Pre-print. See Digital Humanities 2022, Conference Abstracts, July 25-29, 2022, Tokyo, The University of Tokyo, pp. 270-273. https://dh2022.adho.org

Quantitative Analysis of Gendered Assumptions in a Nineteenth-Century Women's Encyclopedia

Erik Ketzan
Trinity College Dublin
ketzane@tcd.ie

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

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

Andreas Witt
IDS Mannheim & University of Cologne
awitt6@uni-koeln.de

1 Introduction

This paper quantifies textual patterns relating to gendered assumptions in a fairly unique text, an entire "women's encyclopedia" from 1830's Germany, which at 10 volumes and 1,461,000 word tokens was of comparable size to contemporary general encyclopedias, but written and marketed for a female audience. We perform experiments on classifying gender of biographical entries and querying a specific textual feature, calendar dates, with context from comparison 19th-20th century encyclopedias from the EncycNet corpus.[1]

Encyclopedias in the European tradition were "an element of culture and peoples' lives," (Loveland, 2019), but while encyclopedias invite interpretation (Einbinder, 1966), their length challenges non-digital scholarship; digital humanities thus "holds promise for the study of encyclopedias" (Loveland, 2019).

The *Damen Conversations Lexikon* ("Ladies' Conversations Encyclopedia," hereafter *DamenLex*) has been the subject of little scholarly analysis. Roßbach (2015) writes that the first edition of the *DamenLex* "primarily aims to act as a behavioral guide for virtuous women",[2] but Schaser (2006) asserts that this "little known" encyclopedia is "a treasure trove for questions of cultural history." Editor Carl Herloßsohn (1804-1849) explained its content selection in gendered assumptions and value judgments, and the extent to which the *DamenLex* actually followed Herloßsohn's stated goals is the starting point of our research questions.

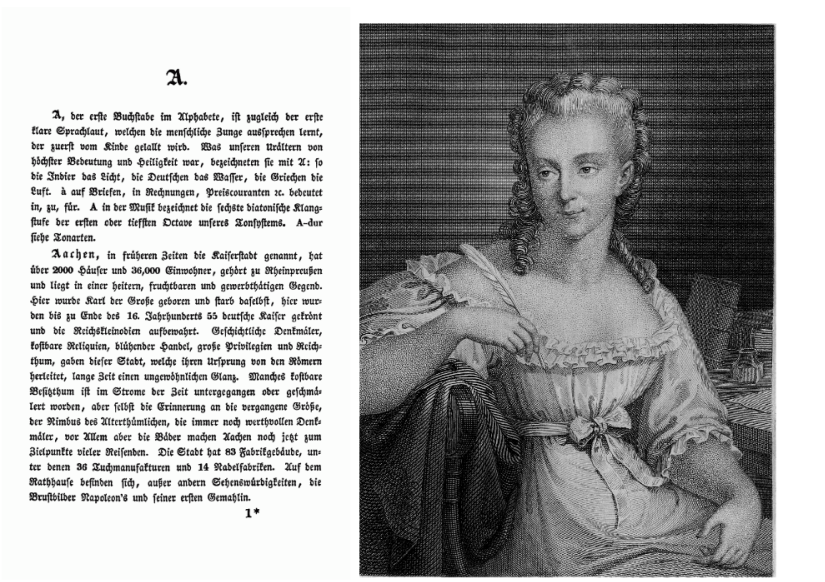
2 Related Work

Distant reading the *DamenLex* for ideological traces relates to issues of women and gender in 19th-century Europe, women's education and achievement, and the history of books not *by*, but marketed *for* women, in which women's access to the written word, by controlling literacy and access to reading material, has been a source of anxiety (Jack, 2012). We would thus expect the *DamenLex* to display evidence of two opposing forces: women's education and controlling ideology through explicit or implicit gender presumptions within the content, stylistics, and selection of topics.

Perceptions of women readers can be traced through such texts as Ovid's books addressed to women (e.g. *Ars Amatoria*) and the Confucian *Four Books for Women* (Mingqi, 1987). Eighteenth- and nineteenth-century texts for women on etiquette and conduct were prescriptive, supporting notions of "ideal womanhood" (Hemlow, 1960; Darby, 2000). The nineteenth century in which the *DamenLex* was published was "a golden age for reading, and for women's reading in particular," per Jack (2012), as the growth of industrialization, printing and publishing were "accompanied by wide-ranging debates about what women [...] should be encouraged to read, or discouraged — even prevented — from reading."

[1] https://encycnet.github.io/

[2] Translation of quotations by Roßbach (2015) and Schaser (2006) in this paragraph are by us.



A.

A, der erste Buchstabe im Alphabete, ist zugleich der erste klare Sprachlaut, welchen die menschliche Zunge aussprechen lernt, der zuerst vom Kinde gelallt wird. Was unseren Urältern von höchster Bedeutung und Heiligkeit war, bezeichneten sie mit A: so die Indier das Licht, die Deutschen das Wasser, die Griechen die Luft. à auf Briefen, in Rechnungen, Preiscouranten ꝛc. bedeutet in, zu, für. A in der Musik bezeichnet die sechste diatonische Klangstufe der ersten oder tiefsten Octave unseres Tonsystems. A-dur siehe Tonarten.

Aachen, in früheren Zeiten die Kaiserstadt genannt, hat über 2000 Häuser und 36,000 Einwohner, gehört zu Rheinpreußen und liegt in einer heitern, fruchtbaren und gewerbthätigen Gegend. Hier wurde Karl der Große geboren und starb daselbst, hier wurden bis zu Ende des 16. Jahrhunderts 55 deutsche Kaiser gekrönt und die Reichskleinodien aufbewahrt. Geschichtliche Denkmäler, kostbare Reliquien, blühender Handel, große Privilegien und Reichthum, gaben dieser Stadt, welche ihren Ursprung von den Römern herleitet, lange Zeit einen ungewöhnlichen Glanz. Manches kostbare Besitzthum ist im Strome der Zeit untergegangen oder geschmälert worden, aber selbst die Erinnerung an die vergangene Größe, der Nimbus des Alterthümlichen, die immer noch werthvollen Denkmäler, vor Allem aber die Bäder machen Aachen noch jetzt zum Zielpunkte vieler Reisenden. Die Stadt hat 83 Fabrikgebäude, unter denen 36 Tuchmanufakturen und 14 Nadelfabriken. Auf dem Rathhause befinden sich, außer andern Sehenswürdigkeiten, die Brustbilder Napoleon's und seiner ersten Gemahlin.

1*

Figure 1: A sample page and illustration from the *Damen Conversations Lexikon*.

3 Gendered assumptions

Gender bias in biographical entry subjects is an unfortunate theme in the history of the encyclopedia, with the 11th Britannica, for instance, including an entry on Pierre Curie but not Marie Curie (Thomas, 1992), and Bamman and Smith (2014) estimated that only 14.8% of biographical entries in Wikipedia have women as subjects.

The *DamenLex*'s editor explicitly included many biographies of women to appeal to women readers, suggesting experiments to classify the gender of the ~800 biographical entries in the *DamenLex* and almost 44,000 in comparison encyclopedias.[3] To classify biographical entries, we trained a bag-of-words-based SVM classifier to label entries as either biography, place, object or abstract concept (with an accuracy of 0.92 for biographies). To classify the gender of each biographical entry, a rule-based approach based on Reagle and Rhue (2011) compares the ratios of male and female personal pronouns in the entry (e.g. sein/his, ihr/her). Only entries longer than 20 tokens were classified, and we only proceeded with entries for which a gender was identified. The amount of unclassified entries is low for *Brockhaus 1837* (about 1%), where entries tend to be relatively long, but higher in *Brockhaus 1911* (about 37%), for example.

	Brockhaus 1809	DamenLex 1834	Brockhaus 1837	Herder 1854	Meyer 1905	Brockhaus 1911	Wikipedia 2014	Wikipedia 2015
Male	948 (95%)	480 (60%)	957 (96%)	5,952 (95%)	26,223 (94%)	7,356 (94%)	85.20%	84.50%
Female	52 (5%)	329 (40%)	42 (4%)	345 (5%)	1,599 (6%)	479 (6%)	14.80%	15.50%

Table 1: Estimate of of male and female biographies in historical German encyclopedias. Wikipedia results reported by Bamman and Smith (2014) and Graells-Garrido et al. (2015).

From these results, *DamenLex* contains a much higher percentage of female biographies than all other comparison texts (Table 1) including Wikipedia, with female biographies around 40%. Two chi-squared tests (Table 4 in the appendix) reveal that there is a highly significant relationship between encyclopedia and the amount of entries on women only when the *DamenLex* is included in the test, confirming our hypothesis. A similar gender disparity is observed in entry lengths of of male and female biographies in the *DamenLex* (Figure 2). We calculated Mann–Whitney U tests for article lengths of female biographies compared to similar sized samples from male biographies for all encyclopedias (see

[3]The encyclopedias for our experiments are part of a larger set of historical reference works converted to TEI (Hagen et al., 2020): http://dx.doi.org/10.5281/zenodo.4039569.

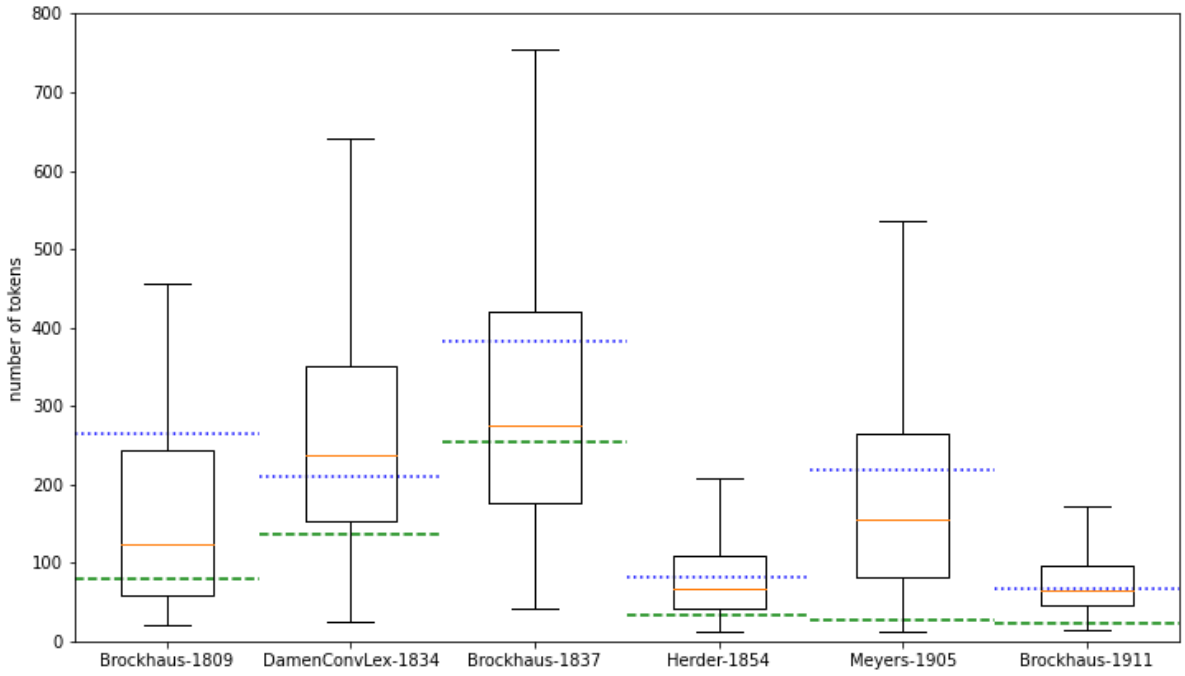


Figure 2: Boxplot visualizing the entry lengths of female biographies in tokens in all encyclopedias. The solid, orange line marks the median. Additionally, the median of male biography entry lengths (dotted, blue line) and the median of all entry lengths, including biographies, (dashed, green line) are given.

table 3 in the appendix). Only the *DamenLex*'s and *Brockhaus*' (1911) lengths of female biographies are not significantly shorter than the male entries.

Did *DamenLex* devote biographies to the same "notable" women as contemporary encyclopedias? The overlap is quite low: of 329 entries on women in *DamenLex*, only about 65 appear in at least one other encyclopedia, confirming that the editors of the comparison encyclopedias had a different perception of who "important women" were.

Finally, most frequent words in *DamenLex* biographical entries provide insight into content differences. Among the 15 most frequent nouns in female biographies are role labels such "daughter," "wife," and "mother," and family relations such as "husband" and "child." In male biographies, in contrast, only "father" and "son" appear in the 20 most frequent nouns, while references to artistic production such as "poet," "poem," "opera," and "writing" fill out the rest. Such artistic terms can also be found in female biographies, only at lower frequency ranks. Among the 50 most frequent adjectives in male entries, only a handful do not appear in the 100 most frequent adjectives in female entries: "tremendous," "glowing," "musical," and "exquisite." The first two words are typical terms to describe sublime aesthetic experiences, which aligns with contemporary gendered assumptions about aesthetics.

Herloßsohn wrote that a "romantic representation" of the subjects in the *DamenLex* was desired: "not a tiresome enumeration of facts and the course of time, but a lively, rapidly gliding painting [...] should be given." We thus hypothesize that the amount of calendar dates will be far lower in its entries. To investigate, we tagged encyclopedia entries with heideltime,[4] a multilingual temponym tagger. As Table 2 shows, the amount of dates is indeed far lower in *DamenLex* than comparison encyclopedias, confirming the gendered assumption that hard facts such as calendar dates were considered undesirable by women readers. Verification via the chi-squared test (Table 4 in the appendix) results that there is no statistical difference between the encyclopedias concerning the amount of dates, however.

[4] https://github.com/HeidelTime/heideltime

	Brockhaus 1809	DamenLex 1834	Brockhaus 1837	Herder 1854	Meyer 1905	Brockhaus 1911
Dates	1.29%	0.73%	1.38%	1.87%	3.81%	4.32%

Table 2: Relative amount of dates found in a sample of 256 entries of similar length (about 100,000 tokens overall) per encyclopedia.

4 Conclusion

By quantifying the ratios of male/female biographical entries in the DamenLex and comparison encyclopedias, comparative length of biographical entries, and a query of calendar dates in the texts, we provide new knowledge and add historical context to vibrant ongoing research on gender bias in encyclopedias (including Wikipedia). We agree with Schaser (2006) that the "little known" encyclopedia of the *DamenLex* is "a treasure trove for questions of cultural history," and have presented evidence that distant reading of gender distribution in biographical entries and content presentation can reveal gendered assumptions in the text. This paper will include these and other experiments to quantify gendered assumptions in encyclopedia texts, and could support future work in gender bias in not only historical but also contemporary encyclopedias.

References

David Bamman and Noah A Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Barbara Darby. 2000. The more things change . . . the rules and late eighteenth-century conduct books for women. *Women's Studies*, 29(3).

Harvey Einbinder. 1966. The myth of the britannica. *Science and Society*, 30(1).

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 165–174.

Thora Hagen, Erik Ketzan, Fotis Jannidis, and Andreas Witt. 2020. Twenty-two Historical Encyclopedias Encoded in TEI: a New Resource for the Digital Humanities. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 112–120, Online. International Committee on Computational Linguistics.

Joyce Hemlow. 1960. Fanny burney and the courtesy books. *PMLA*, 65(5).

Belinda Jack. 2012. *The Woman Reader*. Yale University Press.

Jeff Loveland. 2019. *The European Encyclopedia: From 1650 to the Twenty-first Century*. Cambridge University Press.

Zhang Mingqi. 1987. *The Four Books for Women*: Ancient chinese texts for the education of women. *B.C. Asian Review*, 1(1).

Joseph Reagle and Lauren Rhue. 2011. Gender bias in wikipedia and britannica. *International Journal of Communication*, 5(0).

Nikola Roßbach. 2015. *Wissen, Medium und Geschlecht: Frauenzimmer-Studien zu Lexikographie, Lehrdichtung und Zeitschrift*. Peter Lang International Academic Publishers.

Angelika Schaser. 2006. Rezension zu: Herloßsohn, carl (hrsg.): Damen conversations lexikon. neusatz und faksimile der 10-bändigen ausgabe leipzig 1834 bis 1838. berlin 2005. *H-Soz-Kult*.

Gillian Thomas. 1992. *A position to command respect: women and the eleventh Britannica*. Scarecrow Press, Metuchen, NJ.

A Appendix

Brockhaus 1809	$U = 737.0, n1 = n2 = 52, p < .001$
Brockhaus 1837	$U = 758.5, n1 = n2 = 42, p < .001$
Brockhaus 1911	$U = 112124.5, n1 = n2 = 479, p = 0.27$
Herder 1854	$U = 49367.0, n1 = n2 = 345, p < .001$
Meyer 1905	$U = 728612.0, n1 = n2 = 1599, p < .001$
DamenLex 1834	$U = 57677.5, n1 = n2 = 329, p = 0.93$

Table 3: Results of the one-sided Mann–Whitney U tests (p<.01) to confirm or reject the hypothesis whether male biography entries are significantly longer than female biography entries per encyclopedia.

Number of entries (incl. DamenLex)	$\chi^2(5, N = 44762) = 1635.9, p < .001$
Number of entries (excl. DamenLex)	$\chi^2(4, N = 43953) = 7.7, p = .1$
Number of dates	$\chi^2(5, N = 600) = 5.03, p = .41$

Table 4: Results for the chi-squared tests (p<.01) for the amount of entries on men and women as well as the amount of dates (and non-dates, in percent) per encyclopedia. For the dates, we opted for choosing the percentages over the raw counts, as the sample size makes the interpretation of the otherwise very low p-values difficult.