Budowa i charakterystyka Korpusu Polskich Czasopism Naukoznawczych

Emanuel Kulczycki*,
ORCID: 0000-0001-6530-3609
Yeimer Alexander Zambrano Mena,
Franciszek Krawczyk
ORCID: 0000-0003-1097-9032
Pracownia Komunikacji Naukowej
Uniwersytet im. Adama Mickiewicza w Poznaniu

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 ze 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.

Tekst wpłynął do Redakcji: 2 listopada 2023 r.

1. Wprowadzenie

Celem niniejszego artykułu jest prezentacja celów, założeń i zawartości obecnej wersji Korpusu Polskich Czasopism Naukoznawczych (KPCN). Możliwości korpusu przedstawione są na przykładzie analizy tematycznej czasopisma „Nauka (Polska)". Dodatkowo omawiamy ograniczenia oraz możliwości jego rozwoju.

Korpus Polskich Czasopism Naukoznawczych jest korpusem tematycznym (ang. *topic-specific corpus*), czyli cyfrowym zbiorem dokumentów pochodzących z polskich czasopism naukoznawczych, który ma służyć badaniom naukoznawczym oraz badaniom nad powstaniem i rozwojem naukoznawstwa w Polsce.

Pierwotna idea korpusu oraz jego aktualny kształt zostały wypracowane na potrzeby rekonstrukcji dyskursów nad oceną nauki w Polsce w ramach projektu „Punktoza w czasach systemów ewaluacji nauki" finansowanego przez Narodowe Centrum Nauki. W związku z powyższym wybór 12 czasopism, które tworzą aktualną wersję korpusu, był warunkowany obszarem tematycznym projektu. Niemniej jednak należy podkreślić, że najważniejsze, największe i najstarsze czasopisma naukoznawcze zostały w korpusie uwzględnione już na tym etapie.

W przyszłości korpus może pozwolić na rekonstrukcję dyskursów na temat nauki i szkolnictwa wyższego w Polsce oraz podniesienie rozpoznawalności polskiego naukoznawstwa i jego dorobku na świecie. Obecnie naukoznawstwo (ang. *science of science* jako angielski odpowiednik polskiego terminu „naukoznawstwo" lub „nauka o nauce") uznaje się za *emerging field* (Fortunato i in., 2018; D. Wang & Barabási, 2021), a pomysły pojawiające się w tym zakresie np. w Stanach Zjednoczonych uznaje się niemalże zawsze za nowatorskie. Świadomość o polskim wkładzie w ustanowienie tej dyscypliny na początku XX wieku jest niestety niewielka pomimo wysiłku współczesnych polskich badaczy i ich wyśmienitych prac, również publikowanych w języku angielskim (Kawalec, 2019; Kokowski, 2015, 2016; Walentynowicz, 1975). Niestety recepcja polskiego naukoznawstwa w krajach Zachodu była ograniczona nie tylko późnymi tłumaczeniami – jeden z fundacyjnych tekstów „Nauka o nauce" (1935) Marii i Stanisława Ossowskich został przetłumaczony i zaczął być powszechnie dostępny na zachodzie dopiero trzy dekady później poprzez publikację w czasopiśmie „Minerva" (Ossowska & Ossowski, 1964) – ale również „przykryciem" polskiej myśli naukoznawczej radzieckim modelem nauki i szkolnictwa wyższego oraz sposobami zarządzania tymi sektorami.

Uczynienie z korpusu narzędzia użytecznego do badania dyskursów wymagało zastosowania techniki analizy tekstu, ze szczególnym uwzględnieniem metod Text Mining i Natural Language Processing (NLP) (Jo, 2019; Kao & Poteet, 2007). Metody te umożliwiają automatyczną ekstrakcję informacji z tekstów oraz identyfikację kluczowych tematów, co przyczynia się do lepszego zrozumienia struktury i dynamiki dyskursu naukowego. Przygotowanie do ekstrakcji danych z tekstów wymagało zastosowania kilku kroków, opisanych w dalszej części tekstu, w tym

czyszczenia, tokenizacji i lematyzacji. Analizę tematów można przeprowadzać na różne sposoby, które zależą w dużej mierze od specyfiki korpusu, tj. tematyki, długości tekstów, języka, stylu (Sbalchiero & Eder, 2020). Istnieje wiele podejść do wyłonienia tematów, można jednakże wskazać na dwie najpopularniejsze metody, tj. Non-negative Matrix Factorization (NMF) oraz Latent Dirichlet Allocation (LDA) (Han, 2020; Sugimoto i in., 2011; Y.-X. Wang & Zhang, 2013). W niniejszej analizie posłużyliśmy się metodą NMF (wszystkie elementy procedury zostały wykonane w Pythonie), gdyż w przypadku badanego korpusu pozwala ona na wyłonienie bardziej rozłącznych tematów niż LDA. Należy oczywiście wspomnieć, że wyłonienie tematów zależy w bardzo istotnym stopniu od jakości danych tekstowych. Co więcej, ustalenie najlepszej liczby tematów – to znaczy na ile tematów cały korpus zostanie podzielony – jest dokonywane w sposób iteracyjny, tj. metodą prób i błędów, z użyciem różnych miar statystycznych i ostatecznie, co najważniejsze, decyzji eksperckich.

2. Czasopisma uwzględnione w korpusie

Aktualna wersja KPCN zawiera 12 czasopism zebranych w Tabeli 1. Biorąc pod uwagę zmienność tytułów prasowych, przekształcenia redakcji i czasopism, zdajemy sobie sprawę, że czasopisma te można byłoby rozdzielić na mniejsze części, dla przykładu „Forum Akademickie" traktować jako czasopismo odrębne od „Przeglądu Akademickiego", a „Życia Szkoły Wyższej" nie łączyć z „Życiem Nauki: Miesięcznikiem Naukoznawczym", zamiast którego było publikowane. Jednakże biorąc pod uwagą całą historię analizowanych czasopism uznaliśmy, że właśnie taka agregacja jest nie tylko akceptowalna, ale również użyteczna (pozwala bowiem ukazywać różnice treściowe i wydawnicze na przestrzeni lat). Oczywiście ze względu na budowę korpusu, na potrzeby dowolnej analizy czasopisma te można „rozdzielić".

Tabela 1. Lista czasopism oraz lata ukazywania się (stan na koniec 2020 r.) czasopism zawartych w Korpusie Polskich Czasopism Naukoznawczych.

Lp.	Czasopismo	Lata ukazywania się	Poprzednie tytuły
1	„Forum Akademickie"	1991–2020	„Przegląd Akademicki"
2	„Kwartalnik Historii Nauki i Techniki"	1956–2020	
3	„Nauka"	1954–2020	„Nauka Polska"; w 1957 r. „Nauka Polska" została połączona ze „Sprawozdaniami z Czynności i Prac"
4	„Nauka i Szkolnictwo Wyższe"	1993–2019	

Lp.	Czasopismo	Lata ukazywania się	Poprzednie tytuły
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[1]	Obecna wersja korpusu nie zawiera dodatku samoistnego „Problems of the Science of Science”, który ukazywał się w latach: 1970–1971, 1973, 1974, 1976, 1977/1979
12	„Życie Szkoły Wyższej”	1946–1952; 1953–1991	„Życie Szkoły Wyższej” było publikowane od 1953 r. zamiast „Życie Nauki: miesięcznik naukoznawczy”

Źródło: opracowanie własne.

Niektóre czasopisma naukoznawcze zawarte w korpusie stanowiły już materiał do analizy dyskursu bądź też doczekały się opracowań monograficznych. Dla przykładu, „Forum Akademickie” i „Nauka” stanowiły korpus dla analizy dyskursu na temat parametryzacji (Ostrowicka & Spychalska-Stasiak, 2017). Kawalec analizował „Zagadnienia Naukoznawstwa” pod kątem tematyki, używając danych z Google Scholar (2017) oraz ich umiędzynarodowienia w oparciu o bazę Scopus (2020). Opracowania monograficzne „Kwartalnika Historii Nauki i Techniki” oraz czasopisma „Nauka Polska. Jej Potrzeby Organizacja i Rozwój” zostały przedstawione

1 We wrześniu 2023 r. można było znaleźć spis treści numeru z 2020 r., lecz nie odnaleźliśmy egzemplarza obowiązkowego nawet w Bibliotece Narodowej.

przez Stefana Zameckiego w kilku książkach (Zamecki, 2016, 2017, 2018, 2020). Za takie opracowania można uznać też autoreferencyjne zeszyty czy artykuły opracowywane przez poszczególne czasopisma, jak np. zeszyt z 2006 r. „Kwartalnika Historii Nauki i Techniki" z tekstem wprowadzającym Zameckiego (2006). Należy również zaznaczyć, że na przestrzeni lat powstały teksty podsumowujące rozwój piśmiennictwa naukoznawczego (Rutkowski, 1947) oraz rolę czasopism, takich jak „Nauka Polska. Jej Potrzeby Organizacja i Rozwój" czy „Życie Nauki: Miesięcznik Naukoznawczy" (Choynowski, 1948; Kowalczyk i in., 1969). Cały czas nie podjęto jednak systematycznych analiz, które uwzględniałyby większą liczbę czasopism naukoznawczych w całym okresie ich ukazywania się.

3. Budowa korpusu

Niniejsza sekcja opisuje kolejne kroki tworzenia KCPN, zaczynając od zgromadzenia zeszytów czasopism, przez przygotowanie i przetwarzanie dokumentów, aż po przygotowanie ich do dalszych analiz.

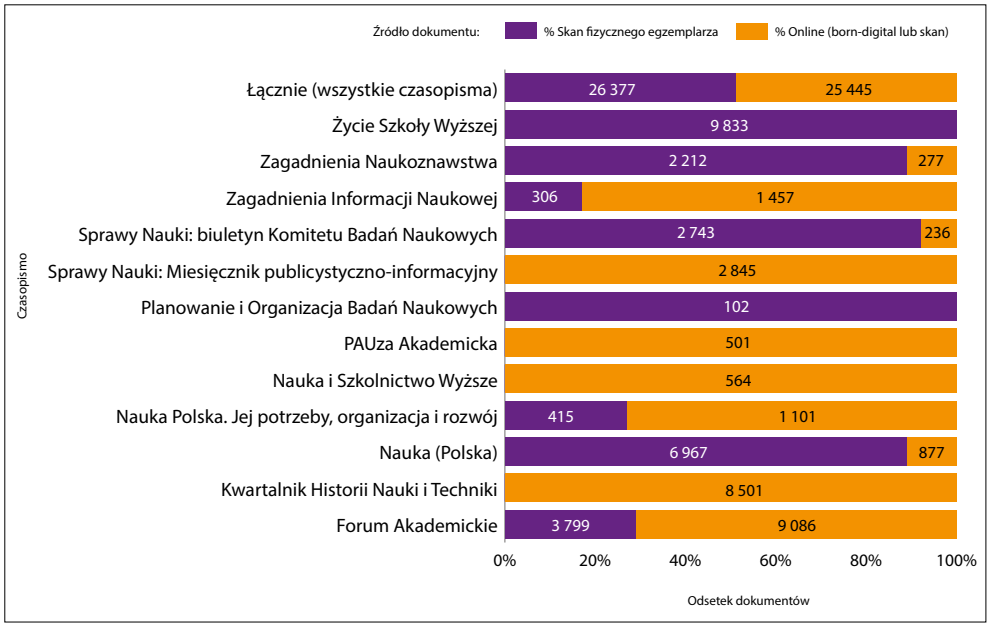
3.1. Pobieranie zawartości czasopism oraz skanowanie wydań papierowych

Zdecydowaliśmy się, tam, gdzie było to możliwe, wykorzystać zeskanowane dokumenty czasopism, które zostały zamieszczone w otwartym dostępie na stronach domowych czasopism lub w bibliotekach cyfrowych. W tym celu wykorzystaliśmy skrypty napisane w Pythonie do masowego pobierania dokumentów wraz z metadanymi publikacji, gdy było to możliwe, takimi jak tytuł artykułu, autorzy. Dla każdego czasopisma określaliśmy jakość skanów (czy jest wystarczająca do dobrego rozpoznania tekstów lub czy warstwa rozpoznanego tekstu w dokumencie jest dobrej jakości, tj. bez licznych przekłamań).

Czasopisma, które nie miały dostępnych wersji cyfrowych, skanowaliśmy zależnie od jakości druku w rozdzielczości 600 lub 300 DPI w skalach szarości do formatu TIFF lub JPG.

Jak pokazuje Rysunek 1, część tytułów posiadała niemalże kompletne archiwum cyfrowe, jak czasopismo „Nauka i Szkolnictwo Wyższe" składające się zarówno z dokumentów zeskanowych, jak i wytworzonych cyfrowo (ang. *born-digital*) dla późniejszych roczników, czy też „PAUza Akademicka", która od początku swojego istnienia publikuje dokumenty wytworzone cyfrowo.

Ostatecznie w Korpusie Polskich Czasopism Naukoznawczych 49,1% dokumentów pochodzi z wersji online (*born-digital* lub skan), a 50,9% zostało zeskanowanych przez nas podczas tworzenia KPCN.



Rysunek 1. Liczba dokumentów online oraz skanów dla każdego czasopisma.

Źródło: Opracowanie własne.

3.2. Podział zeszytów czasopism na dokumenty

Zeszyty czasopism skanowaliśmy w całości, tj. uwzględniając zarówno strony redakcyjne, jak i spisy treści. Zeszyty czasopism, które pobraliśmy ze stron tychże czasopism lub z bibliotek cyfrowych, były publikowane albo jako całe zeskanowane zeszyty (np. w przypadku części „Zagadnień Informacji Naukowej"), albo jako pojedyncze dokumenty. W tym drugim przypadku czasopisma nie zamieszczały skanów stron redakcyjnych czy spisów treści.

Informacja ta była dla nas istotna, gdy ocenialiśmy jakość dokonanej przez nas kategoryzacji dokumentów oraz odsetek dokumentów wyłączanych z analiz (np. analizy synonimów czy analizy tematów).

Ostatecznie w KPCN znajduje 43,5% dokumentów podzielonych oryginalnie przez czasopisma i 55,5% przez zespół projektu. W 1% dokumenty (są to zeszyty czasopisma „PAUza Akademicka") zostały sklasyfikowane jako niepodzielone.

3.3. Kategoryzacja dokumentów

Założyliśmy, że do dalszych analiz będziemy brać pod uwagę przede wszystkim artykuły naukowe, ale również stanowiska organów państwowych czy instytucji

sektora nauki i szkolnictwa wyższego. W związku z tym wykluczyliśmy z analizy dokumenty, które byliśmy w stanie rozpoznać w procesie digitalizacji jako ewidentnie niemieszczące się w interesujących nas kategoriach. Taka ekspercka kategoryzacja nie posiadała jednak pełnej skuteczności i nie mogła być zastosowana na większą skalę w czasopismach skupionych raczej na informowaniu o życiu naukowym niż na publikowaniu artykułów naukowych (dotyczy to np. czasopisma „Sprawy Nauki: Miesięcznik Publicystyczno-Informacyjny"). Jako dokumenty niewchodzące do analizy uznaliśmy:

- strony redakcyjne;
- spisy treści;
- artykuły opublikowane w innym języku niż polski;
- biografie;
- dokumenty składające się jedynie z tabel lub zestawień numerycznych;
- zestawienia bibliograficzne;
- ogłoszenia;
- streszczenia artykułów.

Niemniej jednak wszystkie dokumenty z korpusu były poddawane tym samym działaniom (tj. rozpoznanie tekstów, przetwarzanie tekstu).

Tabela 2 prezentuje zestawienie liczby dokumentów z każdego czasopisma sklasyfikowanych lub nie do dalszej analizy. Dokumentami nazywamy wszystkie publikacje zamieszczone w danym czasopiśmie, natomiast artykułami te dokumenty, które wchodzą do analizy.

Tabela 2. Klasyfikacja dokumentów według czasopisma.

Czasopismo	Tak	% Tak	Nie	% Nie	Łącznie dokumentów
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 Publicystyczno-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

Czasopismo	Tak	% Tak	Nie	% Nie	Łącznie dokumentów
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
Łącznie	44 256	85,4%	7 556	14,6%	51 812

Źródło: Opracowanie własne.

3.4. Rozpoznanie tekstów (OCR) zeskanowanych dokumentów

Testowaliśmy różne rozwiązania, aby uzyskać zadowalającą jakość rozpoznania tekstów. Ostatecznie zdecydowaliśmy na używanie programu ABBYY FineReader 11 Professional Edition, który pozwalał na rozpoznanie nie tylko tekstu, ale również akapitów oraz łączenie ich w przypadku druku dwuszpaltowego.

Plikami wsadowymi były pliki PDF (oryginalnie przygotowane przez czasopisma) lub pliki graficzne TIFF/JPG. Plikami wynikowymi były pliki PDF z dokumentami stworzonymi na podstawie plików graficznych oraz pliki tekstowe TXT, które były podstawą dla dalszej obróbki dokumentów.

3.5. Czyszczenie, tokenizacja i lematyzacja

Wejściowe dokumenty tekstowe w formacie TXT zawierają wiele danych, które stanowią zbędny szum w analizie. W związku z tym procedura przetwarzania danych składa się z następujących etapów:

- Usunięcie liczb, adresów stron internetowych, znaków interpunkcyjnych, adresów e-mail i znaków specjalnych (takich jak !@$%*><+?);
- Konwersja całego tekstu na małe litery;
- Usunięcie z danej publikacji jej tytułu oraz nazwisk autorów (wiele czasopism stosuje żywą paginę górną, w związku z czym tytuł oraz autorzy wielokrotnie występują w danym dokumencie);
- Tokenizacja, czyli podział tekstu na indywidualne jednostki, zwane tokenami, które zazwyczaj reprezentują pojedyncze słowa lub wyrażenia. Celem tokenizacji jest uporządkowanie tekstu i przekształcenie go w strukturę, którą można łatwo przetwarzać;
- Usunięcie StopWords, czyli słów pojawiających się często, lecz nieistotnych z perspektywy podejmowanych analiz (Schofield i in., 2017). Chodzi zarówno o np. przyimki, imiona, jak i słowa, które są powszechne w danym korpusie tematycznym (w przypadku KPCN będzie to również m.in. „polska" czy „nauka").

W ten sposób przetworzone i wyczyszczone teksty poddawane są lematyzacji. Lematyzacja to proces przekształcania słów do ich formy podstawowej, zwanej lematem. Jest to kluczowy etap w analizie tekstu, który ma na celu zredukowanie różnych form gramatycznych słów do jednej, ułatwiając tym samym analizę i porównywanie danych tekstowych. W przypadku języka polskiego lematyzacja jest zadaniem wyjątkowo trudnym, ze względu na bogactwo form fleksyjnych, złożoność gramatyki i liczne wyjątki. W trakcie prac okazało się, że najlepsze efekty i wydajność dostarczają dwie biblioteki Pythona, tj. Lemmagen oraz Morfeusz. Przekształcenie tekstu za pomocą lematyzacji umożliwia skupienie się na kluczowych elementach tekstu, eliminując równocześnie redundancje wynikające z różnorodności form gramatycznych.

3.6. Tworzenie bazy danych o dokumentach

Korpus składa się z dokumentów tekstowych powiązanych z danymi bibliograficznymi. Informacje o dokumentach spisywaliśmy w przypadku materiałów skanowanych z natury ze spisów treści oraz informacji zamieszczonych przy danych dokumentach, a w przypadku plików cyfrowych (pobieranych ze stron internetowych czasopism lub archiwów) z informacji zawartych przy danym dokumencie. W bazie korpusu zawarte są informacje dla każdego dokumentu. Ich zakres odnosi się do: Tytułu, Autorów, Roku publikacji, Kategorii dokumentu (ta kategoria była wytwarzana przez nas), Informacji, czy dany dokument przechodzi do analizy, Tomu, Numeru zeszytu, Identyfikatora DOI, Linku do pliku PDF (lub html) w wersji online, Strony początkowej, Strony końcowej.

Odsetek informacji w danej kategorii zależy od jakości danych, ale również typu dokumentów (artykuły publikowane w wersji html nie mają informacji o paginacji, nie wszystkie dokumenty mają wskazaną informację o autorach).

4. Ilościowa charakterystyka czasopism

4.1. Liczba roczników i artykułów

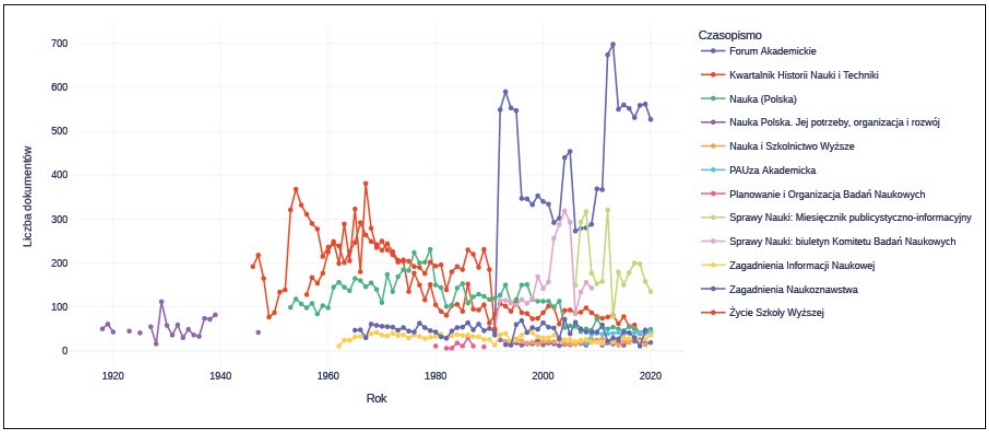
Tabela 4 prezentuje liczbę roczników (do 2020 r.) uwzględnionych w korpusie dla każdego czasopisma, liczbę dokumentów oraz liczbę artykułów (czyli dokumentów zaklasyfikowanych do analizy) na rocznik.

Rysunek 2 prezentuje zmianę liczby dokumentów (nie artykułów) w czasie dla każdego czasopisma z korpusu. Dwa roczniki są połączone linią tylko wtedy, gdy była ciągłość wydawnicza, tj. roczniki wydano rok po roku.

Tabela 3. Podsumowanie informacji bibliograficznych w korpusie.

Czasopismo	Tytuł	Autor	Rok	Kategoria dokumentu	Do analizy	Tom	Zeszyt	DOI	Strona początkowa	Strona końcowa
Forum Akademickie	100,0%	70,4%	100,0%	5,8%	100,0%	100,0%	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%	99,4%	99,3%
Nauka (Polska)	100,0%	84,2%	100,0%	100,0%	100,0%	100,0%	99,7%	0,0%	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%	94,8%	94,7%
Nauka i Szkolnictwo Wyższe	100,0%	99,4%	100,0%	100,0%	100,0%	100,0%	100,0%	0,0%	0,0%	0,0%
PAUza Akademicka	0,0%	0,0%	100,0%	0,0%	100,0%	0,0%	100,0%	0,0%	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%	89,3%	89,3%
Sprawy Nauki: Miesięcznik Publicystyczno-Informacyjny	100,0%	0,0%	100,0%	0,7%	100,0%	100,0%	0,0%	0,0%	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%	87,4%	87,4%
Zagadnienia Informacji Naukowej	100,0%	81,4%	100,0%	21,0%	100,0%	100,0%	100,0%	0,0%	72,2%	71,6%
Zagadnienia Naukoznawstwa	100,0%	88,5%	100,0%	0,0%	99,7%	0,0%	0,0%	0,0%	79,1%	79,1%
Życie Szkoły Wyższej	100,0%	75,3%	100,0%	6,4%	100,0%	100,0%	100,0%	0,1%	91,6%	91,6%

Źródło: Opracowanie własne.



Rysunek 2. Liczba dokumentów w czasie dla każdego czasopisma.

Źródło: Opracowanie własne.

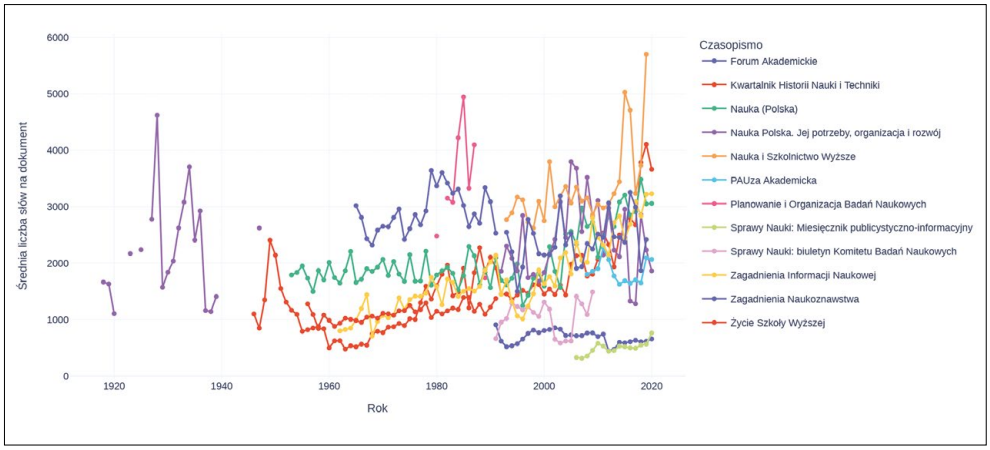
Tabela 4. Podsumowanie liczby roczników oraz artykułów na rocznik.

Czasopismo	**Łącznie dokumentów**	**Liczba artykułów do analizy**	**Liczba roczników**	**Liczba artykułów na rocznik**
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 Publicystyczno-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
Łącznie	51 822	44 256	452	114,7

Źródło: Opracowanie własne.

4.2. Długość artykułów

Policzyliśmy średnią długość dokumentów dla każdego czasopisma. W tej analizie liczyliśmy słowa już po procedurze czyszczenia oraz lematyzacji. Rysunek 3 pokazuje wyniki tej analizy.



Rysunek 3. Liczba słów na dokument w czasie dla każdego czasopisma.

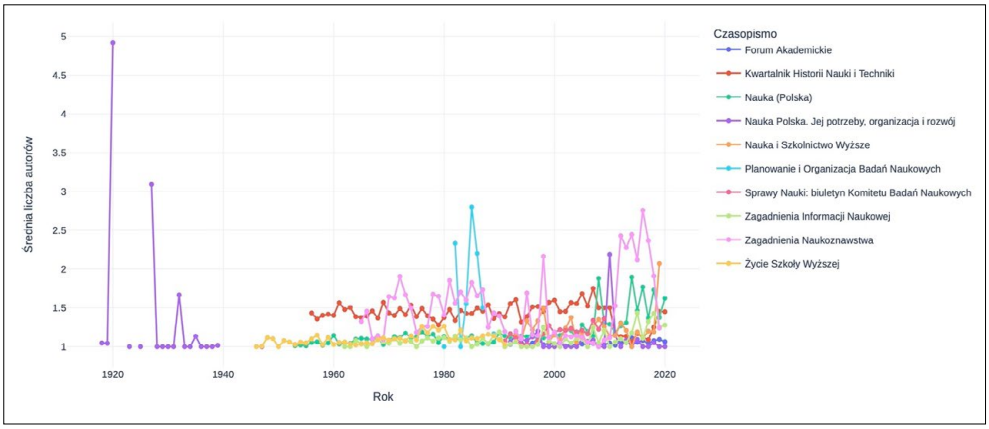
Źródło: Opracowanie własne.

4.3. Liczba autorów w podziale na dokumenty

Przeanalizowaliśmy zmianę średniej liczby autorów na artykuł według lat dla każdego czasopisma. W analizie uwzględnione są tylko dokumenty, które posiadały informacje o autorach. Rysunek 4 pokazuje wyniki tej analizy. Wartości odstające dla czasopisma „Nauka Polska. Jej Potrzeby, Organizacja i Rozwój" wynikają z publikacji tzw. rozpraw, to jest wypowiedzi wielu uczonych w obrębie jednego dokumentu (każda wypowiedź ma wskazane autorstwo).

5. Analiza tematów na przykładzie „Nauki (Polskiej)"

Niniejsza sekcja ma na celu pokazanie, w jaki sposób można wykorzystać KPCN do stworzenia krajobrazu dyskusji (tematów) toczonych na łamach pojedynczego czasopisma. Zdecydowaliśmy się na „Naukę (Polską)", gdyż jest ona wydawana bez przerwy od 1953 do 2020 r. (jest to końcowy rok obecnej wersji KPCN), najpierw pod nazwą „Nauka Polska", później od 1994 r. „Nauka". Dlatego też stosujemy zapis z „Polską" umieszczoną w nawiasie, aby podkreślić scalenie tych dwóch czasopism.



Rysunek 4. Średnia liczba autorów dokumentu w czasie dla każdego czasopisma.

Źródło: Opracowanie własne.

Należy wskazać, że analizę tę można byłoby jeszcze rozszerzyć o tomy wydane od 1918 r. przez czasopismo „Nauka Polska. Jej Potrzeby Organizacja i Rozwój", losy tych wszystkich czasopism bowiem są ze sobą zespolone. Najlepiej to przeplatanie się oddają początkowe akapity pierwszego numeru „Nauki" z 1994 r., które przytaczamy tutaj w całości:

> „Nauka Polska" była wydawana jako rocznik w latach 1918–1939 i 1947 przez Kasę im. Józefa Mianowskiego. Od roku 1953 „Nauka Polska" ukazywała się jako kwartalnik Polskiej Akademii Nauk. W latach 1962–1974 była dwumiesięcznikiem, następnie, w okresie 1975–1981, miesięcznikiem, aby w latach 1982–1993 powrócić do formy dwumiesięcznika. W numerze 5 (270) „Nauki Polskiej" z 1993 r. omówione zostały bardziej szczegółowo losy organizacyjne, profil i zawartość treściowa tego czasopisma. Dalszą, zasadniczą zmianę przynosi rok 1994. „Nauka Polska" zostaje przekształcona w nowy tytuł wydawniczy – kwartalnik „Nauka".
>
> Dwie są podstawowe przyczyny tego przekształcenia. Pierwsza, to reaktywowanie w 1991 r. po czterdziestu latach Kasy im. Józefa Mianowskiego – Fundacji Popierania Nauki. Kasa im. Józefa Mianowskiego powróciła do wydawania swojego dawnego tytułu – publikacji ciągłej, rocznika „Nauka Polska", którego pierwszy, a kolejny XXVI numer ukazał się w 1992 r. W tej sytuacji stało się sprawą oczywistą dla kierownictwa Polskiej Akademii Nauk ustąpienie instytucji zaprzyjaźnionej tytułu, który był jej własnością do 1951 r. („Od Redakcji", 1994).

W latach 1953–2020 „Nauka (Polska)" opublikowała 7844 dokumentów, z czego do analizy zakwalifikowaliśmy 6024 artykuły (1820 zostało wyłączonych). Większy odsetek artykułów wyłączonych z analizy został zlokalizowany przed 2004 r., bowiem od tego roku redakcja posiada cyfrowe archiwum czasopisma na swojej stronie, z zeszytami podzielonymi przez redakcję. W ten sposób w naszym korpusie nie znalazły się np. spisy treści z tego okresu, które były wyłączane z analizy w całym okresie. Dodatkowo w kolejnym kroku wyłączyliśmy z analizy 252 artykuły, które posiadały mniej niż 300 słów (już po lematyzacji Lemmagenem). Ostatecznie nie uwzględniliśmy w analizie 26,42% dokumentów. Zatem finalny zbiór do analizy tematów składa się z 5772 artykułów.

Przypisanie tak wielu artykułów do poszczególnych tematów (nieznanych jeszcze przed lekturą) byłoby zadaniem teoretycznie możliwym, ale niezwykle pracochłonnym. Dlatego też analizę tematów (ang. *topic modeling*) można wykonać, posługując się technikami uczenia maszynowego. W naszym przypadku będzie to nienadzorowane uczenie maszynowe, gdyż przed procesem nie definiujemy tematów ani nie dostarczamy oznaczonych danych treningowych (np. nie wskazujemy, że dany tekst powinien być zaklasyfikowany z konkretnym innym tekstem w ramach tego samego tematu).

Analiza tematów to ilościowa analiza tekstu, która pozwala pogrupować dokumenty korpusu według dominujących tematów. Każdy dokument będzie charakteryzowany przez wiele tematów, jednakże zostanie wyłoniony ten dominujący i dokument zostanie do niego przypisany. W tym podejściu korpus jest traktowany jako zbiór dokumentów, z których każdy jest złożony ze zdefiniowanej liczby tematów, na które składają się słowa z danego korpusu. Oznacza to, że poszczególne słowa w korpusie są powiązane z danym tematem/tematami. Każde słowo posiada swoją wagę, wskazującą jego znaczenie dla danego tematu.

5.1. Opis przyjętej procedury wyłaniania tematów

Zaletą NMF, w przypadku analizowania wielu dłuższych dokumentów tekstowych jest to, że nie pracuje się na całej macierzy TF-IDF (ang. *Term Frequency-Inverse Document Frequency*), lecz w ramach analizy redukujemy złożoność tej macierzy, ograniczając liczbę słów branych pod uwagę. TF-IDF to jedna z metod obliczania wagi słów w oparciu o liczbę ich wystąpień. Ta metoda bierze pod uwagę zarówno częstość występowania danego słowa w dokumencie (TF), jak i jego unikatowość wśród całego zbioru dokumentów (IDF), co umożliwia zrozumienie znaczenia i kontekstu poszczególnych słów w danym korpusie tekstowym. Analizując wagi słów we wszystkich dokumentach w zbiorze, można identyfikować słowa kluczowe, które charakteryzują poszczególne tematy. Dzięki temu możliwe jest tworzenie grup dokumentów na podstawie podobieństwa ważnych słów, co umożliwia modelowanie tematyczne. Możliwość redukcji złożoności TF-IDF w ramach korzystania z NMF jest szczególnie istotne w przypadku korpusu, którego część dokumentów powstała ze zeskanowanych dokumentów, co sprawia, że mogą pojawiać się pojedyncze słowa „niemające sensu", gdyż są to błędy wynikające z przetwarzania danych.

Wyjściowa macierz TF-IDF opisująca nasz korpus składa się z 5772 artykułów oraz 64110 unikatowych słów (uwzględniamy tylko takie słowa, które występują co najmniej 5 razy w całym korpusie, lecz mogą wystąpić w jednym dokumencie). Rzadkość (ang. *sparsity*) macierzy wynosi w tym przypadku 98,59%, co oznacza, że taki odsetek elementów macierzy to zero (to znaczy, że dane słowo nie występuje w danym dokumencie). Dlatego też, aby ulepszyć model, należało zredukować rzadkość macierzy, gdyż przetwarzanie tak wielu zer nie jest efektywne, a macierze

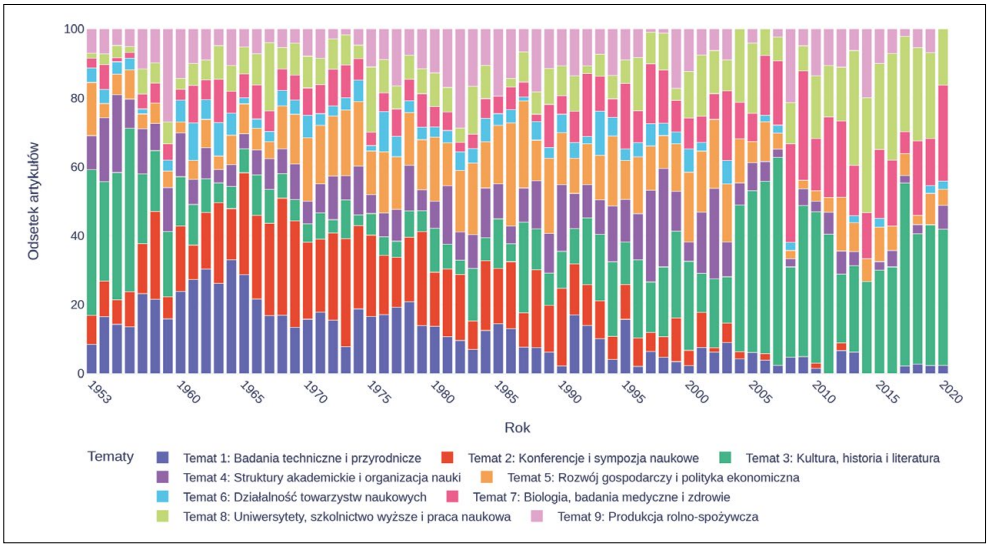
rzadkie są trudniejsze do obliczania, co sprawia, że algorytmy „preferują" działanie na mniej rzadkich danych.

W tym celu testowaliśmy, jakie parametry modelu możemy przyjąć, aby nie tylko zmniejszyć rzadkość macierzy, ale przede wszystkim uzyskać sensowne i rozłączne tematy. Ustaliliśmy, że finalna macierz posiada rzadkość 61,33% i do analizy tematów będzie charakteryzować się parametrami max_features=800 oraz mindf_ig=20 dla dziewięciu tematów, gdzie max_feature oznacza liczbę pierwszych słów kluczowych dla korpusu (uporządkowanych ze względu na ich wagę), a mindf_ig wskazuje próg ignorowania słów kluczowych w modelu (w tym przypadku słów, które występują rzadziej niż 20 razy w korpusie).

5.2. Tematyka „Nauki (Polskiej)"

Przyjmując powyżej opisaną metodę wyłoniliśmy 9 spójnych i rozłącznych tematów, do których zostały zaklasyfikowane wszystkie analizowane artykuły. Należy raz jeszcze podkreślić, że analiza tematów przypisuje do dokumentu temat dominujący, co nie oznacza to, że inne tematy nie są w nim „widoczne" podczas lektury.

Rysunek 5 przedstawia rozkład tematów według lat ukazywania czasopism. Widać, że temat pierwszy, poświęcony opisowi badań technicznych i przyrodniczych, miał istotniejsze znaczenie w czasopiśmie przed rokiem 2010 (w szczególności w latach 60-tych ubiegłego wieku).



Rysunek 5. Rozkład tematów w czasopiśmie „Nauka (Polska)" w latach 1953–2020.

Źródło: Opracowanie własne.

Widać również, że po roku 2005 w zasadzie całkowicie przestano publikować materiały opisujące wydarzenia konferencyjne i sympozja naukowe, podczas gdy jednocześnie bardzo istotnie zwiększyła się liczba artykułów poświęconych naukom humanistycznym (Temat 3: *Kultura, historia i literatura*) oraz sprawom szkolnictwa wyższego i pracy naukowej (Temat 8: *Uniwersytety, szkolnictwo wyższe i praca naukowa*). Istotnie zmniejszyła się również liczba publikacji dotyczących badań technicznych i przyrodniczych (Temat 1: *Badania techniczne i przyrodnicze*). Obserwacja tego, że temat badań technicznych i przyrodniczych zwiększa ilościowo swoją częstotliwość sukcesywnie od lat 50-tych ubiegłego wieku i znacząco spada po transformacji ustrojowej, może okazać się istotna dla bardziej pogłębionych badań historycznych. Jak wskazywał chociażby Hubner (1994, s. 32), już pierwsze inspirowane nauką radziecką próby reform powojennej nauki w Polsce miały na celu zwiększenie roli nauk technicznych i ścisłych względem nauk humanistycznych i przyrodniczych.

5.3. Ewaluacja wyników

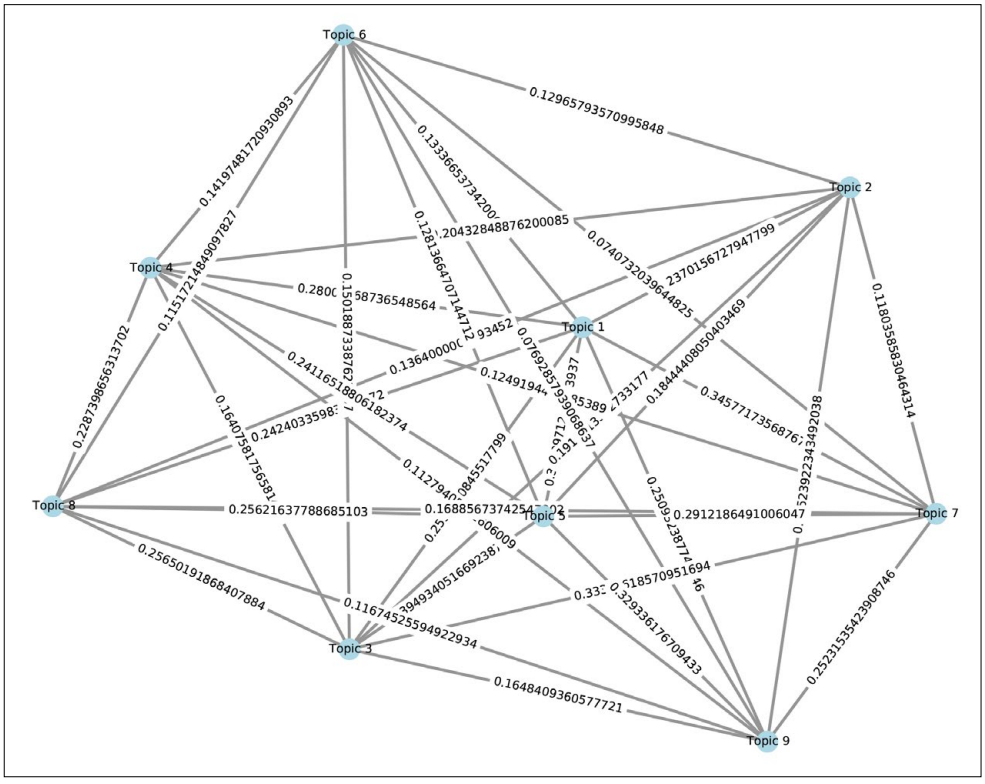
Największym wyzwaniem w analizie tematów jest ewaluacja modelu, tj. ocenienie czy wyłonione tematy dobrze opisują dane (teksty), są sensowne i spójne. Mimo iż można wykorzystać wiele ilościowych miar do oceny liczby wyłonionych tematów oraz ich sensowności, takich jak spójność tematyczna, *perplexity* (miara stosowana do oceny, jak dobrze model jest w stanie przewidzieć nowe, nieznane wcześniej dane tekstowe), wizualizacja tematów (np. z użyciem biblioteki pyLDAvis), to jednakże ocena ekspercka każdego z wyników jest w przypadku takiego korpusu tematycznego niezastąpiona.

Niniejszy model ewaluowaliśmy ostatecznie w następujący sposób: (1) analizowaliśmy relacje między tematami przy użyciu sieci grafów; (2) analizowaliśmy najważniejsze wyłonione słowa dla danego tematu i tworzyliśmy ogólne etykiety dla tematów (tj. ich nazwy), aby dodatkowo zweryfikować potencjalne nachodzenie się tych tematów (w tym procesie każdy z autorów tworzył etykiety, a następnie uzgadnialiśmy wspólną etykietę lub wyjaśnialiśmy rozbieżności); (3) sprawdzaliśmy losowo wybrane artykuły i sensowność ich zaklasyfikowania do poszczególnych tematów; (4) w przypadku analizy tematycznej „Nauki (Polskiej)" byliśmy w stanie posłużyć się porównaniem z Bibliografią Polskiej Naukometrii (BPN)[2], która indeksuje blisko dwa tysiące publikacji polskich badaczy, sklasyfikowanych jako publikacje naukometryczne. W związku z tym wyłoniliśmy z BPN wszystkie artykuły opublikowane w „Nauce (Polskiej)" i sprawdziliśmy, czy zostały przypisane do tematów, które można uznać za „naukometryczne".

2 Bibliografia Polskiej Naukometrii, https://sc.amu.edu.pl/bibliography/, data dostępu: 1 września 2023 r.

5.3.1. Sieć grafów

Sieci grafów mogą być używane do wizualizacji związków pomiędzy tematami w modelu NMF. Sieć taka składa się z węzłów i krawędzi: węzły reprezentują różne tematy, a krawędzie reprezentują siłę związku pomiędzy tematami (im wartość bliższa 0, tym tematy są bardziej rozłączne, a im bliższa 1, tym tematy są bardziej podobne). Rysunek 6 pokazuje wyniki dla przyjętych w niniejszej analizie założeń na podstawie analizy podobieństwa cosinusów (ang. *cosine similarity*). Aby zinterpretować sieć grafów, można spojrzeć na sklasteryzowanie węzłów oraz połączenia między nimi. Klastry węzłów, które są mocno połączone, wskazują na tematy, które są powiązane lub obejmują podobne aspekty korpusu. Węzły, które są słabo połączone lub wcale niepołączone, wskazują na tematy, które są odrębne i obejmują różne aspekty korpusu. Sieć grafów może być również używana do identyfikowania tematów odstających lub tematów, które są słabo zdefiniowane.



Rysunek 6. Sieć grafów między 9 tematami (Topics) „Nauki (Polskiej)" w latach 1953–2020.

Źródło: Opracowanie własne.

5.3.2. Najważniejsze słowa

Poniższa lista zawiera dla każdego tematu 10 najważniejszych słów kluczowych, uporządkowanych od najważniejszego (z najwyższą wagą dla tematu) do najmniej ważnego.

- Dla Tematu 1: *Badania techniczne i przyrodnicze* najważniejsze słowa kluczowe to: badanie, praca, zakład, zakres, badawczy, dziedzina, metoda, zagadnienie, fizyka, rozwój.
- Dla Tematu 2: *Konferencje i sympozja naukowe* najważniejsze słowa kluczowe to: referat, kongres, konferencja, sesja, sympozjum, sekcja, międzynarodowy, odbyć, uczestnik, wygłosić.
- Dla Tematu 3: *Kultura, historia i literatura* najważniejsze słowa kluczowe to: kultura, historia, język, człowiek, literatura, dzieło, świat, wielki, prawo, dzieje.
- Dla Tematu 4: S*truktury akademickie i organizacja nauki* najważniejsze słowa kluczowe to: akademia, placówka, wydział, prezydium, komisja, sekretarz, zgromadzenie, współpraca, działalność, sprawa.
- Dla Tematu 5: *Rozwój gospodarczy i polityka ekonomiczna* najważniejsze słowa kluczowe to: rozwój, kraj, społeczny, gospodarka, socjalistyczny, gospodarczy, społeczeństwo, państwo, polityka, program.
- Dla Tematu 6: *Działalność towarzystw naukowych* najważniejsze słowa kluczowe to: towarzystwo, działalność, oddział, zjazd, biblioteka, upowszechnianie, wydawniczy, regionalny, praca, społeczny.
- Dla Tematu 7: *Biologia, badania medyczne i zdrowie* najważniejsze słowa kluczowe to: komórka, choroba, białko, człowiek, genetyczny, badanie, zwierzę, zdrowie, organizm, biologia.
- Dla Tematu 8: *Uniwersytety, szkolnictwo wyższe i praca naukowa* najważniejsze słowa kluczowe to: uczelnia, uniwersytet, wysoki, profesor, szkoła, student, doktorski, akademicki, szkolnictwo, praca.
- Dla Tematu 9: *Produkcja rolno-spożywcza* najważniejsze słowa kluczowe to: roślina, produkcja, rolnictwo, rolniczy, ochrona, wodny, woda, energia, gospodarka, węgiel.

Na podstawie listy tych słów mogliśmy sprawdzić nie tylko wewnętrzną spójność tematów oraz ich rozłączność, ale również ocenić jakość przetwarzania danych i lematyzacji. Listy słów kluczowych (w iteracyjnym procesie budowania modelu generowaliśmy Top20 słów kluczowych) były istotnym wyznacznikiem etykiet nadawanych tematom.

5.3.3. Przypisanie artykułów do tematów

Proces ten miał charakter czysto ekspercki. Mając już wytworzoną propozycję modelu oraz etykiety dla tematów, sprawdzaliśmy, jaki temat wiodący został przypisany

do losowo wybranych artykułów. Podczas weryfikacji, tj. lektury tekstu i weryfikacji przypisania tematu, mieliśmy na uwadze, że przypisywany jest temat wiodący, a nie temat „jedyny". Proces ten potwierdził jakość finalnego modelu oraz przyjętych parametrów.

5.3.4. Porównanie z Bibliografią Polskiej Naukometrii

W Bibliografii Polskiej Naukometrii odnaleźliśmy 41 artykułów z „Nauki (Polskiej)" do roku 2020. Dwadzieścia pięć z nich, czyli 60% analizowanych, zostało zaklasyfikowanych w naszym modelu jako artykuły z dominującym Tematem 8: *Uniwersytety, szkolnictwo wyższe i praca naukowa*, siedem artykułów z Tematem 3: *Kultura, historia i literatura*, sześć artykułów z Tematem 1: *Badania techniczne i przyrodnicze*, po jednym artykule z tematami 4, 5, 7. Biorąc pod uwagę, że model wskazuje na dominujący (a nie jedyny) temat, należy uznać ten wynik porównania dwóch zupełnie odmiennych podejść, tj. uczenia maszynowego i klasyfikacji eksperckiej w Bibliografii Polskiej Naukometrii, za dobry. Celem analizy tematów nie było takie zreprodukowanie klasyfikacji, aby wszystkie artykuły z BNP opublikowane w „Nauce (Polskiej)" zostały sklasyfikowane do jednego tematu, lecz dodatkowa weryfikacja tego, czy większość artykułów sklasyfikowanych ekspercko znajdzie się w niewielkiej liczbie tematów. Uważamy zatem, że ta analiza potwierdziła wartość przedstawionego modelu analizy tematów.

6. Perspektywy rozwoju korpusu

Niniejszy tekst prezentuje obecny kształt Korpusu Polskich Czasopism Naukoznawczych w 2022 r. po trzech latach pracy. Jest to oczywiście dopiero początek, a nie koniec drogi. Poniżej przedstawiamy kierunki, w których korpus może i będzie się rozwijać.

Przede wszystkim warto rozbudować korpus o kolejne czasopisma, na łamach których rozwijały się dyskusje naukoznawcze, takie jak „Studia Historiae Scientiarum" czy „Organon". Jednym z większych wyzwań jest poprawienie jakości danych tekstowych po cyfrowym rozpoznaniu tekstu i w ten sposób ulepszenie procesu lematyzacji. Stworzenie unikatowych identyfikatorów autorów pozwoli na przeprowadzenie dodatkowych analiz bibliometrycznych. Obecnie w korpusie są zawarte informacje o autorach dokumentów (jeśli takowe były zawarte w spisie treści lub w dokumencie). Należy jednakże wykonać istotną pracę, aby połączyć ze sobą autorów zapisanych w różny sposób, tak aby „F. Znaniecki" był połączony z „Florianem Znanieckim".

Jednym z kierunków rozwoju może być dokonanie ekstrakcji cytowań z dokumentów, zarówno z bibliografii załącznikowej, jak i przypisów dolnych. W tym momencie istnieją pojedyncze narzędzia do ekstrakcji informacji bibliograficznych

z bibliografii, lecz ekstrakcja z przypisów dolnych jest w zasadzie niewykonalna na masową skalę, mimo iż pojedyncze grupy naukowe pracują nad narzędziami.

Finansowanie

Prace nad tą wersją korpusu zostały sfinansowane w ramach projektu „Punktoza w czasach systemów ewaluacji nauki", finansowanego ze środków Narodowego Centrum Nauki, nr decyzji UMO-2017/26EHS2/00019.

Podziękowania

Na wielu etapach uzyskaliśmy pomoc w wyszukiwaniu, skanowaniu i opracowywaniu czasopism i dokumentów. Chcemy podziękować serdecznie Paulinie Dudzińskiej, Jolancie Noskowiak, Wiesławie Krysztofiak, Sarze Rotnickiej, Małgorzacie Rychlik, Krzysztofowi Skibniewskiemu i Michałowi Spaleniakowi, których wsparcie pozwoliło nam ukończyć naszą pracę.

Bibliografia

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.

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

Abstract

Purpose/Thesis: This article introduces the Corpus of Polish Science of Science Journals (CPSSJ), a specialized corpus created to support research in the field of science of science and its development in Poland.

Approach/Methods: The construction of the corpus was based on the digitization of previously non-digitized articles and the retrieval of articles from scientific journal websites and digital libraries. The documents were processed by various natural language processing methods.

Results and Conclusions: The capabilities of the CPSSJ are demonstrated through a topic modeling 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 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 digitization 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 modeling. Topic-specific corpus.

EMANUEL KULCZYCKI – zajmuje się oceną nauki oraz studiami nad nauką. Jest profesorem uczelni w Uniwersytecie im. Adama Mickiewicza w Poznaniu, gdzie kieruje Pracownią Komunikacji Naukowej. W latach 2018–2020 był przewodniczącym European Network for Research Evaluation in the Social Sciences and the Humanities zrzeszającej naukowców z 37 krajów. W 2018 r. otrzymał nagrodę naukową Prezesa Polskiej Akademii Nauk za serię artykułów naukowych poświęconych naukometrii, opublikowanych w uznanych czasopismach międzynarodowych.

YEIMER ALEXANDER ZAMBRANO MENA – jest kolumbijskim fizykiem oraz data scientist. Uzyskał licencjat z fizyki na Narodowym Uniwersytecie Kolumbii. Był członkiem grupy fizyki nowych materiałów na Narodowym Uniwersytecie Kolumbii, gdzie zajmował się analizą danych spektrów rentgenowskich. Następnie zdobył tytuł magistra fizyki na Uniwersytecie im. Adama Mickiewicza w Poznaniu (studia sfinansowane przez program stypendialny Ignacego Łukasiewicza). Jego zainteresowania badawcze obejmują data science, przetwarzanie języka naturalnego oraz uczenie maszynowe.

FRANCISZEK KRAWCZYK – jest doktorantem w Szkole Doktorskiej Uniwersytetu im. Adama Mickiewicza w Poznaniu. W rozprawie doktorskiej skupia się na organizowaniu oporu przeciwko

nierównym relacjom między centrami a peryferiami w naukach. Napisał pracę magisterską poświęconą rozwojowi tzw. czasopism drapieżnych. Zainteresowania badawcze Franciszka Krawczyka obejmują geografię wiedzy, czasopisma drapieżne, ewaluację oraz socjologię nauki.

Kontakt z Autorem:
emek@amu.edu.pl
Emanuel Kulczycki
Scholarly Communication Research Group
ul. Międzychodzka 5, pokój 405
60-371 Poznań