CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

TNA ACTIVITY REPORT

LitCORP: BUILDING A LITERARY CORPUS OF 19TH CENTURY CZECH PROSE

Author: Richard Změlík

Current position: assistant professor

Affiliation: Palacky University, Olomouc, Czech Republic

Host institution: University of Trier, Germany

Mentor(s): prof. Christof Schöch, Evgenia Fileva

Period of stay: 1. 10. - 27. 10. 2023

Introduction of the project

As part of the CLS Infra grant, I participated in a month-long research and study stay at the University of Trier with the project *Literary Cartographic and Quantitative Models of Czech Novels from the 19th to 21st Century*.[1] The aim of this project is to build a representative digital corpus of Czech narrative fiction texts that primarily represent Prague's topography. As the title of the project implies, the lower time limit for the collection of texts is the 19th century, or the second half of the 19th century, when texts explicitly thematizing Prague topography are just beginning to appear in Czech literature; the upper limit is the 21st century. The original intention was to realize literary-cartographic models that would provide a systematic overview of the ways in which Prague's topography was literarized. However, in the next phase of the project, the functionality of the corpus was expanded to offer, in addition to the cartographic models, a number of different quantitative models relating to selected aspects of literary texts (see the Methodology section below). In the next phase of the project's refinement, the language corpus itself, or tools for its mining, was added to the database of cartographic and quantitative models. As a result, the project is a multifunctional tool that will serve both literary scholars and linguists, and which, by combining cartographic and quantitative models together with corpus search tools, will enable multiple and variable types of analyses (see in more detail Corpus Utilization). However, the project does not need to be limited exclusively to Czech prose texts dealing with the Prague environment, but due to the existence of the digital language corpus as a functional part of the project, it is possible to gradually add to the corpus also those prosaic texts that do not deal with the Prague

[1] See https://www.korpusprozy.com/

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

environment, e.g. texts with rural or other themes. This step will enable the creation of a much larger and ultimately more representative database of literary texts (prosaic), which would open the way to the realization of a full-fledged literary corpus of prose, which is not yet represented in the Czech literary and linguistic environment, unlike the corpus of Czech poetry.

The project is being developed at the Department of Bohemian Studies, Faculty of Arts, Palacky University in Olomouc (Czech Republic) and its sole author is Richard Změlík.

Methodology

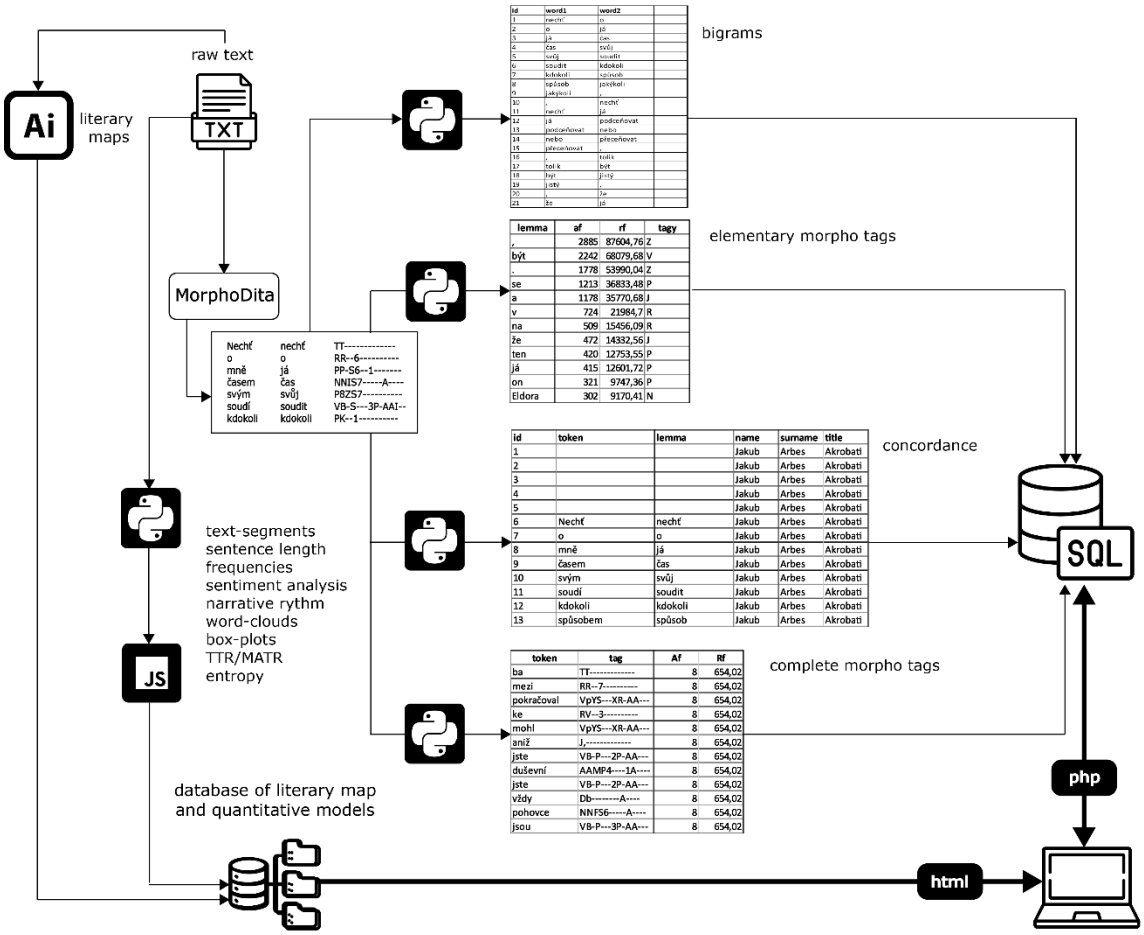

Fig. 1: Structure of corpus and data processing

The initial step in setting up the whole project the architecture of which is designed as shown in Fig. 1, is to obtain reliable digital texts in *.txt format. In the context of Czech literature (specifically prose), the situation in acquiring such formats is somewhat more complex, as there is no central database of them that is properly representative and, above all, open, such as Project Gutenberg. In recent years, the Distant Reading for European Literary History project[2] has been creating databases of digitized literature in various languages, including the Czech language, but in the case of the Czech ELTeC, which contains 100 prose texts, it is unfortunately not sufficiently representative. For example, the 19th century is represented by

[2] See https://www.distant-reading.net/

both canonical and non-canonical texts, which do not have the potential to represent the various developmental stages of Czech prose. Another potential source for digital formats of Czech prose are library databases, of which the most important is the electronic database of the National Library of the Czech Republic, Kramerius,[3] , as it is the largest in terms of its scope. However, some book titles are not openly available there, and some are not even in the records. Another source of digital text data is the catalogue of the Municipal Library in Prague,[4] which offers some selected titles for free download, e.g. in pdf format. Due to the limited possibilities of obtaining digital versions of texts, it was in many cases necessary to scan the texts and then convert them into an editable format using OCR. However, this could not be done without the necessary checking and cleaning of the output OCR format.

The texts obtained by one of the above stated methods were saved in TXT format after manual checking, which was then processed by two basic processes. The first focused on entering information about Prague's fictional topography into the underlying historical maps. The manual input was chosen deliberately, despite the fact that the current trend in literary cartography is primarily the use of GIS, which, however, has its disadvantages, especially when mapping fictional environments. One of the disadvantages is the incompatibility between fictional and real topography. In fact, GIS models use actual map bases (e.g. Google Maps or Open Street Maps) that do not correspond to the situation of a place in the 19th century, for example, which became an inspirational source for a fictional setting. Manually plotting locations or character paths through fictional locations is also preferable because many fictional locations are not clearly topographically defined and thus cannot be projected onto a base map layer using precisely given coordinates. These are just some of the advantages of manual mapping, which in our case was done in Adobe Illustrator. The final cartographic models were then implemented in a web environment (application) that allows switching between different models or layers.

Other ways of processing literary texts depended on the needs of the chosen quantitative analyses. One of these was to develop quantitative models of the frequency load of selected text segments; specifically, a Gennetan typology of narrators (homo- hetero- intra- extra-diegetic narrators), direct speech distributed across different layers of narrative or text segments (e.g., letter, diary). The delimitation of these segments is done partly manually, often automatically using a specially written Python script and a specially created tag set. The tagged texts are further automatically processed in a special program written in Python, the output of which are both tables of frequency values (absolute and relative frequency) of individual text segments (calculated per number of words) and bar charts, which after conversion into Java Script became part of the web presentation.

Other Python scripts were used to create I Words Clouds representing the basic motifs for a given prose in quantitative terms, or graphs showing the narrative rhythm of alternating narrator types, graphs showing the frequency load of Prague toponyms, or Sentiment Analysis graphs (see the Research Internship section).

A large part of the project is a language corpus with selected tools for its mining. Specifically, these are the following options for searching the corpus. For each authoring subcorpus, the following tools are available at this level:

- Boxplot search, TTR/MATTR, entropy
- Sentiment analysis and Word clouds (see below)

[3] See https://kramerius.nkp.cz/kramerius/Welcome.do
[4] See https://www.mlp.cz/cz/katalog-on-line/

- Search for word types by tags and searching word types
- Concordance
- Collocations
- Searching lemmas and their forms, frequency dictionaries, searching tokens by morphological tags

All these functions are implemented in the web environment as php scripts are that linked to the corresponding MYSQL databases (see Fig. 1). The overall structure of the site can be seen in Fig. 2.

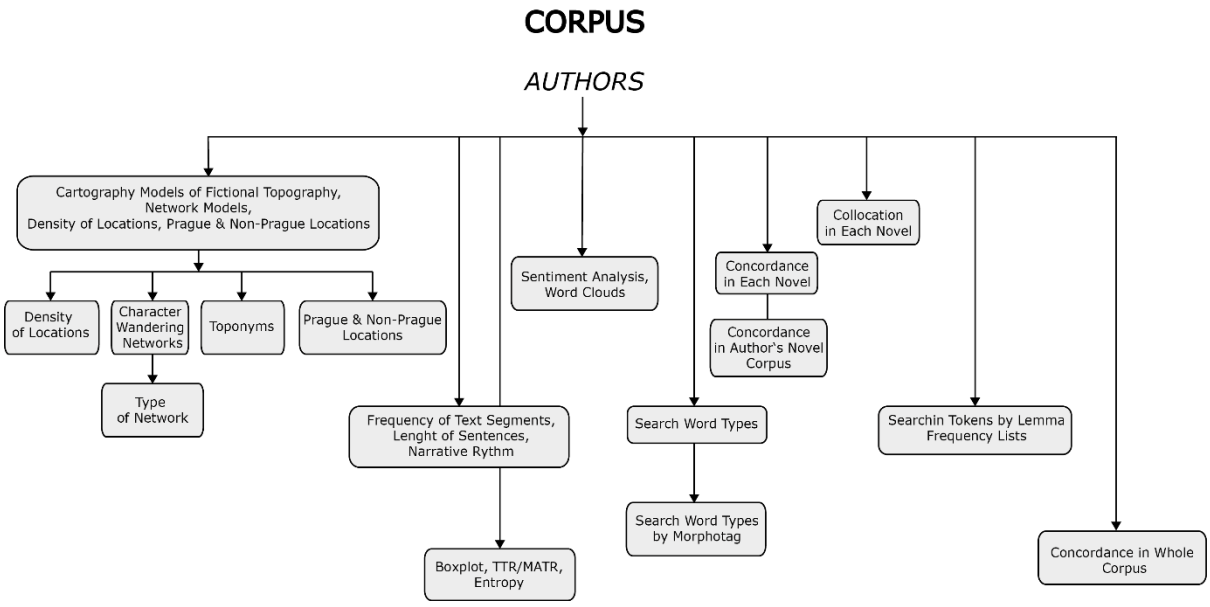

Fig. 2: Structure of web application

Research internship and its benefits

During my research stay in Trier, my goal was to work on 30 novels by Jakub Arbes, a Czech prose writer of the second half of the 19th century who significantly thematized the Prague environment in his prose. Gradually, I not only implemented Arbes's prose into the language corpus, but also created a new sentiment analysis feature in the web application called Clusters of Sentiment. Its essence is the search of literary texts based on a dictionary of emotions, which was compiled on the basis of the entries of the thematic thesaurus of Czech.[5] The overview of the project and its functionalities presented so far shows its multifunctional potential, as well as the possibilities of further not only material but also functional expansion. One of the essential functionalities, which appears to be highly useful in such a project, is stylometry. However, no stylometric tools or models are yet included in the project.
In this respect, it turned out that the host institution (University of Trier) not only has a high level of expertise, which is oriented, among other things, on this issue, but also a wide

[5] See Klégr, Aleš et al. Thesaurus of the Czech language. Prague: NLN - Lidové noviny Publishing House, 2007.

experience in application of stylometric analyses on historical textual materials. During my research stay at the University of Trier, I participated in the workshop *Potentials and Limits of Stylometry for Early Modern Text in Romance Languages* at the invitation of Prof. Christof Schoch, which proved to be a significantly inspiring source for the further direction of my project. Simon Dagenais from the Univesity of Trier in his paper OCR *Problems and Stylometrical Analysis of the Pamphlets of the Pro-Maupeou Pamphlets* talked about the possibilities of OCR of historical texts. Besides ABBY Fine Reader software, he was particularly interested in the possibilities of using GPT for text recognition. He deliberately used shorter text passages to train the model in order to achieve better results. Among other things, S. Deganais talked about the usefulness of the STYLO tool for stylometry, which is primarily programmed in R, but which can also be implemented in Python, which is also my working environment for quantitative analyses of literary texts. This led me to study François Dominic Laramée's text Introduction to stylometry with Python, which discusses some stylometry methods with practical coding examples in Python.
Matthew McDonald from the University of Trier gave a very inspiring talk in his contribution entitled *From Style to Stylometry,* which dealt with measuring stylistic differences in different French texts written in different provenances and years. This research aspect led me to the idea of attempting to apply stylometric analysis to the corpus of Czech prose being built, not for the purpose of determining authorship, but to distinguish the stylistic features that define not only monolithic authors, but more importantly the poetics of different historical stages. For such an analysis, it would probably be appropriate to work not only with the most frequent words (synsemantics), but also with other aspects, e.g. sentence length, syntactic constructions, the way adjectives develop their nouns, motivics, sentiment analysis clusters, etc. The aim of such a procedure would be to find a certain stylistic invariant that could define what texts of a certain period, e.g. Czech Romanticism in prose, etc., have in common on the most general levels. Such research would not only extend and significantly enrich my project, but at the same time would fit into the current tendency of research on historical processes of Czech literature, which are treated in contemporary Czech literary history as non-linear, so-called synoptic-pulsational processes. [6]
Equally inspiring was the presentation by Julia Röttgermann and Johanna Konstanciak from the University of Trier representing the research of the Mining and Modelling Text (MiMoText) research group. Within the framework of stylometry, the authors make use of well-known Genettian narrative categories. With regard to my project, such approach is more than interesting. When I asked them how they extract or delimit these categories, the authors replied that they have already obtained the information in the form of metadata.
The paper of Julian Csaspo from the University of Trier *Diderot's Contributions to the Histoire des deux Indes: A stylometric analysis* introduced me to the basic methods of stylometry: cluster-analysis, bootstrap-analysis network and PCA-analysis.

In addition to a very useful professional discussion about stylometry, I also received full support at Trier, including the possibility to use the office with technology, the university library and the possibility to consult any problem in solving my project at any time.

[6] See e.g. Tureček, Dalibor. Czech Literary Romanticism. Brno: Host 2012.

Project sustainability and future prospects

The above stated findings led me to several possibilities to take my project to the next level, which would include stylometric methods and models. Specifically, the following framework for future project development:

- Segment the corpus according to the authors and works that literary history classifies as representing a certain stage of development, e.g. Czech literary romanticism, realism, the so-called májovci, ruchovci, lumírovci, moderna, etc.
- To observe, by means of stylometric analysis, the similarities and differences between these groups and to try to establish common invariant characteristics for each group.
- To try to find within each group, by means of a stylometric analysis, those works that represent extreme positions and are close to works from groups, thus allowing a more accurate and objective reinterpretation of the synoptic-pulsational map of literary-historical processes.

Evaluation of the research stay

During my one-month research stay at the University of Trier, specifically at the Kompetenzzentrum - Trier Center for Digital Humanities (TCDH) under the supervision of Prof. Christof Schoch, I first of all supplemented the corpus with the announced novels by Jakub Arbes, implemented a new cluster of sentiment function in the web application, which meant generating quantitative values for all existing texts in the corpus, and above all participated in an expert debate on stylometry, its possibilities and methods. Considering the nature of my project, I consider these results as clearly positive and developing the project in a further direction.

Also with regard to the recently established Olomouc Centre for DH (CDHLBS)[7] a possible cooperation between TCDH and CDHLBS is proposed for the future. I have invited Prof. Schoch to visit Palacky University in Olomouc in 2024, where he could present the work of his center to the wider professional community. The planned lecture is a matter of coordination at the moment, which will be based on the time and organizational possibilities of Prof. Schoch and Palacký University in Olomouc.

Overall, I rate my stay within the CLS INFRA fellowship very positively. The grant allowed me to meet experts in the field of research that I have been doing in the Czech Republic for the last more than ten years. It allowed me to visit a modern workplace where such research is conducted and where its members show excellent outputs, either in the form of publications or concrete applications.

In Olomouc 18. 11. 2023

[7] See https://cdhlbs.upol.cz/