Published in: Bański, Piotr; Biber, Hanno; Breiteneder, Evelyn; Kupietz, Marc; Lüngen, Harald; Witt, Andreas (eds.) (2015): Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3). Mannheim: Institut für Deutsche Sprache, pp. 1-4

Recent Developments in the Czech National Corpus

Michal Křen
Charles University in Prague
Institute of the Czech National Corpus
michal.kren@ff.cuni.cz

1 Introduction

The Czech National Corpus (CNC) is a long-term project striving for extensive and continuous mapping of the Czech language. This effort results mostly in compilation, maintenance and providing free 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.

Since 2012, the CNC is officially recognized as a research infrastructure funded by the Czech Ministry of Education, Youth and Sports which has caused a recent shift towards user service-oriented operation of the project. All project-related resources are now integrated into the CNC research portal at http://www.korpus.cz/.

Currently, the CNC has an established and growing user community of more than 4,500 active users in the Czech Republic and abroad who put almost 1,900 queries per day using one of the user interfaces. The paper discusses the main CNC objectives for each particular domain, aiming at an overview of the current situation supplemented by an outline of future plans.

2 Corpus compilation

Most of the CNC corpora can be characterized as traditional (as opposed to the web-crawled corpora), with emphasis on cleared copyright issues, well-defined composition, reliable metadata and high-quality data processing.

Synchronic written corpora of the SYN series (Hnátková et al., 2014) with current overall size 2.2 billion word tokens (i.e. tokens not including punctuation). The series consists of three general-language representative corpora (containing a large variety of fiction, newspapers and professional texts) published every five years that cover consecutive time periods, and large newspaper corpora. The annotation of the SYN-series corpora includes detailed bibliographical information, lemmatization and morphological tagging.

Synchronic spoken corpora of the ORAL series with current overall size 4.8 million word tokens; the corpora include only unscripted informal dialogical speech. The newest corpus of the series, ORAL2013 (Válková et al., 2012), is designed as a representation of contemporary spontaneous spoken language used in informal situations on the area of the whole Czech Republic; it features manual one-layer transcription aligned with audio. A new ORTOFON series with two-layer transcription (orthographic and phonetic) has been recently established (Kopřivová et al., 2014).

Multilingual parallel corpus InterCorp (Čermák and Rosen, 2012; Rosen and Vavřín, 2012) with Czech texts aligned on sentence level with their translations to or from 30+ languages (some 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. The total size of foreign-language texts is almost 1.4 billion word tokens, out of which 173 million make up the core (version 7 published in December 2014).

Diachronic corpus of historical Czech DIAKORP (Kučera and Stluka, 2014) with current size 2 million word tokens includes texts from the 14th century onwards. However, the current focus of DIAKORP development is on the 19th century.

Specialized corpora of various kinds and for specific research purposes that supplement the variety of hosted corpora. The specialized corpora include most prominently a dialectal corpus and a corpus of Czech texts written by the deaf (neither of them published yet).

3 Data processing and annotation

Apart from the corpus compilation, the CNC develops or adapts software technologies for data processing and annotation that supplement standard project-independent tools.

- Software environments for internal project **work flow management** of data collection (large networks of external collaborators for the spoken corpora and the InterCorp) and processing of various corpora. For the most part, the environments function as a web-based "wrapper" that combines both CNC and third-party tools.

  - **SynKorp** – database and data processing toolchain for the SYN-series corpora of written language (text conversion, clean-up, metadata annotation and text classification).

  - **Mluvka** – database and integrated project management system for coordination of spoken and dialectal data collection, manual two-layer annotation (orthographic and phonetic), expert revision and balancing.

  - Database of parallel texts and integrated project management system for coordination of the **InterCorp**, manual verification and revision of the alignment (implements a three-level project coordination hierarchy similar to Mluvka). The work flow includes **InterText** (Vondřička, 2014), a project-independent editor of aligned parallel texts.

- Tools for **linguistic annotation** of Czech language data on morphological and syntactic level. For this purpose, the CNC mostly adapts language-independent software tools and develops Czech-specific ones.

  - The **morphological level** includes Czech morphological analyser and lexicon (Hajič, 2004) (both provided by LINDAT/CLARIN; `http://lindat.mff.cuni.cz/`) that is being continuously administered in collaboration with the CNC. Subsequent morphological disambiguation involves a combination of language-independent stochastic tagger with rule-based components developed specifically for Czech (Hnátková et al., 2014; Jelínek, 2008; Petkevič, 2006; Spoustová et al., 2007). Works on extension of the current morphological annotation to spoken and diachronic data are already under way.

  - **Syntactic level** annotation is – similarly to the morphological one – carried out by Czech-specific adaptation of existing stochastic language-independent third-party tools for syntactic parsing and enhancement of their results by various methods, including rule-based corrections (Jelínek, 2014). The first syntactically parsed CNC corpus will be published by the end of this year.

4 Application development

Design and development of new intuitive analytical web-based applications as well as continuous enhancement of the existing ones are an integral part of the effort to promote empirical linguistic research. All the applications are open-source and all of them (except for KWords) currently use Manatee (Rychlý, 2007) as their backend.

KonText (`http://kontext.korpus.cz/`), a web-based general-purpose corpus concordancer (CNC fork of the NoSketch Engine; Rychlý, 2007) with built-in basic statistical functions, subcorpus manager, filtering, word-to-sound alignment support etc. It is the only application that requires user registration to switch from restricted functionality to regular access.

SyD (`http://syd.korpus.cz/`; Cvrček and Vondřička, 2011), a web application for 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.

Morfio (`http://morfio.korpus.cz/`; Cvrček and Vondřička, 2012), a web application for study of word formation and derivational morphology. It searches the corpus to identify and analyze selected derivational patterns, specified by prefixes, suffixes or word roots. It can be used to analyze morphological productivity of affixes and to estimate the accuracy of a selected derivational model in Czech.

KWords (`http://kwords.korpus.cz/`), a web application for corpus-based keyword and discourse analysis of Czech and English. It enables users to upload their own texts to be compared against one of the reference corpora available or against a selected text. It also supports the analysis and visualization of distance-based relations of keywords.

5 User services

User support and services are concentrated at the CNC research portal at `http://www.korpus.cz/`, a common platform for language research aimed at both the research community and the general public that integrates web applications mentioned above with active support. In addition to the research portal, the CNC offers also organization of workshops and lectures, involvement in academic training, expert consultations and tutoring etc.

- **User Forum**: a virtual platform accessible to all registered users. It features an advisory centre (with Q&A) that also handles all web requests for new application features and bug reports, which serve as a valuable source of user feedback.

- **CNC Wiki** (corpus linguistics knowledge base) with an on-line manual is freely available on the portal without registration. It contains an introduction into corpus linguistics, details about the CNC resources, and an on-line tutorial in seven lessons aimed at both beginners and advanced users (for the time being in Czech only).

- **Biblio**: a repository of CNC-based research outputs; users are encouraged not only to submit references about their research papers, books or theses based on CNC resources, but also to upload them directly to make them accessible to all visitors of the CNC portal.

- **Corpus hosting**: the CNC provides hosting service of – mostly small and/or specialized – corpora created at other institutions which do not have the possibility or know-how to ensure adequate final technical processing of their data (including quality checks with possible labour-intensive corrections). This is offered by the CNC, as well as maintenance of the resulting corpora, providing public access to them and related services; appropriate credit of the hosted corpus is always given, including a link to the relevant publication. Hosted corpora constitute a valuable enrichment of the CNC-compiled corpora and include learner corpora, web corpora and foreign-language corpora (including Upper and Lower Sorbian).

- **Data packages**: the CNC strives to be as open as possible also in terms of language data. On the other hand, restrictions arising from the laws in force have to be observed and this is one of the reasons why the CNC has introduced the service of providing data packages. This service enables users to obtain corpus-derived data with less restrictive licensing than the licensing of the original corpus texts. The data packages are either available through LINDAT/CLARIN repository, or they can be prepared in accordance with individual requirements of the particular user or institution. The licensing depends on the nature of the data and it ranges between the CC BY license (for word lists or n-grams for small n) to proprietary license that permits neither commercial use nor redistribution (for full texts shuffled at the sentence level).

6 Future plans

The applications and user services are planned to be maintained continuously, with new functionality added to the existing applications and new ones developed while responding to user requirements. To mention just a few planned enhancements:

- better visualization of query results, especially their diachronic development;
- multi-word unit identification and extraction component based on alternative approaches;
- interface enhancements leading the users to more appropriate interpretations and comparative statistical evaluation of corpus search results.

The spectrum of collected data will be broadened in the near future by adding semi-formal spoken language and by establishing a new corpus series that would contain selected specific semi-official language used on the internet, including blogs, discussion forums etc. (i.e. not yet another web corpus). In the long-term perspective, one of the main goals is to compile a monitor corpus of written Czech that would cover the period from 1850 to the present and enable a systematic and sophisticated study of language change. This corpus will help to eventually bridge the gap between the diachronic and synchronic data in the CNC, while taking full advantage of the CNC's twenty year tradition of data collection.

Acknowledgements

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

This paper resulted from the implementation of the Czech National Corpus project (LM2011023) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

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

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

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

J. Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Praha.

M. Hnátková, M. Křen, P. Procházka, and H. Skoumalová. 2014. The SYN-series corpora of written Czech. In *Proceedings of LREC2014*, pages 160–164, Reykjavík. ELRA.

T. Jelínek. 2008. Nové značkování v Českém národním korpusu. *Naše řeč*, 91(1):13–20.

T. Jelínek. 2014. Improvements to dependency parsing using automatic simplification of data. In *Proceedings of LREC2014*, pages 73–77, Reykjavík. ELRA.

M. Kopřivová, H. Goláňová, P. Klimešová, and D. Lukeš. 2014. Mapping diatopic and diachronic variation in spoken Czech: the Ortofon and Dialekt corpora. In *Proceedings of LREC2014*, pages 376–382, Reykjavík. ELRA.

K. Kučera and M. Stluka. 2014. Corpus of 19th-century Czech texts: Problems and solutions. In *Proceedings of LREC2014*, pages 165–168, Reykjavík. ELRA.

V. Petkevič. 2006. Reliable morphological disambiguation of Czech: Rule-based approach is necessary. In M. Šimková, editor, *Insight into the Slovak and Czech Corpus Linguistics*, pages 26–44. Veda, Bratislava.

A. Rosen and M. Vavřín. 2012. Building a multilingual parallel corpus for human users. In *Proceedings of LREC2012*, pages 2447–2452, İstanbul. ELRA.

P. Rychlý. 2007. Manatee/Bonito – a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno.

J. Spoustová, J. Hajič, J. Votrubec, P. Krbec, and P. Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha.

L. Válková, M. Waclawičová, and M. Křen. 2012. Balanced data repository of spontaneous spoken Czech. In *Proceedings of LREC2012*, pages 3345–3349, İstanbul. ELRA.

P. Vondřička. 2014. Aligning parallel texts with InterText. In *Proceedings of LREC2014*, pages 1875–1879, Reykjavík. ELRA.