Edited by Darja Fišer and Jakob Lenardič
Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN National Consortia and Knowledge Centres with the aim to increase their visibility, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia and knowledge centres as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

The brochure presents the Czech CLARIN Knowledge Centre for Corpus Linguistics and is organized in two sections:

- Section One presents the members of the Knowledge Centre and their work
- Section Two includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the Knowledge Centre’s infrastructure in their research
The CNC activities can be divided into the three main areas:

- **Data collection.** Focusing on quantity, quality, and variety, the CNC corpora feature careful text selection, reliable annotation and rich metadata. The following areas are currently covered:
  - contemporary written Czech: SYN-series corpora (total size 4.2 billion running words) which also include representative 100-million corpora released every five years;
  - contemporary spoken Czech: corpora consisting solely of spontaneous informal conversation of the ORAL and ORTOFON series (total size 5.3 million running words);
  - InterCorp multilingual parallel corpus: manually aligned and proofread fiction texts supplemented by collections of automatically processed texts from various domains (total size 1.5 billion running words in all 40+ languages);
  - specialized corpora include historical Czech (DIAKORP), Czech dialects (DIALEKT), and many more.

- **Annotation involves data curation,** metadata annotation, morphological tagging and syntactic parsing. For all these procedures, CNC uses open-source software, third-party tools, as well as specialized tools developed in-house. The tools developed in-house include mainly the Phras module for identification of idioms, Mluvka for management of distributed spoken data collection, and the parallel text alignment editor InterText.

- **User application development.** We recognize the key importance of presenting corpora in an intuitive way that makes them accessible to researchers from various fields of social sciences and humanities. The following web applications are currently offered:
  - KonText – a general-purpose corpus query interface and concordancer with an advanced subcorpus manager, parallel corpus support and support for word-to-sound alignment;
  - SyD – corpus-based analysis of language variants, both synchronic and diachronic;
  - Morfio – identification of derivational models in Czech including estimation of their morphological productivity;
  - KWords – corpus-based keyword analysis;
  - Treq – translation equivalent search interface based on the InterCorp parallel corpus.

Czech CLARIN Knowledge Centre for Corpus Linguistics was recognized by CLARIN on December 4, 2018.1

The centre is based at the institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague. Czech National Corpus (CNC) is a long-term academic project with the main aim to continuously map the Czech language by building, annotating and providing access to a variety of large general-purpose corpora. CNC also develops specialized web-based applications for user-friendly access to the corpora and offers wide-ranging user support which includes a user forum with Q&A, bug reporting, detailed documentation and a knowledge base.

CNC is recognized by the Ministry of Education, Youth and Sports of the Czech Republic as a research infrastructure and included on the Roadmap of Large Research Infrastructures of the Czech Republic for 2016-2022. CNC is an associated member of the CLARIN-CZ consortium with established long-term collaboration with LINDAT/CLARIN. CNC is also a CLARIN FCS endpoint and it supports single sign-on (Shibboleth) as one of the options for accessing the CNC resources. In addition to this service-oriented line of work, CNC is also a research centre that promotes an empirical approach to language and runs a PhD programme in Corpus Linguistics.

---

Documentation and knowledge base for CNC applications, data and services. It provides interdisciplinary guidance and promotes empirical methods in language research. It also features an online tutorial aimed at both beginners and advanced users.

Repository of CNC-based resources and research output that can also serve as a bibliography for looking up information concerning the corpus research on Czech.

Repository of corpus-based exercises (Czech only) for L1 and L2 language teaching.

Consulting, education, and training: in addition to the general user workshop held on a regular basis twice a year, workshops on various topics are held upon request.

Corpus hosting: the service includes final technical processing of user-compiled corpora, quality checks, and public access with related services.

Customized data packages: data sets prepared on demand and extracted from the CNC corpora while observing the legal limitations that may not allow for redistribution of the texts per se.

Our expertise includes not only data formats, text curation, annotation, metadata encoding and corpus querying, but also empirical research on the Czech language, corpus linguistics methodology and statistical methods. The centre can also provide external pointers to other institutions regarding any aspect of Czech language including language resources and natural language processing.

Currently, CNC has 7,500+ registered active users who perform (on average) 3,000+ corpus queries per day. The repository of CNC-based research outputs has yielded more than 150 theses (bachelor, master or doctoral) defended a year.

The CNC user support and related services are available also through the CLARIN K-Centre. This includes:

- K-Centre Helpdesk and CNC User Forum, virtual platforms for active user support and feedback. The CNC user forum features also a discussion forum (with Q&A) that can handle requests for new application features as well as bug reports.

Figure 1: KonText user application being actively developed by the CNC

CNC User Workshop
What is your involvement with the CNC K-Centre?

I am both a dedicated user of their infrastructure and a collaborating researcher. I have been invited by the centre to give talks on diachronic corpus topics (for instance, on lexical obsolescence in Late Modern English or on the quantification of orthographical variation in Early Modern English, which are two of my current research interests), I have consulted on some of these projects with a number of colleagues at the centre and I hope our fruitful collaboration to continue in the future as well. But mainly, I use the centre’s infrastructure, tools and expertise to host and analyse corpora I need for my own projects. Many of these corpora are not in the public domain (either by the decision of their compilers or due to the licensing restrictions of their source material) and are only hosted for licensed users for research and teaching, but in cooperation with the centre we have also started publicly hosting data from the Early English Books Online (EEBO) project, and are about to host the Old Bailey Corpus, which is based on a selection of the Proceedings of the Old Bailey, the published version of the trials at London’s Central Criminal Court.

Which data collections in CNC do you use in your own research? Could you present and discuss some of your research that has resulted from your use of the CNC corpora?

I mostly use English diachronic corpora that the centre specifically processed and hosts for our department and students, but I have also used the DiaKorp, InterCorp and the SYN corpora for a contrastive angle.

One example of the research I do using the centre’s infrastructure is my recent work on spelling variation in Early Modern English based on the Parsed Corpus of Early English Correspondence. I introduced a novel methodology for the quantification of spelling regularity, which allowed a more objective assessment of its progression in time and which also makes use of the metadata provided by the CEEC such as gender, letter authenticity or relationship/kinship between the author and the recipient. I have explored interactions of such variables from the diachronic perspective using quantified levels of spelling regularity.

The measure introduced for this purpose is based on weighted information (Shannon) entropy, as a measure of predictability of a spelling of individual functionally defined types, and its calculation is partly based on the morphological tagging of the parsed version of the corpus.
I have also tackled the problem of underrepresentation in certain periods by establishing a size-based sampling for scalar variables like time. For instance, I was able to show that letters written by women showed a greater degree of entropy – so a greater degree of variability – in spelling regularity than letters written by men through the whole period (roughly from 1410 to 1680). However, this difference turned out to be a function of another sociolinguistic variable that I was accounting for besides the author’s gender; namely, the relationship between the author and the recipient. Female authors corresponded significantly more with other members in the family than male authors who mostly corresponded with acquaintances outside the family. In a familial context, there might be less pressure to conform with spelling standards, hence the greater degree of variation.

Another example is an older study on measuring the typological change in English that was based on the parsed versions of the Helsinki corpora. In this paper, my colleague Jan Čermák and I proposed a quantitative, but also holistic, methodology for establishing the level of morphological syntheticity within a language – that is, how much a language relies on morphological markings to convey syntactic information. The methodology is based on a series of corpus-based probes into the morphological behaviour of selected high-frequency nouns, adjectives and verbs from Old English to Present-Day English in corpora hosted by the CNC. We thereby managed to establish several levels of syntheticity that correspond to the well-known typological re-shaping that happened in the history of English, which shifted from a heavily synthetic language in its early days to an analytic one in the present day. For instance, Old English was highly synthetic, its nouns ending in seven different inflections corresponding to the complex case system, whereas Present-Day English nouns only use the -s affix to mark plurality, and our proposed methodology was able to capture this quite precisely. It should be also noted that CNC often consults and helps out indirectly, not with their corpora or tools, but with their scientific and technical expertise. For example, in my research into the obsolescence of multi-word expressions in the history of English, it was only thanks to a colleague at CNC and the centre’s computing resources that I was able to pre-process most of the Google Ngram dataset (about 2 terabytes of data).
A lot of the theses are based on the contrastive approach focusing on features of Present-Day English and Czech, but there have been a number of diachronic theses and papers as well. One of my PhD students is now working with the CNC-hosted EEBO data to research lexical losses in Early Modern English, another of our PhD students is developing a parallel corpus of Old English and Latin translations that will again be hosted by CNC that has already extended its support in this. Finally, one of my PhD students prepared lessons in English available on the CNC wiki for using the diachronic EEBO corpus, which show how KonText can be used to account for spelling variation, looking at diachronically competing word forms, analysing morphology, among other uses. We hope that some of our students will develop a similar online course for the Old Bailey Corpus.

What kind of feedback have you provided on the CNC corpora and its user interface? What is your experience with the CNC User Forum? Why is it important for the CNC K-Centre to offer such user support?

Since CNC often accommodates me by hosting all kinds of corpora that tend to be different than the Czech corpora they are predominantly focused on, I often request changes or new features – mostly by e-mail to specific colleagues but also through GitHub. While the CNC may not always immediately implement all my outlandish ideas it has in general been very forthcoming about my requests. Here is one example, where I requested that headers be added to the .csv and .xlsx files exported from KonText, and the CNC team quickly implemented the change.

How do you use the CNC corpora in your teaching? Have your students obtained any interesting results from the CNC corpora?

I use KonText in most of my classes focused on the History of English to showcase specific changes, and I also teach how to use the interface in my English Diachronic Corpora course. Almost all of the students at our department learn to use KonText and InterCorp, and the majority of theses in our linguistic programmes are corpus-based, so most of the final theses (several dozen a year) are based on CNC-hosted corpora.

https://github.com/czcorpus/kontext/issues/2038