texts to be compared against a reference corpus or a user-selected text. The output is a list of keywords that includes collocations and the keywords are highlighted in the text. KWords also supports the analysis and visualisation of distance-based relations of keywords. It is targeted mainly at scholars and students in text-linguistic and discourse studies.

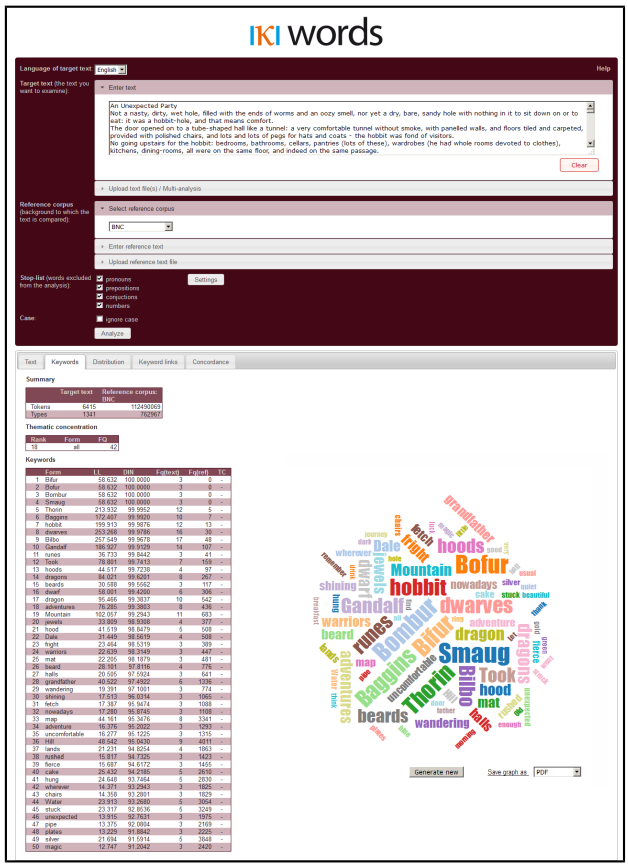

Figure 4: KWords.

• Treq (http://treq.korpus.cz/; Škrabal and Vavřín, 2017): intuitive web interface to automatically-generated translation dictionaries derived from the InterCorp parallel corpus. The users only need to specify their desired language pair for querying either individual word forms or lemmas. The result is a list of translation candidates of the given item sorted by decreasing frequency. By clicking on a particular translation candidate, its occurrences in InterCorp open up in KonText and can be further examined. Similarly to SyD, Treq is very straightforward and easy to use, so it is very popular among students and general public.

Figure 5: Treq.

• Word at a Glance (https://www.korpus.cz/slovo-v-kostce/; Machálek 2020b) is a brand new web application that has been designed as the main CNC word search service. There are three main operation modes: single word search, two or more words comparison, and word translation mode. In all of them, Word at a Glance (WaG) creates an aggregated word profile that is based on existing language resources (possibly also remote ones) and displayed as a structured and comprehensive overview of various properties of the given word. WaG is an application where many important decisions (which (sub)corpus or statistics to use, its parametrization etc.) have already been made for the user, in order to facilitate (relatively) safe generalizations. Furthermore, WaG has been developed[3] with reusability in mind: deployment and customization by other projects is very easy, adaptation of pre-packaged tiles requires only editing of configuration files.

[3] https://github.com/czcorpus/wdglance

Figure 6: Word at a Glance.

• Lists (https://www.korpus.cz/lists/): simple web application for browsing and comparing frequency lists where they can be inspected, sorted and filtered based on a frequency cut-off, substring and/or part of speech.

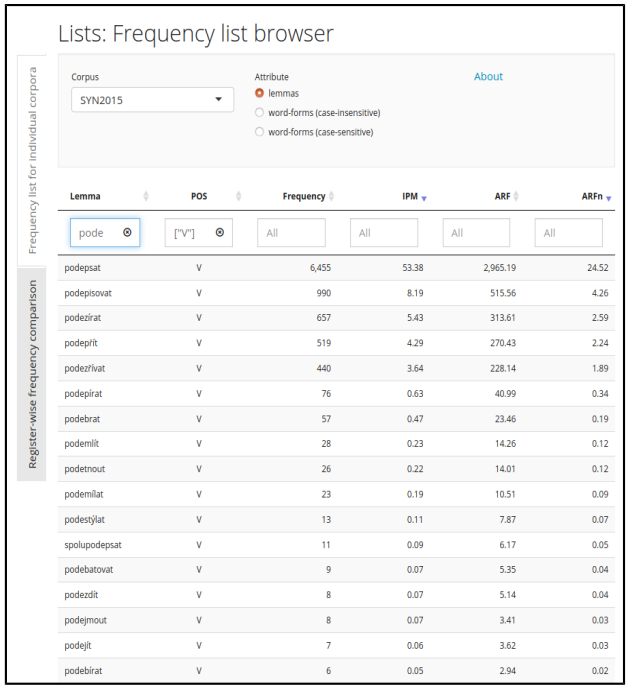

Figure 7: Lists.

• Calc (https://www.korpus.cz/calc/): corpus calculator designed to help the corpus users calculate basic statistical tasks most commonly encountered in corpus research. Currently, there are seven such tasks supported by Calc. Unlike most other statistical calculators, Calc is task-based, which means that appropriate statistical tests are already pre-selected so users don't need to think about their suitability for the given task.

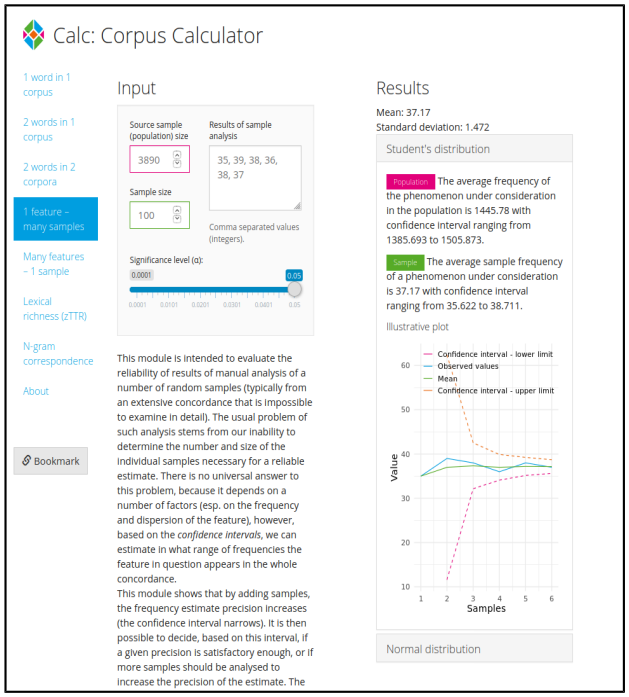

Figure 8: Calc.

5. User Services

User services are concentrated on the CNC research portal at http://www.korpus.cz/ that integrates web applications with user support. The individual services have already been described in more detail in (Křen, 2015) and have been quite stable since then. Therefore, the following overview only summarizes them briefly:

• on-line helpdesk with Q&A that also handles requests for new application features and bug reports;

• web documentation and manuals;

• repository of CNC-based research outputs (currently more than 2,500 entries);

• corpus-based exercises for language teaching at schools;

• corpus hosting (technical processing, quality checks, publication and maintenance of third-party corpora) as a valuable enrichment of the in-house corpora offered by the CNC; currently, the hosted corpora include comparable web corpora of the Aranea series for 14 languages (Benko, 2014), several learner corpora, author corpus of Jan Čep (complemented by the CNC-compiled author corpus of Karel Čapek), Early English Books Online, corpora of Upper and Lower Sorbian etc.

• data packages: corpus-based data sets prepared on demand in case of legal limitations on the redistribution of the original texts;

• consulting, workshops and academic training on various levels.

6. Future Plans

The data collection shall continue along the established lines. As already noted in Section 2, we plan to further extend its coverage to other areas of spoken language, and also to concentrate on building a representative monitor corpus of Czech from the 19th century to the present. However, this goal very much depends on the availability of 20th century texts (until 1989) that are – or at least may be – still subject to the copyright, and thus not generally available from libraries.

As for the annotation, the priority (and also a challenge) is now the development of uniform annotation for all language varieties covered by CNC (cf. Section 3). We are also working on UD annotation of the InterCorp parallel corpus that is currently tagged only by national taggers. The national tagsets usually provide quite rich description in terms of the morphological features covered, but they are not compatible with each other. Therefore, we plan to compile a new version of InterCorp fully annotated in UD (including the syntax), and at the same time, to enhance the UD support in KonText.

The emphasis of the future development of CNC will be put on web applications. In addition to the continuous maintenance and improvement of those mentioned in Section 4, a brand new Map application will be released in 2020. It will display summary information on Czech language dialects on the map, including a description of the individual dialectal features, illustrative corpus-based examples and localization of the corpus probes.

As an output of other project, we are currently building a variation database that records variants (especially stylistic, phonological, orthographic and morphological) of all individual word form and lemma types as evidenced in the CNC corpora. The database will be made available for searching via a dedicated web user interface, and it will also provide valuable paradigmatic information to be added to WaG.

Last but not least, we plan to develop a special application that would examine public discourse based on the data of the ONLINE corpus with its daily updates (cf. Section 2). Its design is still to be discussed, but we believe that it will prove to be a valuable tool for researchers from many domains beyond linguistics.

7. Acknowledgements

The data, tools and services described in this paper are the result of a team work. Many thanks to all for their ideas, hard work and endurance that make the CNC project possible.

This paper resulted from the implementation of the Czech National Corpus 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.

8. Bibliographical References

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček and K. Pala (Eds.), *TSD 2014, LNAI 8655*. Springer, pp. 257–264.

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

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A. and Zasina, A. (2018). From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory*, doi:10.1515/cllt-2018-0020.

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.

Jelínek, T. (2014). Improvements to Dependency Parsing Using Automatic Simplification of Data. In *Proceedings of LREC 2014*. Reykjavík: ELRA, pp. 73–77.

Jelínek, T. (2017). FicTree: a Manually Annotated Treebank of Czech Fiction. In J. Hlaváčová (Ed.), *ITAT 2017: Information Technologies – Applications and Theory*. Praha: Aachen & Charleston, pp. 181–185.

Jelínek, T. (2019). Using a database of multiword expressions in dependency parsing. In K. Ekštein (Ed.), *Text, Speech, and Dialogue. TSD 2019*. Springer, pp. 19–31.

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. (2015). Recent Developments in the Czech National Corpus. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*. Mannheim: Institut für Deutsche Sprache, pp. 1–4.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P. and Vondřička, P. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of LREC 2016*. Portorož: ELRA, pp. 2522–2528.

Kučera, K., Najbrtová, K., Pivoňková, K., Řehořková, A. and Stluka, M. (2019). Korpus českého jazyka 2. poloviny 19. století. *Časopis pro moderní filologii*, 101(1): 92–98.

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

Machálek, T. (2020b). 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.

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, pp. 65–70.

Straková, J., Straka, M. and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, pp. 13–18.

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.