Traveling through Space and Time, or: Making Historical Travelogues Accessible.

Jan Rörden[1][0000−0002−5824−8397], Bernhard Haslhofer[1][0000−0002−0415−4491], Rainer Simon[1][0000−0002−4116−9684], and Sven Schlarb[1]

AIT Austrian Institute of Technology, Vienna, Austria
{first}.{last}@ait.ac.at

Abstract. Investigating perceptions of Otherness is the overarching goal of the Travelogues project. It studies a corpus comprising of thousands of recently digitized travelogues dating back to the 16th century held by the Austrian National Library. Driven by an interdisciplinary team of historians and data scientists, it aims at making knowledge that is now hidden in a huge text corpus accessible to researchers. In the current, initial project phase, we explore how statistical methods, such as word embeddings, can be used to assess the structure and semantics of large text corpora in order to make those resources accessible. We developed an initial methodology that combines visual and statistical cues for identifying possible starting points for a more fine-grained text corpus exploration. Ultimately, this data-driven approach is expected to result in new and possibly unexpected insights stemming from resources that were previously de-facto inaccessible.

Keywords: Digital Humanities · Machine learning · Information extraction.

1 Introduction

Close reading (the careful and exhaustive interpretation of a passage of text) of source material is a key methodology and daily routine for historians. However, this methodology has its natural limits when being applied on large digitized book holdings such as the Austrian Books Online[1], which comprises more than 400,000 volumes including titles from the early 16th century up to the second half of the 19th century.

The Travelogues[2] project focuses on historical travel reports ranging from the 16th to late 19th century, published in German language, and works towards developing a toolbox that helps exploring source materials and understanding their semantics. Aiming for qualitative insights into historical sources, we describe a set of methods and their combination that can help domain experts (in our case, historians) in finding answers to research questions based on source material hidden in huge text corpora.

[1] https://www.onb.ac.at/digitale-bibliothek-kataloge/austrian-books-online-abo/
[2] http://www.travelogues-project.info

In addition to applying well-known information retrieval methods, we also aim at enabling the presentation of results in ways that are easily accessible by non-experts as well. Uncovering previously hidden knowledge is the priority of this project.

2 The Dataset

The corpus in question consists of more than 150,000 recently digitized works in German language (out of a total 419,000 for the given time frame), which are part of the Austrian National Library's inventory, with all works published between 1500 and 1875.[3] Out of those, an estimated 1,000 to 2,000 books can be classified as travel reports, either completely or containing chapters describing travels. Travel reports are an important source genre in the historical science [3], and serial analysis has been suggested as a way to analyze them [1].

Innate and challenging properties of this corpus include, for example, the linguistic variety, which comes from works spanning 450 years of German language, most of it without standardized orthography. Additionally, problems arise due to the digitization process such as OCR errors (especially with texts printed in Fraktur) and loss of information due to missing layout information. For example, there can be the unclean appearance of individual characters due to old printing techniques, letters from the inverse page can be visible because of thin book page paper (bleed-through), or the book pages might be damaged due to storage conditions (e.g. humidity caused page warping) or by catastrophic events (e.g. fire). On top of that, the correct classification of travel reports is a problem to be solved. This is also due to incomplete or incorrect meta data from the existing catalog, which was in parts created in the 19th century.

3 Methodology

In a first step, we make use of the word2vec algorithm [5, 4] and visualizations through t-SNE [2].

To develop our methodology, we worked on a small subset of the corpus. This subset consists of 100 books, published between 1800-1875 and selected through basic criteria - the title had to contain the German word for travel, 'Reise', in one of four different spellings (to cover language variations: 'reis*', 'reyß*', 'reiß*', 'reys*'). The result was a mid-sized corpus of just over 10 million words.

We kept the preprocessing of the source material to a minimum, to preserve as much of the original structure and wording of the text as possible. Those aspects were deemed important for further language analysis.

We then applied word2vec[4] to extract semantic relationships contained in the test corpus, and plotted word clusters using tSNE [2].

[3] Letterpress printing was widely available from the early 16th century, and after 1875, copyright issues might arise - hence this time frame.

[4] The gensim[6] implementation was used.

4 Preliminary results

The first results and observations are promising, as we were quickly able to get an intuition of the main topics of the source material. Plotting the 500 most frequent words immediately drew attention to thematic clusters, such as settlements (different types of buildings from huts to fortifications to cathedrals), nobility (rulers, tribes etc.), food (types of meat, methods of preparation), seasons and also abstract topics like emotions. By plotting the semantically closest words to 'danger' (see figure 1), we are able to draw a picture of dangers associated with traveling, as far as our subset is concerned. This includes seafaring, bandits and bad roads; illness; but also feelings like fear, anger and sadness. Interestingly, many adjectives were also included - hinting that this topic is especially connected to vivid descriptions and emotions. Other topics centered on food or cities offer insights into contemporary descriptions.

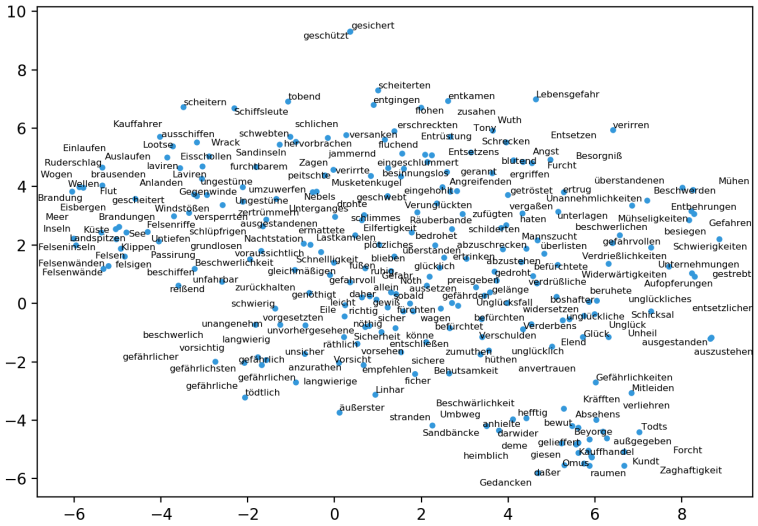

Fig. 1. Focus on the 'danger' cluster.

5 Conclusion and further work

We have learned that text-exploration techniques like word embeddings can be applied to a basically unknown text corpus in a domain that is challenging

through linguistic variety and OCR errors. In the long run, we plan to update the library catalogue with additional and corrected meta data. For the next steps, three historians picked a variety of representative travelogues, published between 1700-1875. They consist of just over 2.5 million words, and we will use them as the starting point in creating a classifier to identify additional, non-obvious travelogues in the corpus.

We also plan to introduce additional approaches that can be used for information extraction, such as topic models and tf-idf. Furthermore, the ground truth will be manually annotated using the software platform Recogito[5] to mark certain semantic structures. The goal of this approach is the identification of the semantic concepts that describe notions of Otherness and its evolution through time.

Acknowledgements

This work is supported by the Austrian FWF as project I 3795 and the German DFG as project 398697847.

References

1. Harbsmeier, M.: Reisebeschreibungen als mentalitätsgeschichtliche quellen: Überlegungen zu einer historisch-anthropologischen untersuchung frühneuzeitlicher deutscher reisebeschreibungen. Reiseberichte als Quellen europäischer Kulturgeschichte, hg. v. A. Mączak, HJ Teuteberg pp. 1–31 (1982)
2. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
3. Maczak, A., Teuteberg, H.J.: Reiseberichte als Quellen europäischer Kulturgeschichte: Aufgaben und Möglichkeiten der historischen Reiseforschung:[Vorträge gehalten anlässlich des 9. Wolfenbütteler Symposions vom 22. bis 25. Juni 1981 in der Herzog August Bibliothek]. Herzog August Bibliothek (1982)
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
6. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), `http://is.muni.cz/publication/884893/en`

[5] https://recogito.pelagios.org/