The structure and characteristics of the Corpus of Polish Science of Science Journals

Emanuel Kulczycki*,
ORCID: 0000-0001-6530-3609
Yeimer Alexander Zambrano Mena,
Franciszek Krawczyk
ORCID: 0000-0003-1097-9032
Scholarly Communication Research Group
Adam Mickiewicz University in Poznań

Abstract

Purpose/Thesis: This article introduces the Corpus of Polish Science of Science Journals (CPSSJ), a specialised corpus created to support research in the field of science of science and its development in Poland.
Approach/Methods: The corpus was constructed by digitising previously non-digitized articles and retrieving articles from scientific journal websites and digital libraries. The documents were processed using various natural language processing methods.
Results and Conclusions: The capabilities of the CPSSJ are demonstrated through a topic modelling analysis of the "Nauka Polska" journal. The current iteration of the CPSSJ incorporates 12 Polish science of science journals published between 1918 and 2020, comprising a total of 51,822 documents.
Research Limitations: The study acknowledges the limitations of the corpus, particularly in the context of natural language processing and optical text recognition. While acknowledging some limitations, the article also explores opportunities for the future development of corpus.
Practical Implications: In the future, the corpus could facilitate the reconstruction of discourses related to science and higher education in Poland, thus enhancing the recognition of Polish science of science globally.
Originality/Value: The construction of this corpus represents an original undertaking involving the digitisation and processing of science of science papers. This effort resulted in the creation of a unique tool for discourse analysis.

Keywords

Bibliometrics. Polish science. Science of science. Topic modelling. Topic-specific corpus.

Text received: 2nd of November 2023

ZIN 2023, 61(2), 1–22
DOI: 10.36702/zin2023.02.01.1

Translation into English by Magdalena Paul

1. Introduction

This article presents the aims, objectives, and content of the current version of the Corpus of Polish Science of Science Journals (CPSSJ). It also presents the corpus's possibilities based on thematic analysis of the "Nauka (Polska)" journal (English title: "(Polish): Science")[1]. Additionally, it discusses its limitations and possibilities for development.

The Corpus of Polish Science of Science Journals is a topic-specific corpus and digital collection of documents from Polish science of science journals. It is intended to serve the science of science research and support the study of the emergence and development of the science of science in Poland.

The original idea of the corpus and its current shape was developed with the intent of reconstructing discourses on the evaluation of science in Poland within the framework of the project "Evaluation Game", funded by the National Science Centre in Poland. Therefore, the selection of the 12 journals that make up the current version of the corpus was conditioned by the thematic area of the project. Nevertheless, it should be emphasised that the most important, largest, and oldest science journals were already included in the corpus at this stage.

In the future, the corpus may allow the reconstruction of discourses on science and higher education in Poland and raise the visibility of Polish science of science research and its output worldwide. Currently, the science of science research (the equivalent of the Polish term *naukoznawstwo* or *nauka o nauce*) is recognised as an emerging field (Fortunato et al., 2018; Wang D. & Barabási, 2021), and ideas emerging in this field in the United States, for example, are almost always considered innovative. It is caused by the low awareness of the Polish contribution to the establishment of the discipline at the beginning of the twentieth century despite the efforts of contemporary Polish researchers and their excellent works, including those published in English (Kawalec, 2019; Kokowski, 2015, 2016; Walentynowicz, 1975). Unfortunately, the reception of Polish science of science research in Western countries was limited not only by late translations – one of the foundation texts, "Science of Science" (1935) by Maria and Stanisław Ossowski, was translated and started to be widely available in the West only three decades later through publication in the journal "Minerva" (Ossowska & Ossowski, 1964), but also by the 'overshadowing' of Polish tradition of science of science by the Soviet model of science and higher education and their management.

Making the corpus a valuable tool for studying discourses requires text analysis techniques, with a particular focus on Text Mining and Natural Language Processing (NLP) methods (Jo, 2019; Kao & Poteet, 2007). These methods enable the

[1] Editor's note: As a rule, we provide journal titles in Polish and, in the case of references to specific journals, also in English. English translations made by the translator.

automatic extraction of information from texts and the identification of key themes, which contributes to a better understanding of the structure and dynamics of scientific discourse. The preparation for data extraction from texts required applying several steps described further in the text, including cleaning, tokenisation, and lemmatisation. Topic analysis can be carried out in different ways, which depend largely on the specific characteristics of the corpus, i.e. subject matter, length of texts, language, and style (Sbalchiero & Eder, 2020). There are many approaches to topic emergence. One can, however, point to the two most popular methods, i.e. Non-negative Matrix Factorisation (NMF) and Latent Dirichlet Allocation (LDA) (Han, 2020; Sugimoto et al., 2011; Wang Y.-X. & Zhang, 2013). In the present analysis, we used the NMF method (all parts of the procedure were performed in Python), as it allows for the emergence of more disjunctive topics than LDA in the case of our corpus. It should, of course, be mentioned that the emergence of topics depends very much on the quality of the textual data. Moreover, determining the best number of topics – how many topics the whole corpus will be split into – is done iteratively, i.e. by trial and error, using various statistical measures and, ultimately, most importantly, expert decisions.

2. Journals included in the corpus

The current version of the CPSSJ contains 12 journals, as shown in Table 1. Given the variability of journal titles, and editorial boards and journal transformations, we realise that these journals could be separated into smaller parts. For example, "Forum Akademickie" ("Academic Forum") could be treated as a separate journal from "Przegląd Akademicki" ("Academic Review"), and "Życie Szkoły Wyższej" ("Life of Higher Education") could be separated from „Życie Nauki: Miesięcznik Naukoznawczy" ("Life of Science: Science of Science Monthly"), which proceeded with the former. However, considering the entire history of the analysed journals, we decided this aggregation is not only acceptable but also useful (as it allows us to show differences in content and publishing over the years). Of course, due to the construction of the corpus, these journals can be 'separated' for other analyses.

Table 1. List of journals included in the Corpus of Polish Science of Science Journals and their years of publication (as of the end of 2020).

No.	Journal's title	Years of publication	Previous titles
1	"Forum Akademickie"	1991–2020	"Przegląd Akademicki"
2	"Kwartalnik Historii Nauki i Techniki"	1956–2020	

No.	Journal's title	Years of publication	Previous titles
3	"Nauka"	1954–2020	"Nauka Polska"; in 1957 "Nauka Polska" was combined with "Sprawozdania z Czynności i Prac".
4	"Nauka i Szkolnictwo Wyższe"	1993–2019	
5	"Nauka Polska. Jej Potrzeby Organizacja i Rozwój"	1918–1920, 1923, 1925, 1927–1939, 1947, 1992–2020	"Nauka Polska. Jej Potrzeby, Organizacja i Rozwój. Rocznik Kasy Pomocy dla Osób Pracujących na Polu Naukowym Imienia Doktora Józefa Mianowskiego"
6	"PAUza Akademicka"	2008–2020	"PAUza"
7	"Planowanie i Organizacja Badań Naukowych"	1980, 1982–1987, 1989	
8	"Sprawy Nauki: Biuletyn Komitetu Badań Naukowych"	1991–2009	"Biuletyn Komitetu Badań Naukowych"
9	"Sprawy Nauki: Miesięcznik Publicystyczny-Informacyjny"	2006–2020	
10	"Zagadnienia Informacji Naukowej"	1962–2020	"Biuletyn Ośrodka Dokumentacji i Informacji Naukowej PAN"
11	"Zagadnienia Naukoznawstwa"	1965–2019[2]	The current version of the corpus does not contain a self-contained supplement, "Problems of the Science of Science," published in 1970–1971, 1973, 1974, 1976, and 1977/1979.
12	"Życie Szkoły Wyższej"	1946–1952; 1953–1991	"Życie Szkoły Wyższej" was published from 1953 onwards instead of "Życie Nauki: miesięcznik naukoznawczy"

Source: compiled by the Authors.

Some of the science of science journals included in the corpus had already been provided material for discourse analyses or monographic studies. For example, "Forum Akademickie" and "Nauka" been provided the corpus for discourse analysis on parameterisation (Ostrowicka & Spychalska-Stasiak, 2017). Kawalec analysed "Zagadnienia Naukoznawstwa" ("Issues in Science of Science") in terms of topics using data from Google Scholar (2017) and its internationalisation based on the Scopus database (2020). The monographic studies of "Kwartalnik Historii Nauki i Techniki" ("Quarterly Journal of the History of Science and Technology") and the journal "Nauka Polska. Jej Potrzeby Organizacja i Rozwój" ("Nauka Polska.

[2] In September 2023, the table of contents of the 2020 issue could be found, but the Authors could not find even a legal deposit copy in the National Library.

Its Needs Organisation and Development") have been presented by Stefan Zamecki in several books (Zamecki, 2016, 2017, 2018, 2020). There were also self-referantial issues or articles compiled by individual journals, such as the 2006 issue with an introductory text by Zamecki (2006). It should also be noted that over the years, texts summarising the development of science writing (Rutkowski, 1947) and the role of journals such as "Nauka Polska. Jej Potrzeby Organizacja i Rozwój" or "Życie Nauki: Miesięcznik Naukoznawczy" (Choynowski, 1948; Kowalczyk et al., 1969) have been produced. However, no systematic analysis has been undertaken that would consider the more significant number of scientific journals over the entire publication period.

3. Corpus development

The following section describes the subsequent steps in creating CPSSJ, starting with collecting journal issues and proceeding through preparing and processing documents and preparing them for further analyses.

3.1. Retrieval of journal content and scanning of paper issues

It was decided, where possible, to use scanned journal documents distributed in open access on the journals' home pages or in digital libraries. Python scripts were adopted to download the documents in bulk, along with publication metadata – if possible – such as article title and authors. For each journal, the quality of the scans was determined – whether it was sufficient to recognise the texts well or whether the layer of recognised text in the document was of good quality (i.e. without numerous distortions).

Journals without digital versions were scanned at 600 or 300 DPI grayscale resolution to TIFF or JPG format, depending on the print quality.

Figure 1 shows that some titles had an almost complete digital archive, such as the journal "Życie Szkoły Wyższej", which consists of scanned and born-digital documents for later annuals, or "PAUza Akademicka", which has published digitally produced documents since its beginning.

To summarise, 49.1% of the documents in the Corpus of Polish Science of Science Journals are from the online version (born-digital or scanned), and 50.9% were scanned by the research team during the creation of the CPSSJ.

3.2. Division of journal issues into documents

The journal issues were scanned in their entirety, including editorial pages and tables of contents. The issues downloaded directly from the webpages of journals or digital libraries were published either as whole scanned issues (e.g., in the case of the “Zagadnienia Informacji Naukowej” (“Issues in Scientific Information”) or as individual documents. In the latter case, the journals did not include scans of editorial pages or tables of contents.

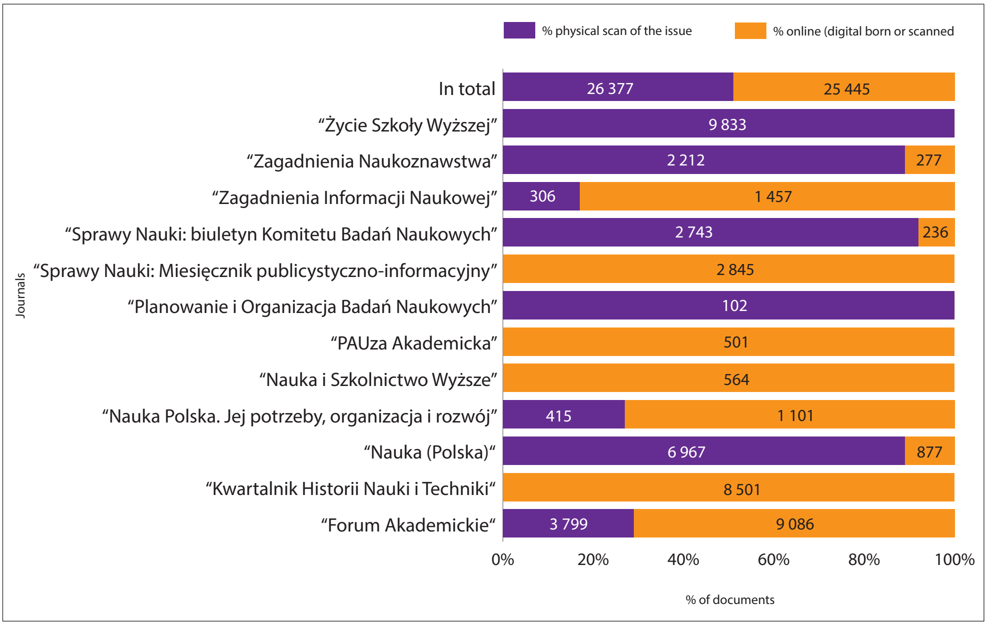

Figure 1: Number of online documents and scans for each journal.

Source: compiled by the Authors.

This information was important when assessing the quality of document categorisation and the percentage of documents excluded from the analyses (e.g., synonym analysis or topic analysis).

Finally, 43.5% of the documents in CPSSJ were originally divided by journals and 55.5% by the research team. One per cent of the documents (these are issues of the journal “PAUza Akademicka”) were classified as undivided.

3.3. Categorisation of documents

For further analysis, scholarly articles and positions of state bodies or institutions of the science and higher education sector were primarily considered. Therefore, documents recognised during the digitisation process as clearly not fitting into the scope were excluded from the analysis. However, such an expert categorisation was not fully effective. It could not be applied on a larger scale to journals focused on reporting on scientific life rather than publishing scientific articles (this applies, for example, to the journal "Sprawy Nauki: Miesięcznik Publicystyczny-Informacyjny" ("Matters of Science: Monthly Journal of Opinion and Information")). The following types of documents were not included in the analysis:

- editorial pages;
- tables of contents;
- articles published in a language other than Polish;
- biographies;
- papers consisting only of tables or numerical lists;
- bibliographical lists;
- advertisements;
- abstracts of articles.

Nevertheless, all documents from the corpus were subjected to the same operations (i.e., text recognition and text processing).

Table 2 summarises the number of documents from each journal classified or not for further analysis. The term 'documents' refers to all publications in a journal, while articles are those documents that enter the analysis.

Table 2: Documents classification by journal.

Journal	**Classified**	**% of classified**	**Non-classified**	**% of non-classified**	**Documents in total**
"Forum Akademickie"	12 053	93,5%	832	6,5%	12 885
"Kwartalnik Historii Nauki i Techniki"	7 038	82,8%	1 458	17,2%	8 496
"Nauka (Polska)"	6 024	76,8%	1 820	23,2%	7 844
"Nauka Polska. Jej Potrzeby Organizacja i Rozwój"	1 095	72,2%	421	27,8%	1 516
"Nauka i Szkolnictwo Wyższe"	549	97,3%	15	2,7%	564
"PAUza Akademicka"	501	100,0%	0	0,0%	501
"Planowanie i Organizacja Badań Naukowych"	52	51,0%	50	49,0%	102
"Sprawy Nauki: Miesięcznik Publicystyczny-Informacyjny"	2 825	99,3%	20	0,7%	2 845
"Sprawy Nauki: Biuletyn Komitetu Badań Naukowych"	2 144	72,0%	835	28,0%	2 979
"Zagadnienia Informacji Naukowej"	1 362	77,3%	401	22,7%	1 763
"Zagadnienia Naukoznawstwa"	2 113	85,1%	371	14,9%	2 484
"Życie Szkoły Wyższej"	8 500	86,4%	1 333	13,6%	9 833
In total	44 256	85,4%	7 556	14,6%	51 812

Source: compiled by the Authors.

3.4. Text recognition (OCR) of scanned documents

Various solutions to achieve satisfactory text recognition quality were tested, with the final decision to use ABBYY FineReader 11 Professional Edition. This tool allowed text and paragraph recognition and linking in the case of double-page printing.

The input files were PDF files (originally prepared by journals) or TIFF/JPG image files. The output files were PDF files with documents created from the image and TXT files, which were the basis for further document processing.

3.5. Data cleaning, tokenisation, and lemmatisation

The input text documents in TXT format contain data representing redundant noise in the analysis. Therefore, the data processing procedure consists of the following steps:

- Removal of numbers, webpage addresses, punctuation marks, email addresses and special characters (such as !@$%*><+?);
- Converting all text to lowercase;
- Removal of a publication’s title and authors’ names (many journals use a running head so that the title and authors appear multiple times in a given document);
- Tokenisation, dividing the text into individual units, called tokens, which usually represent single words or phrases. The purpose of tokenisation is to organise the text and turn it into a structure that can be quickly processed;
- Removing StopWords, i.e., words that appear frequently but are irrelevant from the perspective of the analyses undertaken (Schofield et al., 2017). This includes, e.g., prepositions, first names, and words common in a given topic-specific corpus (in the case of CPSSJ, this would also include ‘Polish’ or ‘science’, among others).

The texts thus processed and cleaned data are subjected to lemmatisation. Lemmatisation is converting words to their basic form, known as a lemma. It is a crucial step in text analysis, which aims to reduce the different grammatical forms of words to one, thus facilitating the analysis and comparison of textual data. In the case of the Polish language, lemmatisation is extremely difficult due to the richness of inflectional forms, the complexity of grammar and the numerous exceptions. During the research work, it turned out that two Python libraries, namely Lemmagen and Morpheus, provide the best results and performance. Transforming the text employing lemmatisation makes it possible to focus on the critical elements of the text while eliminating redundancies due to the variety of grammatical forms.

3.6. Creating a database of documents

The corpus consists of text documents linked to bibliographic data about the documents (functioning as metadata). Metadata was transcribed in the case of documents scanned from the tables of contents. In the case of digital documents (downloaded from the web pages of journals or archives), the information provided next to the document was used. The corpus database contains information for each document. Their scope refers to Title, Authors, Year of publication, Category of the document (category was produced by the research team), Information on whether the document is suitable for analysis, Volume, Issue number, DOI identifier, Link to PDF (or HTML) online version, Start page, End page.

The percentage of information in each Table 3 column depends on the quality of the data and the type of documents (articles published in HTML do not have pagination information, and not all documents indicate author information).

4. Quantitative characteristics of the magazines

4.1. Number of volumes and articles

Table 4 shows the number of volumes (up to 2020) included in the corpus for each journal, the number of documents and the number of articles (i.e. papers classified for analysis) per volume.

Figure 2 shows the change in the number of documents (note: not articles) over time for each journal in the corpus. Two volumes are only connected by a line if there is editorial continuity, i.e. the volumes were published year after year.

4.2. The length of the documents

The average length of the documents for each journal was calculated (as shown in Figure 3). In this analysis, words after the cleaning and lemmatisation procedure we counted words.

4.3. Number of authors by document

The change in the average number of authors per article by year was analysed for each journal (as shown in Figure 4). Only papers with author information are included in the analysis. The outliers for the journal „Nauka Polska. Jej Potrzeby Organizacja i Rozwój" are due to the publication of so-called dissertations, i.e. statements by multiple scholars within a single paper (each statement has an authorship indicated).

Table 3. Summary of bibliographic information in the corpus.

Journal	Title	Author	Year	Document's category	To be analysed	Vol-ume	Issue	DOI	PDF	Start page	End page
"Forum Akademickie"	100,0%	70,4%	100,0%	5,8%	100,0%	100,0%	0,0%	0,0%	0,0%	27,4%	27,3%
"Kwartalnik Historii Nauki i Techniki"	100,0%	86,3%	100,0%	100,0%	100,0%	0,0%	0,0%	0,0%	0,0%	99,4%	99,3%
"Nauka (Polska)"	100,0%	84,2%	100,0%	100,0%	100,0%	100,0%	99,7%	0,0%	14,0%	93,9%	93,9%
"Nauka Polska. Jej Potrzeby Organizacja i Rozwój"	100,0%	66,5%	100,0%	100,0%	100,0%	100,0%	0,6%	0,0%	0,0%	94,8%	94,7%
"Nauka i Szkolnictwo Wyższe"	100,0%	99,4%	100,0%	100,0%	100,0%	100,0%	100,0%	0,0%	97,5%	0,0%	0,0%
"PAUza Akademicka"	0,0%	0,0%	100,0%	0,0%	100,0%	0,0%	100,0%	0,0%	100,0%	0,0%	0,0%
"Planowanie i Organizacja Badań Naukowych"	100,0%	61,4%	100,0%	0,0%	100,0%	100,0%	0,0%	0,0%	0,0%	89,3%	89,3%
"Sprawy Nauki: Miesięcznik Publicystyczny-Informacyjny"	100,0%	0,0%	100,0%	0,7%	100,0%	100,0%	0,0%	0,0%	100,0%	0,0%	0,0%
"Sprawy Nauki: Biuletyn Komitetu Badań Naukowych"	100,0%	55,3%	100,0%	0,5%	100,0%	100,0%	100,0%	0,0%	0,0%	87,4%	87,4%
"Zagadnienia Informacji Naukowej"	100,0%	81,4%	100,0%	21,0%	100,0%	100,0%	100,0%	0,0%	0,0%	72,2%	71,6%
"Zagadnienia Naukoznawstwa"	100,0%	88,5%	100,0%	0,0%	99,7%	0,0%	0,0%	0,0%	0,0%	79,1%	79,1%
"Życie Szkoły Wyższej"	100,0%	75,3%	100,0%	6,4%	100,0%	100,0%	100,0%	0,1%	0,0%	91,6%	91,6%

Source: compiled by the Authors.

Table 4. Summary of the number of years of publication and articles per year.

Journal	Documents in total	Number of articles to be analysed	Number of years	Number of articles per year
"Forum Akademickie"	12 885	12 053	30	429,5
"Kwartalnik Historii Nauki i Techniki"	8 501	7 038	65	130,8
"Nauka (Polska)"	7 844	6 024	68	115,4
"Nauka Polska. Jej Potrzeby Organizacja i Rozwój"	1 516	1 095	48	31,6
"Nauka i Szkolnictwo Wyższe"	564	549	27	20,9
"PAUza Akademicka"	501	501	13	38,5
"Planowanie i Organizacja Badań Naukowych"	102	52	8	12,8
"Sprawy Nauki: Miesięcznik Publicystyczny-Informacyjny"	2 845	2 825	15	189,7
"Sprawy Nauki: Biuletyn Komitetu Badań Naukowych"	2 979	2 144	19	156,8
"Zagadnienia Informacji Naukowej"	1 763	1 362	59	29,9
"Zagadnienia Naukoznawstwa"	2 489	2 113	54	46,1
"Życie Szkoły Wyższej"	9 833	8 500	46	213,8
In total	51 822	44 256	452	114,7

Source: compiled by the Authors.

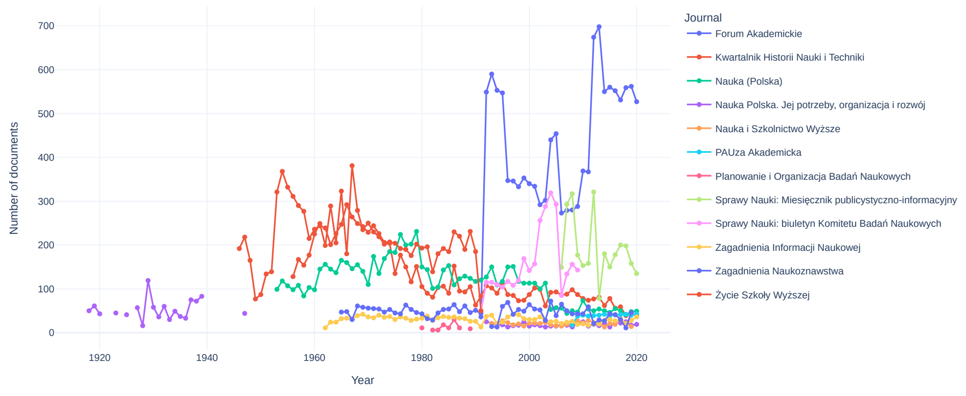

Figure 2: Number of documents for each journal over time.

Source: compiled by the Authors.

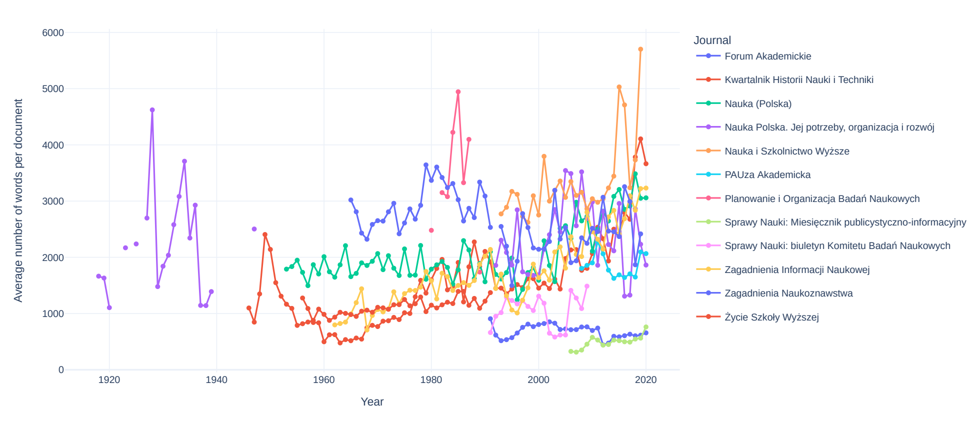

Figure 3: Number of words per document for each journal over time.

Source: compiled by the Authors.

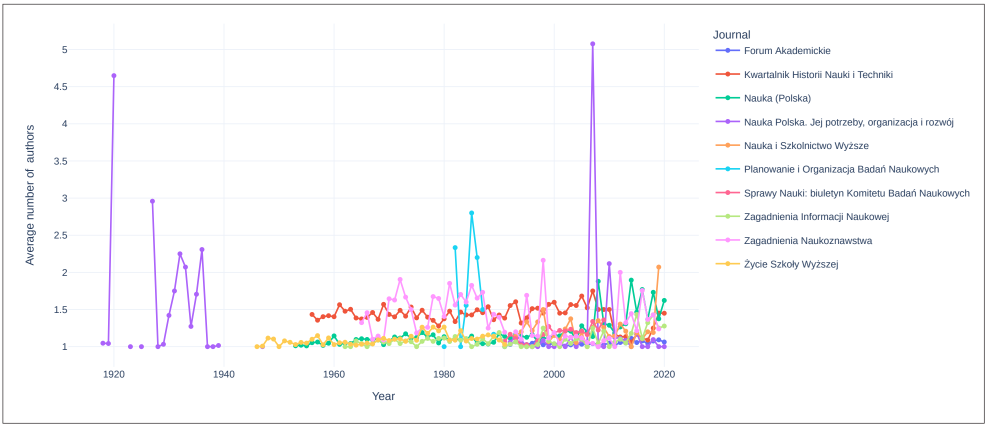

Figure 4: Average number of authors of a document for each journal over time.

Source: compiled by the Authors.

5. Analysis of the topics on the example of "Nauka (Polska)"

This section aims to show how CPSSJ can be used to create a landscape of discussions (topics) within the pages of a single journal. "Nauka (Polska)" was chosen because it has been published continuously from 1953 to 2020 (this is the final year of the current version of CPSSJ), first under the name "Nauka Polska" and then from 1994 "Nauka." Therefore, we use the notation with "Polska" placed in brackets to emphasise the merger of the two journals.

It's worth to acknowledge that this analysis could still be extended to include the volumes published from 1918 onwards by the journal "Nauka Polska. Jej Potrzeby Organizacja i Rozwój", as the history of all these journals is connected and intertwined. This intertwining is best illustrated by the opening paragraphs of the first issue of "Nauka" from 1994, quoted here in full:

> "Nauka Polska" was published annually in 1918–1939 and 1947 by the Józef Mianowski Fund. From 1953, "Nauka Polska" was published quarterly by the Polish Academy of Sciences. In 1962–1974, it was published bimonthly, then, in the period 1975–1981, monthly, to return to the bimonthly form in 1982–1993. In issue 5 (270) of "Nauka Polska" of 1993, the organisational issues, profile and scope of the journal were discussed in more detail. A further fundamental change was brought in 1994, with "Nauka Polska" being transformed into a new title – the quarterly "Nauka".
>
> There are two main reasons for this transformation. The first was the reactivation of the Józef Mianowski Fund – the Foundation for the Promotion of Science in 1991, after forty years. The Józef Mianowski Fund returned to the publication of its former title – an annual journal, "Nauka Polska," the first and subsequent 26th issue, which appeared in 1992. In this situation, it became obvious for the management of the Polish Academy of Sciences to give up the title it had held until 1951 ("Od Redakcji" [From the Editor], 1994).

Between 1953 and 2020, "Nauka (Polska)" published 7,844 documents, of which 6024 articles qualified for analysis (1820 were excluded). A more significant percentage of articles were excluded from the analysis before 2004 because, since that year, the editors have a digital archive of the journal on their webpage, with the issues divided by the editors. Thus, our corpus did not include tables of contents from 2004, for example, which were generally excluded from the analysis. Additionally, in a further step, 252 articles that had fewer than 300 words (after Lemmagen lemmatisation) were excluded from the analysis. Ultimately, 26.42% of the documents were excluded from the analysis. Thus, the final set for topic analysis consists of 5,772 articles.

Assigning this many articles to individual topics (unknown before reading) would be theoretically possible but highly labour-intensive. Therefore, topic modelling can be performed using machine learning techniques. In this case, this was unsupervised machine learning, as the topics were not defined beforehand or labelled training data provided before the process (e.g., there was no indication that a text should be classified with a specific other text within the same topic).

Topic modelling is a quantitative textual analysis that allows the corpus documents to be grouped according to dominant themes. Multiple themes will characterise each document. However, the dominant one will be identified, and the document will be assigned to it. In this approach, the corpus is treated as a collection of documents, each of which is composed of a defined number of topics, which topics are composed of words from the corpus. This means that the individual words in the corpus are associated with a given topic(s). Each word has its weight, indicating its relevance to the topic.

5.1. The procedure adopted for the selection of topics

When analysing multiple longer text documents, the advantage of Non-negative Matrix Factorisation is that one does not work on the entire TF-IDF (Term Frequency-Inverse Document Frequency) matrix. However, as part of the analysis, one reduces the complexity of this matrix by reducing the number of words considered. TF-IDF is one method of calculating word weights based on the number of occurrences of words. This method considers the frequency of occurrence of a word in a document (TF) and its uniqueness among the entire set of documents (IDF), enabling one to understand the meaning and context of individual words in a given text corpus. Analysing the word weights of all the documents in the collection makes it possible to identify keywords that characterise particular topics. This makes it possible to create groups of documents based on the similarity of important words, enabling topic modelling. The ability to reduce the complexity of the TF-IDF when using NMF is essential for a corpus if some documents have been created from scanned documents, with the result that there may be individual words that 'do not make sense', as these are errors resulting from data processing.

The output TF-IDF matrix describing CPSSJ consists of 5,772 articles and 64, 110 unique words (taking into account words that occur at least five times in the entire corpus but can occur in a single document). The sparsity of the matrix in this case is 98.59%, which means that such a percentage of matrix elements is zero (that is, the word does not occur in the document). Therefore, to improve the model, the sparsity of the matrix had to be reduced, as processing so many zeros is not efficient and sparse matrices are more challenging to compute, making the algorithms 'prefer' to operate on less sparse data.

To this end, we tested what model parameters we could adopt to reduce the matrix's sparsity and, more importantly, obtain coherent and disjunctive topics. It was determined that the final matrix has a sparsity of 61.33% and, for the topic analysis, will be characterised by the parameters max_features=800 and mindf_ig=20 for the nine topics, where max_feature indicates the number of first keywords for the corpus (ordered by their importance) and mindf_ig indicates the threshold for ignoring keywords in the model (in this case words that occur less than 20 times in the corpus).

5.2. The topics of "Nauka (Polska)"

Adopting the method described above, nine coherent and disjunctive topics were identified. All analysed articles were classified according to these topics. It should be emphasised again that through topic analysis, the leading topic of a document is assigned, which does not mean that other themes are not 'visible' in the document.

Figure 5 shows the distribution of topics by year of journal publication. The first topic, dedicated to the description of technical and natural research, was more important in the journal before 2010 (particularly in the 1960s).

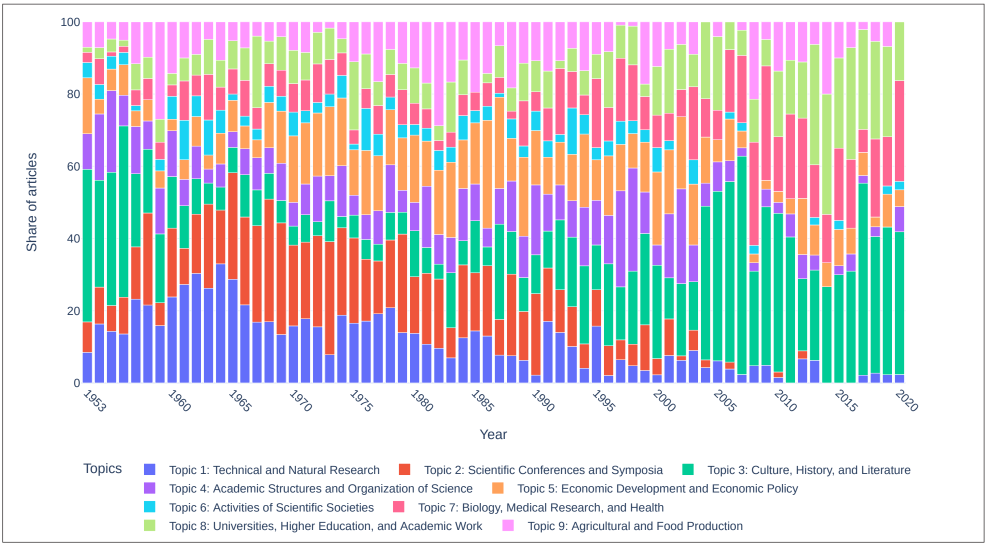

Figure 5: Distribution of topics in the journal "Nauka (Polska)" between 1953 and 2020.

Source: compiled by the Authors.

It can also be seen that after 2005, the publication of material describing conference events and scientific symposia ceased altogether, while at the same time, the number of articles on the humanities (Topic 3: *Culture, history, and literature*) and on higher education and scientific work (Topic 8: *Universities, higher education, and scientific work*) increased significantly. There was also a significant decrease in the number of publications on technical and natural science research (Topic 1: *Technical and natural science research*). The observation that the technical and natural science research theme has been increasing quantitatively in frequency successively since the 1950s and declining significantly after the political transformation of the 1990s may prove relevant for more in-depth historical research. As Hubner (1994, p. 32) discussed, the first Soviet-inspired attempts at reforming post-war science in Poland aimed to increase the role of technical and 'hard' sciences versus the humanities and natural sciences.

5.3. Results evaluation

The biggest challenge in topic analysis is evaluating the model, i.e. assessing whether the emerged themes describe the data (texts) well and are meaningful and consistent. Although several quantitative measures can be used to assess the number of emerged topics and their meaningfulness, such as topic coherence, perplexity (a measure used to assess how well the model is able to predict new, previously unknown textual data), and visualisation of topics (e.g. using the pyLDAvis library), expert evaluation of each outcome is irreplaceable for such a topic-specific corpus.

The model was evaluated as follows: (1) the relationships between topics using network graphs were analysed; (2) the most important emergent words for a given topic were created and analysed, and generic labels for the topics were established to verify potential overlaps between these topics further (in this process, each research team member created labels, and then it was agreed on a common label or discrepancies were explained); (3) the accuracy of classification into particular topics was checked for randomly selected articles; (4) in the case of the topic analysis of "Nauka (Polska)", a comparison was possible with the *Bibliografia Polskiej Naukometrii* (*Bibliography of Polish Scientometrics*, BPN)[3], which indexes nearly two thousand publications by Polish researchers, classified as scientometric papers. Therefore, all articles published in "Nauka (Polska)" were retrieved from BPN and checked whether they were assigned to topics that could be considered 'scientometric'.

5.3.1. Network graphs

Network graphs can be used to visualise the relationships between topics in the NMF model. Such a network consists of nodes and edges: nodes represent different topics, and edges represent the strength of the relationship between topics (the closer the value is to 0, the more disconnected the topics are, and the closer the value is to 1, the more similar the topics are). Figure 6 shows the results for the assumptions made in the analysis based on cosine similarity. To interpret the graph, one can look at the clustering of nodes and the connections between them. Clusters of nodes that are strongly connected indicate topics that are related or cover similar aspects of the corpus. Weakly connected or not connected nodes indicate disjunctive topics and cover different aspects of the corpus. The network graph can also be used to identify outlier topics or poorly defined topics.

5.3.2. The most significant keywords

The following list contains the ten most significant keywords (translated to English) for each topic, ordered from most important (with the highest weighting) to least important.

[3] *Bibliografia Polskiej Naukometrii*, https://sc.amu.edu.pl/bibliography/, accessed: 1st of September 2023.

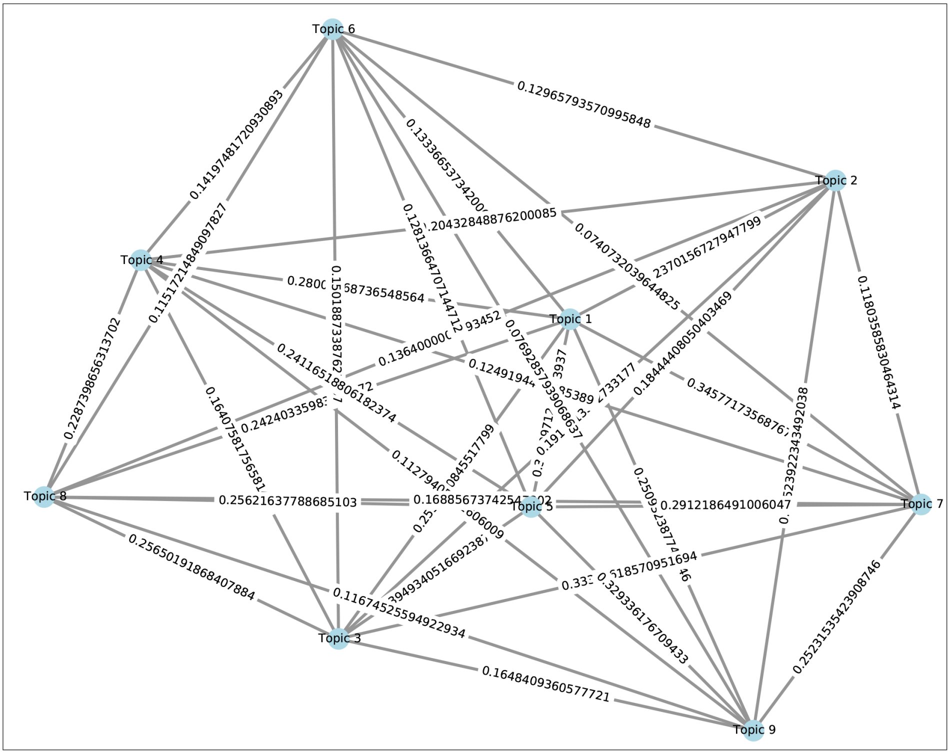

Figure 6: The network graph between the 9 Topics of "Nauka (Polska)" from 1953 to 2020.

Source: compiled by the Authors.

- For Topic 1: *Technical and natural science research*, the most important keywords are research, work, facility, scope, (adj.) research, scientific field, method, issue, physics, and development.
- For Topic 2: *Scientific conferences and symposia*, the most important keywords are paper, congress, conference, session, symposium, section, international, to hold, participant, and to deliver.
- For Topic 3: *Culture, history, and literature*, the most important keywords are culture, history, language, human, literature, piece, world, great, law, and history.
- For Topic 4: *Academic structures and science organisation*, the most important keywords are academy, institution, department, praesidium, committee, secretary, assembly, cooperation, activity, and matter.
- For Topic 5: *Economic development and policy*, the most important keywords are development, country, social, economy, socialist, economical, society, state, policy, and programme.

- For Topic 6: *Scientific associations activity*, the most important keywords are association, activity, branch, scientific meeting, library, dissemination, (adj.) publishing, regional, work, and social.
- For Topic 7: *Biology, medical research, and health*, the most important keywords are cell, disease, protein, human, genetic, research, animal, health, organism, and biology.
- For Topic 8: *Universities, higher education, and scientific activity*, the most important keywords are college, university, high, professor, school, student, doctoral, academic, education system, and work.
- For Topic 9: *Agri-food production*, the most important keywords are plant, production, agriculture, agricultural, conservation, aquatic, water, energy, economy, and carbon.

Based on the lists of these keywords, the internal consistency of the topics and their disjunctions was checked, and the quality of data processing and lemmatisation was assessed. The keyword lists (the Top 20 keywords in an iterative model-building process) were an important determinant of the labels given to the topics.

5.3.3. Assignment of articles to topics

This process was purely based on expert assessment. Having already produced a model proposal and labels for the topics, the leading topic assigned to randomly selected articles was verified. During the verification process, it was necessary to note that an assigned topic was leading but not the 'only' topic. This process confirmed the quality of the final model and the parameters adopted.

5.3.4. Comparison with the Bibliography of Polish Scientometrics

Forty-one articles from "Nauka (Polska)" were found in the *Bibliography of Polish Scientometrics* till 2020. Twenty-five of them, i.e. 60% of those analysed, were classified in the model as articles with the leading Topic 8: *Universities, higher education, and scientific activity*, seven articles with Topic 3: *Culture, history, and literature*, six articles with Topic 1: *Technical and natural science research*, one article each with topics 4, 5, 7. Given that the model indicates a leading (and not the only) topic, comparing two quite different approaches, i.e., machine learning and expert classification in the *Bibliography of Polish Scientometrics*, should be considered good. The aim of the topic analysis was not to reproduce the classification in such a way that all BPN articles published in "Nauka (Polska)" would be classified into a single topic but to additionally verify that the majority of expertly classified articles would be in a small number of topics. Therefore, we believe this analysis has confirmed the value of the presented topic analysis model.

6. Corpus development perspectives

This paper presents the current state of the Corpus of Polish Science of Science Journals in 2022, after three years of work. This is, of course, only the beginning and not the end of the road. Below are the directions in which the corpus can and will be developed.

Above all, it would be worthwhile to expand the corpus to include more journals which hosted science of science discussions, such as "Studia Historiae Scientiarum" or "Organon". One of the significant challenges is improving the text data quality after the OCR process, thus improving the lemmatisation. Creating unique author identifiers will allow additional bibliometric analyses to be carried out. Currently, the corpus includes information about the authors of documents (if any were included in the table of contents or the document). However, significant work needs to be done to link authors' names recorded in different ways so that 'F. Znaniecki' is merged with 'Florian Znaniecki'.

One development direction could be extracting citations from documents, both from the references and the footnotes. Currently, there are individual tools for extracting bibliographic information from references. However, extraction from footnotes is not feasible on a mass scale, although individual research teams are working on developing the tools.

Funding

The development of this version of the corpus was supported by the project "Evaluation Game in Academia", funded by the National Science Centre in Poland, decision number UMO-2017/26EHS2/00019.

Acknowledgement

At many stages, we received assistance in searching, scanning and compiling journals and documents. We would like to sincerely thank Paulina Dudzińska, Jolanta Noskowiak, Wiesława Krysztofiak, Sara Rotnicka, Małgorzata Rychlik, Krzysztof Skibniewski, and Michał Spaleniak, whose support enabled us to complete our work.

References

Choynowski, M. (1948). Life of Science. *Synthese*, 6(5/6), 248–251.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, 359(6379), eaao0185–eaao0185. https://doi.org/10.1126/science.aao0185

Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model. *Scientometrics*, 125(3), 2561–2595. https://doi.org/10.1007/s11192-020-03721-0

Hübner, P. (1994). *Siła przeciw rozumowi: Losy Polskiej Akademii Umiejętności w latach 1939–1989*. Kraków: Wydawn. i Druk. „Secesja".

Jo, T. (2019). *Text Mining* (Vol. 45). Springer International Publishing. https://doi.org/10.1007/978-3-319-91815-0

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Kawalec, P. (2017). Wizualizacja publikacji naukoznawczych – na przykładzie wybranych artykułów z „Zagadnień Naukoznawstwa". *Zagadnienia Naukoznawstwa*, 53(4), 373–388.

Kawalec, P. (2019). Najnowsze postępy naukoznawstwa. *Ruch Filozoficzny*, 75(2), 33. https://doi.org/10.12775/RF.2019.019

Kawalec, P. (2020). Analiza poziomu umiędzynarodowienia Zagadnień Naukoznawstwa w kontekście światowych studiów nad nauką i szkolnictwem wyższym. *Zagadnienia Naukoznawstwa*, 55(1(219)), 33. https://doi.org/10.12775/ZN.2019.002

Kokowski, M. (2015). The Science of Science (Naukoznawstwo) in Poland: The Changing Theoretical Perspectives and Political Contexts – A Historical Sketch from the 1910s to 1993. *Organon*, 47, 147–237.

Kokowski, M. (2016). The Science of Science (naukoznawstwo) in Poland: Defending and Removing the Past in the Cold War. In: W E. Aronova & S. Turchetti (Eds.), *Science Studies during the Cold War and Beyond* (pp. 149–176). Palgrave Macmillan US. https://doi.org/10.1057/978-1-137-55943-2_7

Kowalczyk, K., Paszkowska, A., & Wójcik, J. (1969). *Bibliografia zawartości „Życia Nauki" 1946–1952*. Wrocław: Zakład Narodowy im. Ossolińskich.

Od Redakcji. (1994). *Nauka*, 1, 3–4.

Ossowska, M., & Ossowski, S. (1935). Nauka o nauce. *Nauka Polska*, 20, 1–12.

Ossowska, M., & Ossowski, S. (1964). The science of science. *Minerva*, 3(1), 72–82.

Ostrowicka, H., & Spychalska-Stasiak, J. (2017). Uodpowiedzialnianie akademii – formacje wiedzy i władza parametryzacji w dyskursie akademickim. *Nauka i Szkolnictwo Wyższe*, 49(1(49)), 105–132. https://doi.org/10.14746/NISW.2017.1.6

Rutkowski, J. (1947). O zadaniach Kół Naukoznawczych. *Nauka Polska. Jej Potrzeby, Organizacja i Rozwój*, 25, 303–309.

Sbalchiero, S., & Eder, M. (2020). Topic modeling, long texts and the best number of topics. Some Problems and solutions. *Quality & Quantity*, 54(4), 1095–1108. https://doi.org/10.1007/s11135-020-00976-w

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling Out the Stops: Rethinking Stopword Removal for Topic Models. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, Short Papers, 432–436. https://doi.org/10.18653/v1/E17-2069

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation. *Journal of the American Society for Information Science and Technology*, 62(1), 185–204. https://doi.org/10.1002/asi.21435

Walentynowicz, B. (1975). The Science of Science in Poland: Present State and Prospects of Development. *Social Studies of Science*, 5(2), 213–222.

Wang, D., & Barabási, A.-L. (2021). *The Science of Science* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108610834

Wang, Y.-X., & Zhang, Y.-J. (2013). Nonnegative Matrix Factorization: A Comprehensive Review. In: *IEEE Transactions on Knowledge and Data Engineering*, 25(6), 1336–1353. https://doi.org/10.1109/TKDE.2012.51

Zamecki, S. (2006). Wprowadzenie. *Kwartalnik Historii Nauki i Techniki*, 51(1), 5–7.

Zamecki, S. (2016). *Problematyka naukoznawcza na łamach periodyku „Nauka Polska. Jej Potrzeby, Organizacja i Rozwój". Studium historyczno-metodologiczne. Lata 1918–1947: T. CLXVI*. Warszawa: Wydawnictwo IHN PAN.

Zamecki, S. (2017). *O niektórych potrzebach nauki polskiej omawianych na łamach periodyku „Nauka Polska. Jej Potrzeby, Organizacja i Rozwój". Lata 1918–1947. Aneks*. Warszawa: Wydawnictwo IHN PAN.

Zamecki, S. (2018). *Problematyka naukoznawcza na łamach periodyku „Nauka Polska. Jej Potrzeby, Organizacja i Rozwój". Studium historyczno-metodologiczne. Lata 1992–2016*. Warszawa: Wydawnictwo IHN PAN.

Zamecki, S. (2020). *„Kwartalnik Historii Nauki i Techniki"—Ludzie i problemy. Lata 1956–1993*. Warszawa: Wydawnictwo IHN PAN.

Budowa i charakterystyka Korpusu Polskich Czasopism Naukoznawczych

Abstrakt

Cel/teza: Artykuł przedstawia Korpus Polskich Czasopism Naukoznawczych (KPCN), to jest specjalistyczny korpus stworzony w celu wsparcia badań w dziedzinie naukoznawstwa oraz jego rozwoju w Polsce.

Koncepcja/Metody badań: Budowa korpusu oparta była na digitalizacji wcześniej niezdigitalizowanych artykułów oraz pobieraniu tekstów z stron internetowych czasopism naukowych i bibliotek cyfrowych, które zostały poddane metodom przetwarzania języka naturalnego.

Wyniki i wnioski: Możliwości KPCN zademonstrowano poprzez analizę modelowania tematycznego czasopisma „Nauka Polska". Obecna wersja KPCN obejmuje 12 polskich czasopism naukowych z lat 1918–2020, zawierających łącznie 51 822 dokumenty.

Ograniczenia badań: Badanie uznaje pewne ograniczenia korpusu, zwłaszcza w kontekście przetwarzania języka naturalnego i optycznego rozpoznawania tekstu. Pomimo zauważonych ograniczeń, artykuł bada również możliwości przyszłego rozwoju korpusu.

Zastosowania praktyczne: W przyszłości korpus może ułatwić rekonstrukcję dyskursów związanych z nauką i szkolnictwem wyższym w Polsce, przyczyniając się do zwiększenia rozpoznawalności polskiego naukoznawstwa na arenie międzynarodowej.

Oryginalność/wartość: Budowa tego korpusu stanowi oryginalne przedsięwzięcie, obejmujące digitalizację i przetwarzanie artykułów naukowych z dziedziny naukoznawstwa. Ten wysiłek zaowocował stworzeniem unikatowego narzędzia do analizy dyskursów.

Słowa kluczowe

Bibliometria. Korpus tematyczny. Modelowanie tematyczne. Naukoznawstwo. Polska nauka.

EMANUEL KULCZYCKI – a university professor at Adam Mickiewicz University in Poznan, where he heads the Scholarly Communication Research Group. From 2018 to 2020, the chair of the European Network for Research Evaluation in the Social Sciences and the Humanities, bringing together scientists from 37 countries. In 2018, he received the Scientific Award of the President of the Polish Academy of Sciences for a series of scientific articles on scientometrics published in international journals.

YEIMER ALEXANDER ZAMBRANO MENA – a Colombian physicist and data scientist. He obtained his bachelor's degree in physics from the National University of Colombia. A member of the Physics of New Materials group at the National University of Colombia, where he worked on the analysis of X-ray spectra data. Then, obtained a master's degree in physics from the Adam Mickiewicz University in Poznan (studies funded by the Ignacy Lukasiewicz Scholarship Programme). His research interests include data science, natural language processing, and machine learning.

FRANCISZEK KRAWCZYK – a doctoral student at the Doctoral School of the Adam Mickiewicz University in Poznan. His dissertation focuses on resistance against the unequal relations between centres and peripheries in the sciences. He wrote his master's thesis on the development of so-called predatory journals. His research interests include geography of knowledge, predatory journals, evaluation and sociology of science.

Contact with Author:
emek@amu.edu.pl
Emanuel Kulczycki
Scholarly Communication Research Group
ul. Międzychodzka 5, pokój 405
60-371 Poznań