TNA ACTIVITY REPORT

Digital Mapping of the Historical Landscape of Baltic Germans Using Named-Entity Recognition (NER) and geographical mapping

Author: Anna Baryshnikova

Current position: Research Assistant at the Chair of Computational Corpus Linguistics

Affiliation: University of Erlangen-Nuremberg

Host institution: Charles University, Prague

Mentor(s): Michal Křen

Period of stay: 1 September 2024 until 26 October 2024

Introduction

The project “Digital Mapping of the Historical Landscape of Baltic Germans” focuses on exploring and visualizing the history and culture of the Baltic Germans using modern digital tools. The primary goal was to create a digital map highlighting important locations and geographical patterns documented in the newspaper *Baltische Briefe*. This project aimed to investigate the cultural identity of the Baltic Germans and their historical significance. It was conducted in close collaboration with the Institute of the National Czech Corpus, whose expertise in linguistic analysis and Named-Entity Recognition (NER) played a vital role in the project’s success.

Background

The Baltic Germans were a German-speaking ethnic group that shaped the social and cultural life of the Baltic region for centuries. Following their resettlement during World War II, they left behind a rich legacy of history and traditions. The newspaper *Baltische Briefe*, published from 1949 to the early 2000s, serves as a crucial source for documenting this history. By analyzing and mapping the locations mentioned in this newspaper, the project provided new insights into the movements and cultural identity of the Baltic Germans.

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

Research Questions

The project was driven by the following key questions:
1. How can digital geographical analysis tools be applied to historical texts like *Baltische Briefe*?
2. What places are significant in these texts, and what roles did they play in the history of the Baltic Germans?
3. How can the connection between geographical patterns and the cultural identity of the Baltic Germans be visualized?

Data Collection and Corpus Compilation

Five key editions of *Baltische Briefe* were selected for analysis, covering the years 1949, 1950, 1988, 1991, and 2004. The 1949 and 1950 issues reflect the early postwar period and the beginning of the Baltic German diaspora's efforts to preserve its cultural identity. The 1988 and 1991 issues capture the cultural awakening and geopolitical changes associated with the independence of the Baltic states from the Soviet Union. The 2004 issue marks a new chapter with the Baltic States' integration into the European Union, symbolizing a reconnection with broader European identity and politics. The editions were digitized from the Bavarian State Library in Munich and the Berlin State Library.

The corpus was compiled with the technical support of the Institute of the National Czech Corpus. Annotation followed the Universal Dependencies framework and included:

- **Parts of Speech and Morphology:** Annotated to reflect the grammatical structure of the text. (Universal Dependencies)
- **Syntax:** Annotated for syntactic dependencies, allowing deeper insights into the structure and relationships within sentences. (Universal Dependencies)
- **Named Entities:** Annotated using NameTag3 to identify people, organizations, and locations. Location annotations were enriched with geographical coordinates and linked to Wikidata for reference and contextual information.

This richly annotated corpus provided a solid foundation for analysis and geographical mapping.

Methodology

The methodology consisted of several phases, where collaboration with the Institute of the National Czech Corpus was crucial:
1. Digitizing the Texts: The OCR software ABBYY FineReader was used to convert printed editions into machine-readable text. The cleaned data was then converted into XML formats to enable structured processing.
2. Named-Entity Recognition (NER): Supported by the National Czech Corpus, geographic entities such as cities, regions, and countries were extracted. Tools like NameTag3 and the SpaCy library were used, with linguistic expertise from the institute aiding in adapting the models to historical language variants.

3. Corpus Compilation and Annotation**:** The corpus was annotated to include parts of speech, morphology, syntax, and named entities. Annotated locations were geocoded with coordinates and linked to Wikidata, ensuring that each entity could be connected to a wider knowledge base.
4. Geographical Mapping: The identified places were geocoded into geographical coordinates (latitude and longitude). Maps were created using OpenStreetMap and historical maps, with customized markers to distinguish different types of locations.
5. Digital Map Development**:** To enhance visualization and interactivity, digital maps were developed using OpenStreetMap and Python libraries such as Folium. These maps displayed geographical locations and enriched the user experience by including linked Wikidata URLs, short descriptions of the places, and accompanying photographs.

Challenges and Solutions

A major challenge involved standardizing historical place names. Historical names such as “Dorpat” (modern Tartu) or “Ritterstraße” (now Rüütli) required cross-referencing with historical maps and gazetteers. The National Czech Corpus supported this process by providing linguistic resources specifically designed for analyzing historical texts.

Results

- The map revealed geographical patterns that traced the history of the Baltic Germans, including key cities like Riga, Tallinn, and Lüneburg.
- The annotated corpus provided a rich linguistic resource for understanding the historical language and cultural context of Baltische Briefe.
- The digital map highlighted the close connection between the Baltic Germans and their geographical environment. The analysis of the newspaper content also uncovered a rich cultural heritage made tangible through the mapped locations.
- The inclusion of Wikidata links, descriptions, and photos added depth to the map, transforming it into an interactive resource for historical research.
- The tools and methods developed can be readily applied to other historical texts, offering potential for further studies.

Outlook

The project opens up numerous avenues for future research:

- Expanding the analysis to include additional historical documents for a broader understanding of the Baltic Germans’ history.
- Incorporating additional datasets, such as demographic or economic information, to enrich the map and enable deeper analysis.
- Developing new features for the digital map, such as interactive layers that visualize historical events or cultural characteristics.

Conclusion

By applying modern digital tools such as OCR, NER, and geographical mapping techniques, the project successfully analyzed the content of *Baltische Briefe* in an innovative way. Collaboration with the Institute of the National Czech Corpus was instrumental, particularly in adapting linguistic tools to the specific requirements of historical texts.

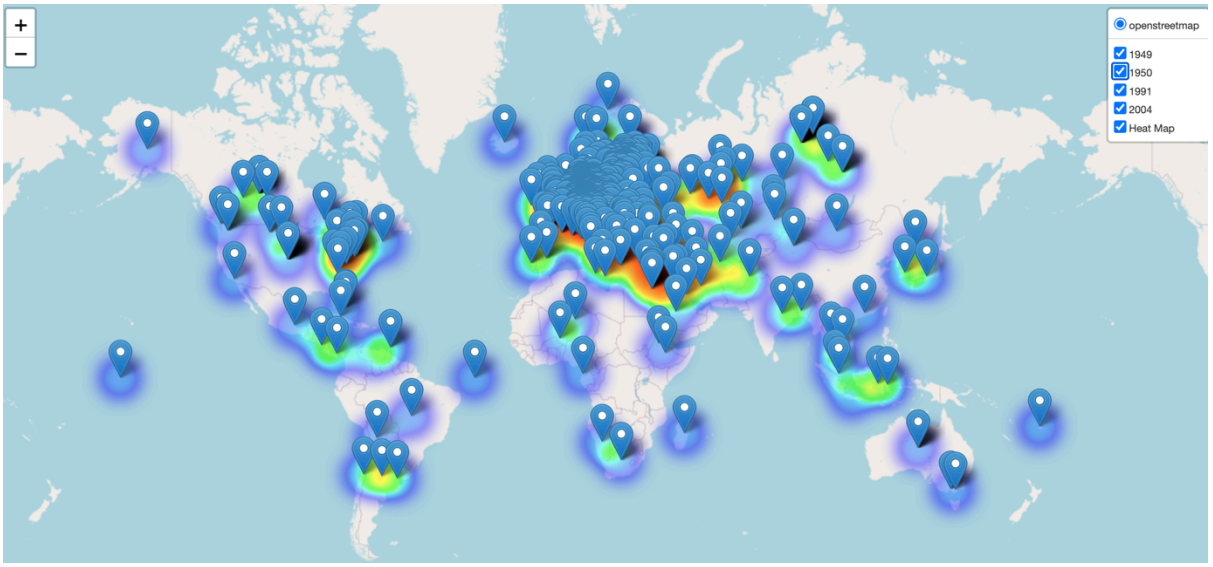
This project demonstrates how interdisciplinary approaches in the digital humanities can not only facilitate research on historical topics but also open up new perspectives. The developed map is a valuable resource for documenting and visualizing the cultural and historical identity of the Baltic Germans.

Through this work, I not only deepened my skills in digital humanities but also laid a solid foundation for future research endeavours.

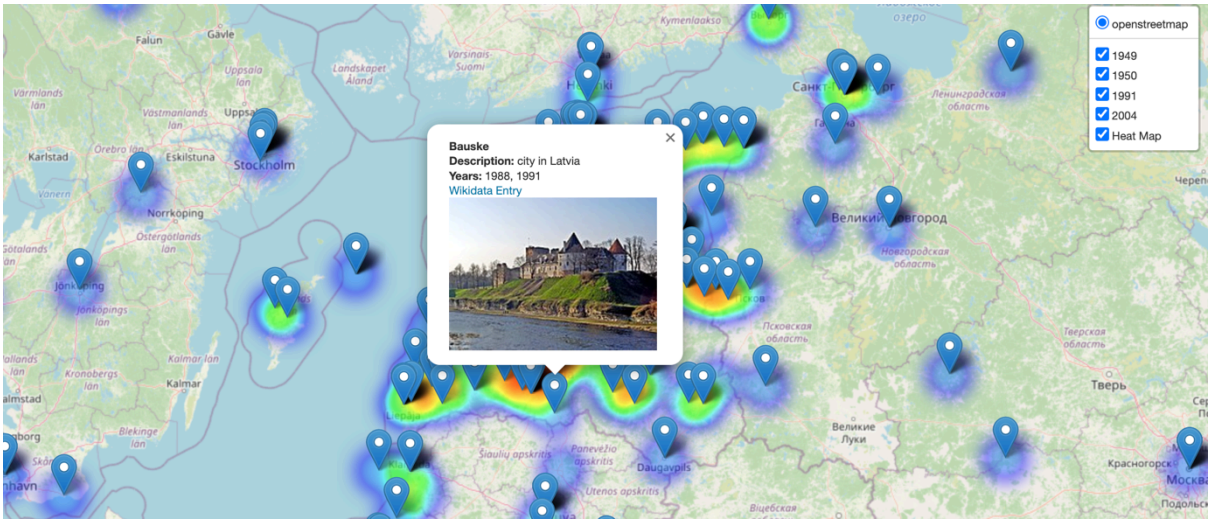
Supplementary Materials

1. Examples of the Digital Map

Below are two pictures showcasing the interactive digital map developed during this project. The map visualizes the geographical locations extracted from *Baltische Briefe*, enriched with metadata such as linked Wikidata URLs, short descriptions, and photos.



The final heat map with filtering option based on the publication years.



Example of the location (Bauska – city in Latvia) visualization enriched with a short description, a link to wikidata and a picture.

2. Publicly Available Corpus

The corpus compiled from *Baltische Briefe* is publicly accessible and can be explored at the following link:
Baltische Briefe Corpus on the Czech National Corpus
https://wiki.korpus.cz/doku.php/en:cnk:baltischebriefe