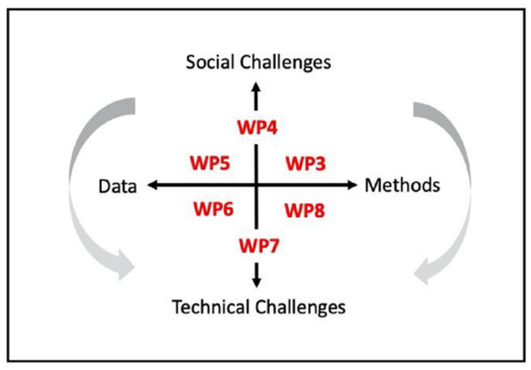
BEYOND BABYLONIAN CONFUSION: A CASE-STUDY BASED APPROACH FOR MULTILINGUAL NLP ON HISTORICAL LITERATURE

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

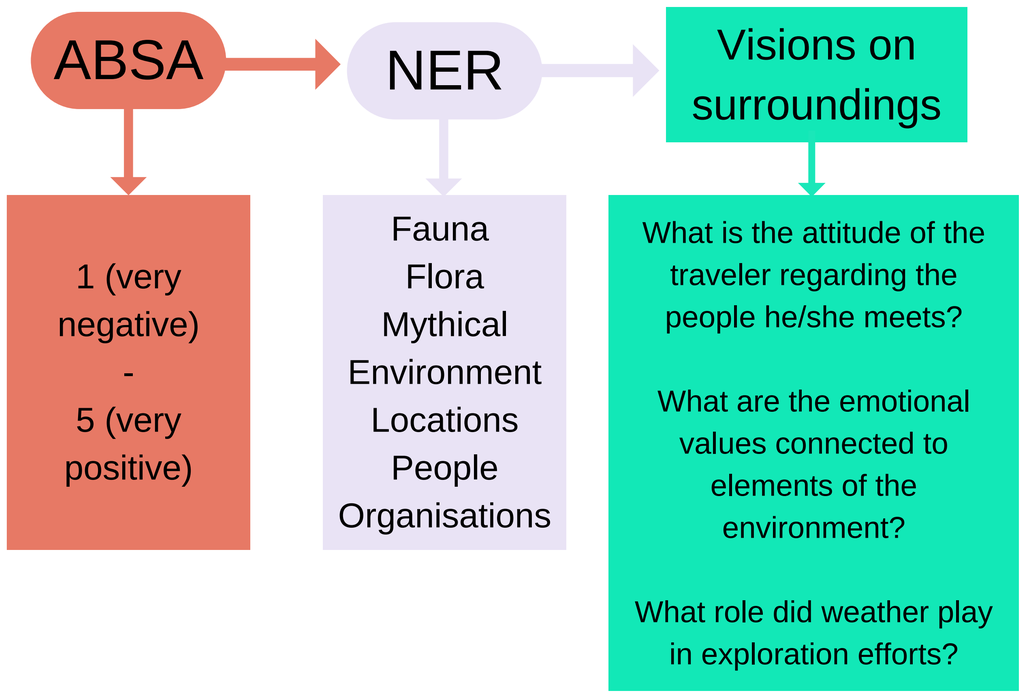
Excerpts

"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 stenopter […]

"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.

OBJECTIVES

1 CLS DELIVERABLES

1. Machine learning pipeline for named entity recognition.
2. Prototype to extract relations between entities.
3. A lexicon-based pipeline for sentiment analysis.
4. Machine learning pipeline for sentiment analysis.

2 PHD RESEARCH

Our research aims to generate much-needed insights regarding the potential and limitations of NER and SA in literary-historical research and intends to foster a tool- and data-critical attitude among digital humanists through the development of step-by-step guidelines regarding open-source tool selection and evaluation, tool adaptation and mitigating the challenges inherent to literary-historical and multilingual corpus processing, benefiting the CLARIN-infrastructure and DH-community alike.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No° 101004984.

GHENT UNIVERSITY

CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE