ELTEC-pol

Reporting Joanna Byszuk

Organization

Not that much of it, mostly talks between Jan Rybicki, Maciej Eder and me

Text selection: Jan Rybicki

Encoding: Jan Rybicki, Joanna Byszuk

Thanks to Lou for suggestions!

ELTEC-pol – current state

- 99 texts, 3 per author
- Selected from
  - our benchmark corpora (https://computationalstylistics.github.io/)
  - https://wolnelektury.pl/ platform digitizing obligatory and recommended school readings into their own XML
- 13 female authors (so 39 texts), popularity and length distribution roughly according to ELTEC rules
- Level 1 mostly fulfilled
- E5C was ~80% last week

TEI knowledge, software

- Budapest TS
- Modelling after collections in other languages

- Processing in Notepad++ and Oxygen

Problems

- Tiny team with time constraints and little TEI knowledge
- Relying on already-digitized texts so no direct access to prints (e.g. to render title pages)

Now working on:

- Chapter division where chapters not clearly marked
- We have tested tools for further automatic annotation but quality verification will be difficult given the size of the team