Development of the ELTeC corpus by/at ELTE

Gábor Palkó, Tímea Bajzát, Zsófia Fellegi

- how was the project organized ? (personnel, management…)
- what is the current state of the collection ? (E5C, encoding, open issues, etc.)
- how were texts selected? (source/s used; balancing methods etc,)
- how did you learn how to use TEI ? (self-study, training, previous knowledge?)
- how were texts processed? (software used etc.)
- what problems did you encounter?

Project team

- TEI XML knowledge: DigiPhil.hu project (digital and digitized critical editions of Hungarian literary works online) - currently 7500 XML files
  - Editor-in-Chief: Gábor Palkó, TEI XML specification: Zsófia Fellegi
- Human resources: Centre for Digital Humanities (project based funding)
- MA Students at ELTE
  - Emma Takács => Bence Vétek => Tímea Bajzát

State of the art

https://distantreading.github.io/ELTeC/

No known issues, E5C is 100

Encoding: Level_0

For one single feature the validation is done with the level_1 schema

Subchapter structure: <milestone>

Balance?

Text sources/selection

- No novels in markup at all from the period
- Hungarian Electronic Library (MEK: mek.oszk.hu)
- Agreement: TEI XML files will be part of the MEK collection
- http://mek.oszk.hu/05500/05573/
- More than 2000 “novels” (including translations)
- Metadata is not detailed enough (original publication date missing)
- OPAC for further metadata:
- nektar.oszk.hu (Hungarian National Library)
- http://www.mokka.hu/ (Hungarian National Common Catalogue)

Learning TEI

- TEI XML knowledge from the DigiPhil team
  - XSLT Zsófia Fellegi
- For the encoders (MA Students)
  - First step: https://teach.dariah.eu/course/view.php?id=40
- Translation of the course by the DigiPhil team (B. Bobák, Zs. Fellegi, G. Palkó)
- Tutorials made for encoders earlier (simple PDF-s):
  - Oxygen XML Editor, GIT repo usage
- XPath, regular expressions - self-study

Text procession 1.

Text level:

1. Rich Text Format => Oxygen XML Editor Author mode (smart paste)
    a. Desktop automation to accelerate the process
    b. Batch operations (Find & replace, regex)
2. HTML => Oxygen XML Editor
3. OCR-red PDF’s
    a. too many errors: Abbyy FineReader new OCR process => export to .docx => Oxygen
4. Google Books

Metadata:

Dublin Core XML + XSLT => TEI XML Header => manual correction

Text processing 2.

OCR correction (error list: valuable source!)

- Special characters: /, <, *
- Common errors: ii => n
- Text structure (page breaks, chapters)

Validation via schema in Oxygen XML Editor

Closed GIT repository (Bitbucket)

Main issues

- Corpus composition
  - Changing criteria (!!!!)
- Combination of certain criteria
- Reprint count metadata

Further plans

Including the corpus into curricula (workshops)

Including more novels (300 for now)

Stylometric analysis (authorship attribution)

NLP tagging and publication via a search engine:

http://verskorpusz.elte-dh.hu/