CZECH NATIONAL CORPUS

Czech repository report

Michal Křen

ELTeC ZOOM party
7 December 2020

Czech National Corpus project

- `http://www.korpus.cz`
- long-term, continuous mapping of the Czech language
- compilation, maintenance and providing public access to a variety of language corpora
- historical Czech: focus is now on the 19th century
- MONITOR corpus
  - interconnection of the 19th century and contemporary language (1990)
  - balanced in terms of time period and register (fiction, non-fiction, newspapers)
- mapping the available texts and processing them (careful proofreading, structural markup, etc.)

ELTeC-cze collection

Text selection and processing

- no documents in XML markup
- scanned texts available from several libraries in plain text format (automatic OCR)
- we collected approx 150 texts that meet ELTeC eiligibility criteria
- text processing: Python conversion scripts, Atom editor

Credits

- text corrections for MONITOR:
  Kateřina Najbrtová, Klára Pivoňková, Anna Řehořková, Martin Stluka
- processing for ELTeC:
  Anna Řehořková (text selection and encoding)
  Pavel Procházka (conversions)

ELTeC-cze collection

Current status of the GitHub repository

- overall: 23 texts uploaded, level1 encoding
- E5C score: 33.85
- takes into account only 16 texts, as 7 texts had issues during update to the new schema
- prevalence of short texts by male writers with low reprint count

Challenges

- text availability: relying on already digitized texts
- composition: Czech texts from this period are typically short and written by male authors
- reprint count: not easy to find out reliable numbers

ELTeC-cze collection

Current situation on our side

- 7 more texts with completed proofreading, ready to be added
- pool of approx 120 more texts after automatic OCR
  - all of them meet the eligibility criteria for ELTeC
  - without proofreading
  - not needed for MONITOR
- this means that we will not be able to do the proofreading

ELTeC-cze collection

Plans for spring 2021

- final selection of the remaining 70 texts
- conversion to XML (marking the front matter, chapters, etc.)
- adding metadata (reprint count, VIAF codes etc.)
- contribution to the next ELTeC release

ELTeC-cze collection

criterion	estimate
number of texts	100
time slot	OK
author gender	slightly more than 10% of female authors
texts per author	OK
reprint count	hopefully OK (no exact data yet)
size	short texts will prevail, lack of long texts

Balance of the final composition.

Thank you for your attention!