BOOK OF ABSTRACTS

DIGITAL HUMANITIES 2023

DH 2023
Collaboration
as Opportunity

ALLIANCE OF
DIGITAL
HUMANITIES
ORGANIZATIONS

Digital Humanities 2023: Book of Abstracts

Editors
Anne Baillot, Le Mans Université, ORCID: 0000-0002-4593-059X
Walter Scholger, University of Graz, ORCID: 0000-0002-9256-0958
Toma Tasovac, Belgrade Center for Digital Humanities, ORCID: 0000-0002-3919-993X
Georg Vogeler, University of Graz, ORCID: 0000-0002-1726-1712

Data Management
Elisabeth Raunig, University of Graz, ORCID: 0009-0007-9019-8215
Martina Scholger, University of Graz, ORCID: 0000-0003-1438-3236
Elisabeth Steiner, University of Graz, ORCID: 0000-0001-9116-0402

Publisher
Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities,
University of Graz

Citation Recommendation
Baillot, Anne / Scholger, Walter / Tasovac, Toma / Vogeler, Georg / Raunig, Elisabeth / Scholger, Martina / Steiner, Elisabeth (eds.). Digital Humanities 2023: Book of Abstracts. Graz 2023. DOI: 10.5281/zenodo.7961822

Tracing the Shift to "Objectivity" in German Encyclopedias of the Long Nineteenth Century

Hagen, Thora
thora.hagen@uni-wuerzburg.de
University of Würzburg

Konle, Leonard
leonard.konle@uni-wuerzburg.de
University of Würzburg

Ketzan, Erik
KETZANE@tcd.ie
Trinity College Dublin

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg

Witt, Andreas
witt@ids-mannheim.de
University of Cologne, Leibniz Institute for the German Language

Introduction

European encyclopedias of the long nineteenth century can be seen as transitional texts between those of the Enlightenment, and encyclopedias of the 20th century and today, which strive to complete the Enlightenment's project of "impartial scientific objectivity." Though it may be unattainable, objectivity is and has been an important goal of the European encyclopedia (Loveland 2019).

This article seeks to examine historical encyclopedias in the light of objectivity in order to contribute to the broader field of encyclopedia studies. We present experiments to observe this shift toward greater objectivity in six TEI-annotated German encyclopedias from 1809 to 1911 (299,907 encyclopedia entries, 27,378,000 word tokens) from the EncycNet project (Hagen et al. 2020). We consider two dimensions of non-objectivity in particular: emotions and personal interpretation (Chen 2008). Table 1 presents the corpus, with entries classified into four groups (Hagen et al. 2022). [1] The DH study of encyclopedias is growing as the texts become available digitally (e.g. Grabus et al. 2019), with these experiments adding new knowledge.

Title	Word tokens	Entries	Mean entry lengths in tokens	People	Places	Abstract	Objects
Brockhaus Conversations-Lexikon oder kurzgefaßtes Handwörterbuch (1809-1811)	1,186,000	6,960	168	17.4%	22.2%	36.2%	24.2%
Damen Conversations Lexikon (1834-1838)	1,461,000	7,099	204	14.4%	20.3%	34.6%	30.7%
Brockhaus Bilder-Conversations-Lexikon (1837-1841)	2,604,000	7,049	371	15.4%	15.8%	37.6%	31.3%
Herders Conversations-Lexikon (1854-1857)	2,256,000	39,755	56	32.6%	24.2%	21.8%	21.4%
Meyers Großes Konversations-Lexikon (1905-1909)	17,437,000	156,264	111	21.6%	40.5%	19.3%	18.5%
Brockhaus Kleines Konversations-Lexikon (1911)	2,434,000	82,780	28	30.2%	25.1%	20.3%	24.4%

Table 1: Encyclopedia corpus.

The encyclopedias in our corpus are different in terms of content selection, depth of the concepts discussed (entry length), and presentation of content. All of these internal differences are potentially confounding variables in our experiments. For emotion detection, we used a sample of around 250 entries per encyclopedia that are 2000 to 4000 tokens long, as shorter entries tend to follow a different style. Randomly sampling or using all entries could have skewed the results due to different mean entry lengths.

Experiments

Surface Level Features

As Loveland writes, "In the seventeenth and eighteenth centuries, many researchers used first-person singular pronouns to narrate their scientific observations and experiments" (76). Loveland meant scientists generally, but later, "[b]y around 1900 [...] a consensus had emerged making encyclopedias off-limits to personal disclosures, explicit opinions, passionate language, and playfulness" (81). These suggest textual surface features which may be queried. For all queries, we attempted to remove direct and indirect speech from the results, as it caused a number of false positives which do not reflect the sentiment of the authors.

We first query the first person pronoun *ich*(German: I), which can be seen as a marker of an opinionated statement and, therefore, objectivity (Chen 2008). The pronoun emphasizes the individual authorship of entries, but that eventually disappeared as encyclopedias became written in a style which erased individual "voices" as part of a more objective tone. Examples from entries are "shall I say," "I conclude with" (*Brockhaus 1809*). Figure 1 implies a declining trend in the disappearance of first-person narration. Still, the earliest and highest results are only ~1%, suggesting that I was not a strong feature of long nineteenth-century German encyclopedias.

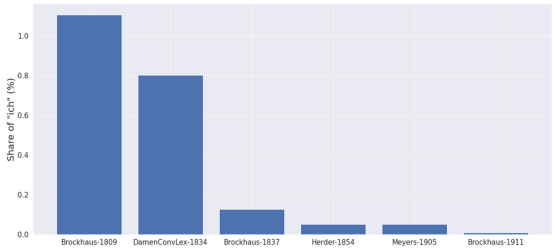

Figure 1: Occurrences of the first person singular, ich, relative to the amount of entries per encyclopedia.

Exclamation points suggest emotion and individual expression, rather than anonymous objectivity. Querying this punctuation mark also indicates a steady decline (Figure 2). The *DamenConvLex*, written for women, is a clear outlier, which is not surprising due to the "liveliness" and "romantic representation" intended by its editors (Ketzan et al. 2022). False positives, identified through manual inspection, were excluded with additional queries to approximate a precision of 1.0 for all results (e.g. to indicate surface areas: "100 ! M.").

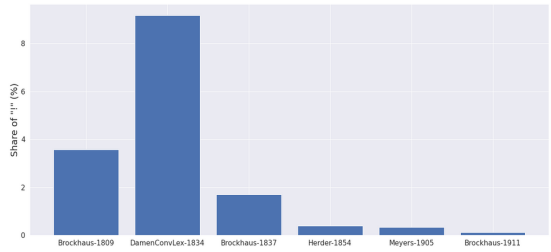

Figure 2: Occurrences of exclamation points relative to the amount of entries per encyclopedia.

Interjections are a class of words whose function, per Crystal (2008), is "purely emotive, e.g. Yuk! Strewth!" For identifying interjections, we tagged all entries with spaCy, [2] and counted tokens tagged as "ITJ". False positives took the form of abbreviations, which were excluded with regular expressions, as in the previous experiment. Examples of interjections include *yes* or *indeed*: "one tenth, one eighth, indeed probably one fourth" (*Brockhaus 1809*). A decline can be observed for interjections, but it is not clear whether they can indicate subjectiveness in this case, since the overall occurrences are quite low.

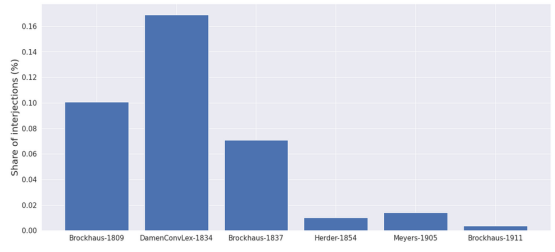

Figure 3: Occurrences of interjections relative to the amount of entries per encyclopedia.

Emotion Detection

The more literary tone of earlier encyclopedias which Loveland alludes to, full of "explicit opinions, passionate language, and playfulness," may be investigated by sentiment analysis, and indeed, we report that emotional language goes down in the texts over the course of the century.

In our sentiment analysis of the encyclopedia texts, we rely on existing models, choosing an emotion focused model trained on German poetry from 1859-1911 (Konle et al. 2022). [3] Lacking annotations for evaluation, we manually checked samples for their plausibility and found comprehensible predictions. Although poems and encyclopedias are quite different text types, their publication periods roughly coincide, which may explain the reasonable results. Figure 4 reports our results as a constant decline of emotion in encyclopedias over time; Figure 5 shows the distribution of emotions.

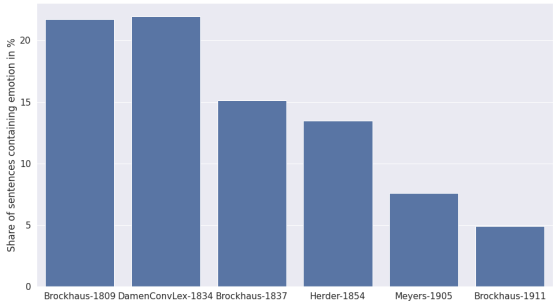

Figure 4: Share of sentences containing at least one emotion per encyclopedia.

Figure 5: Share of sentences containing an emotion per encyclopedia.

The extensive *Meyers* encyclopedia deviates from the others through an increased proportion of Joy. The most prominent outlier is the *DamenConvLex* with reduced Anger and large proportions of Love, Joy, and Sadness. Again, as this is an encyclopedia intended for women, this result merits further investigation.

The model erroneously associates certain word groups with emotions because they have high co-occurrences in the training material (e.g. "mother", "child" indicate Love). It also does not distinguish between carriers of emotions. As our goal is to investigate objectivity, it is highly relevant whether the expressed emotion is attached to the author of the text or the entity discussed.

While these problems are well known in sentiment analysis, the third issue is of conceptual nature and arises from the tension between emotion detection and objectivity. For example, there are

sentences in various encyclopedias which describe the disease Cholera. In one text, the symptoms of the disease are described objectively, [4] yet in another text, the effects of Cholera are described in figurative language. [5] Both were tagged as "Sadness" by the model, but this small example illustrates the limitations of our emotion detection method for the goal of tracing objectivity.

Conclusions and Future Work

We presented methods to detect objectivity in encyclopedia entries through textual markers and the detection of emotions with a pre-trained model. While both methods can reveal subjective statements, they both have room for improvements. The former will require more reliable exclusion of speech acts, and the latter a clear distinction between descriptions of a third person's emotions and emotional statements by the authors. This challenge especially could pose a new task in the field of sentiment analysis. In conclusion, the methods presented so far only suggest broad trends towards detecting objectiveness in encyclopedias, however more experiments and statistical validation are necessary to strengthen the results. More features of a text that might identify polarizing statements should be explored; this could be making cognitive bias in text visible, for example through hedge words. Other aspects could be the usage of tenses, tables, or figurative speech. Lastly, future work could explore the triangular constellation between science, public discourse, and the encyclopedia as a mediator from the former to the latter, by applying objectivity detection to other sources such as scientific texts and newspaper articles.

Notes

1. Classified entries include entries with more than 20 tokens. As not all entries could be found in Wikidata, around 30-40% of entries per encyclopedia remain unclassified for now. The percentages in Table 1 only include classified entries.
2. https://spacy.io/
3. We also tested german-sentiment-bert (Guhr et al. 2020) with negative results.
4. *Damen Conversations Lexikon*: "Das Krankheitsbild ist im Allgemeinen folgender Art : Nach kurzem Uebelbefinden und Störungen der Verdauung, besonders mehrtägigem unbedeutenden und leichten Abweichen entsteht heftiges Erbrechen und Absonderung schleimiger Theile in bedeutender Menge."
5. *Brockhaus 1837*: "Zeigt das Übel gleich von Anfang einen bösartigen Charakter, so macht es den Menschen schon im Beginn zum lebendigen Leichname; der ganze Körper wird starr und steif, eisig kalt und blau gefärbt, das bläulichrothe oder blaugraue Antlitz verräth das tiefste Leiden, und die obigen Erscheinungen treten mit erhöhter Heftigkeit ein."

Bibliography

Chen, W. (2008): "Dimensions of Subjectivity in Natural Language", in: *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*: 13-16.

Crystal, D. (2008): *A Dictionary of Linguistics and Phonetics*. Sixth Edition. Malden, MA: Blackwell Publishing.

Grabus, S. / **Greenberg, J.** / **Logan, P.** / **Boone, J.** (2019): "Representing Aboutness: Automatically Indexing 19th-Century Encyclopedia Britannica Entries", in: *NASKO* 7: 138-148.

Guhr, O. / **Schumann, A. K.** / **Bahrmann, F.** / **Böhme, H. J.** (2020): "Training a broad-coverage German sentiment classification model for dialog systems", in: *Proceedings of The 12th Language Resources and Evaluation Conference*: 1627-1632.

Hagen, T. / **Ketzan, E.** / **Jannidis, F.** / **Witt, A.** (2020): "Twenty-Two Historical Encyclopedias Encoded in TEI: A New Resource for the Digital Humanities", in: *Proceedings of LaTeCH-CLfL 2020: 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*.

Hagen, T. / **Jannidis, F.** / **Witt, A.** (2022): "Word Sense Alignment and Disambiguation for Historical Encyclopedias", in: *Proceedings of the 6th international conference on Graphs and Networks in the Humanities*.

Ketzan, E. / **Hagen, T.** / **Jannidis, F.** / **Witt, A.** (2022): "Quantitative Analysis of Gendered Assumptions in a Nineteenth-Century Women's Encyclopedia", in: *Book of Abstracts, 2022 Digital Humanities conference*.

Konle, L. / **Krönke, M.** / **Winko, S.** / **Jannidis, F.** (2022): "Emotions and Literary Periods", in: *Book of Abstracts, 2022 Digital Humanities conference*.

Loveland, J. (2019): *The European Encyclopedia*. Cambridge: Cambridge University Press.

Transformer-Based Named Entity Recognition for Ancient Greek

Yousef, Tariq
info@tariq-yousef.com
Leipzig University

Palladino, Chiara
chiara.palladino@furman.edu
Furman University

Jänicke, Stefan
stjaenicke@imada.sdu.dk
University of Southern Denmark

Introduction

The identification and classification of names in ancient texts, especially Classical ones like Homer's Iliad , is an essential task to support further text processing, but also simple reading facilitation and the creation of reading environments (Blackwell / Crane 2009). Typically, readers of Ancient Greek or Syriac face the challenge of a complex and sometimes obscure language, but also of very distant cultural references, which go back to events and traditions that may not be immediately comprehensible to a non-specialist. Names are a substantial part of this challenge. Texts like the Bible or the Iliad contain hundreds, if not thousands, of names,