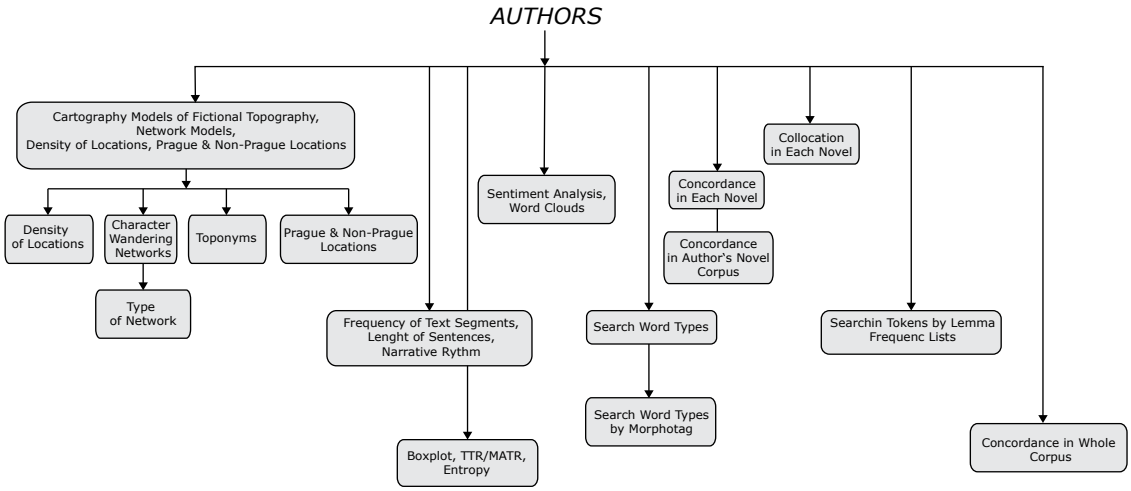
Manual for Digital Corpus „Literary Cartographic and Quantitative Models of Czech Novels from the 19th to 21st Century„

This manual introduces a special digital corpus called Literary-Cartographic and Quantitative Models of Czech Prose of the 19th-21st Centuries, which includes Czech prose of a defined period that significantly thematizes the Prague environment. Although this is the criterion of the original selection, the possibilities of the corpus are broader, which means that in the future the corpus will also include those prose works that do not thematise the Prague topography. The manual informs about the structure of the project's website, explains how the data is processed and describes in detail the individual functionalities of the project and how to work with them.

1. Structure of Corpus



Cartograpy Models of Fictional Topography, Network Models, Denisty of Locations, Prague & Non-Prague Locations

The diagram above shows the structure of the web presentation. On the left side there are pages on which literary-cartographic models of the fictional Prague topography, i.e. the ways in which the Prague environment is structured in individual prose works, are or will be gradually presented.

The map displayed on the page is a vector model (blind map) of a real historical map, which can be displayed by clicking on the basic map box. Below this button there are links to uniform works. Clicking on each of these will display the literary-cartographic model for the selected work.

In each model, the places, areas, premises or paths of the characters are marked with different colours that refer to the type of place and to the pronunciation in the narrative that refers to the place. Explanatory notes on the typology of sites can be accessed here.

In the bar above the map there are links to other pages, where you can find information about the density of locations in selected episodes (Desnity of Locations), trajectories of character movement (Character Wandering Networks), a graph of the frequency of toponyms in individual episodes (Toponyms) and a frequency graph showing which non-Prague locations (later Prague districts) are represented in specific episodes.

Under the Character Wandering Networks link, you will find another link to a page where network models showing general network type character travel trajectories are located.

Frequency of Text Segments, Lenght of Text Segments, Narrative Rythm

This website provides: sentence length charts for selected text segments of the narrative, which show the frequency of a given segment in the single-volume works of a given author; an overview chart of the frequency of segments throughout the work; and text segment frequency charts for specific works; and a narrative rhythm frequency chart, which shows the frequency disrtibution of narrator categories in specific works.

Boxplots, TTR/MATR, Entropy

On this web site you founf te graphs show: boxplots, moving average type-token ratio (MATTR) and entropy especialy relative entropy for each work. A high MATTR indicates a high degree of lexical variation while a low MATTR indicates the opposite. This means that a higher MATTR indicates more different types and vice versa. Entropy indicates the degree of uncertainty of the system. A window size of 100 is used to calculate MATTR.

Sentiment Analysis & Word Clouds

The sentiment analysis graph shows the so-called balancing values, which are determined by the difference between the measured absolute positive and negative values. In the bottom menu you can select between the different parts. Once the cluster model is displayed, it can be clicked on to view it in a larger format in a new window.

Search for Word Types by Tags & Search Word Type

To search for each word type, type tag (see table on the left) in the appropriate row with the title of the work. You can specify the extent of the listing by specifying number of lines. To see the percentage distribution of word types in a given work, choose from the menu on the right.

Concordance

Type lemma in the appropriate line. The concordance listing contains five positions on the left and right (including punctuation). When searching repeatedly or searching in other fields, first press the Refresh button.

On the web page you find thew link to another web page where you can search in all authors corpus.

Collocations

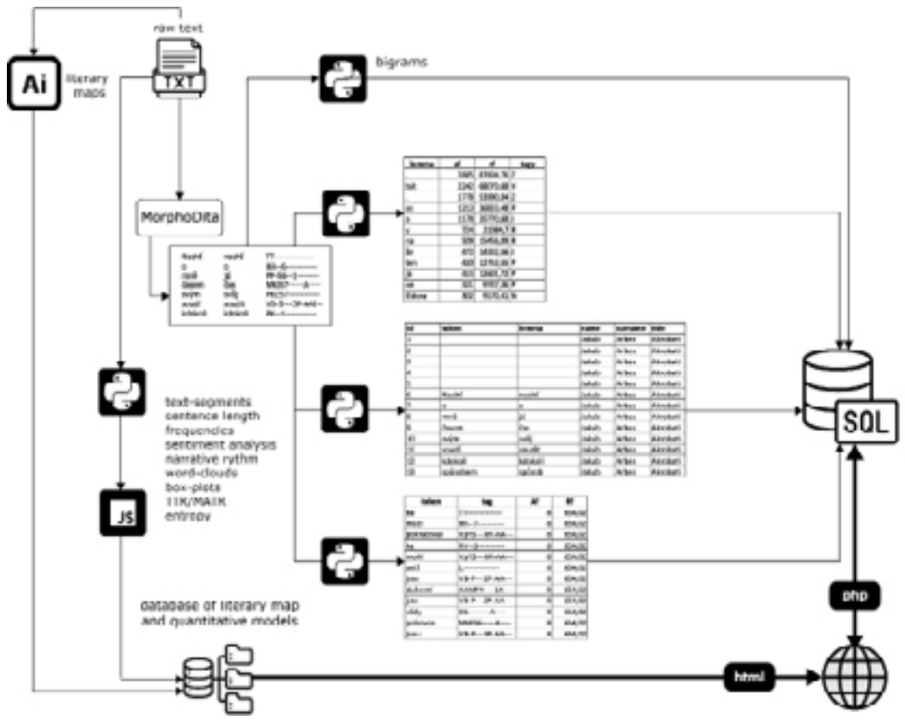
Type lemma in the search bar. The tables will list the collocations sorted by logDice, MI-score and T-score. For the calculation, see . The → symbol indicates the right context, the ← symbol indicates the left context. Collocations are calculated from the list of bigrams.

Searching Lemmas and Their Forms, Frequency Dictionaries, Search Tokens by Morphological Tags

In the line, type the lemma whose frequency and shapes you want to search for in the piece. When the results are displayed, at the end of the line of found shapes for that lemma, you will see a PDF file with a complete frequency list of all the lemmas in that part.

---

2. Structure of Data Processing



The digital text of the work or OCR text is cleaned and saved in a TXT file (UTF-8). It is then lemmatized using Morphodita. From the lemmatized file, the corresponding tables are created using Python scripts and exported to SQL databases, from where they are called via php scripts to the web browser.

The literary-cartographic maps are processed in Adobe Illustrator from historical map templates and form a base (blind map) onto which individual models of fictional topographies are projected in layers.

On the project homepage you will find links to news &archived inforamtions, open data, tutorials, publications and other activities related to the presentation of this project and a list of links to similarly themed projects.