A quantitative analysis of the semantics of fictional spatial entities in the German literature of the 19th century using SUNSET

UNIVERSITÄT BIELEFELD

Keywords: Distant Reading, Language Models, Machine Learning, Spatial Literary Studies, Computational Literary Studies

Daniel Kababgi , Berenike Herrmann

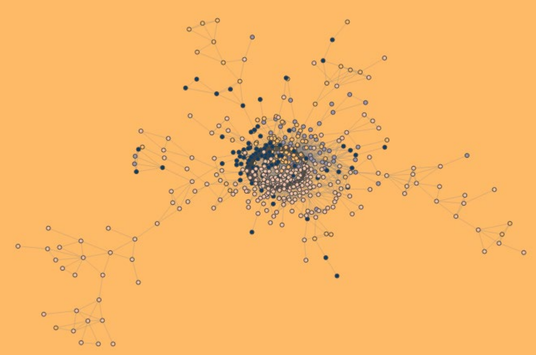
Goals and Theory

Research Aims:

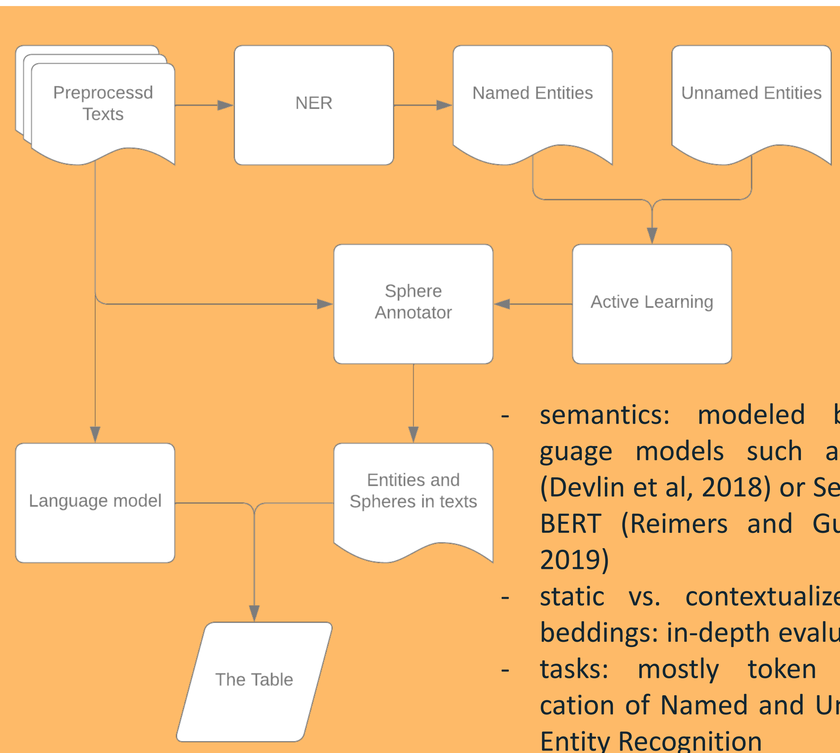
- get a deeper understanding of the semantic role of space in narratives
- operationalize **spatial semantics** (Lotman, 1972) and apply to analysis of literary texts of the long 19th century written in German
- delineate concept of spatial semantics
- introduce concept of **spatial spheres**
  - three spheres: **"nature"**, **"village"** and **"city"**
  - analyze in depth for diachronic change
  - analyze role of the "village" sphere: co-evaluation/co-variation?
  - hypothesis: “village” as an aggregation of the semantics of "nature" and "city"
- possible future research avenues: differences in semantics of spatial spheres regarding gender and nationality of authors, as well as literary periods, genres, and high vs. low literature

Spatial Spheres:

- definition: a spatial sphere is an amalgamation of different, but semantically close, spatial semantic fields
- **“nature”** is encoded by semantic entities largely independent by civilization
- as Lobsien (1981) and Zelle (1989) note, “**nature”** is semantically diverse (e.g., the sublime vs. the relaxing)
- **"village"** aggregates rural spatial entities, seen in the popular “*Dorfgeschichten*" (Twellmann, 2019)
- **"city"** is an amalgamation of fully urbanized spatial entities, that rose to prominence in the later half of the 19th century, as Corbineau-Hoffmann (2003) notes



SUNSET Pipeline



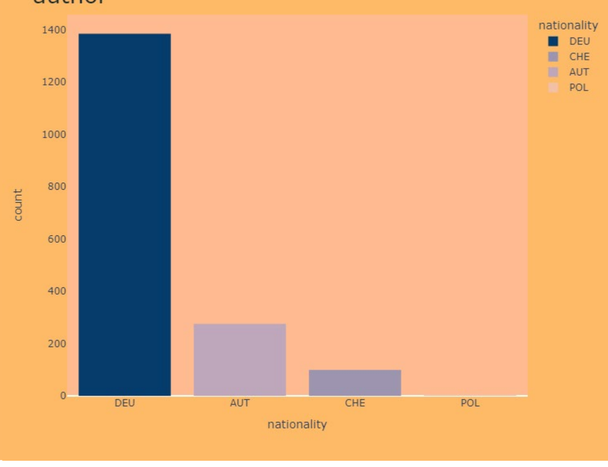
- **SUNSET** = **S**emantics of **U**nnamed and **N**amed **S**patial **E**ntities in **T**exts
- builds on theoretical and methodological foundation introduced by Schumacher (2023)
- semantics: modeled by language models such as BERT (Devlin et al, 2018) or Sentence-BERT (Reimers and Gurevych, 2019)
- static vs. contextualized embeddings: in-depth evaluation
- tasks: mostly token classification of Named and Unnamed Entity Recognition
- Active Learning for a better understanding of the texts and the model through closeness to data
- semantics of spatial entities extracted by finding the most closely related tokens

Corpus

Corpus Construction:

- first iteration of corpus
- German narrative prose texts (novels, novellas and short stories), high and low literature between 1784 - 1918
- texts by N = 372 authors born 1720 – 1895
- 4 different nationalities (see Fig 1)
- gathered from sources DTA, Zeno.org and Projekt-Gutenberg-DE
- duplicates: preferred text was selected by source
- there are more authors and resources (Kolimo+, ELTeC, TextGrid, etc.) available

Fig 1: Distribution of texts by nationality of author



Corpus Description:

Fig 2: Distribution of texts by period and gender authors

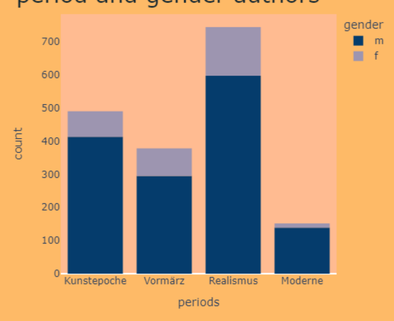


Fig 3: Distribution of text lengths per period

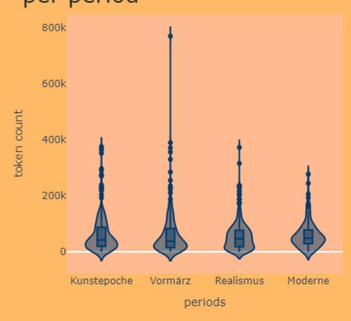
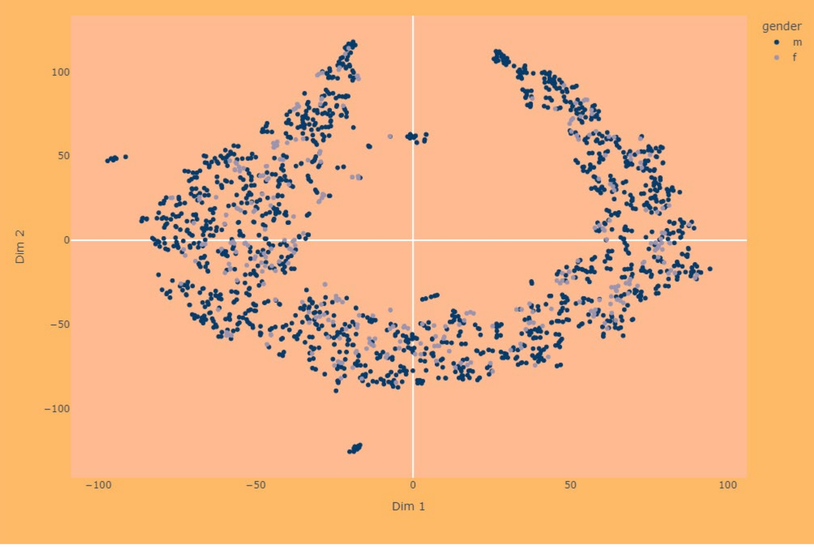


Fig 4: A first exploration of the corpus based on MFW and gender



Overall Oberservations

- complete corpus: N = 101 mio. tokens
- longest text: Karl Gutzkow’s *Die Ritter vom Geiste* (n= 770,456 tokens)
- most texts (N = 1,384 or 78.45%) by German authors (see Fig 1)
- median length of texts in corpus: 44,094 tokens, but grows over time (see Fig 3)
- N = 1,444 texts by male authors (81.85%), especially in Moderne
- N = 320 texts by female authors (18.14%)

First Clustering Experiment

- results shown in Fig 4
- texts colored by author gender
- generated by tSNE (perplexity: 10, iterations 5000) based on MFW
- stopwords were excluded; lowercased & normalized with CAB (Jurish, 2012)
- first results:
  - few clusters clearly separated
  - no clear separation by gender
  - seemingly one small cluster at the left tip
  - Rudolf von Tavel‘s („Mundartdichter“) texts group together (Dim 1 = -15, Dim 2 = -120)
  - two different mini clusters for the *Kinder- und Hausmärchen* (several editions 1817-1852)
  - Karl Gutzkow’s *Ritter vom Geiste* at the spearhead of the separated cluster at left tip

Next Steps and Challenges

- metadata: lack of standardization of three used online resources. What are possible ways (next to manual correction) to handle this problem?
- bias: are there ways to address the corpus’ imbalance regarding author gender and/or nationality?
- messiness: “messy” texts especially from Zeno can influence possible quantitative analysis. The problem is advertisement on the website.
- features & phenomena: how to tackle spatial metaphors (e.g. *“sich auf den Weg machen”)?*
- features & phenomena: how to tackle the problem of spatial entities in several different spheres? (e.g., “house” in “village” or “city”)
- evaluation: how can we evaluate the proposed method?
- next steps:
  - further prototyping of the pipeline
  - extending text collection (Gutenberg, retrodigitization, other resources)
  - complementary close reading of relevant texts for possible evaluation and selection of criteria?

References

- Corbineau-Hoffmann, Angelika (2003). Kleine Literaturgeschichte der Grossstadt. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. /abs/1810.04805.
- Jurish, Bryan. 2012. „Finite-state Canonicalization Techniques for Historical German“. Universität Potsdam. https://www.deutschestextarchiv.de/public/cab/.
- Lobsien, Eckhard (1981). Landschaft in Texten: Zu Geschichte und Phänomenologie der literarischen Beschreibung. Studien zur allgemeinen und vergleichenden Literaturwissenschaft 23. Stuttgart: J.B. Metzler.
- Lotman, Juri (1972). Die Struktur literarischer Texte. Uni-Taschenbücher 103. München: W. Fink
- Reimers, Nils, und Iryna Gurevych. 2019. „Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks“.
- Schumacher, Mareike K. (2023). Orte und Räume in Romanen. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Twellmann, Marcus (2019). Dorfgeschichten: Wie die Welt zur Literatur kommt. Göttingen: Wallstein Verlag.
- Zelle, Carsten (1989). “Schönheit Und Erhabenheit: Der Anfang doppelter Ästhetik bei Boileau, Dennis, Bodemer und Breitinger”. In: Das Erhabene: Zwischen Grenzerfahrung Und Größenwahn. Ed. by Christine Pries. DE GRUYTER, pp. 55–73.

Daniel Kababgi: daniel.kababgi@uni-bielefeld.de, ORCID iD: 0009-0002-0990-6418
Berenike Herrmann: berenike.herrmann@uni-bielefeld.de, ORCID iD: 0000-0002-5256-0566