Student Scientometrics – What do German Students of the Humanities Cite in their Term Papers?

Henning, Tim
th58qazu@studserv.uni-leipzig.de
Computational Humanities Group, Leipzig University

Gutiérrez De la Torre, Silvia E.
silviaegt@uni-leipzig.de
Computational Humanities Group, Leipzig University

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Computational Humanities Group, Leipzig University

Introduction

Citation analysis (Nicolaisen, 2007) is a common tool in scientometrics (Mingers & Leydesdorff, 2005; Ivancheva, 2008; Mester, 2015), which deals with the exploration and evaluation of scientific disciplines. Scientometrics typically deals with published research (books, journal articles, conference papers) and thus naturally entails a strong focus on researchers at the PhD level or higher. This is not surprising, since scientometrics is also an important tool for evaluating individual scientists, which is an important aspect in the academic career of any scientist. At the same time there is a growing interest in the citation practices of undergraduates, although they are generally researched to understand the impact of library instruction sessions on students' behavior (Barratt et al., 2009; Davis, 2002; Davis & Cohen, 2001; Gratch, 1985; Hovde, 2000; Mohler, 2005; Robinson & Schlegl, 2004; Ursin et al., 2004). We believe this area deserves more attention, as students' papers can be seen as a mirror of different disciplines' curricula, which can affect later academic careers and have canonization effects. In this paper, we add to the existing research by showcasing a mix of scientometrics and DH methods to investigate a novel corpus of German term papers and shed some light on the citation practices of students in German studies.

Data Acquisition and Methods

GRIN Publishing offers more than 17,848 papers accessible free of charge (data from 03.05.2022). These papers are voluntarily uploaded to the website by students. We scraped both full texts and metadata. Filtering by discipline, university, and document type, our corpus comprises 502 term papers in German studies written by Bachelor students at German universities. [1] The bibliography section was extracted using a regular expression [2], as this method proved to be resource-efficient, time-saving, and fairly accurate (Councill et al., 2008; Wu et al., 2015). For the task of reference extraction, some out-of-the-box parsers exist. Most parsers use ML approaches and yield very good evaluation results for English papers from the natural sciences (Tkaczyk et al., 2018). However, as has been shown in previous experiments (Gutiérrez De la Torre et al., 2022; Rodrigues Alves et al., 2018), they are not suitable for a corpus of German term papers. We thus evaluated a selection of different parsers, namely: *Anystyle*[3], *EXparser* (Hosseini et. al., 2019), and GROBID (Lopez, 2009) with F1 scores of 0.76, 0.79, and 0.67, respectively. [4] For the evaluation, a ground truth of 13 papers was created, taking into account the degree of formal, stylistic, structural, and orthographic diversity. [5] With this method, 7,133 references and 5,207 author names were extracted. Using a simple heuristic, name variations such as "Friedrich Nietzsche" and "F. W. Nietzsche" were mapped, resulting in a total of 3,610 different authors of which 3,076 appear only once in the corpus. [6] A co-authorship network is used for the analysis, where an edge represents a simultaneous mention of two authors in a bibliography. Using the OpenOrd (Martin et al., 2011) visualizing algorithm, 15 larger clusters emerged. [7] The intellectual proximity between authors in a cluster is confirmed with topic modeling using Gensim's Mallet LDA wrapper (Rehurek & Sojka, 2011; Sbalchiero & Eder, 2020). The coherence value calculation suggested the use of 12-14 topics, and after a revision 11 topics were selected (see Table 1).

Table 1: Topic Distribution. Note: The Topic Modelling script can be found at https://github.com/Henning-round/Student_Scientometrics/blob/main/Topic_Modelling_paper.ipynb. For an overview of (German) keyword per topic see https://github.com/Henning-round/Student_Scientometrics/blob/main/analysis/keywords.PNG.

Dominant Topic	Doc_count	Total_Docs_Percentage
Linguistics	80	15.94%
Narratology	65	12.95%
Families, genders and their conflicts	54	10.76%
Political systems, war, and post-war literature	53	10.56%
Philosophy	49	9.76%
Medieval Studies	48	9.56%
Lyrics	48	9.56%
Study of historical sources (especially religious)	33	6.57%
Theatre, Drama and Tragedy	32	6.37%
Communication Sciences	25	4.98%
Addressing the school and teaching system	15	2.99%

Results

The result of the combined co-author-topic network in Figure 1 reveals new trends in German studies when compared to previous research (Riddell, 2014): (1) A hitherto unknown engagement with medieval studies with a 9.56% share in the cluster. (2) An interest in "gender studies" is particularly confirmed with a 10.76% share and multiple clusters. (3) However, the downward trend of Goethe-related topics (Riddell, 2014) cannot be confirmed. Rather, by appearing in the overall corpus with a relative share of 6.6% of the papers, Goethe seems to be a central and important component of the Bachelor's programme in German Studies. [8]

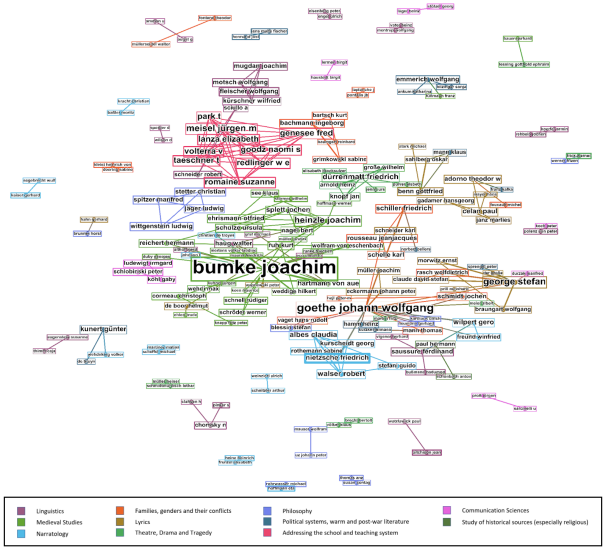

Figure 1: Co-Author-Topic-Network. Note: for better readability download the high-resolution image from https://github.com/Henning-round/Student_Scientometrics/blob/main/analysis/topic_author_model.png

Discussion

Two main limitations of our approach – which still is work in progress – must be taken into account: (1) Around 18% of the extracted reference strings were discarded either because they were too short to be a reference string (<200 characters); too long (>200); pure ISBNs (26 cases); or with no dates nor other typical reference keywords such as *S* or *Hg* (abbreviations for pages and editor in German). (2) Our normalization method allowed us to explore rich patterns, however, in further iterations we would like to apply more robust algorithms and document the differences with our heuristic. Nevertheless, these first experiments are very promising. Not only were we able to gather a vast number of authors and their co-citation networks, but we were also able to show that they correlated with certain fields and thus show unknown trends among German studies undergraduates. To our knowledge, this is the first work to computationally study term papers in German from a scientometric perspective, which is why we want to gradually eliminate the above-mentioned limitations and extend the corpus to other disciplines than just German studies.

Notes

1. https://github.com/Henning-round/Student_Scientometrics/blob/main/metadata_works.csv
2. https://github.com/Henning-round/Student_Scientometrics/blob/main/Extract_bibliography.ipynb
3. https://github.com/inukshuk/anystyle
4. https://github.com/Henning-round/Student_Scientometrics/blob/main/Evaluation_of_parsers.ipynb
5. https://github.com/Henning-round/Student_Scientometrics/blob/main/goldstandard_paper.pdf
6. https://github.com/Henning-round/Student_Scientometrics/blob/main/Name_normalization.ipynb
7. https://github.com/Henning-round/Student_Scientometrics/blob/main/analysis/topic_author_model.png
8. https://github.com/Henning-round/Student_Scientometrics/blob/main/Author_distribution.PNG

Bibliography

Barratt, C. C., Nielsen, K., Desmet, C., & Balthazor, R. (2009). Collaboration is Key: Librarians and Composition Instructors Analyze Student Research and Writing. portal: Libraries and the Academy, 9(1), 37–56.

Councill, I., Giles, C. L., & Kan, M.-Y. (2008). ParsCit: An Open-source CRF Reference String Parsing Package. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco.

Davis, P. M. (2002). The effect of the Web on undergraduate citation behavior: A 2000 update. College & Research Libraries, 63(1), 53–60.

Davis, P. M., & Cohen, S. A. (2001). The effect of the web on undergraduate citation behavior 1996-1999. Journal of the Association for Information Science and Technology, 52(4), 309–314.

Gratch, B. G. (1985). Toward a Methodology for Evaluating Research Paper Bibliographies. Research Strategies, 3(4).

Gutiérrez De la Torre, S. E., Equihua, J., Niekler, A., & Burghardt, M. (2022). Into the bibliography jungle: Using random forests to predict dissertations' reference section. Understanding LIterature References in Academic Full TExt at JCDL 2022, 3220, 7.

Hosseini, A., Ghavimi, B., Boukhers, Z., & Mayr, P. (2019). EXCITE – A Toolchain to Extract, Match and Publish Open Literature References. 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 432–433.

Hovde, K. (2000). Check the citation: Library instruction and student paper bibliographies. Research Strategies, 17(1), 3–9.

Ivancheva, L. (2008). Scientometrics today: A methodological overview. Collnet Journal of Scientometrics and Information Management, 2(2), 47-56.

Lopez, P. (2009). GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, & G. Tsakonas (Hrsg.), Research and Advanced Technology for Digital Libraries (Bd. 5714, S. 473–474). Springer Berlin Heidelberg.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: An open-source toolbox for large graph layout. 786806.

Mester, G. (2015). New Trends in Scientometrics. In Papers of 33rd International Scientific Conference" Science in Practice" (pp. 22-27).

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. European journal of operational research, 246(1), 1-19.

Mohler, B. A. (2005). Citation analysis as an assessment tool. Science & Technology Libraries, 25(4), 57–64.

Nicolaisen, J. (2007). Citation analysis. Annual review of information science and technology, 41(1), 609-641.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2), 2.

Riddell, A. B. (2014). How to read 22,198 journal articles: Studying the history of German studies with topic models. In M. Erlin & L. Tatlock (Hrsg.), Distantreadings: Topologies of German culture in the long nineteenth century (S. 91–114). Boydell & Brewer.

Robinson, A. M., & Schlegl, K. (2004). Student bibliographies improve when professors provide enforceable guidelines for citations. portal: Libraries and the Academy, 4(2), 275–290.

Rodrigues Alves, D., Colavizza, G., & Kaplan, F. (2018). Deep Reference Mining From Scholarly Literature in the Arts and Humanities. Frontiers in Research Metrics and Analytics, 3. https://doi.org/10.3389/frma.2018.00021

Sbalchiero, S., & Eder, M. (2020). Topic modeling, long texts and the best number of topics. Some Problems and solutions. Quality & Quantity, 54, 1095-1108.

Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018). Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers (arXiv:1802.01168). arXiv.

Ursin, L. A., Lindsay, E. B., & Johnson, C. M. (2004). Assessing library instruction in the freshman seminar: A citation analysis study. Reference Services Review, 32(3), 284–292.

Wu, J., Williams, K. M., Chen, H.-H., Khabsa, M., Caragea, C., Tuarob, S., Ororbia, A. G., Jordan, D., Mitra, P., & Giles, C. L. (2015). CiteSeerX: AI in a Digital Library Search Engine. AI Magazine, 36(3), 35–48.