Large Bibliographies as a Source of Data for the Humanities – NLP in the Analysis of Gender of Book Authors in German Countries and in Poland (1801-2021)

Adam Pawłowski
University of Wrocław
pl. Uniwersytecki 1
50-137 Wrocław, Poland
adam.pawlowski@uwr.edu.pl

Tomasz Walkowiak
Wrocław University of Science and Technology
27 Wybrzeże Wyspiańskiego St.
50-370 Wrocław, Poland
tomasz.walkowiak@pwr.edu.pl

Abstract

The subject of this article is the application of NLP and text-mining methods to the analysis of two large bibliographies: a Polish one, based on the catalogs of the National Library in Warsaw, and a German one, created by the Deutsche Nationalbibliothek. The data in both collections are stored in MARC 21 format, allowing the selection of relevant fields that are used for further processing (basically author, title, and date). The volume of the Polish corpus (after filtering out non-relevant or incomplete items) includes 1.4 mln of records, and that of the German corpus 7.5 mln records. The time span of both bibliographies extends from 1801 to 2021. The aim of the study is to compare the gender distribution of book authors in Polish and German databases over more than two centuries. The proportions of male and female authors since 1801 were calculated automatically, and NLP methods such as document vector embeddings based on deep BERT networks were used to extract topics from titles. The gender of the Polish authors was recognized based on the morphology of the first names, and that of the German authors based on a predefined list. The study found that the proportion of female authors has been steadily increasing both in Poland and in German countries (currently around 43%). However, the topics of women's and men's writings invariably remain different since 1801.

1 Introduction

The research conducted straddles two broad areas. One relates to the issue of the resources and methodologies (NLP and text-mining tools applied to **large bibliographies** – hereafter LB), and the other concerns a certain problem that belongs to the field of cultural anthropology and/or social sciences (equal gender participation in various social activities). We will first address the latter issue, while the former (data and methods) will be presented in the subsequent sections.

The complex and long-standing processes leading to a fairer participation of both genders in social, scientific, economic, and cultural life have been ongoing in Europe for at least two centuries. Their most visible public representation was for decades the feminist movement initiated in the USA in the mid-19th century (Women's Convention in Seneca Falls, USA, 1848). The actual gender status in European countries, however, is different in specific regions and largely independent of official policies and spectacular events publicized by the media. The real scope of the ongoing changes in this area can only be revealed by analyzing big data that consistently and synthetically reflect the state of affairs over a long period of time. In particular, a convincing analysis of this phenomenon should not highlight the most mediated, individual exponents of women's status (e.g. awards, prominent positions in politics or business). It should rather rely on information resources aggregating large amounts of data dispersed in various sources that we can consider the most objective possible.

Large bibliographies can be considered as sources that meet these conditions. They consist of data which represent a very important segment of intellectual life and are collected systematically according to the same principles. Also of significance is the fact that for decades, LB have remained insufficiently exploited by scientific methods. The present work is therefore of a pioneering nature. It should be noted that, so far, in empirical studies of complex processes related to gender equality, economic and legal measures have been used. However, they can hardly be trusted in the case of a politically fragile region with a turbulent history, such as Central Europe. Its multilingual diversity, the intensity of political change, different monetary systems, and border shifts over the past two centuries make publication stream analysis a more reliable approach that is highly resistant to the influence of random factors or systemic disruptions.

Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL2023), pages 63–71

Writing books is, among other things, the result of one's prosperity, exposure to knowledge, and education. In addition, social acceptance of the author's gender and social origin was necessary for the publication of a book. For example, medicine in the nineteenth century was almost entirely masculinized. Therefore, a female author of a work on, for example, surgery would not have been accepted by the readers, ergo no publisher would publish such a work.

2 Research Objective and Hypothesis

The first goal of this study is to show the proportions of women and men among the authors of books published from 1801 to 2021 in two European communities of communication, associated in various periods with the concept of a nation or of a state. For that purpose, we generated historical histograms showing the proportions of male, female, and unrecognized authors. A chronological (vertical) survey allowed us to show the dynamics of cultural change, occurring in the most populous and influential countries of Central Europe over a long period of time.

The second goal of the research was to extract topics specific to male and female authors from book titles. Its aim was to identify the main contents or topics that characterize the writing of both genders. We assumed that the authorship of a book aggregates a number of socially and psychologically salient variables that are difficult to capture, especially over a long period of time. Topics were generated from time sections, representing main historical periods. We hypothesized that in the 19th century (until ca 1910) there would be little, if any, similarity between the two sets both in Polish and in German data. We also expected to observe a growing overlap of topics generated from titles by male and female authors after the interwar period 1918-1945 for both Polish and German data.

The question as to whether this parameter is likely to become a universal and effective measure of gender equality remains open. Yet, it certainly provides a quantitative and possibly objective estimate of this phenomenon. Its advantage, from the point of view of methodology, is the use of big and "clean" data and the ensuing independence from random factors. Another issue was the choice of Poland and Germany as the objects of the study. These are countries geographically and culturally close but historically different in size and status. Despite the instability of borders and political systems, German culture maintained continuity in the 19th and 20th centuries based on various state organisms (including the Rhine Union, the Second Reich, the Weimar Republic, the Third Reich, and the Federal Republic of Germany). Additionally, in the 19th century German Countries (especially Prussia) were among the world leaders in science and culture. The situation of Poland was radically different. Poland lost its statehood at the end of the 18th century, and in effect existed until World War I only as an entity identified with its culture, history, religion, and, above all, its language. And even this status of an "imagined state" was constantly challenged by the occupation regimes. The country regained its independence in 1918, but then again was occupied between 1939 and 1945 by the Third Reich and the Soviet Union. After 1945, Poland became a vassal state subordinated to the Soviet Union; not regaining full sovereignty until 1989. Given the above circumstances, the discovery of similarities (or differences) related to the gender of book authors in both corpora will provide a result that seems objective, and scientifically relevant.

3 Related work

Much more research is now focused on the exploratory analysis of library catalogs around the world, for example, Lahti et al. (2019); Tolonen et al. (2019). However, surprisingly little research has been published on the use of artificial intelligence techniques in bibliography data (Wheatley and Hervieux, 2019; Pawłowski and Walkowiak, 2020; Pawłowski and Walkowiak, 2021). The problem of topic recognition from short texts is studied in the literature (Albalawi et al., 2020; Grootendorst, 2022), but application of such a method to a large number of book titles, that is, very short texts, is an original approach. Moreover, as discussed in Section 5 the applied method differs from methods known from literature (Grootendorst, 2022). Moreover, the paper deals with the problem of merging two bibliographies and deduplication of records (see Section 4). The problem is discussed in the literature (Wysota and Trzaska, 2021; Sitas and Kapidakis, 2008; Heron et al., 2013).

4 Research Material

The study was carried out on large bibliographies produced by the National Libraries of Poland[1]

[1] https://data.bn.org.pl/databases

and Germany (Deutsche Nationalbibliothek)[2]. Although these are not the official "national bibliographies", they functionally fulfill the conditions set for such monumental repositories. In particular, they have a predictable structure of data, the permanent care of a central institution, and the aspiration of covering the entire body of writings. It is worth mentioning that due to the lack of a Polish state in the 19th century (until 1918), there was no central institution to keep track of publications during this period. Records extracted from the Estreicher Polish Bibliography[3], which registers Polish works and Polonica from a period very poorly represented in the collection of the National Library, were therefore taken into account. For this purpose, the online part of the bibliography was used, which covers approximately 40% of the entire collection.

The records of the central libraries are stored in MARC format (Thomale, 2010) which allows for field searches and elimination of works that do not meet the analysis conditions. In the first iteration, records lacking authors or titles were filtered out. All nontext works (maps, notes, gramophone records, other sound recordings, etc.) were omitted. Periodical publications were also discarded. In the case of works extracted from the Estreicher Polish Bibliography, 5 pages were accepted as the lower limit of acceptable volume (this source includes very short documents too). In principle, most works in languages other than Polish or German were eliminated from both bibliographies. However, this criterion was problematic, as there is no easy way to automatically distinguish works by German (or Polish) authors writing in other languages from authors of other nationalities, but publishing in Germany or Poland (a frequent example from the 20th century are German doctorates in English). Several works by Polish authors from the 19th century that for political reasons were published in the languages of states that occupied Polish territory (mainly in German or Russian) or in some other other international language (e.g., Latin, French, and Ruthenian), were excerpted by hand and included.

The whole research material for German comprised 25.9 mln records, out of which 72% were rejected as nonrelevant, while for Polish a total of 2.3 mln records were processed (95% from the Polish National Library database, 5% from the Estreicher Bibliography), and 38% were rejected (38% among those from the National Library database, 50% from the Estreicher Bibliography). The distribution of data over time was not balanced, but this does not impede the results of our study (with incomplete representation, inductive inference is applicable).

The association of first names with gender is almost unambiguous and, in addition, the feminine gender in Polish is always indicated by the name ending *-a*. In the case of the Polish base, recognition was, therefore, based on this rule. The automatically generated database (mapping of name to gender) was then manually checked. Some names, especially foreign ones, are ambiguous, so they were marked as unknown. The German language, on the other hand, is not so consistent, as some Old High German names end in a consonant (e.g., Annetrud, Edeltraut, Gudrun etc.). An additional difficulty - especially after 1945 - in German, is the large number of borrowed names. For this reason, a reference catalog of German male and female names was prepared using open resources. It was manually verified again and completed manually for missing names that had an occurrence larger than 20. Finally, the gender of the authors was determined by automatically comparing their names with the list.

5 Methods of Data Processing

The MARC database is structured, but often requires additional procedures for information retrieval. For example, the publication date often contains additional characters that need to be removed. The author's first name is not a separate field and has to be automatically separated from the last name. Another problem was the use of umlaut character encoding (the umlaut was encoded as two characters) in German data that are not compliant with standard UTF-8. For processing MARC-21 files we have used a Python 3 PyMarc[4] library.

In the case of the Estreicher Polish Bibliography, the data are available as HTML tables, and the date of publication is an element of a single text containing the place of publication and the name of the publisher. This required the development of a set of heuristic rules to extract the publication data from the text. Another problem with the Estreicher Bibliography was the need for deduplication with the National Library of Poland. A special method

[2] https://data.dnb.de/DNB/

[3] https://www.estreicher.uj.edu.pl

[4] https://pymarc.readthedocs.io/

was developed that first splits data based on publication date and then uses Jaccard's token similarity between titles and authors to detect duplicates in a group of papers with the same publication date.

Detecting topics from short texts requires a dedicated procedure (Albalawi et al., 2020; Grootendorst, 2022), which differs from classic approaches to topic modeling, such as LDA (Blei et al., 2003). This is because titles are very short and we cannot rely on the co-occurrence of words in the same text and on the assumption that the text is a mixture of several topics. Therefore, the semantic analysis of titles (the second objective of the work) was based on a clustering procedure. This procedure requires a measure of similarity between titles. Therefore, each title was transformed into a vector space using deep BERT networks (Devlin et al., 2018). Since pre-trained BERT networks are not suitable for solving semantic similarity problems (Reimers and Gurevych, 2019), we used a Sentence-BERT approach (Reimers and Gurevych, 2019) based on metric learning (Bellet et al., 2015). It uses Siamese and triplet network structures to derive semantically meaningful sentence embeddings. In the case of Polish, we tuned the Sentence-BERT network on a publicly available corpus[5] of human-annotated sentence pairs in Polish for their semantic relatedness starting from the HerBERT (Mroczkowski et al., 2021) pretrained model. In the case of German, we used an already predeveloped Sentence-BERT model dedicated to English and German text[6]. Having embedded text, we applied classical K-Means (Hastie et al., 2013) clustering to obtain what we hoped were clusters of semantically distinct groups of titles. We then described the clusters using a set of words that identifies the content of the clusters. This is done using a modified TF-IDF (Salton G, 1988) procedure that takes into account class information (c-TF-IDF), which was proposed in Grootendorst (2022). This method yields list of words with their probabilities for each cluster; that is, topic as a concept used in topic modeling approaches (Blei et al., 2003). Our approach differs from BERTopic (Grootendorst, 2022) by omitting UMAP (McInnes et al., 2020) used for document vector reduction and replacing HDBSCAN (McInnes and Healy, 2017) clustering with K-Means. The authors tested BERTopic, but the results were not satisfactory because HDBSCAN's outlier detection function caused about 75% of the data to go off-topic, and UMAP's preservation of only local similarities caused semantically different titles to be mixed within a single topic.

Detected topics can be linked to gender by counting the number of titles assigned to a topic/cluster with authors of that gender. In the case of multiauthor books, we required that each author have the same gender. To express the relevance of a topic to a given gender, the importance index is defined as the ratio of books authored by women to men (or vice versa) about a given topic (cluster) normalized to the ratio of books for each gender. This allows the detection of a topic relevant to both men and women, irrelevant of the gender distribution among authors.

6 Results

6.1 The Volume of German and Polish Data

Analysis of the data volume has confirmed our expectations, but there were also some surprises. Namely, it turned out that the data coverage of the period roughly referred to as the 19th century, both in the case of Poland and Germany, is poor (until 1910). However, this did not significantly impede the analyses. We have assumed that data from the 19th century should be treated as a representative sample (consisting of the most significant works), while data from the period 1911-2021 are almost complete. Inspection of the histogram (time series) of the number of titles in successive years confirms that the German culture was and remains very productive (red plots in Figure 1). As a matter of fact each year the number of book titles published in Germany has exceeded the corresponding parameter in Poland by at least 5 times, and in some years even more (see Figure 1c). This is an unfavourable result from the Polish point of view because the difference in the size of the populations of the two countries would justify an advantage of only three times. This difference in volume can be explained, however, to some extent. Firstly, in Germany, all PhD dissertations must appear in print (with an ISBN code), whereas in Poland there is no such obligation. Secondly, the databases of the National Library of Germany include works belonging to the German Countries, including also those from Austria and Switzerland. Thirdly, in Poland during the communist period, due to paper shortages and the lack of a normal publishing market, multiple

[5] https://huggingface.co/datasets/allegro/klej-cdsc-r

[6] https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer

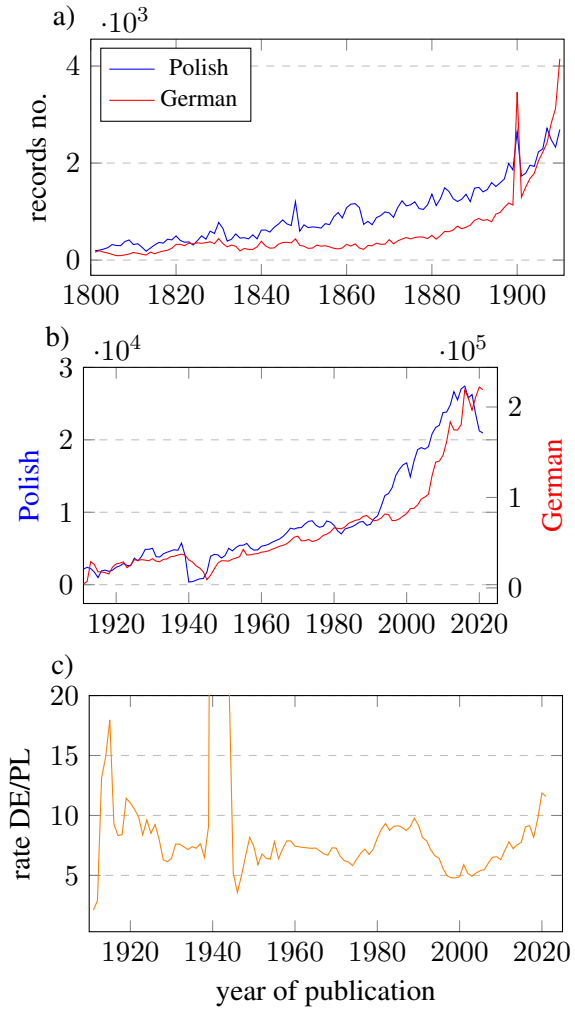

Figure 1: Number of records meeting the analysis criteria (non-empty field of title and authors, publication date between 1801 and 2021, and Polish/German language); a) Polish and German for years 1801-1910; b) 1911-2021 (notice that scale for German is 10 times larger then for Poland); c) Number of German records in relation to Polish records in 1911-2021 (in 1940 it goes over 80).

editions of the same work were very rare, while in Germany many books appeared at the same time as softcover and as hardcover.

A closer analysis of these contemporary data also reveals interesting fluctuations due to World War 2 or political changes. What is conspicuous is the period from 1980 to 1990 in Poland (martial law), marked by a general collapse in culture and economy (Figure 1b in blue), and the period after 1990. Events at that time in Poland and partly in Germany were a co-occurrence of three factors: the fall of the totalitarian system and the abolition of censorship that released the creative energy of the society, the technological revolution that lowered the costs of publishing, and the spread of the personal computer that increased the speed of writing texts by authors. During that period, the DE/PL ratio systematically decreased (Figure 1c). The reason for the reversal of this tendency after 2005 is a change in the long-term trend in the German data (Figure 1b in red). This is, however, not due to the sudden increase of the number of German books published (a stable trend observed since the 1920s cannot change from year to year) but to the change of book coding method. Most of the new titles started to be counted twice or three times despite being the same work: printed version, e-books in various formats, and audio-books had different ISBN codes. Digital editions were also frequent in Poland at that period but were not considered as separate books (see sharp falls of the curve in Figure 1b).

The material from the years 1801-1911, although incomplete, is of interest too (Figure 1a). It shows, for example, negative effects of catastrophic events (e.g., in the Polish data there are visible traces of national uprisings in 1830 and 1861). On the other hand, the "round date effect", i.e., the tendency of people to accumulate special interest around points on the time line that are deemed some sort of symbolic borderlines, should be considered very interesting. This is the explanation for a strong peak of the curve in 1900 (Germany and Poland), and a smaller one in 1850. The question arises why such a peak did not appear in the millennium year (2000). Most likely this anomaly can be explained by fundamental changes of the leading medium in public communication. The peak was observed in electronic media including TV, and not in printed books, which in 1900 covered a much larger scope of social life.

6.2 Gender Distribution in German and Polish Data

An impressive visual representation of the enormous historical changes that have taken place in the societies of Poland and German Countries, and probably also throughout Europe, revealed here as Figure 2, showing two symbolic lines: the share of male and female authors over the period of the last 220 years (upper and lower line, respectively). The percentage of unrecognized data, as can be seen, is small and stable - fluctuating at approximately 6%. On the contrary, the female and male share lines run similarly, showing a very slow increasing

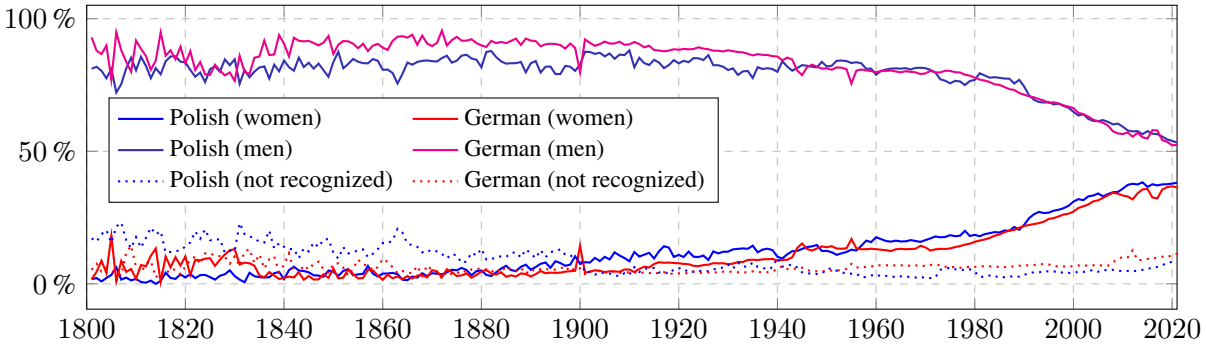

Figure 2: Gender breakdown of book authors over the years 1801-2021.

/ decreasing trend almost until the middle of the 20th century. In German Countries, the cut-off date indicating a change in the trend is around 1970 (one could hypothesize if this shift was related to the consequence of the mass unrest of 1968). In Poland, a strong acceleration in the development of women's writing occurs after 1990. Both curves react to one-off events (different for Poland and German Countries), but both have similar shapes. The disproportion in the number of records between Poland and Germany, as well as between different historical periods, is considerable (Figure 1). However, this does not affect the result presented here, as we are using proportions and not absolute values. The number of recognized male and female author names is high. In the German database, out of a total number of 8.1 mln individuals, the percentage of unrecognized items (or rejected for some other reason) was 7.7%, and in the Polish database the corresponding percentage was 5.3% out of a total number of 1.3 mln authors. All this makes the obtained result reliable. The process of the increasing participation of women writers is noticeable in the data (Figure 2). This indirectly indicates a growing gender equality, leading to a more balanced participation of men and women in intellectual and cultural life. Interestingly, despite the great differences between Poland and Germany in terms of culture, politics, and economy, both curves show virtually the same upward trend, which demonstrates the universality of the observed phenomenon, at least in this part of Europe. It is only on closer inspection that small differences are noticeable. Some of them seem long-lasting, while others are temporary disturbances in the overall trend.

The tools commonly applied in time series analysis (trend estimation, ACF function, etc.) were not used at this stage because there is no indication of periodic oscillations. However, the similarity between the two tendencies (PL, DE) was evaluated using a cross-correlation measure. The Pearson correlation coefficient of the Polish and German women's share series is equal to 0.93 with a p-value for testing for lack of correlation equal to $1e^{-97}$ thus verifying the hypothesis that the two processes are similar.

The shares of male and female authors were also compared separately for the Polish and German data. Overall, the average share of Polish female authors in the total publication stream is higher than that of German female authors (Figure 2). When the area under the curve is calculated over the entire period, the Polish data account for 12.1% and after 1900 it is 18.8%. For German data, it is, respectively, 11.1% and 16.1%. An exhaustive explanation of this result in terms of historical research and cultural anthropology would require a separate, comprehensive qualitative study that would take into account the following factors: the impact of the Polish struggle for the rebirth of a sovereign state between 1801 and 1918, the ideology of feminism and political struggle for equal rights for women throughout Europe, the influence of religion (Catholic, Protestant, Orthodox, Judaism) on women's status, and last but not least, specific social traditions in Polish and German culture.

6.3 Gender Specific Topics in German and Polish Titles

Table 1 shows the three most gender-differential topics (topics with the highest values of the importance ratio) for each gender in six time periods for Polish titles. The results for German titles are presented in Table 2. Topic detection was carried out independently for each period analyzed and each language. For the first time slot (1801-1910), we generated 20 topics for each language, for the next (1911-1945) 20 for Polish and 40 topics for Ger-

Period	Women		Men	
	Topic	C	Topic	C
1801-1910	woman	5.7	thesis defence	8.8
	youth	4.8	society report	5.6
	novel	2.8	academic	4.3
1911-1945	youth	4.9	judiciary	3.9
	romance	3.1	lecture	3.1
	novel	2.1	academic	2.2
1946-1980	child	4.7	electrics	4.7
	language	2.6	construction	4.1
	school	2.6	transport	4.0
1981-1999	school	4.2	transport	4.8
	child	3.8	electrics	4.4
	romance	3.3	construction	3.8
2000-2021	school	6.3	machine	4.8
	children	3.6	war	4.7
	woman	3.6	software	3.9

Table 1: The most gender specific topics detected in Polish titles in selected time periods. C denotes the importance ratio, calculated as the ratio of books by women to men (or vice versa) with a given theme (topic) normalized to the ratio of books for each gender.

Period	Women		Men	
	Topic	C	Topic	C
1801-1910	novel	4.6	Germany	3.3
	Rome	2.5	report, speech	3.1
	letter	1.9	religion	3.0
1911-1945	novel	3.5	tax	5.7
	story	3.3	economy	3.0
	child	3.2	judicary	2.9
1946-1980	romance	3.3	science	3.6
	woman	3.1	judicary	3.2
	child	2.8	electrics	2.8
1981-1999	child	2.8	software	3.2
	romance	2.3	mathematics	2.8
	medicine	2.0	applied science	2.8
2000-2021	romance	3.1	investigation	2.3
	child	2.8	finances	2.1
	cooking	2.1	religion	1.8

Table 2: The most gender specific topics detected in German titles.

Period	Polish		German	
	importance coefficient treshold			
	2	1.5	2	1.5
1801-1910	43.1%	55.7%	29.1%	71.9%
1911-1945	33.3%	57.4%	18.8%	49.6%
1946-1980	32.1%	50.8%	32.0%	49.5%
1981-1999	20.0%	41.5%	18.5%	41.9%
2000-2021	25.1%	42.6%	16.2%	42.8%

Table 3: Coverage (percentage of books) of gender-specific topics detected in Polish and German bibliographies. We count the ratio of books belonging to topics with a coefficient of importance above the given threshold (1.5, 2) to all books analyzed time periods.

man (since there are much more records) and 40 for the next time slots. For clarity of presentation, the topics were labelled by the authors on the basis of the list of key words generated by the c-TF-IDF algorithm. For example, the first four keywords (with highest probabilities) for topic "software" (for Polish data, male, years 2000-2021) are: programming (9.4%), windows (5.3%), excel (4.6%), and Microsoft (3.8%). And for topic "school" (in female group) they are: classroom (20.4%), school (18.5%), primary (9.3%), and textbook (7.8%). In the case of female authors, the topics are very stable over time and there is little difference between German and Polish titles. They mainly cover areas such as romance, novels, children, and women. In the group of male authors, the most gender-specific topics differ slightly between German and Polish texts, but there are many common elements, such as judiciary, science, and various technical fields. The limited volume of the article does not allow for more topics (and the corresponding keywords) to be presented, but their overtones are very similar: there are elements specific to male authors and others that are specific to female authors.

We have also used Fisher's exact test to analyze the topics shown in Tables 1 and 2. Technically, we built contingency tables for each topic and verified the null hypothesis that men and women are equally likely to write books on a particular topic. All the tests returned a p-value very close to 0, hence the listed topics are specific for gender. In addition, we analyze the volume of books covering gender-specific topics, that is, topics with an importance factor greater than a given threshold. The results for the Polish and German bibliographies and the thresholds equal to 1.5 and 2 are presented in Table 3. It shows that the share of books with gender-specific topics is slowly decreasing over time but is still very high (more than 40%).

All experiments can be replicated using standard workstations. We used an Nvidia GeForce GTX 1080/2080 Ti card to train Sentence-BERT and generate embeddings; all other analyses do not require a GPU. The only computational problem was

the process of generating topics from German data for the period 2000-2021, which required about 100GB of memory.

7 Conclusions

The research presented here was developed by combining advanced NLP techniques, mathematical statistics, programming, and large bibliographic data. It has demonstrated that the ratio of male and female authors in book publishing, when measured over a long period of time, should be considered one of the most reliable indicators of women's empowerment in the society. In the context of politically unstable regions, it has two main advantages: it is relatively **time-proof** in the period of late modernity (i.e. approximately since 1800), and it is **synthetic**. The former characteristic implies that the data are comparable over a long period of time. They were created under similar conditions (open publishing market) and the object of measurement remains the same (books from the 19th and the 20th century do not differ in essence). The latter feature means that it includes some specific measures that economics, history, or cultural anthropology used to apply separately (access to education, financial standing, social status, etc.). It also showed to be sensitive to one-time events such as wars, political or technological breakthroughs.

The study confirmed the hypothesis that in German Countries and in Poland similar upward trend in gender equality may be observed (Figure 2). However, the question remains open as to whether this phenomenon would have a similar dynamics throughout Europe. The current share of the authors of the book is approximately 43% women and 57% men (note that these are values after deducting unrecognized items, so slightly different from those in Figure 2). The value of the cross-correlation coefficient, i.e., 0.93, confirmed that statistically the two processes (gender equality in German Countries and in Poland) may be observed are not identical, although very similar. An interesting issue is whether participation of women in public life was higher in Poland or in German Countries. The overall ratio of female authors in Poland and Germany is slightly higher in Poland (12.1%) than in German Countries (11.1%). Analyzing Figure 2, one can also ask whether there is a target state of optimal social balance between both genders. For example, would the ideal be an equal share of male and female authors? Perfect symmetries are a product of human imagination and expectations, rather than empirical observable phenomena (Fleck et al., 1981). Equal parities should be treated with distrust in the social sphere as well as attempts to realize new utopias, not different in essence from those once conceived by philosophers, e.g. Thomas Morus (*Utopia*) or Tommaso Campanella (*City of the Sun*). Differences in the psychological profiles and interests of men and women have always existed and – as we demonstrate in our study – translate into various types of book content published. Therefore, a balanced and socially favorable level of participation of both genders among book authors would have to be considered 50% with a large margin, even 10%. It seems that this point will soon be reached both in Germany and in Poland.

The result of the research on the topics confirms some of the above statements. Multiple analyses of the entire corpus, as well as of its horizontal sections (19th and 20th centuries, contemporary period), conducted on German and Polish data, confirmed that the areas of interest of male and female authors are different. Their thematic profiles, generated using machine learning methods (BERT language models), shows a wide number of almost non-shared topics. This result does not, of course, resolve the issue of gender, and thus whether it should be seen as a purely biological or culturally conditioned phenomenon. However, it is an important contribution to the discussion on this topic, as it is based on sound methodology and a massive factual resource from two languages and cultures. The concluding remark applies to all the research conducted here. It shows that the analysis of large bibliographies by methods of data science, text mining, corpus linguistics, and NLP is a new, fully-fledged, promising strand of research.

7.1 Limitations

The study raised some debatable methodological issues. The first was the comparison of sets with significantly different numbers (7.1 mln compared with 1.4 mln records). However, in the case of gender, we are analyzing proportions of numbers and not absolute values. This makes the results of comparison, despite the different volumes of the Polish and German corpus, fully reliable. Another difficult issue was the automatic identification of the gender of the authors. This information cannot be found in MARC records, so it is necessary either

to retrieve it from another source (data linking) or to recognize the gender automatically. Automatic gender recognition by name is not 100% effective, but it has been proven practically feasible.

Acknowledgements

Financed by the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN - Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19.

References

Rania Albalawi, Tet Hin Yeap, and Morad Benyoucef. 2020. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. 2015. *Metric Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ludwik Fleck, Thaddeus Joseph Trenn, R. Merton, Fred Bradley, and Thomas Kuhn. 1981. *Genesis and development of a scientific fact*. University of Chicago Press, Chicago.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. 2013. *The elements of statistical learning: data mining, inference, and prediction*. Springer Series in Statistics. Springer, New York.

Susan Jane Heron, Betsy Simpson, Amy K. Weiss, and Jean Phillips. 2013. Merging catalogs: Creating a shared bibliographic environment for the State University Libraries of Florida. *Cataloging & Classification Quarterly*, 51(1-3):139–155.

Leo Lahti, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. 2019. Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly*, 57(1):5–23.

Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*, pages 33–42. IEEE.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Adam Pawłowski and Tomasz Walkowiak. 2020. Automatic recognition of gender and genre in a corpus of microtexts. In *Theory and Applications of Dependable Computer Systems*, pages 472–481, Cham. Springer International Publishing.

Adam Pawłowski and Tomasz Walkowiak. 2021. Analysis of toponyms from the Polish National Bibliography. In *Proceedings of the 6th International Workshop on Computational History (HistoInformatics 2021) co-located with ACM/IEEE Joint Conference on Digital Libraries 2021 (JCDL 2021), Online event, September 30-October 1, 2021*, volume 2981 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Buckley C. Salton G. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Anestis Sitas and Sarantos Kapidakis. 2008. Duplicate detection algorithms of bibliographic descriptions. *Library Hi Tech*, 26.

Jason Thomale. 2010. Interpreting MARC: Where's the bibliographic data? *Code4Lib Journal*, 11.

Mikko Tolonen, Leo Lahti, Hege Roivainen, and Jani Marjanen. 2019. A quantitative approach to book-printing in Sweden and Finland, 1640–1828. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 52(1):57–78.

Amanda Wheatley and Sandy Hervieux. 2019. Artificial intelligence in academic libraries: An environmental scan. *Information Services & Use*, 39:1–10.

Witold Wysota and Kacper Trzaska. 2021. Correlation of bibliographic records for OMNIS project. In *Theory and Engineering of Dependable Computer Systems and Networks*, pages 487–495, Cham. Springer International Publishing.