BEYOND BABYLONIAN CONFUSION: A CASE-STUDY BASED APPROACH FOR MULTILINGUAL NLP ON HISTORICAL LITERATURE

Author: Tess Dejaeghere
Supervisors: Julie Birkholz, Els Lefever, Christophe Verbruggen

CLSINFRA
Computational Literary Studies Infrastructure (CLS INFRA) is a four-year partnership to build a shared resource of high-quality data, tools and knowledge to aid new approaches to studying literature in the digital age.

CHALLENGES
NLP-tools such as Named Entity Recognition (NER) and sentiment analysis (SA) could support literary-historical research, but research on the topic is limited.

- Different user cultures and end goals.
- Differences in technical knowledge.
- NLP-tools are not adapted to literary-historical data.
- Need for NLP-based research infrastructures for Digital Humanities research.

DATA
Travel literature
The exceptional characteristic of travelogues as highly idiosyncratic lenses into the past accounts for a wide range of linguistic and historical variance. Travelogues are a rendition of an author’s personal travel experiences, thus allowing the researcher to reconstruct writer identities, historic environments and cultural traditions.

Corpus characteristics
- Different genres
  - nature writing, travel memoirs, journals, poetry, letters, ...
- Multilingual
  - NL, DEU, FR, EN
- Linguistic & historical variance
  - 16thC-20thC
- OCR mistakes

Excerpts
"While wandering about the banks of these gold-besprinkled streams, looking at the plants and mines and miners, I was so fortunate as to meet an interesting French Canadian, an old coureur de bois, who, after a few minutes’ conversation invited me to accompany him to his gold-mine on the head of Defot Creek [...]

... oranges, peaches, and other fruit trees, ferns, especially Gleichenia linearis, weeds of cultivation, miscellaneous shrubs and trees, including Pterocarya stenoptera [...]

"Remember that righteousness and our real ultimate self-interest demand that the blacks be treated justly."

METHODS

1. Create gold standard data with an aspect-based sentiment analysis layer and a named entity recognition layer.
2. Use annotated dataset to evaluate and adapt open-source systems.
3. Use output of NER- and SA-tools to support answering literary-historical questions.
4. Create NLP-workflows to support literary-historical research.